

Bioestatística sem segredos

Annibal Muniz Silvany Neto

Bioestatística sem segredos

Bioestatística sem segredos

Annibal Muniz Silvany Neto

Médico, Epidemiologista e Mestre em Saúde Comunitária.
Professor Adjunto do Departamento de Medicina Preventiva e Social
da Faculdade de Medicina da Bahia da Universidade Federal da Bahia.

1^a edição

Edição do autor

Salvador – Bahia
2008

Copyright © 2008 pela humanidade.

Nenhum direito reservado. Qualquer parte deste livro pode ser reproduzida ou transcrita, sob qualquer forma ou por qualquer meio – eletrônico, mecânico, por fotocópia, por gravação – sem necessidade de prévia autorização, para fins não comerciais, desde que o autor e a fonte sejam citados e esta nota seja incluída.

Escrito e impresso no Brasil.

S586 Silvany Neto, Annibal Muniz.
Bioestatística sem segredos / Annibal Muniz Silvany Neto. – Salvador,
2008.
321p.: il.

ISBN 978-85-907970-0-5

1.Biometria. 2.Bioestatística. I.Título.

CDU – 57.087.1

Descrição desta publicação:

Formato 21,6 x 27,9 cm;

Fontes “arial”, “arial black” “symbol” e “times new roman”;

Miolo em papel sulfite 90 g/m²;

Criação da capa Purê Design;

Capa em papel supremo 250 g/m²;

Fotolito e impressão da capa Cian e acabamento Finish;

Tiragem 100 exemplares.

☐ Aos meus pais, ao meu filho e a todas as outras pessoas virtuosas ☐

▣ APRESENTAÇÃO ▣

O livro Bioestatística Sem Segredos do Prof. Annibal Muniz Silvany Neto é dirigido aos estudantes da área de Saúde. O estilo em forma de diálogo com os leitores é uma tentativa de motivá-los e também de vencer a resistência que muitos deles trazem para o estudo e a discussão de conceitos estatísticos.

O livro é uma ótima referência para um primeiro curso de Bioestatística, pela riqueza de detalhes com que os tópicos básicos de Estatística Descritiva, Modelos Probabilísticos e Inferência Estatística são tratados, pelo estilo atraente e agradável e pela cuidadosa apresentação e discussão dos conceitos, introduzidos através de exemplos e aplicações oriundos, na maioria, da Epidemiologia, área de atuação do Prof. Neto. Pesquisadores da área também podem se beneficiar com a leitura, principalmente dos capítulos sobre amostragem e cálculo do tamanho da amostra.

Trata-se, assim, de um livro que deverá contribuir para a difusão do ensino da Bioestatística e que ajudará na formação dos futuros pesquisadores na área de Saúde.

Salvador, 25/11/2007
Nelson Fernandes de Oliveira

▣ PREFÁCIO ▣

Ao longo da minha vida profissional estudei Estatística em vários livros-texto e artigos em revistas especializadas. Em todos senti a falta de muitas explicações necessárias ao entendimento dos assuntos abordados. Acho que isso tem provocado muita resistência, confusão, impaciência e desânimo, levando muitos(as) estudantes a perderem a motivação para aprender esta disciplina.

Isso ocorre, em minha opinião, em todos os níveis de formação, e o problema tem sido “resolvido”, na prática, da seguinte maneira: quando as pessoas precisam obter créditos obrigatórios na disciplina “Estatística (ou Bioestatística)” para os cursos que estão fazendo, desistem logo de aprender e relaxam e, mais adiante, quando precisarem utilizar procedimentos estatísticos, aqueles que tiverem condições financeiras para tanto, pagam a um estatístico para executar essa tarefa. Isso significa que a maioria das pessoas vai passando de um nível a outro de sua formação, sem acrescentar novos conhecimentos e capacitações em Bioestatística. Não defendo que cada pessoa se torne um estatístico profissional, mas que vá aumentando sua capacitação nessa área de modo a poder dialogar eficientemente com o estatístico. Esse diálogo é essencial para que sejam evitadas incorreções graves na aplicação da Estatística, por falhas na comunicação entre o “pesquisador” e o “estatístico”.

O desejo de fazer um livro de Bioestatística que explicasse mais cada tópico abordado, tornando mais fácil o seu entendimento e aprendizagem, foi a motivação principal para que eu escrevesse este livro. Todo seu conteúdo é apresentado através de um diálogo entre o autor e o(a) leitor(a). Espero ter conseguido tornar sua leitura uma tarefa estimulante, interessante e proveitosa. Este é, portanto, um livro de Bioestatística básica, com a pretensão de pegar na mão do(a) estudante e ajudá-lo(a) a “abrir as portas” da Bioestatística.

Outra ambição foi a de ser um educador além de professor. Isso se reflete em breves comentários com minhas opiniões sobre a natureza humana, o modo de organização das sociedades humanas, e na sugestão de leituras fora da área da Bioestatística. É evidente que os(as) leitores(as) podem discordar completamente das posições por mim defendidas, mas minha intenção não foi de modo algum “fazer a cabeça” de ninguém, e sim evitar uma postura alienada neste livro, falando somente de Bioestatística, como se o uso desta ferramenta fosse uma atividade neutra nas sociedades onde vivemos. Por isso, este livro não deve ser considerado como apenas “técnico”.

Do ponto de vista econômico, uma característica deste livro é ter sido feito de modo completamente artesanal. Isso, por um lado, me permitiu uma liberdade total de expressão e de definição da extensão e nível de profundidade dos capítulos. Em momento algum fiquei preocupado com o número de páginas do livro, porque sabia que, para explicar melhor os assuntos, era inevitável “gastar” mais tempo dialogando com o(a) estudante. Além disso, consegui não ceder os direitos autorais deste livro. Imagine o absurdo: você escreve um livro; é, portanto, o(a) autor(a) do mesmo; mas, se quiser obter financiamento para a edição do livro, a probabilidade de ter que ceder seus direitos autorais é muito grande. Existe absurdo maior do que cedermos um direito que, por definição, é “incedível”? Algumas editoras resolvem essa questão requerendo que o autor lhes ceda o direito de publicação e não o de autoria, mas isso não soluciona um outro problema que é a repartição desigual do valor da venda dos livros, ficando uma proporção mínima desse valor com o autor. Não me submeti também a essa exploração do meu trabalho. O caráter artesanal foi, então, importante para me livrar da tirania das leis de mercado, que certamente interfeririam nas características do livro. Mas, em

decorrência desse seu caráter, o livro deve conter “erros” gramaticais, lingüísticos, de normalização e de editoração, porque não foi revisado por profissionais dessas áreas, já que esses serviços são caríssimos.

Coloquei no livro apenas as referências bibliográficas essenciais ao atendimento dos objetivos propostos. Mostrar erudição estatística não era um desses objetivos. Além disso, os temas abordados já são consagrados no âmbito da Estatística Clássica. As referências são apresentadas no momento em que são referidas, para evitar que o(a) leitor(a) tenha de se dirigir ao final dos capítulos ou do livro para consultá-las.

A revisão dos fundamentos estatísticos que norteiam toda a apresentação dos assuntos foi feita por Nelson Fernandes de Oliveira, professor de Estatística aposentado do Instituto de Matemática da Universidade Federal da Bahia. Seu esmero e competência nessa tarefa conferiram ao livro uma qualidade técnica que, sem ele, não teria sido obtida.

Agradeço a todos(as) que me incentivaram a escrever. Ao meu filho que me ajudou decisivamente com as demonstrações algébricas, essenciais para explicar melhor vários assuntos. Ao meu pai pelo entusiasmo que demonstrou quando soube que eu estava escrevendo um livro, e pelas várias sugestões que ele ainda teve tempo de me dar para melhorá-lo. Aos Estudantes de Medicina Alba Cristina Sousa Oliveira, Ana Cláudia Oliveira Silva, Átila Cerveira Lueska, Carlos Eduardo Cerqueira Rolim, Dalton Willy Santos Oliveira, Lucas Santos Argolo, Luciana Santos Pimentel, Rafaela Sousa dos Santos, Rodrigo Santos Matos e Sandra Sousa Santos, e aos Professores Marco Antônio Vasconcelos Rêgo e Meirelayne Borges Duarte, pela revisão cuidadosa de vários capítulos. Ao Professor José Romélio Cordeiro e Aquino, pela revisão do texto e por sua amizade e incentivo.

Sou imensamente grato à minha mulher e seu filho pela paciência que tiveram com o envolvimento de tempo e energia que me foi exigido para enfrentar esse desafio. À minha mãe e meus irmãos por poder contar sempre com seu carinho e atenção.

Fica claro, então, que o livro resultou de um trabalho coletivo no qual, para minha alegria, um grupo de pessoas em cooperação e sem a necessidade de competir com outros, se dedicou a um projeto, sem nenhum interesse além da demonstração mútua de amizade, generosidade, carinho ou respeito.

Espero ter conseguido realizar os objetivos propostos e desejo a todos(as) uma boa jornada ao lerem este livro.

Salvador, janeiro de 2.008.

Annibal M. Silvany Neto.

▣ ÍNDICE ▣

CAPÍTULO 1.....	1
Quais as diferenças entre Estatística Descritiva, Analítica e Inferencial?.....	2
Quais as técnicas estatísticas mais utilizadas?.....	3
O que as denominações estatística paramétrica e não-paramétrica significam?.....	7
Como surgiu a Estatística Moderna?.....	8
E o que é Bioestatística?.....	9
CAPÍTULO 2.....	13
O que são variáveis?.....	14
Como classificar as variáveis?.....	14
Quanto à natureza qualitativa ou quantitativa.....	14
Quanto à posição no quadro de hipóteses da pesquisa.....	15
Quanto à sua expressão em valores contínuos ou não.....	18
Quanto ao número de categorias.....	19
Quanto à fixação prévia das freqüências nas categorias.....	19
Quanto à individualização da informação.....	20
Quanto à modalidade de escala.....	20
CAPÍTULO 3.....	23
O que é amostragem e por que realizá-la?.....	24
Quais os tipos de amostragens mais utilizados?.....	26
Amostragem aleatória simples.....	26
Amostragem aleatória sistemática.....	26
Amostragem aleatória por conglomerados.....	27
Amostragem aleatória estratificada e proporcional.....	27
Amostragem por conveniência.....	29
Amostragem de voluntários.....	29
CAPÍTULO 4.....	31
Quais os dados necessários para utilizarmos a Bioestatística?.....	32
Quais as técnicas mais aplicadas nas primeiras etapas de descrição de dados quantitativos?.....	34
Organização dos dados.....	35
Cálculo de freqüências.....	36
CAPÍTULO 5.....	41
O que são medidas de tendência central e quais as suas aplicações?.....	42
Moda.....	42
Média aritmética.....	44
Média ponderada.....	46
Mediana.....	47
Em quais circunstâncias deveremos usar a moda, a média ou a mediana?.....	52

CAPÍTULO 6.....	57
O que são medidas de dispersão e quais as suas aplicações?.....	58
Amplitude.....	59
Desvio médio.....	60
Variância.....	62
Desvio-padrão.....	65
Coeficiente de variação.....	66
CAPÍTULO 7.....	69
Quais as principais medidas de posição?.....	70
Porcentil.....	71
Quartil.....	72
Como os quartis são calculados?.....	73
Quais as principais aplicações dos percentis?.....	78
Amplitude interquartil.....	80
CAPÍTULO 8.....	85
Quadro.....	86
Tabela.....	86
Gráfico.....	93
Cartograma.....	93
Diagrama.....	93
De setores.....	93
De barras.....	95
De barras de erro.....	98
Histograma.....	100
Polígono de freqüências.....	102
De talo e folha.....	104
De pontos.....	105
De linhas.....	105
De dispersão.....	107
De caixa.....	109
De linhas de afastamento.....	113
CAPÍTULO 9.....	117
O que são distribuições de freqüências e distribuições probabilísticas e quais as suas aplicações?.....	118
Distribuição Binomial.....	122
Distribuição de Poisson.....	122
Distribuições reais.....	124
Distribuição normal.....	125
Definição estatística de normalidade.....	125
Outros critérios para definição de normalidade.....	127
Propriedades matemáticas da distribuição normal.....	127
Distribuição normal padrão.....	130
Obtenção de áreas sob a curva normal padrão.....	133

CAPÍTULO 10.....	139
PRIMEIRA PARTE.....	140
Por que precisamos fazer inferência estatística?.....	140
O que é afinal inferência estatística?.....	142
O que é inferência não-estatística?.....	142
Como se distribuem as freqüências dos resultados de diferentes amostras?.....	143
Erro-padrão.....	147
Teorema central do limite.....	149
SEGUNDA PARTE.....	150
Como a inferência estatística é feita?.....	150
Teste de hipóteses estatísticas / Inferência sobre uma média / Teste z	150
Erros envolvidos na inferência estatística.....	167
TERCEIRA PARTE.....	177
O que é um intervalo de confiança?.....	177
CAPÍTULO 11.....	185
Quando devemos aplicar o teste z ou o t ?.....	186
Como realizamos o teste t ?.....	193
Cálculo de intervalo de confiança usando valor de T	195
CAPÍTULO 12.....	201
Quando a inferência estatística é sobre duas médias e não sobre apenas uma?.....	202
Teste z	206
Cálculo de intervalo de confiança utilizando a distribuição normal padrão.....	210
Teste da razão de variâncias.....	212
Teste t	217
Cálculo de intervalo de confiança usando valor de T	222
Teste t'	223
Cálculo de intervalo de confiança usando valor de T'	226
CAPÍTULO 13.....	231
Qual o teste a ser aplicado quando as amostras não forem independentes?.....	232
Teste t para amostras não independentes.....	233
Cálculo de intervalo de confiança para amostras não independentes.....	239

CAPÍTULO 14.....	243
E se estivermos comparando proporções e não médias?.....	244
Por que podemos usar o teste z também para inferência sobre proporções?.....	244
Como fazemos inferência sobre uma proporção utilizando o teste z ?.....	247
Como fazemos inferência sobre duas proporções utilizando o teste z ?.....	252
CAPÍTULO 15.....	261
Quais os fundamentos estatísticos para os cálculos do tamanho da amostra?.....	262
Como calculamos o tamanho da amostra para estimar uma média?.....	262
Como calculamos o tamanho da amostra para estimar uma proporção?.....	269
CAPÍTULO 16.....	273
Qual a aplicação mais comum do teste qui-quadrado? Por que esse teste recebe essa denominação?.....	274
Como realizamos o teste qui-quadrado para avaliar a independência entre variáveis?.....	274
Existem outras aplicações para o teste qui-quadrado?.....	287
CAPÍTULO 17.....	291
Como realizamos o teste exato de Fisher?.....	292
Por que este teste é chamado de exato?.....	294
APÊNDICE 1.....	301
APÊNDICE 2.....	313
APÊNDICE 3.....	317

CAPÍTULO 1

- Quais as diferenças entre Estatística Descritiva, Analítica e Inferencial?
 - Quais as técnicas estatísticas mais utilizadas?
 - O que são contagens?
 - O que são medições?
 - Como escolher a técnica estatística mais adequada a cada situação?
 - O que significam as denominações “estatística paramétrica” e “não-paramétrica”?
 - Como surgiu a Estatística Moderna?
 - O que é Bioestatística?
 - Em que etapas da pesquisa epidemiológica utilizamos a Estatística?
 - Em que etapas e em quais tipos de estudos epidemiológicos utilizamos a Estatística?
 - Como saber qual a técnica estatística mais adequada a cada situação?
 - O que é Estatística Bayesiana?
 - É possível gerar conhecimento científico sem a Estatística?
-



— **Quais as diferenças entre Estatística Descritiva, Analítica e Inferencial?**

— No âmbito deste livro dividiremos a Estatística em três partes, de acordo com a finalidade de cada uma. Se o seu objetivo for descrever quantitativamente uma determinada realidade você deverá utilizar as técnicas da **Estatística Descritiva**. Se quiser analisar quantitativamente essa realidade, ou seja, investigar as relações entre os fatores descritos, usará os procedimentos da **Estatística Analítica**. Mas, se o seu objetivo for inferir, isto é, avaliar se os resultados obtidos em uma amostra aleatória podem ser generalizados para a população da qual a amostra foi retirada, utilizará as técnicas da **Estatística Inferencial**.

A Estatística pode ser dividida em três partes:		
Estatística Descritiva	Descreve	Caracterização dos indivíduos estudados
Estatística Analítica	Analisa	Investigação das relações entre as características estudadas
Estatística Inferencial	Inferre	Avaliação da possibilidade de generalização

Se essa divisão da Estatística ainda não ficou clara para você, tenha paciência e aguarde um pouco, porque com certeza isso ficará mais claro ao longo deste livro.

— **Mas, se essas partes da Estatística são mesmo diferentes, de que consiste afinal a Estatística Analítica? Esta não é a mesma Estatística Inferencial que usa os famosos testes de significância estatística?**

— Muitas vezes a Estatística Analítica e a Estatística Inferencial são consideradas como uma só modalidade, mas achamos essa equiparação inadequada, pois podemos utilizar as técnicas da primeira sem o uso de procedimentos inferenciais, e vice-versa. A primeira situação ocorrerá, p. ex., quando estivermos considerando dados obtidos de toda a população ou de amostras não-aleatórias. Nessas situações não faz sentido avaliarmos se o resultado obtido é estatisticamente significativo (Estatística Inferencial), mas seria inteiramente necessário utilizarmos indicadores quantitativos para a análise desses dados (Estatística Analítica). Mais adiante, no capítulo 10 (páginas 142 e 143), explicaremos o por quê de não fazer sentido aplicarmos testes de significância estatística quando investigamos toda uma população ou amostra não-aleatória dessa população.

Outra maneira de lhe responder é com um exemplo:

Suponha que você esteja realizando um **estudo transversal**¹ para investigar uma possível associação

¹ Estudo transversal: estudo epidemiológico no qual as informações sobre a(s) exposição(ões) de interesse e sobre a(s) doença(s) estudada(s) são coletadas simultaneamente, de modo a obtermos a situação de saúde existente em um certo

entre dieta e câncer da boca. Na descrição (caracterização) dos indivíduos estudados você utilizaria procedimentos da Estatística Descritiva, verificando quantos são homens ou mulheres, quantos negros ou brancos, etc.; na avaliação da existência, direção e magnitude da associação de interesse, lançaria mão das técnicas da Estatística Analítica, comparando, p. ex., a proporção de doentes em indivíduos com um tipo de dieta à proporção de doentes naqueles com outro tipo de dieta e, supondo que o seu estudo tenha sido realizado em uma amostra e que o método de amostragem tenha sido aleatório, aplicaria os testes de significância da Estatística Inferencial. Estes testes serviriam para verificar se os resultados encontrados no estudo realizado seriam válidos para representar os verdadeiros resultados da população de onde a única amostra que você estudou foi retirada. Informações populacionais quase sempre não são conhecidas, pois demandam muito tempo e trabalho, sendo muito caro obtê-las. Geralmente, então, estimamos informações populacionais com base em resultados obtidos em uma ou em poucas amostras, através dos procedimentos da Estatística Inferencial.

Assim, poderíamos descrever os indivíduos estudados segundo o sexo, a raça, o estado civil, etc., analisar a associação entre dieta e câncer da boca calculando a prevalência deste câncer nos indivíduos que consomem uma determinada dieta e comparando-a à prevalência deste mesmo câncer naqueles que consomem um outro tipo de dieta, obtendo, p. ex., uma razão entre estas prevalências (*RP*), e poderíamos também aplicar um teste de significância estatística (nesse caso o teste qui-quadrado, que será abordado no capítulo 16) ou calcular um intervalo de confiança (capítulo 10), para avaliarmos se seria possível inferir para a população inteira os resultados obtidos na única amostra ou nas poucas amostras retiradas dessa população.

— Quais as técnicas estatísticas mais utilizadas?

— Antes de lhe respondermos, será importante destacarmos que os dados a partir dos quais toda a Estatística é produzida consistem de contagens e/ou medições.

— Contagens, medições?

— Sim. Contagens, como sua denominação indica, são números que resultam de contagens feitas pelos estatísticos nos indivíduos estudados. Podemos contar quantos eram do sexo masculino ou do feminino, da raça negra ou branca, etc. Essas contagens nos permitirão descrever, analisar e/ou inferir, a depender dos objetivos da nossa pesquisa.

As medições, como também sua denominação indica, são medidas de interesse para o estudo, feitas nos indivíduos estudados. Medidas da altura, da glicemia, da concentração de chumbo no sangue, etc., são exemplos desse tipo de informação quantitativa. Essas medições também nos permitirão descrever, analisar e/ou inferir. No capítulo 4 (páginas 32 a 34) explicaremos mais detalhadamente as contagens e medições.

Agora, vamos listar abaixo as principais técnicas das Estatísticas Descritiva, Analítica e Inferencial,

momento em uma determinada população. Se desejar revise esse tema em: *Medronho RA, Carvalho DM, Bloch KV, Luiz RR, Werneck GL, editores. Epidemiologia. São Paulo (SP): Atheneu; 2002.*

com o intuito, por enquanto, de lhe dar uma idéia dos procedimentos que poderão ser utilizados em suas pesquisas, destacando os que serão abordados neste livro e que são de aplicação mais constante. Não se preocupe com o grande número de técnicas que verá, pois, nos esforçaremos para explicá-las da forma mais clara possível. Não se preocupe também com o grande número de técnicas que não verá neste livro. Muitas dessas não serão necessárias em sua vida profissional, e outras, você aprenderá ao longo de outros níveis de formação, como Cursos de Especialização, Mestrado, Doutorado e Pós-doutorado, ou quando for necessário aplicá-las, durante a realização de suas pesquisas.

TÉCNICAS MAIS UTILIZADAS NA ESTATÍSTICA DESCRITIVA
• Cálculo de frequências simples, simples acumulada, relativa e relativa acumulada
• Cálculo de medidas de tendência central (moda, média aritmética, média ponderada, mediana)
• Cálculo de medidas de dispersão (amplitude, desvio médio, variância, desvio-padrão, coeficiente de variação)
• Cálculo de medidas de posição (porcentis)
• Elaboração de tabelas univariáveis (veja a definição de variável na página 14)
• Elaboração de gráficos (cartograma, histograma, diagrama de talo e folha, diagrama de caixa, diagrama de setores, diagrama de barras, etc.)
• Avaliação da forma como as frequências de uma variável se distribuem

Se você quiser detalhar mais e/ou tornar sua descrição mais robusta (mais fidedigna, mais válida), poderá utilizar um conjunto de procedimentos denominados **análise exploratória de dados**. Essas técnicas são abordadas em livros específicos e não serão abordadas neste livro. A análise exploratória de dados também inclui a elaboração de diagramas. Apenas dois deles, o diagrama de talo e folha, e o de caixa serão aqui apresentados. Se você estiver interessado na análise exploratória de dados sugerimos que estude esse tema nos livros *Exploratory data analysis*, de John W. Tukey, Reading (MA): Addison-Wesley; 1976 e *Understanding robust and exploratory data analysis*, de David C. Hoaglin, Frederick Mosteller e John W. Tukey, editores, New York (NY): John Wiley; 1983.

Depois de descrever os indivíduos estudados, se você também tiver o objetivo de analisar seus resultados, poderá aplicar algumas das técnicas relacionadas abaixo:

TÉCNICAS MAIS UTILIZADAS NA ESTATÍSTICA ANALÍTICA
• Elaboração de diagramas (os mesmos da Estatística Descritiva, porém considerando mais de uma variável; diagrama de dispersão, p. ex.)
• Elaboração de tabelas de contingência bivariáveis (com duas variáveis) ou multivariáveis (com mais de duas variáveis)
• Cálculo de medidas de associação entre variáveis (razão ou diferença entre prevalências; entre incidências ou risco relativo ou atribuível; entre chances; coeficientes de correlação (de Pearson, de Spearman, parcial, parcial múltiplo, etc.); coeficientes de regressão)
• Análise estratificada
• Análise multivariável

Em cada uma das técnicas estatísticas acima mencionadas calculamos um ou mais indicadores quantitativos que nos ajudam a avaliar como e com que força duas ou mais variáveis estão associadas. Esses indicadores constituem os procedimentos da Estatística Analítica. Em seguida, são feitos testes apropriados de significância estatística (que já são procedimentos da Estatística Inferencial), para verificar se os valores obtidos para as estatísticas descritivas ou analíticas no estudo realizado são válidos para a população.

Veja uma listagem das principais técnicas de inferência estatística no quadro abaixo:

TÉCNICAS MAIS UTILIZADAS NA ESTATÍSTICA INFERENCIAL	
• Teste z para uma ou duas médias	• Cálculo do índice capa (Teste z)
• Teste t para uma ou duas médias	• Análise de regressão linear (Teste F ou Teste z)
• Teste t para amostras emparelhadas	• Teste exato de Fisher
• Teste z para uma ou duas proporções	• Teste do sinal
• Teste qui-quadrado para duas ou mais proporções	• Teste de Wilcoxon
• Teste qui-quadrado de Mantel e Haenszel	• Teste da mediana
• Teste para uma variância	• Teste de Mann-Whitney
• Teste F para duas variâncias	• Teste de Kruskal-Wallis
• Análise de variância (Teste F)	• Teste de Friedman
• Análise de correlação intraclass (Teste F)	• Análise de correlação de Spearman
• Análise de correlação de Pearson (Teste t)	• Teste de McNemar
• Cálculo do alfa de Cronbach (Teste F)	• Elaboração de diagrama de barra de erro

Várias das técnicas mencionadas serão explicadas neste livro e, como já foi prometido, faremos o maior esforço possível para que você compreenda cada uma e seja capaz de utilizá-las facilmente quando precisar.

Na descrição dos resultados consideramos apenas uma variável de cada vez. Na análise estatística temos que utilizar duas variáveis (análise bivariável) ou mais de duas (análise estratificada e análise multivariável). A inferência estatística é realizada tanto para uma variável isoladamente, quanto para duas ou mais.

Note no quadro acima que a elaboração de um diagrama foi citada como técnica de inferência estatística. Do mesmo modo que os diagramas podem ser usados na descrição e análise quantitativa de dados, alguns podem também ser utilizados para inferência. Isto será explicado no capítulo sobre elaboração de diagramas (capítulo 8, páginas 98 a 100).

O quadro apresentado na próxima página mostra as técnicas que não serão abordadas neste livro.

TÉCNICAS ESTATÍSTICAS NÃO ABORDADAS
• A grande maioria das técnicas da “análise exploratória de dados”
• Cálculo de medidas de associação (risco relativo, razão de chances, etc.)
• Cálculo do índice de concordância Capa
• Cálculo do alfa de Cronbach
• Teste qui-quadrado de Mantel e Haenszel
• Teste para uma variância
• Análise de variância / Análise de correlação intraclasse
• Teste do sinal
• Teste de Wilcoxon
• Teste da mediana
• Teste de Mann-Whitney
• Teste de Kruskal-Wallis
• Teste de Friedman
• Teste de McNemar
• Análise de correlação de Spearman
• Análise de correlação de Pearson
• Análise de regressão linear
• Análise de regressão logística
• Análise de regressão de Cox
• Análise de regressão de Weibull
• Análise de regressão de Poisson
• Análise de regressão binomial negativa
• Análise de regressão log-linear
• Análise de regressão hierárquica
• Análise discriminante
• Análise de variância multinomial (MANOVA)
• Análise de correlação de Kendall
• Análise de contingência
• Análise de correlação canônica
• Análise de correlação parcial múltipla
• Análise de escala multidimensional
• Análise de componentes principais
• Análise de fator
• Análise de correspondência
• Análise de homogeneidade
• Análise de agrupamento (“cluster analysis”)
• Análise por redes neurais artificiais

Existem ainda outras técnicas estatísticas que, por serem menos utilizadas, não foram mencionadas no quadro acima. Quando for necessário você poderá estudá-las e aplicá-las.

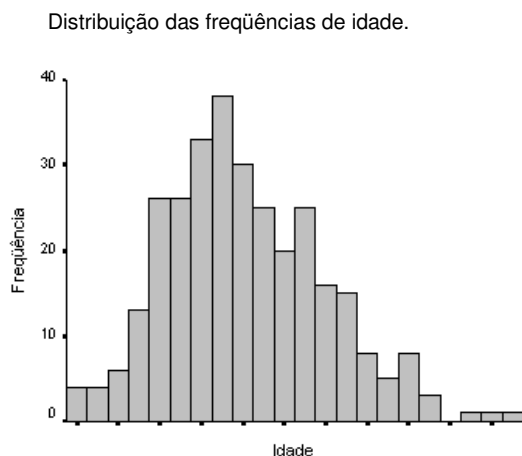
— Mas, há uma “montanha” de técnicas que não serão abordadas! Vou continuar sabendo poucas técnicas estatísticas!

— Tenha paciência! Este livro foi “bolado” para servir como livro-texto em cursos de Estatística Básica. É um primeiro degrau. É impossível aprender todas as técnicas existentes em um espaço de tempo curto. Essa é uma tarefa para ser feita ao longo de toda sua vida profissional. Além disto, não se esqueça de que você não precisará utilizar todas essas técnicas em suas pesquisas. O importante neste momento é que você se capacite a empregar as técnicas mais simples e de uso mais freqüente.

— O que as denominações estatística paramétrica e não-paramétrica significam?

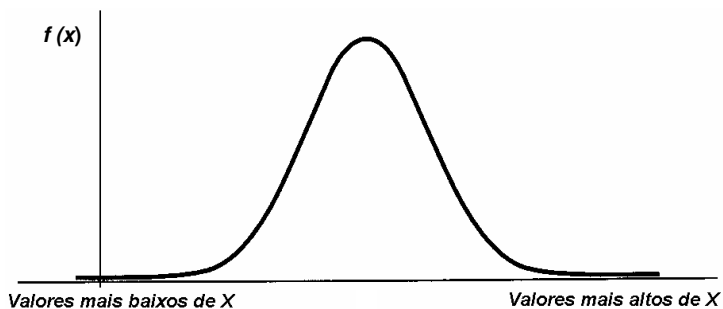
— Boa pergunta! Para você entender isso será necessário explicarmos o que é uma **distribuição de freqüências** ou uma **distribuição de probabilidades**.

Se estudarmos uma amostra, que é uma parte de uma população, e coletarmos, para cada indivíduo dessa amostra, informação sobre uma determinada característica, a idade, p.ex., podemos contar quantas vezes cada valor de idade apareceu nessa amostra. Para avaliarmos como essas freqüências de valores de idade se distribuíram nessa amostra, elaboramos um diagrama de freqüências, também chamado de distribuição de freqüências. Na ordenada dessa distribuição apresentamos as freqüências com que os valores de idade ocorreram naquela amostra e na abscissa os valores de idade. Muitos dos fenômenos estudados por nós na área biomédica apresentam baixas freqüências dos valores mais baixos e também dos mais altos, e altas freqüências dos valores mais intermediários, conformando um diagrama de distribuição de freqüências semelhante ao desenhado abaixo:



As partes mais altas da distribuição representam os valores mais freqüentes e as partes mais baixas os valores menos freqüentes, já que na ordenada representamos as freqüências. Quanto mais alta a coluna, mais freqüentes os valores correspondentes de idade contidos na abscissa, e vice-versa.

Com base na teoria estatística, se o número de indivíduos tender para infinito, será usado um modelo ou equação matemática para representar essa distribuição, como mostramos a seguir:



Nesse diagrama, na abscissa estão representados os diversos infinitos valores de uma característica de interesse, a idade, p.ex., denotada por X , e na ordenada valores de uma função matemática de X . Você

verá no capítulo 9, que a distribuição acima recebe a denominação de **distribuição normal**, e que as áreas entre essa curva e a abscissa equivalem às probabilidades dos valores de idade ocorrerem naquela população infinita. Por isso, esta e outras distribuições recebem também a denominação genérica de **distribuições de probabilidades**. No capítulo 10 (páginas 143 a 150), explicaremos que, muitas vezes, vamos poder assumir que a distribuição de uma determinada característica na população estudada é normal. Sendo assim, poderemos utilizar essa distribuição como modelo para verificar se os resultados obtidos em uma parte (amostra) dessa população são estatisticamente iguais ou diferentes dos valores que porventura obteríamos, se tivéssemos estudado toda a população e não apenas uma parte da mesma. Por enquanto, nossa intenção é destacar que podemos utilizar distribuições de probabilidades para fazermos inferência estatística.

Quando, ao fazermos essa inferência, tivermos de assumir, na população de onde o grupo investigado foi retirado, que a característica estudada tem uma determinada distribuição de probabilidades previamente conhecida, classificaremos o procedimento estatístico como **paramétrico**, porque utilizaremos os parâmetros dessa distribuição já conhecida. Quando, p. ex., pudermos assumir previamente que a distribuição na população é do tipo normal, a média dessa distribuição, μ , será um dos parâmetros considerados para realizarmos inferência estatística.

Quando não necessitarmos assumir previamente um determinado formato da distribuição na população para realizarmos o procedimento estatístico, este será denominado **não-paramétrico**. Como nenhuma distribuição já conhecida é utilizada, esse tipo de técnica estatística é também chamado de **“livre de distribuição”**.

Outra razão para usarmos um procedimento não-paramétrico é a natureza da característica estudada. Se esta é medida em valores que podem ser postos em ordem crescente ou decrescente, mas seus valores não compreendem todos os possíveis valores em uma escala quantitativa contínua, não poderemos calcular sua média e, conseqüentemente, a distribuição normal não poderá ser utilizada para fazermos inferência sobre essa característica. Isto será explicado nas páginas 54 e 55.

— Como surgiu a Estatística Moderna?

— Fazendo uma abordagem bem sucinta, podemos começar destacando que, desde os seus primórdios, os seres humanos sentiram a necessidade de e, efetivamente, fizeram contagens e medições. E, à medida que a matemática se desenvolveu, a quantificação de eventos de interesse foi também evoluindo.

Mas, a Estatística Moderna é relativamente recente. Surgiu na transição entre o feudalismo e o capitalismo, em um período denominado de mercantilismo. Durante a criação dos Estados Absolutistas na Europa, intensificou-se a necessidade de se saber quantos indivíduos nasciam ou morriam, quantos eram sadios ou doentes.

— Por quê?

— Porque naquele período eram freqüentes as guerras de conquista e, conseqüentemente, as de defesa de território, e então, foi se tornando cada vez mais necessário saber-se mais precisamente com quantas pessoas o Estado poderia contar para conquistar novos territórios ou para defender-se de agressores. A denominação Estatística é, por isso, derivada da palavra “Estado” e abrangia, originalmente, o conhecimento resultante de contagens e/ou medições de eventos de interesse do Estado.

Se quiser ler sobre a história da Estatística, recomendamos o livro de *Stigler SM. The history of Statistics. The measurement of uncertainty before 1900. Cambridge (MA): The Belknap Press of Harvard University Press; 1986.*

— **E o que é Bioestatística?**

— Como você já sabe, chamamos de Bioestatística o ramo da Estatística aplicado ao agrupamento metódico e ao estudo de fenômenos biológicos passíveis de avaliação quantitativa.

— **Utilizaremos a Bioestatística apenas na descrição, análise e inferência dos resultados obtidos no nosso estudo?**

— Não. Se considerarmos resumidamente as etapas de uma pesquisa epidemiológica, veremos que a Estatística pode ser usada na maioria das mesmas. Em termos gerais, uma investigação desta natureza comporta as seguintes etapas:

- Definição do tema
- Planejamento do estudo
- Coleta dos dados
- Digitação e processamento
- Descrição, análise e interpretação dos resultados
- Avaliação crítica do estudo
- Redação
- Apresentação / Divulgação

A Estatística será utilizada em todas essas etapas, e em algumas terá um papel indispensável. Na definição do tema, ajudando-nos a avaliar onde existem lacunas no conhecimento devido a falhas na análise dos dados de estudos realizados anteriormente; no planejamento, orientando-nos na seleção dos indivíduos e fatores a serem estudados, na escolha das técnicas adequadas à descrição e análise desses fatores e à generalização dos resultados; esse planejamento será importante para uma realização correta do estudo, estando aí incluídos a coleta, a digitação e o processamento dos dados; na descrição, análise e interpretação, orientando-nos no uso correto das técnicas estatísticas escolhidas durante o planejamento; na avaliação crítica do estudo, fornecendo-nos elementos para um melhor julgamento sobre nosso próprio trabalho, permitindo-nos identificar aspectos positivos e negativos do mesmo, verificando sua validade científica; e na redação, apresentação e divulgação dos resultados, auxiliando-nos com procedimentos práticos, como a elaboração de tabelas e diagramas.

— **Em que tipos de estudos epidemiológicos poderemos usar técnicas estatísticas?**

— Relembrando, listamos abaixo os sete tipos básicos de estudos epidemiológicos:

- | | |
|------------------|-----------------|
| • De prevalência | • Caso-controle |
| • De incidência | • De coorte |
| • De agregados | • Experimental |
| • Transversal | |

Em todos esses tipos a Estatística terá uma contribuição fundamental para que o estudo seja bem planejado e produza resultados cientificamente válidos.

– Como saber qual a técnica estatística mais adequada a cada situação?

– Para escolher corretamente a técnica a ser utilizada é fundamental que você leve em conta o tipo de estudo epidemiológico que realizará e a natureza estatística dos fatores a serem investigados. No próximo capítulo veremos como esses fatores, chamados de **variáveis**, são classificados e, ao longo dos demais capítulos, ficará claro qual(is) técnica(s) será(ão) a(s) mais apropriada(s) em função dos tipos de variáveis envolvidas.

– Reconheço o esforço que vocês estão fazendo para motivar-me a estudar bioestatística, mas há algo que não conseguirei superar que é o entendimento de fórmulas matemáticas.

– Foi bom você ter mencionado isso. Existem fórmulas matemáticas mais simples e outras mais complexas. Você com certeza conseguirá entender as mais simples com a ajuda de uma explicação clara e isto nós tentaremos fazer sempre. Quanto às mais complexas, nossa opinião é que simplesmente não é possível olhar para elas e entendê-las.

– Então não somos obrigados a entender fórmulas mais complexas?

– Não. Essas fórmulas resultaram de várias etapas de manipulação algébrica. É praticamente impossível olharmos para uma fórmula complexa e querermos entendê-la. O máximo que um professor poderá exigir de você será entender o desenvolvimento algébrico e/ou a demonstração empírica dessas fórmulas. Sempre que julgarmos necessário, faremos essas demonstrações ao longo deste livro.

Achamos que muitas pessoas não gostam de Matemática e, portanto, de Estatística, porque se sentem na obrigação de olhar para uma fórmula complexa e entendê-la. Isto é um completo absurdo. Não se cobre isto no decorrer deste livro. Combinado?

– Ouvi falar em uma Estatística Bayesiana. O que isso significa?

– Alguns estatísticos propuseram uma outra maneira de se fazer inferência estatística, diferente da que apresentaremos neste livro. Esta outra maneira é chamada de bayesiana porque utiliza o famoso teorema de Bayes, assim denominado em homenagem ao seu formulador, o matemático e religioso inglês Thomas Bayes (1702-1761).

Se desejar saber como a inferência bayesiana é realizada, sugiro que estude as páginas 20-22, 27, 197-199, 220, 221, e 336 do livro *Rothman KJ e Greenland S, editores. Modern epidemiology. 2ª ed. Philadelphia (PA): Lippincott Williams e Wilkins; 1998.*

Como em qualquer área do pensamento humano, esses dois distintos métodos de inferência, o da estatística bayesiana e o da estatística clássica (sendo este o que será utilizado neste livro), despertou intensa discussão, com posições apaixonadas a favor ou contra uma ou outra. Em nossa opinião, não devemos opor um método ao outro, pois ambos se fundamentam em argumentos científicos e estatísticos válidos, embora diferentes. O ideal seria que em cada estudo aplicássemos os dois métodos, porque um desempenharia um papel confirmatório ou não do outro, dando-nos maior certeza estatística sobre nossos

achados. Esta posição está bem defendida no artigo de *Bradley E.: Bayesians, frequentists, and scientists*, que pode ser acessado na seguinte página da “Internet”: www-stat.stanford.edu/~brad/papers/Bay-Freq_2005.pdf.

— É possível pesquisarmos cientificamente sem a Estatística?

— É claro que sim! Não poderíamos concluir este primeiro capítulo sem discutir esse assunto com você.

A Estatística é uma poderosa ferramenta para pesquisas quantitativas em vários campos do conhecimento, e não somente na área de saúde na qual atuamos. Mas isso não quer dizer que essa ferramenta não tenha limitações, algumas intransponíveis a nosso ver. As pesquisas qualitativas têm um papel também importantíssimo e insubstituível. Elas investigam um número menor de indivíduos, mas com uma profundidade, um detalhamento muito maior do que as pesquisas quantitativas que, por sua vez, propiciam o estudo de um número maior de indivíduos, mas de modo extensivo e superficial.

Defendemos a necessidade de que as abordagens quantitativa e qualitativa se complementem. Para nós, um bom estudo epidemiológico deve conter essas duas abordagens porque nenhuma isoladamente consegue dar conta da totalidade que se pretende investigar, deixando lacunas indesejáveis na investigação. Se você se interessou por este tema tão instigante, busque alguns dos vários livros e artigos disponíveis sobre ele na literatura, pois não o abordaremos novamente, para não nos afastarmos dos objetivos deste livro.

Faça uma pausa e curta um pouco a vida antes de retomar a leitura deste livro. Curtir a vida é tão importante quanto trabalhar. Sobre a defesa desse ponto de vista sugerimos as seguintes leituras: a) *Russel B. O elogio ao ócio. Rio de Janeiro (RJ): Sextante; 2002*; b) *Lafargue P. O direito à preguiça*, que pode ser acessado no seguinte sítio da “Internet”: www.ebooksbrasil.org/eLibris/direitopreguica.html; c) *Kurz R. Manifesto contra o trabalho*, que pode ser acessado no sítio: www.dhnet.org.br/desejos/textos/krisis.htm; d) *De Masi D. O ócio criativo. Rio de Janeiro (RJ): Sextante; 2000*; e e) *Sennett R. A corrosão do caráter. 5ª ed. Rio de Janeiro (RJ): Record; 1999*.

CAPÍTULO 2

- O que são variáveis?
 - Como classificar as variáveis?
 - Qual dessas classificações devemos utilizar?
 - Para que todo esse esforço em classificar variáveis?
-



— O que são variáveis?

— Nosso ponto de partida neste capítulo será definir o que é uma **variável**. Depois discutiremos com você as diversas maneiras de classificá-la. Finalmente, abordaremos rapidamente a importância de classificarmos as variáveis utilizadas em nossas pesquisas.

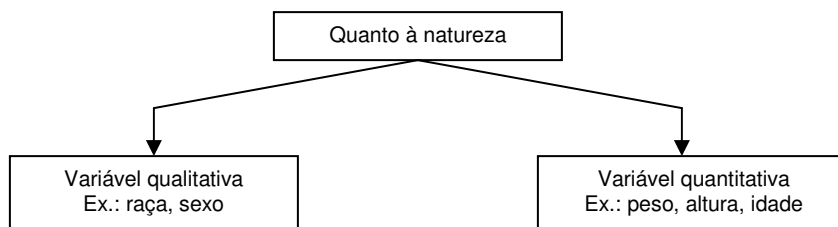
Uma variável, como a denominação já deixa claro, é uma característica que varia entre os indivíduos estudados. A idade, o peso, a altura, o sexo, a raça, entre outras, são características que variam entre os indivíduos a serem estudados; uns são mais jovens, outros mais velhos; mais leves ou mais pesados; mais baixos ou mais altos; homens ou mulheres; brancos, negros, mulatos, índios ou amarelos. Se uma característica não varia em uma determinada população, a rigor não deveria ser chamada de “variável”. Por exemplo, se todos os indivíduos que estivéssemos estudando tivessem a mesma idade não deveríamos considerar a idade como uma “variável” nesse estudo. Na prática, contudo, variáveis que foram neutralizadas ou controladas e, por isso, não variam, continuam sendo chamadas de “variáveis”, porque até o momento ninguém se deu ao trabalho, inclusive nós, de propor uma outra denominação para ser usada nessas situações. Os epidemiologistas já se acostumaram a continuar chamando essas características de “variável”, mesmo quando não variam. Mas, tudo bem! Podemos conviver com tal contradição, porque na prática, você verá que isso não nos atrapalhará.

— Como classificar as variáveis?

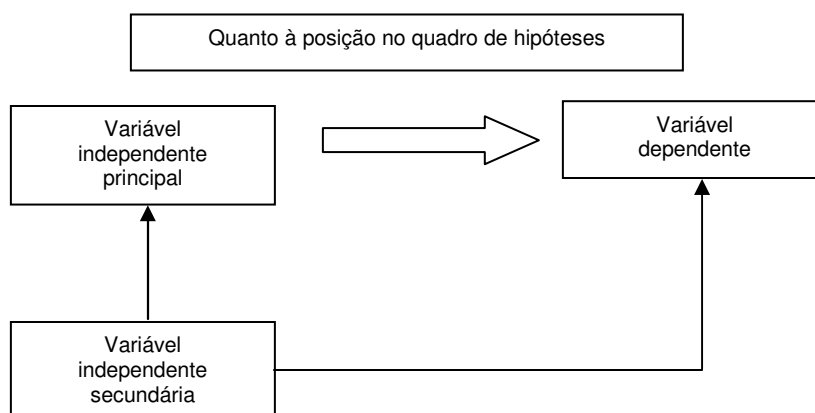
— Reconhecemos que as classificações das variáveis podem ficar muito confusas, mas achamos que isso só ocorre quando os estatísticos não deixam logo claro, desde o início, que existem várias classificações e que cada uma baseia-se em um critério diferente. Assim, uma mesma variável pode ser classificada e, portanto, denominada de diversas maneiras. E é exatamente isso que gera tanta confusão.

Antes de prosseguirmos, é importante você aprender que os valores passíveis de serem assumidos por uma variável são chamados de **categorias** da variável. Sexo (definido biologicamente), p. ex., é uma variável com duas categorias: masculino e feminino; ou homem e mulher. É claro que alguém pode discordar dessa classificação, arguindo que essas duas categorias não incluem todo o espectro de variação dessa variável. Essa crítica é procedente, sendo mais adequado mantermos a variável “sexo biológico” como foi definida acima, e utilizarmos uma outra que poderia ser chamada de “orientação sexual” para englobar um amplo leque de opções.

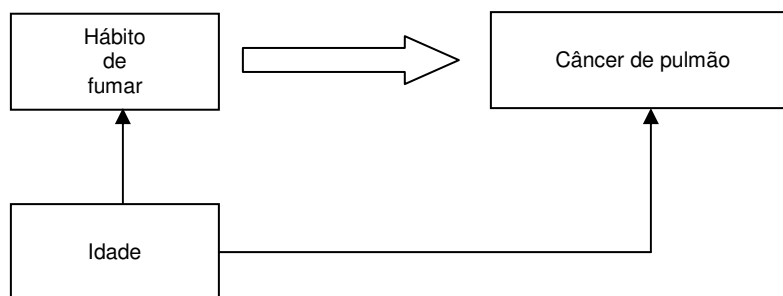
O primeiro critério de classificação leva em conta a **natureza qualitativa** (indicando uma qualidade) ou **quantitativa** (indicando uma quantidade) da variável. Por esse critério, obviamente, podemos classificar uma variável como qualitativa ou quantitativa. A variável “raça” é um exemplo de variável qualitativa. Cada uma de suas categorias (negro, mulato, branco, índio ou amarelo) indica um indivíduo com qualidades (características) diferentes dos demais. Outro exemplo, entre muitos, é o de uma variável que indique se o indivíduo está ou não doente. Já a variável “peso”, é quantitativa, pois suas categorias indicam a quantidade de peso (em kg) para cada indivíduo. Outros exemplos desse último tipo são as variáveis “altura” e “idade”.



Outro critério utilizado é a posição da variável no quadro de hipóteses da pesquisa. Por tal critério uma variável pode ser classificada em dependente ou independente. O primeiro tipo indica o efeito, a resposta, o desfecho que está sendo estudado, sendo na maior parte das vezes uma doença. O segundo representa um possível determinante (causa ou fator associado) ao efeito estudado. As variáveis independentes podem ser subdivididas em independente principal (ou de estudo, ou de interesse, ou causal), ou independente secundária (ou covariável, ou variável de controle ou confundidora). Veja a figura abaixo:

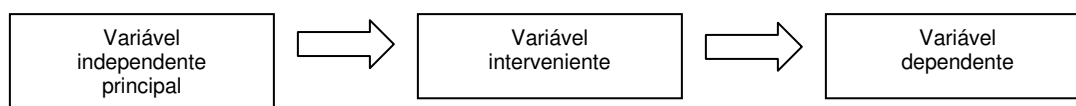


Em um estudo para investigar se o hábito de fumar provoca câncer de pulmão, a variável dependente seria o câncer de pulmão porque a hipótese da pesquisa assumiria que a ocorrência deste câncer dependeria do hábito de fumar, e este hábito seria a variável independente porque, conforme o quadro teórico da pesquisa, sua ocorrência não dependeria (seria independente) da presença do câncer. A variável independente "hábito de fumar" seria denominada de principal porque neste exemplo seria a variável na qual os investigadores estariam especificamente interessados. Logo, a associação entre hábito de fumar e câncer de pulmão seria considerada como a associação principal do estudo. Os pesquisadores também deveriam investigar a influência das variáveis independentes secundárias, de modo a neutralizar o efeito dessas sobre a associação principal estudada. No exemplo dado, teríamos de neutralizar, p. ex., o efeito da variável "idade", porque para estarmos mais seguros de que havia uma associação entre hábito de fumar e câncer de pulmão, seria necessário afastarmos a possibilidade de que essa associação resultasse apenas do fato dos indivíduos mais idosos fumarem mais e, ao mesmo tempo, por sua idade avançada, estarem mais sujeitos a apresentar câncer de pulmão em decorrência de fenômenos degenerativos (que se acentuam com a idade), e não por efeito de substâncias cancerígenas presentes no cigarro. Assim, os fumantes pareceriam ter um risco maior de câncer de pulmão apenas porque eram mais idosos do que os não-fumantes. É justamente pela possibilidade de confundirem a associação principal estudada, que se torna indispensável a neutralização de variáveis independentes secundárias. Veja a figura a seguir:

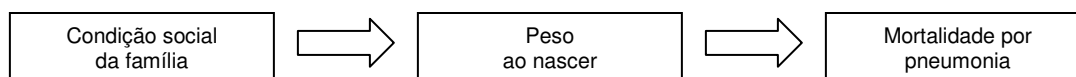


Não explicaremos conceitual nem tecnicamente em maior detalhe esse fenômeno da “confusão” ou “confundimento” em estudos epidemiológicos, pois isso nos afastaria dos objetivos propostos neste livro. Se precisar desse aprofundamento, sugerimos que consulte outras fontes (Pereira MG. *Epidemiologia: teoria e prática*. Rio de Janeiro (RJ): Guanabara Koogan; 1995; Kleinbaum DG, Kupper LL, Morgenstern H. *Epidemiologic Research*. Belmont (CA): Lifetime Learning; 1982; Hothman KJ, Greenland S. *Modern epidemiology*, editores. 2ª ed. Philadelphia (PA): Lippincott Williams e Wilkins; 1998; e Hennekens CH, Buring JE. *Epidemiology in Medicine*. Boston (MA): Little, Brown; 1987).

Outro tipo de variável independente é a interveniente (ou intermediária), assim denominada porque se encontra no caminho causal entre a variável independente principal e a variável dependente. Veja a figura abaixo:



Um exemplo de variável interveniente pode ser encontrado no artigo: Niobey FML, Duchiade MP, Vasconcelos AGG, Carvalho ML, Leal MC, Valente JG. *Fatores de risco para morte por pneumonia em menores de um ano em uma região metropolitana do sudeste do Brasil. Um estudo tipo caso-controle. Rev Saúde Pública* 1992 Ago;26(4):229-38. Os autores sugerem que a condição social da família determina o peso ao nascer, que por sua vez influencia a mortalidade por pneumonia em menores de um ano. Assim, para eles, é através da ocorrência de baixo peso que a condição social se associa a uma maior mortalidade por pneumonia, como apresentado na figura abaixo:



É importante identificarmos as variáveis intervenientes, de modo a evitar que essas sejam neutralizadas, pois, sendo intervenientes, ao neutralizá-las os pesquisadores estariam cometendo uma falha analítica grave, já que sua neutralização anularia também o efeito da variável independente principal.

— Por quê?

— Ora! Se uma variável é interveniente, seu surgimento decorre da ação da variável independente principal, e esta, por sua vez, só poderá exercer sua influência sobre a variável dependente através da variável interveniente. Se esta última, portanto, for neutralizada, o efeito da variável independente principal também será anulado. E a última coisa que você pode desejar que aconteça em seu estudo é que sua

variável independente principal seja neutralizada, porque é preciso que essa variável varie livremente nos grupos que estão sendo comparados para que possamos analisar os resultados e chegar a alguma conclusão a respeito da influência dessa variável sobre a variável dependente do estudo. No exemplo acima, se neutralizássemos a influência do peso ao nascer, estaríamos também anulando o efeito da condição social sobre a mortalidade por pneumonias, o que comprometeria completamente o estudo.

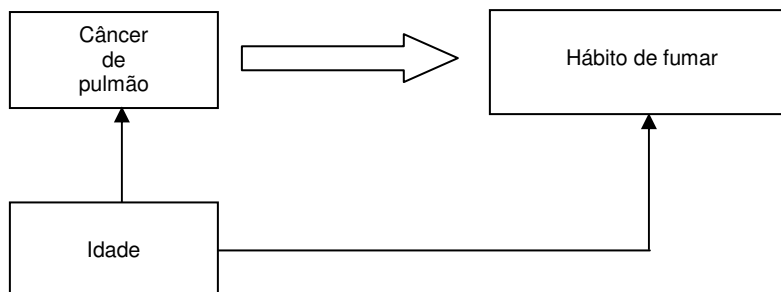
Você deve também estar atento(a) ao fato de que uma variável independente secundária pode modificar o efeito de uma variável independente principal sobre a variável dependente estudada. Vimos anteriormente que uma variável pode confundir uma associação entre outras duas variáveis. Agora estamos vendo que uma variável pode modificar essa associação. Um dos métodos estatísticos mais simples para verificação tanto da existência de confusão como de modificação de efeito (também chamada de interação entre variáveis) é denominado “análise estratificada”. Você pode estudar esse método nos livros já sugeridos.

Se você estivesse estudando a associação entre dieta rica em carnes e verduras frescas e câncer da boca ou orofaringe, e encontrasse associações estatisticamente diferentes entre essas variáveis ao analisar separadamente os indivíduos que consumiam e os que não consumiam freqüentemente bebidas alcoólicas, isso deveria ser considerado por você como evidência de existência de interação entre o consumo de bebidas alcoólicas e dieta rica em carnes e verduras frescas. Haveria interação entre consumo de bebidas alcoólicas e dieta, porque consumir ou não essas bebidas alteraria (modificaria) a associação (o efeito) da dieta sobre aqueles cânceres. Por isso, a “interação” entre variáveis é também chamada de “modificação de efeito”.

A interação pode ser positiva (também chamada de sinergismo) quando a presença de uma aumenta o efeito da outra, ou negativa (também chamada de antagonismo) quando a presença de uma diminui o efeito da outra.

Além das leituras já sugeridas, se estiver interessado(a) nos assuntos “confundimento” e “interação”, pode também estudar as páginas 618 a 623 do livro *Daniel WW. Biostatistics: a foundation for analysis in the health sciences. 7ª ed. New York (NY): John Wiley e Sons; 1999* e/ou as páginas 591 a 605 do livro *Rosner B. Fundamentals of Biostatistics. 5ª ed. Pacific Grove (CA): Duxbury; 2000*.

Outro aspecto importante é que uma variável em um determinado estudo pode ser independente principal, mas em outro, pode ser dependente. Um pesquisador pode, p. ex., estar interessado em investigar se o fato de um indivíduo saber que tem câncer de pulmão em um estágio avançado aumenta a probabilidade dele adquirir o hábito de fumar, devido às tensões psicológicas decorrentes da situação difícil em que se encontra. Nesse estudo, a variável independente principal seria o câncer de pulmão e o hábito de fumar a variável dependente. Observe a seguir que agora as duas variáveis estão em posições completamente diferentes daquelas que ocupavam no exemplo mencionado anteriormente:



Segundo o quadro de hipóteses atual, hábito de fumar é a variável dependente, câncer de pulmão passou a ser a variável independente principal e a idade continua sendo uma variável independente secundária ou covariável.

Há outras denominações para variáveis considerando-se o seu lugar no quadro de hipóteses do estudo, mas detalhar melhor esse tema foge aos objetivos deste livro. Se estiver interessado em aprofundar o tópico, sugerimos que procure outra fonte (*Forattini OP. Epidemiologia Geral. 2ª ed. São Paulo (SP): Artes Médicas; 1996*).

Veja a seguir um resumo dos tipos de variáveis, segundo sua posição no quadro de hipóteses da pesquisa:

CLASSIFICAÇÃO DAS VARIÁVEIS SEGUNDO SUA POSIÇÃO NO QUADRO DE HIPÓTESES DA PESQUISA		
Dependente	Supõe-se que sua ocorrência depende da influência das variáveis independentes	
Independente	Principal (de estudo, de interesse, causal)	É (ou são) a(s) variável(is) de interesse do estudo
	Secundária (covariável, confundidora ou de interação, a ser neutralizada ou controlada)	É (ou são) a(s) variável(is) que pode(m) influenciar a associação principal do estudo
Interveniente	É (ou são) a(s) variável(is) que se encontram no caminho causal entre a variável independente principal e a variável dependente do estudo; Não devem ser neutralizadas	

Você poderá classificar variáveis também levando em conta se estão medidas em um espectro de valores contínuos ou não. Por esse critério, as variáveis são denominadas de contínuas ou discretas. Poderiam tê-las denominado como contínuas ou descontínuas, mas se os estatísticos podem complicar para que simplificar? As variáveis “peso”, “altura”, “idade”, “nível de glicemia”, são exemplos de variáveis contínuas porque os valores que podem ocorrer para essas variáveis variam em uma escala contínua. Portanto, não há intervalos, saltos, entre os possíveis valores dessas variáveis. A idade, p. ex., pode ser medida em anos, meses, semanas, dias, horas, minutos, segundos, de modo a praticamente não haver intervalo entre um valor possível e outro dessa variável.

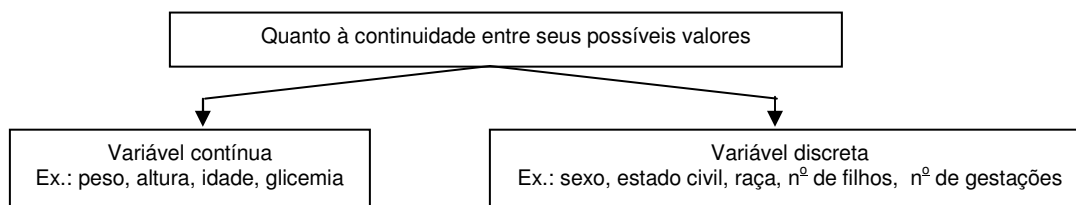
— **Vocês não classificaram anteriormente as variáveis “peso”, “altura” e “idade” como variáveis quantitativas? E a idade também já não foi chamada de variável independente secundária ou covariável? Como é que agora a idade é classificada também como variável contínua?**

— É exatamente isso! Uma mesma variável pode ser denominada por várias maneiras diferentes, porque existem diversos critérios para sua classificação. Então, a variável “idade” pode ser classificada como quantitativa porque expressa quantidades; como dependente, independente principal, covariável ou interveniente, a depender de sua posição no quadro de hipóteses do estudo; como contínua porque é expressa em uma escala contínua de valores; e assim por diante.

As variáveis discretas são expressas em valores descontínuos. Esses, por serem descontínuos, são chamados de categorias da variável. Ou seja, há um intervalo entre uma categoria e outra da variável. Alguns exemplos desse tipo de variável são: sexo, estado civil, raça, número de filhos e número de gestações. Estas duas últimas, embora quantitativas, são variáveis discretas porque não podem ser expressas em valores contínuos. Seria absurdo admitirmos um filho e meio ou uma gestação e meia, não é? Uma mulher não tem metade de um filho, metade de uma gestação (um aborto não deve ser considerado como metade de uma gestação; um aborto é um aborto). Assim tem-se um filho, ou dois, ou três, etc.; uma gestação, ou duas, ou três, etc. Essas variáveis, portanto, não variam em valores contínuos, não podendo ser expressas em

números fracionários.

As variáveis “sexo”, “estado civil” e “raça”, são ainda mais claramente descontínuas (discretas), já que há intervalos evidentes entre suas possíveis categorias.

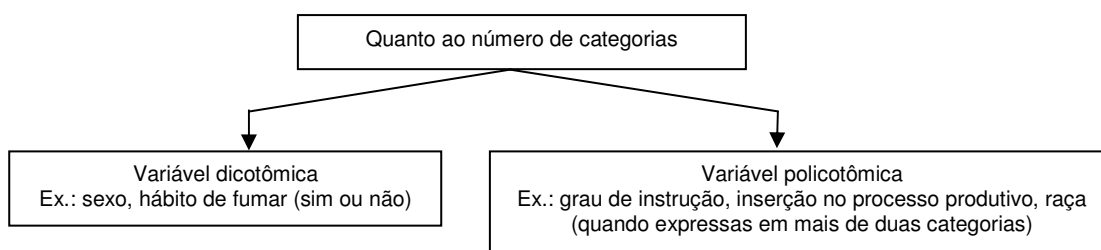


Observe que a variável “sexo” também já recebeu, até o momento, diferentes denominações. É qualitativa porque suas categorias (masculino e feminino) expressam qualidades distintas e não quantidades; geralmente, mas não obrigatoriamente, se posiciona como variável independente secundária no quadro de hipóteses, porque queremos saber qual a sua influência na ocorrência das doenças (se quisermos estudar que características ou hábitos das gestantes influenciam o sexo da criança que irá nascer, a variável “sexo” será dependente); e é discreta porque seus valores (categorias) são separados por intervalos, ou seja, não podem ser expressos em valores contínuos.

Podemos classificar as variáveis também de acordo com o número de categorias que possuem. Por esse critério, classificaremos uma variável como dicotômica se essa admitir apenas duas categorias. A variável “sexo”, p. ex., tal como utilizada comumente, admite apenas duas classificações: sexo masculino ou feminino, sendo portanto, uma variável dicotômica. As variáveis com respostas sim ou não, ou presente ou ausente, são também exemplos de variáveis dicotômicas.

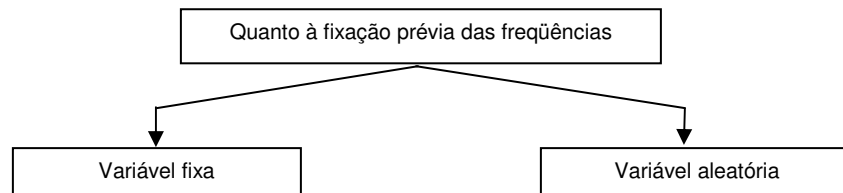
— E se o número de categorias for maior do que dois?

— Denominaremos a variável de policotômica. Alguns exemplos: “Grau de instrução” (analfabeto, primeiro grau incompleto, primeiro grau completo, segundo grau incompleto, segundo grau completo, terceiro grau incompleto, terceiro grau completo, pós-graduação incompleta, pós-graduação completa); “inserção no processo produtivo” (aposentado, assalariado, autônomo, pequeno proprietário, grande proprietário); e “raça” (negro, branco, mulato escuro, mulato médio, mulato claro, amarelo, índio).



Outro critério de classificação leva em conta se as freqüências de indivíduos nas diferentes categorias da variável foram fixadas previamente pelo investigador ou não. Na primeira situação a variável é chamada de fixa, porque teve o número de indivíduos em cada categoria fixado no planejamento do estudo, e na segunda é denominada aleatória, porque suas freqüências puderam variar aleatoriamente, sem interferência do pesquisador. Se, em um estudo epidemiológico do tipo caso-controle, decidíssemos estudar igual número de casos e controles, a presença ou ausência da doença estudada, que seria nossa variável

dependente, e que especificaria os dois grupos a serem comparados, seria uma variável fixa, já que nós fixaríamos previamente quantos seriam os casos e quantos os controles. Se, nesse mesmo exemplo, nossa variável independente principal fosse “hábito de fumar”, e não fixássemos previamente o número de fumantes e não fumantes a serem investigados, deixando que esses números expressassem livremente (sem nossa interferência) quantos fumantes ou não-fumantes realmente existissem na amostra ou população estudada, essa variável seria classificada como aleatória.

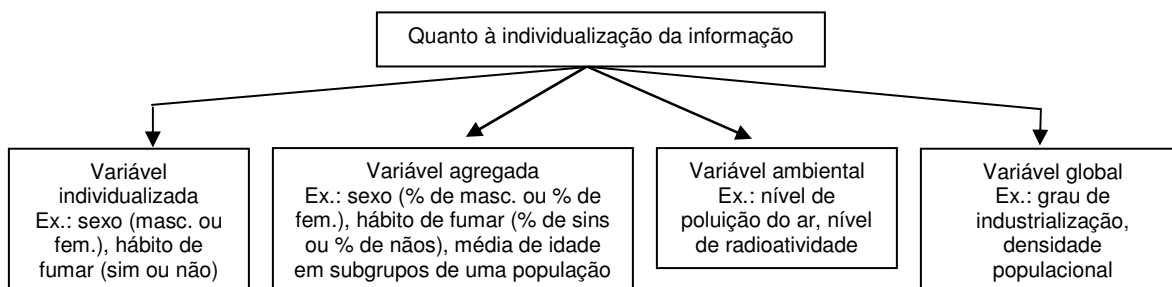


Outro critério para classificarmos uma variável é o nível de individualização da informação contida na mesma. Por esse critério uma variável pode ser: individualizada, agregada, ambiental ou global. Exemplificando: se coletarmos e analisarmos a variável “tabagismo” considerando as categorias “sim” e “não”, deveremos classificá-la como individualizada, porque para cada indivíduo estudado teremos a informação sobre se ele(a) fuma ou não. Mas, se a analisarmos como o percentual de fumantes em determinadas localidades que estivermos comparando, a variável “tabagismo” deve ser classificada como agregada, pois, embora inicialmente tivesse sido coletada ao nível individual, estará sendo considerada ao nível agregado. O percentual de fumantes expressará uma característica de cada grupo (agregado) de indivíduos investigado, e não de cada indivíduo.

Uma variável será chamada de ambiental, se expressar características físicas de um lugar no qual os grupos estudados vivem e/ou trabalham, tais como: nível de poluição do ar, nível de radioatividade, e número de horas de luz solar no local. Essas variáveis podem ser mensuradas ao nível individual, mas, geralmente, isso não é feito.

Uma variável será chamada de global, se tiver intrinsecamente uma natureza coletiva, isto é, se expressar uma informação que não possa ser individualizada. É o exemplo das variáveis: “densidade populacional” e “grau de industrialização”, cuja obtenção não faz sentido para indivíduos isoladamente.

Você pode ler mais sobre esse critério em *Rothman KJ e Greenland S, editores. Modern epidemiology. 2ª ed. Philadelphia (PA): Lippincott Williams e Wilkins; 1998*, nas páginas 460 e 461.



As variáveis podem ser classificadas ainda de acordo com a modalidade da escala em que são medidas. Por esse critério, as variáveis podem ser classificadas como nominais, ordinais, intervalares ou de razão. Sexo, com base nesse critério, é uma variável nominal porque além de cada uma de suas categorias indicar uma qualidade bem distinta da outra, não é possível colocarmos suas categorias em ordem crescente

ou decrescente, por algum critério de ordenamento aceitável. Você sugere algum critério aceitável para colocarmos as categorias da variável sexo em ordem? Nenhum critério seria aceitável, não é?

— E para as categorias da variável “raça”?

— Se fôssemos racistas poderíamos ordenar as raças pela suposta superioridade de umas sobre outras. Na nossa opinião, como não existe um critério aceitável para ordenar as categorias de raça, classificaremos essa variável também como nominal. O nome de cada categoria da variável raça (negro, branco, mulato escuro, mulato médio, mulato claro, amarelo, índio) indica uma qualidade racial, e isso é o máximo que sua modalidade de escala consegue expressar.

As variáveis ordinais, por sua vez, são aquelas cujas categorias podem ser postas em ordem crescente ou decrescente, por algum critério justificável. Este tipo de variável já contém um certo grau de quantificação e é isso que permite colocarmos suas categorias em ordem. “Grau de instrução” (analfabeto, primeiro grau incompleto, primeiro grau completo, segundo grau incompleto, segundo grau completo, terceiro grau incompleto, terceiro grau completo, pós-graduação incompleta, pós-graduação completa), p. ex., é uma variável ordinal porque podemos colocar suas categorias ordenadas do menor ao maior grau de escolaridade (como fizemos acima) ou vice-versa.

Mas, note que não há intervalos regulares entre uma categoria e outra dessa variável. Você pode me garantir que o intervalo entre analfabeto e primeiro grau incompleto tem a mesma amplitude que o intervalo entre primeiro grau incompleto e primeiro grau completo? É claro que não! Portanto, para uma variável ser classificada como ordinal é necessário que possamos colocar suas categorias em ordem e que os intervalos entre essas categorias não sejam regulares. Se as categorias de uma variável puderem ser postas em ordem e, além disto, os intervalos entre essas categorias forem regulares, a variável deverá ser classificada como intervalar ou de razão.

— E então, o que distingue uma variável intervalar de uma de razão?

— Já vimos que tanto uma variável intervalar como uma de razão podem ter suas categorias colocadas em ordem e os intervalos entre essas categorias são regulares. O critério para distingui-las é se o valor zero dessas variáveis representa ou não a ausência do fenômeno indicado pela variável. O valor zero de uma variável intervalar não indica ausência do fenômeno medido. O valor zero para a variável temperatura, p. ex., seja na escala Celsius ou na Fahrenheit, não indica ausência de temperatura. Um determinado grau de temperatura é arbitrariamente designado como zero nessas escalas. Assim, a variável “temperatura”, muito utilizada em pesquisas clínicas para indicar a variável “febre do paciente”, é um exemplo de variável intervalar. Esse tipo de variável é muito raro em pesquisas clínicas e epidemiológicas. Se você nos solicitasse um outro exemplo de variável intervalar usada em pesquisas em saúde, teríamos que procurar e sabemos que essa não seria uma tarefa fácil.

Diferentemente, o zero de uma variável de razão indica ausência do fenômeno medido. Considere as variáveis “número de filhos”, “número de gestações”, “idade”, “peso” e “altura”. Observe que essas podem ter seus valores postos em ordem crescente ou decrescente e, além disso, apresentam intervalos regulares entre os valores que podem assumir. A variável “número de filhos” varia regularmente a intervalos iguais de um filho. O “número de gestações” varia sempre em unidades de uma gestação. A “idade” varia em unidades regulares de anos, meses, semanas ou dias, etc., a depender do grau de precisão com que esteja sendo medida. O mesmo ocorre para o “peso” (unidades de quilos, gramas, centigramas, etc.) e a altura (unidades em metros, centímetros, milímetros, etc.). Todas as variáveis exemplificadas acima são “de razão” porque, além das características apontadas acima, seu valor zero indica ausência do fenômeno representado. Assim, o valor zero para o “número de filhos” ou “de gestações”, indica a ausência de filhos ou de gestações. Note

que para as variáveis “idade”, “peso” ou “altura”, o valor zero não se refere, obviamente, a uma realidade biológica individual impossível, mas “mede” teoricamente a própria inexistência do indivíduo. Então, a equação ausência do fenômeno = ausência do indivíduo = valor zero nos leva a classificar essas variáveis como de razão. Observe que a idade considerada nesse exemplo é aquela freqüentemente utilizada em estudos epidemiológicos, cuja medida se inicia com o nascimento. Se estivéssemos considerando a variável “idade gestacional”, o zero indicaria não-ocorrência da concepção.

— Mas, por que é importante o fato do valor zero de uma variável representar ou não a ausência do fenômeno indicado por aquela variável?

— Porque se o zero expressa ausência, podemos afirmar que, p. ex., o peso de um indivíduo com 100 kg é o dobro daquele de um indivíduo com 50 kg, ou que a idade de uma pessoa com 40 anos é o dobro daquela com 20 anos. Assim, poderemos dizer que a razão dos pesos ou das idades entre esses dois indivíduos é dois. Por isso, esse tipo de variável é denominado de razão. Tal comparação não pode ser feita se a variável for intervalar, pois como o zero desse tipo de variável não indica ausência do fenômeno, não há um valor de referência (o zero) que torne possível o cálculo dessas razões.

Resumindo:

CLASSIFICAÇÃO DAS VARIÁVEIS SEGUNDO SEU GRAU DE EXPRESSÃO QUANTITATIVA (A MODALIDADE DE SUA ESCALA)	
Nominal	Não expressa quantidade; suas categorias não podem ser colocadas em ordem Ex.: sexo, raça
Ordinal	Expressa quantidade embora ainda de forma limitada; suas categorias podem ser postas em ordem; seus intervalos são irregulares Ex.: grau de instrução, grau de desnutrição (leve, moderada, grave)
Intervalar	Expressa quantidade de modo mais exuberante; suas categorias podem ser postas em ordem; seus intervalos são regulares; valor zero <u>não indica ausência</u> do fenômeno medido Ex.: temperatura
De razão	Expressa quantidade de modo mais exuberante; suas categorias podem ser postas em ordem; seus intervalos são regulares; valor zero <u>indica ausência</u> do fenômeno medido Ex.: nº de filhos, nº de gestações, nº de cigarros, idade, peso, altura

— Para que todo esse esforço em classificar variáveis? Isso é realmente necessário?

— Ao longo do livro você verá que é indispensável identificarmos quais os tipos de variáveis que estudaremos na nossa pesquisa, porque será com base neste, e em outros aspectos discutidos mais adiante, que escolheremos a técnica estatística adequada para a análise.

— Tudo bem, mas se são tantas classificações, qual devemos utilizar?

— A última classificação, aquela baseada na modalidade da escala de medição da variável, é a mais utilizada pelos estatísticos.

— E o que veremos em seguida?

— No próximo capítulo, discutiremos com você qual o motivo para estudarmos uma amostra e não toda a população de interesse. Veremos também como selecionar adequadamente a amostra, e quais as implicações estatísticas decorrentes da obtenção de dados em uma amostra.

CAPÍTULO 3

- O que é amostragem e por que realizá-la?
 - Qual a importância da amostra ser representativa?
 - O que é população-alvo?
 - O que é população amostrada?
 - Como obter uma amostra representativa?
 - Quais os tipos de amostragens mais utilizados?
 - Quando escolher um ou outro tipo de amostragem?
 - O que é amostragem com ou sem reposição?
 - Quais as implicações de uma amostragem sem reposição?
 - O que é uma população finita?
-



— O que é amostragem e por que realizá-la?

— Para obter resultados cientificamente válidos em suas pesquisas será necessário que estas sejam bem planejadas e realizadas eficientemente, evitando vieses, que são erros metodológicos que podem produzir distorções nos seus resultados. Você deverá empenhar-se para eliminar ou reduzir ao máximo os vieses na seleção dos indivíduos que irão participar da pesquisa, e na realização das contagens e medições necessárias. Deverá também evitar ou minimizar vieses decorrentes da ausência ou insuficiência de neutralização de variáveis confundidoras.

Se você ainda não estiver bem seguro acerca da conceituação de viés e das técnicas para sua redução ou eliminação, sugerimos que consulte um dos seguintes livros de Epidemiologia: *Pereira MG. Epidemiologia: teoria e prática. Rio de Janeiro (RJ): Guanabara Koogan; 1995; Medronho RA, Carvalho DM, Bloch KV, Luiz RR, Werneck GL, editores. Epidemiologia. São Paulo (SP): Atheneu; 2002; Kleinbaum DG, Kupper LL, Morgenstern H. Epidemiologic Research. Belmont (CA): Lifetime Learning; 1982; Hothman KJ, Greenland S. Modern epidemiology. 2ª ed. Philadelphia (PA): Lippincott-Raven; 1998; ou Hennekens CH, Buring JE. Epidemiology in Medicine. Boston (MA): Little, Brown; 1987.*

Caso você tenha planejado e realizado adequadamente sua pesquisa, muito provavelmente ela poderá ser considerada como suficientemente livre de vieses, ou seja, com vieses que não alteram significativamente os resultados, produzindo, assim, conclusões cientificamente válidas.

Entre os cuidados que precisaremos tomar para assegurar a máxima validade aos nossos resultados, estão: a escolha adequada da população de interesse; do método de seleção da parte dessa população que será estudada; e a seleção de um número suficiente de pessoas para a pesquisa. Os dois primeiros aspectos serão abordados neste capítulo, e o último no capítulo 15.

REQUISITOS PARA OBTER-SE A AMOSTRA MAIS REPRESENTATIVA POSSÍVEL DA POPULAÇÃO:	Identificação de população adequada
	Utilização de um método adequado de escolha dos indivíduos a serem estudados
	Seleção de um número adequado de indivíduos para o estudo

Parece improvável, mas um pesquisador iniciante e sem orientação adequada pode escolher para seu estudo, inadvertidamente, uma população cujas características podem comprometer completamente seus resultados. Se desejássemos investigar, p. ex., a relação entre hábito de fumar e câncer de pulmão, seria desastroso, em termos científicos, se estudássemos uma população na qual predominassem adeptos de uma determinada religião que proibisse tal hábito.

— Por quê?

— Porque isso faria com que a exposição ao fumo (variável independente principal) fosse muito rara (ou mesmo inexistente) nos indivíduos investigados, inviabilizando uma abordagem quantitativa da relação

entre essa exposição e a doença de interesse (câncer de pulmão). Além disso, mesmo que o número de expostos fosse suficiente para uma abordagem quantitativa do problema, a população escolhida por você seria tão selecionada, tão diferente da população geral, que dificilmente poderia ser considerada representativa.

— E qual o problema de não haver representatividade?

— O problema é que todo pesquisador esforça-se para obter resultados que possam ser generalizados para populações mais amplas, pois dessa maneira suas conclusões se tornam úteis a um número maior de pessoas.

Imagine, por um instante, um estudo realizado em um grupo de indivíduos com características bem específicas, aquele mencionado acima, p. ex., cuja maioria pertencesse a uma religião que fizesse com que os indivíduos apresentassem características bem peculiares. Os resultados obtidos pelo estudo só seriam válidos para aquelas pessoas. Não poderiam ser generalizados para a população geral, nem para outros grupos com características distintas, porque estas poderiam modificar substancialmente os resultados.

Portanto, o processo de escolha da população a ser estudada é importantíssimo para a pesquisa. O ideal seria estudarmos sempre toda a população para a qual desejamos que os resultados sejam válidos, porque dessa maneira estaríamos considerando informações sobre todos os indivíduos daquela população e não de uma parte que poderia ter sido inadequadamente escolhida. Mas isso geralmente é muito demorado, trabalhoso, caro e, portanto, inviável. Por isso, a maioria dos pesquisadores estuda apenas uma parte da população, que é denominada de **amostra**, e o processo de selecionar uma amostra de uma população é chamado de **amostragem**.

Se desejarmos estudar uma amostra, com a intenção de generalizarmos os resultados obtidos nessa amostra para a população da qual ela foi retirada, devemos estar cientes da diferença entre dois tipos de populações: a **população amostrada** e a **população-alvo**. População amostrada é aquela da qual retiramos uma amostra. População-alvo é aquela para a qual queremos generalizar os resultados obtidos em uma amostra. Essas duas populações podem ou não ser as mesmas. Os procedimentos estatísticos para generalização de resultados amostrais (testes de significância estatística) permitem apenas generalização da amostra para a população amostrada. Somente quando a população amostrada e a população-alvo forem as mesmas poderemos utilizar esses procedimentos estatísticos para tirar conclusões sobre a população-alvo. Se a população amostrada e a população-alvo forem diferentes só poderemos tirar conclusões sobre a população-alvo utilizando procedimentos não-estatísticos.

Se estivéssemos investigando a eficácia de uma nova droga para o tratamento da hipertensão arterial, nossa população-alvo seria constituída por todos os pacientes com hipertensão arterial residentes na localidade onde realizássemos a pesquisa. Suponha que não fosse viável para nossa equipe estudar uma amostra retirada da população-alvo. Uma alternativa freqüentemente utilizada seria investigarmos uma amostra retirada do conjunto de pacientes hipertensos de um hospital específico existente naquela localidade. Se selecionarmos adequadamente essa amostra e o número de indivíduos estudados for suficiente, poderíamos tentar generalizar os resultados obtidos na amostra para a população amostrada (aquela do hospital estudado), através de procedimentos estatísticos. Não poderíamos, contudo, tentar generalizar esses resultados para a população-alvo, através desses mesmos procedimentos estatísticos, porque a amostra estudada não foi retirada levando em conta toda essa população. Para isso, teríamos que utilizar

procedimentos não-estatísticos, que consistem em simplesmente reunirmos nossa equipe de pesquisa, tantas vezes quantas sejam necessárias, para avaliarmos se a amostra estudada, mesmo não tendo sido selecionada considerando-se toda a população-alvo, tem ou não características semelhantes às desta.

POPULAÇÃO-ALVO	População para a qual se deseja generalizar os resultados
POPULAÇÃO AMOSTRADA	População da qual retiramos a amostra
AMOSTRA	Parte da população selecionada para o estudo
AMOSTRAGEM	Processo pelo qual a amostra é selecionada

– Quais os tipos de amostragens mais utilizados?

– A amostragem pode ser feita por sorteio (aleatória) ou não (não-aleatória). A primeira é preferível à última, porque o sorteio evita o uso de critérios subjetivos na escolha dos indivíduos, o que poderia resultar em não-representatividade da amostra. Uma amostragem aleatória é aquela que dá a cada indivíduo da população-alvo ou da população amostrada igual oportunidade (probabilidade) de ser escolhido para o estudo, porque a seleção é feita por sorteio, evitando a influência de fatores subjetivos, sendo todos os indivíduos igualmente considerados para o sorteio.

Lembre-se de que ao retirarmos uma amostra de cinco bolas de uma caixa contendo cinquenta bolas, podemos repor cada bola sorteada de volta na caixa, antes de realizar novo sorteio. Dizemos então que a amostragem foi feita com reposição. Quando repomos as bolas já sorteadas, essas, obviamente, poderão ser sorteadas novamente. Se cada bola sorteada não for repostada estaremos realizando uma amostragem sem reposição. Nesse caso, cada bola só poderá ser sorteada uma vez. Na pesquisa em saúde as amostragens são realizadas sem reposição, de modo que cada indivíduo só pode ser sorteado uma vez, porque não nos interessa obtermos informações sobre o mesmo indivíduo mais de uma vez. Nosso interesse é coletar informações sobre o maior número possível de diferentes indivíduos, para darmos conta da diversidade existente na população-alvo.

Entre as amostragens aleatórias, a mais simples, como indica sua denominação, é a amostragem aleatória simples. Nesse tipo, identificamos uma população-alvo adequada, p. ex., diabéticos residentes em uma determinada localidade, numeramos cada um deles, e sorteamos um certo número mínimo (“*n* mínimo”) de diabéticos para compor nossa amostra. Esse número deve ser previamente calculado por você, utilizando fórmulas adequadas (veja capítulo 15). Nesse tipo de amostragem, e em quase todos os outros, como o sorteio é de cada indivíduo, dizemos que nossa unidade amostral é o indivíduo.

Outro tipo é a amostragem aleatória sistemática, na qual o pesquisador irá escolher, sistematicamente, ou seja, um em cada cinco indivíduos, ou um em cada dez, etc., para compor sua amostra, até completar o “*n* mínimo”. Se você quisesse estudar, p. ex., doença isquêmica do coração em uma determinada localidade, poderia obter informações visitando os residentes naquela localidade, selecionando um domicílio a cada dez, se ali existissem 1.000 domicílios e o “*n* mínimo” calculado fosse de 100 indivíduos. Em cada domicílio sorteado você escolheria, também aleatoriamente, apenas um indivíduo para compor sua amostra. Outra opção seria você ter tido acesso a todos os registros das unidades de saúde do local e, sistematicamente, selecionado um a cada “*x*” prontuários (pacientes), até completar o “*n* mínimo”. Nessa opção, provavelmente, a amostra obtida, apesar de ser escolhida por sorteio, seria muito selecionada, podendo haver viés de seleção no estudo, porque as pessoas que foram atendidas nas unidades de saúde naquela localidade poderiam não ser representativas da população-alvo. Poderiam ser mais doentes, p. ex., a

ponto de terem buscado atendimento médico naquelas unidades. A amostragem sistemática pode provocar um outro viés de seleção, se a organização do arquivo médico (ou dos domicílios) seguir um determinado padrão que faça com que algumas unidades amostrais tenham uma probabilidade maior de participar da amostra. Se, p. ex., você vai selecionar um a cada cinco prontuários, e o arquivo for organizado de tal modo que um a cada cinco prontuários seja sempre de um paciente atendido no ambulatório de cardiologia, sua amostra não será representativa do conjunto de pacientes atendidos nas unidades de saúde incluídas no estudo, pois, se iniciarmos a seleção por um prontuário de paciente desse ambulatório, ao retirarmos um prontuário a cada cinco, ao final todos os prontuários serão de pacientes do mesmo serviço.

— Mas, isso deve ser muito difícil de acontecer!

— Realmente, fizemos apenas um alerta sobre a possibilidade disso ocorrer, mas concordamos que tal circunstância é muito improvável.

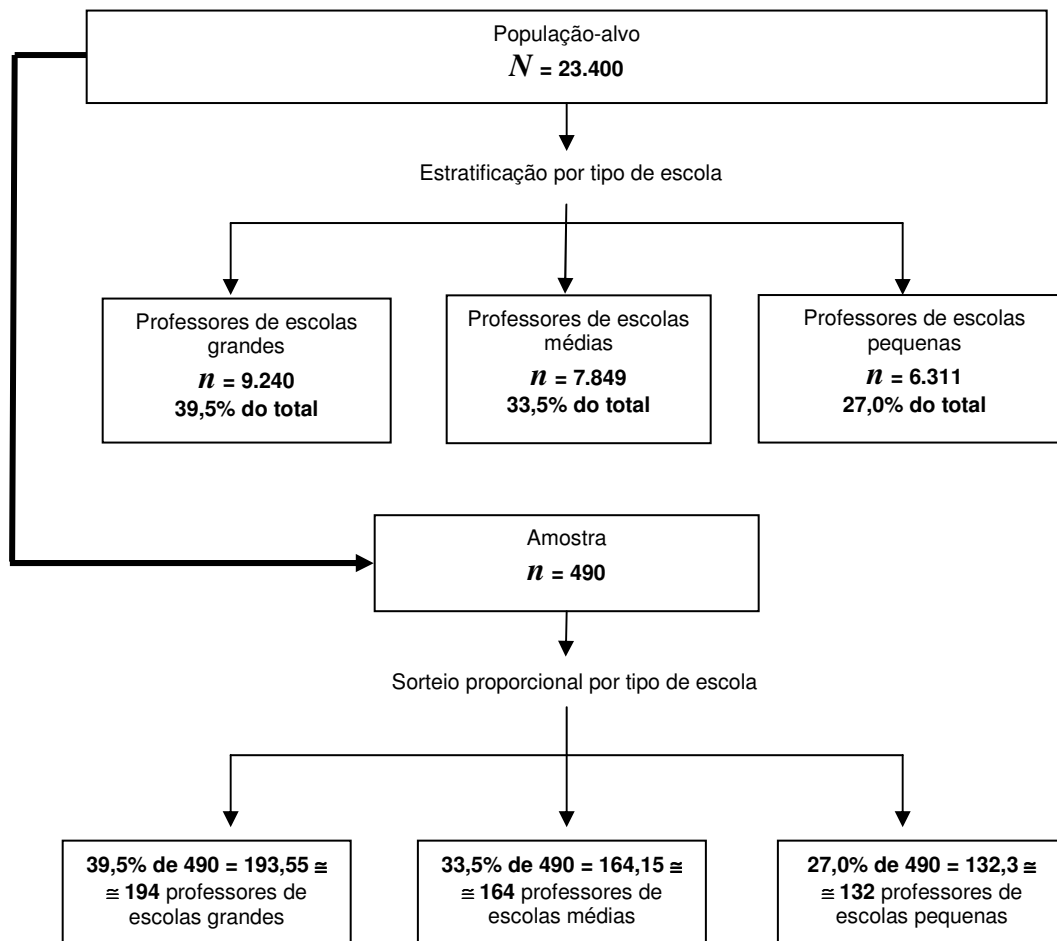
A amostragem aleatória por conglomerados é também muito utilizada. Consiste na seleção de grupos de indivíduos e não de indivíduos isoladamente. Esses agrupamentos de indivíduos (conglomerados) podem ser escolas, indústrias, hospitais, quartéis, bairros de uma cidade, etc. Suponha que queiramos realizar uma pesquisa sobre condições de trabalho e saúde de professores de primeiro e segundo graus da rede particular de ensino de uma cidade. Já sabemos que não seria necessário estudarmos todos os professores. Assim, selecionaríamos uma amostra. Você também já sabe que seria mais correto utilizar um método de amostragem aleatório. Agora você vai perceber que seria também mais prático e eficiente fazer essa amostragem em várias etapas. A primeira dessas seria uma amostragem aleatória por conglomerados (escolas), porque essa lhe permitiria tirar vantagem do fato de os indivíduos estarem reunidos por várias horas de cada dia, trabalhando em local comum, o que tornaria mais fácil o acesso a eles. Então, sua primeira etapa de amostragem consistiria em sortear certo número de escolas dentro do total de escolas. Para isso, seria necessária a obtenção de uma listagem com todas as escolas em funcionamento naquela localidade. Quanto ao número de escolas a serem selecionadas, não existem fórmulas específicas para sua definição, devendo o pesquisador estabelecê-lo tendo em vista os recursos da pesquisa e a busca de representatividade. Depois dessa etapa, você seguiria com a sua amostragem, até finalmente ter os professores selecionados para o seu estudo. Essas etapas adicionais serão abordadas no próximo parágrafo. Note que na amostragem por conglomerados a unidade amostral é o conglomerado e não o indivíduo. Você sorteia conglomerados (grupos de indivíduos; no nosso exemplo, grupos de professores em escolas) e não indivíduos.

Outros dois tipos são a amostragem aleatória estratificada e a aleatória proporcional. Esses serão apresentados ao mesmo tempo porque quando um pesquisador decide realizar uma amostragem estratificada ele geralmente realiza também uma amostragem proporcional. Não se preocupe, pois isto ficará claro com um exemplo. Considere novamente o estudo sobre professores mencionado acima. Na amostragem estratificada, como sua denominação indica, a população-alvo seria inicialmente classificada de acordo com categorias (estratos) de uma determinada variável. Vamos considerar, p. ex., a variável “tamanho da escola onde o professor leciona”, indicada pelo número de professores. Escolas com menos de 20 professores seriam classificadas como pequenas, com 21 a 50 como médias, e aquelas com mais de 50 como grandes. Assim, considerando a população-alvo, constituída por todos os professores de primeiro ou segundo graus da localidade estudada, cada professor seria classificado como atuando em escola pequena, média ou grande, levando em conta a escola na qual ele despendesse sua maior carga horária semanal de trabalho. Essa

informação seria usada para agruparmos os professores em três estratos (de escola pequena, média ou grande), antes de os sortearmos para comporem a amostra, que seria o nosso próximo passo, surgindo então a questão de quantos professores de cada tipo de escola (grande, média ou pequena) colocaríamos na amostra. Temos certeza de que você nos dirá que o ideal seria estabelecermos para nossa amostra determinados números de professores de cada tipo de escola, de tal modo que a proporcionalidade entre os professores dos três tipos na amostra fosse a mesma existente na população-alvo, e é isso mesmo que deveríamos fazer. Por isso, na maior parte das vezes, quando se faz uma amostragem estratificada, essa é também proporcional.

— **Tudo bem! Mas, para que fazer uma amostragem estratificada e proporcional?**

— Quando fazemos uma amostragem aleatória simples e com n suficiente, é muito grande a probabilidade da amostra obtida ser representativa da população-alvo. Mas, se quisermos garantir que determinada variável considerada muito importante para o tema que estivermos investigando, tenha a mesma distribuição na população-alvo e na amostra, utilizaremos uma amostragem aleatória estratificada proporcional. Para esse tipo de amostragem ficar ainda mais claro, apresentamos a seguir como seria seu fluxograma no exemplo atual, supondo que o “ n mínimo” necessário para o estudo seja 490:



Resumindo, selecionaríamos 194 professores nas escolas grandes, 164 nas médias e 132 nas pequenas.

Finalmente, vamos discutir sobre as amostragens não-aleatórias. Não se esqueça de que você deverá se esforçar para não usar esse tipo de amostragem. Entretanto, esteja ciente de que várias vezes teremos de utilizá-las. A maioria dos ensaios clínicos (estudo experimental) e dos estudos caso-controle investigam subgrupos da população-alvo selecionados não-aleatoriamente, o que traz limitações importantes para a generalização dos resultados obtidos.

Em um dos tipos de amostragem não-aleatória, a amostragem por conveniência, como sua denominação indica, o pesquisador inclui no seu estudo indivíduos aos quais ele tenha um acesso mais fácil, como os pacientes de hospitais ou trabalhadores de uma indústria onde ele atua. Os indivíduos são selecionados para a investigação de acordo com a conveniência do pesquisador.

É evidente que um estudo com esse tipo de amostragem fica muito vulnerável a viés de seleção, porque há uma grande probabilidade da amostra investigada não representar a população-alvo de pacientes ou de trabalhadores. Os componentes da amostra devem possuir características próprias dos indivíduos que se internaram naqueles poucos hospitais ou indústrias incluídas na pesquisa. Eles podem constituir uma clientela ou trabalhadores de um nível sócio-econômico ou outras características, distintas da população-alvo. Ao optar por uma amostragem desse tipo o pesquisador deverá ser bastante cauteloso ao generalizar os resultados do estudo. Ele poderá considerar suas conclusões como válidas para clientelas com características semelhantes àquelas que estudou, mas só poderá generalizá-las para a população-alvo se ele e sua equipe estiverem convencidos de que a clientela estudada seja semelhante à população-alvo.

Em outro tipo de amostragem não-aleatória, a amostragem de voluntários ou por auto-seleção, os pesquisadores incluem no estudo indivíduos que voluntariamente se oferecem para participar. Esse tipo de amostragem também é muito vulnerável ao viés de seleção. Indivíduos voluntários podem ser mais preocupados com sua saúde, mais conscientes sobre a importância de colaborar com a produção de conhecimento científico, ou mais doentes e, como tal, podem estar mais interessados na atenção médica oferecida pela pesquisa. Isso poderá torná-los não-representativos da população-alvo. Note que não seria incorreto considerarmos a auto-seleção como um tipo de amostragem por conveniência, pois é mais conveniente ao pesquisador utilizá-la, pela maior facilidade de obtenção de indivíduos para o estudo.

Resumindo os tipos de amostragem:

TIPOS DE AMOSTRAGEM	
Aleatória	Simple
	Sistemática
	Por conglomerados
	Estratificada
	Proporcional
Não-aleatória	Por conveniência
	Por auto-seleção

Já vimos que as amostragens podem ser feitas com ou sem reposição. Outro aspecto importante que devemos avaliar ao realizarmos o processo de amostragem é se a população-alvo ou amostrada é finita ou não.

— **População finita?**

— Sim. O critério para definirmos uma população como finita ou não, é a verificação do quão grande ou pequena é a proporção do n mínimo calculado para o estudo, em relação ao número total de indivíduos na população, N , ou seja, a verificação da quantidade n/N . Se $n/N > 0,05$, consideraremos a população como finita. O n mínimo é calculado através de fórmulas específicas (veja capítulo 15), e o N é obtido em diferentes fontes, a depender de qual população esteja sendo considerada para estudo.

— **E por que precisamos saber isso?**

— Porque se nossa amostragem for sem reposição e de população finita, uma das medidas mais utilizadas de variabilidade dos nossos dados, o erro-padrão, que abordaremos no capítulo 10 (páginas 147 a 149), precisa ser multiplicada por um fator de correção, para que obtenhamos o resultado correto. Essa correção é chamada de **correção para população finita** e é feita multiplicando-se o erro-padrão por $\sqrt{(N-n)/(N-1)}$. Podemos resumir o exposto acima da seguinte maneira, já considerando que a amostragem seja feita sem reposição:

Quando $n/N > 0,05$, ou seja, quando o tamanho da população não é muito maior do que o da amostra, a população é finita e, portanto, a correção para população finita deve ser feita.

Você pode ver uma demonstração empírica da necessidade dessa correção em *Daniel WW. Biostatistics: a foundation for analysis in the health sciences. 7ª ed. New York (NY): John Wiley; 1999*, nas páginas 130 e 131. Retornaremos a esse tema no capítulo 15.

— **Qual o próximo assunto?**

— No próximo capítulo veremos alguns dos procedimentos mais utilizados para a descrição de dados quantitativos.

CAPÍTULO 4

- Quais os dados necessários para utilizarmos a Bioestatística?
 - Quais as técnicas mais aplicadas nas primeiras etapas de descrição de dados quantitativos?
 - O que são freqüências?
 - Quais os tipos de freqüências e suas aplicações?
-



— **Quais os dados necessários para utilizarmos a Bioestatística?**

— Como vimos no capítulo 1 (página 3), os dados básicos necessários são as **contagens** e **medições**.

TIPOS DE DADOS ESTATÍSTICOS
Contagens
Medições

Ao estudarmos uma população ou parte desta (amostra), poderemos classificar cada indivíduo segundo uma determinada variável: “sexo”, por exemplo, fazendo posteriormente as contagens de quantos são homens e quantos são mulheres. Feitas essas contagens, poderemos trabalhar com o cálculo e a comparação de proporções. Proporção de homens, de mulheres, de ricos, de pobres, etc. Com tais proporções, poderemos utilizar diversos procedimentos da Estatística (Descritiva, Analítica ou Inferencial), próprios para se trabalhar com proporções, considerando também o tipo de estudo epidemiológico escolhido e nosso(s) objetivo(s).

As variáveis quantificadas através de contagens são geralmente expressas sob a forma de codificações. Continuando com a variável “sexo” como nosso exemplo, poderemos codificá-la com o valor 1 significando masculino e o valor 2 representando feminino. Ao contrário de digitarmos no computador a palavra “masculino” ou “feminino”, digitariamos o código 1 ou 2, respectivamente. Uma vantagem evidente de fazermos isso seria diminuirmos o número de digitações necessárias para colocarmos os dados no computador, porque digitariamos apenas uma vez para inserir o código 1 ou 2, enquanto o número de digitações seria maior se optássemos por digitar “masculino” ou “feminino”.

— **Mas, não poderíamos digitar simplesmente as letras “M” ou “F” em substituição a “masculino” ou “feminino” diminuindo também o número de digitações?**

— Sim, poderíamos, mas há uma outra vantagem importante em utilizarmos códigos numéricos. Considere a variável “raça”, p. ex., que tem um número de categorias maior do que a variável “sexo”, e suponha que tenha sido necessário reagruparmos posteriormente suas categorias. Se tivéssemos usado palavras ou letras iniciais para designar cada categoria, seria menos prático recodificá-la. Ao contrário, se usássemos códigos numéricos (1=negro; 2=mulato escuro; 3=mulato médio; 4=mulato claro; 5=branco; 6=índio; e 7=amarelo) poderíamos facilmente recodificar essa variável, agrupando, p. ex., os códigos 2, 3 e 4 em um único grupo de “mulatos”. A razão pela qual não conseguimos fazer isso com letras ou palavras é simplesmente a de que os melhores programas estatísticos existentes não foram programados dessa maneira. Assim, ao criar seu banco de dados em um computador, recomendamos que defina suas variáveis como “numéricas”, mesmo que essas sejam nominais, dando códigos numéricos às categorias de cada variável, para que depois, se necessário, você possa recodificá-las mais facilmente.

Outros dados podem resultar de medições de um determinado parâmetro em cada indivíduo, como, p. ex., a medição da tensão arterial (em mmHg), da glicemia (em mg/100ml) e muitos outros. Nesses casos, não estamos fazendo uma contagem. Estamos medindo o valor de um determinado parâmetro em cada indivíduo.

Entretanto, isso não nos impede de estabelecer um certo valor de glicemia como limite de normalidade (“ponto de corte” ou “escore de corte”); de classificar, em seguida, cada indivíduo em diabético ou não-diabético, a depender do seu valor de glicemia encontrar-se acima ou abaixo desse “escore de corte”; e de contar quantos são e quantos não são diabéticos. Assim, poderíamos calcular a proporção de diabéticos e de não-diabéticos, e trabalhar com os procedimentos estatísticos apropriados para utilização de proporções. Assim, a medição da glicemia se transformaria em uma contagem. Mas, quando transformamos uma medição em contagem, devemos estar cientes de que perdemos informações. Com as medições temos, para os indivíduos estudados, um espectro de valores muito mais amplo para aquela variável do que se utilizarmos um “escore de corte”, pois, neste último caso, os valores a serem assumidos por cada indivíduo para aquela variável resumem-se a apenas duas categorias: ser diabético ou não. Por isso, dizemos que perdemos informações quando transformamos medições em contagens.

— E qual o problema em perdermos informações?

— Todo pesquisador se esforça ao máximo para obter o maior número de informações e o maior detalhamento possível, porque dessa maneira ele poderá estudar o seu tema de forma mais completa e aprofundada. Voltando ao nosso exemplo, quando trabalhamos com medições, um número muito maior de valores de glicemia será considerado, permitindo uma análise que levará em conta uma representação mais completa da realidade estudada.

— Então não devo transformar medições em contagens?

— Teoricamente você deve empenhar-se em manter as informações mais detalhadas possíveis. Entretanto, em estudos analíticos, muito provavelmente você será obrigado(a) a transformar medições em contagens, para viabilizar a utilização dos métodos mais comuns de análise estatística de dados. Em algumas etapas da análise estatística, os procedimentos existentes podem ter sua aplicação inviabilizada porque os indivíduos estudados são distribuídos em numerosos subgrupos, ao serem classificados por duas ou mais variáveis simultaneamente. Para viabilizarmos numericamente essas etapas de análise, temos de transformar medições em contagens, porque os procedimentos existentes só serão viáveis se agirmos assim.

Sugerimos a utilização de medições sempre que estivermos fazendo descrição de dados, transformando-as em contagens apenas quando isso se tornar absolutamente necessário em certas etapas da análise.

— Tudo bem! Uma medição pode ser transformada em contagem. E esta pode ser transformada em medição?

— Não. Ao coletarmos informações sobre o hábito de ingerir bebidas alcoólicas, classificando essa variável em “bebe” ou “não bebe”, trabalharemos com contagens. Será impossível transformarmos a informação de que um indivíduo bebe, na quantidade de bebida alcoólica (em ml) consumida por ele, simplesmente porque não coletamos essa informação. Seria necessário coletarmos uma nova variável através da qual, originalmente, obteríamos informações sobre a quantidade de bebida alcoólica consumida por cada indivíduo.

— Quando fazemos contagens utilizamos principalmente o cálculo de proporções para sua descrição, análise e inferência. E quando fazemos medições?

— Nos próximos capítulos veremos vários procedimentos estatísticos utilizados para a descrição, análise e inferência de medições, tais como o cálculo da média aritmética e desvio-padrão.

Resumindo:

TIPOS DE DADOS ESTATÍSTICOS:	
Contagens de variáveis codificadas	Cada categoria da variável recebe um código numérico que a representa; exemplo: sexo: 1= masculino e 2 = feminino; faz-se a contagem de quantos indivíduos foram classificados em cada categoria dessa variável
Medições	A variável expressa uma medida de uma determinada característica; exemplo: medida da tensão arterial, em mmHg, feita para cada indivíduo estudado

— Quais as técnicas mais aplicadas nas primeiras etapas de descrição de dados quantitativos?

— Vamos supor que já tenhamos coletado os dados da pesquisa que estamos realizando, sejam contagens ou medições.

Depois de coletar as informações necessárias para o estudo, você se verá diante de um emaranhado de números. Lembre-se que um pesquisador em saúde pode ter que estudar centenas ou milhares de indivíduos para produzir resultados generalizáveis e que ele precisa considerar muitas variáveis, algumas que são do seu interesse estudar, e outras que deverão ser levadas em conta para que o efeito delas sobre a associação estudada possa ser neutralizado ou controlado. Imagine-se agora diante das medições, contagens e codificações feitas por você e sua equipe. Observe a planilha de dados a seguir, que contém os resultados obtidos para apenas uma das variáveis que você pretende estudar e para apenas 75 indivíduos, lembrando-se de que na prática muito mais variáveis e indivíduos estariam sendo pesquisados:

Número do indivíduo na pesquisa	Idade (em anos)	Número do indivíduo na pesquisa	Idade (em anos)	Número do indivíduo na pesquisa	Idade (em anos)
1	35,8	26	37,7	51	38,2
2	44,8	27	36,8	52	40,7
3	39,9	28	37,5	53	34,3
4	34,0	29	43,4	54	46,9
5	41,4	30	43,7	55	37,6
6	40,0	31	38,7	56	32,6
7	41,0	32	46,1	57	34,3
8	37,6	33	42,5	58	46,3
9	31,5	34	25,5	59	35,5
10	45,5	35	31,5	60	55,6
11	25,1	36	29,0	61	46,7
12	51,3	37	35,5	62	36,9
13	46,4	38	31,0	63	61,7
14	34,4	39	37,0	64	33,2
15	36,2	40	29,2	65	27,7
16	37,5	41	36,3	66	24,2
17	40,9	42	47,9	67	38,7
18	39,0	43	45,7	68	38,3
19	40,5	44	35,4	69	34,5
20	32,2	45	41,6	70	47,2
21	32,5	46	49,2	71	65,8
22	52,4	47	32,0	72	39,3
23	47,3	48	39,7	73	24,0
24	40,0	49	30,5	74	40,9
25	34,5	50	41,8	75	41,8

Ao olhar para essa planilha vemos que são tantos os números que não sabemos por onde e como começar a descrevê-los. Experimente essa sensação olhando a planilha e tentando descrever os resultados. Verifique como é difícil descrever os valores de idade obtidos, porque eles se encontram organizados de um modo que não facilita sua descrição. Com base nessa planilha tente, apenas por alguns instantes, responder às seguintes questões: a) qual o valor mínimo de idade? b) qual o valor máximo? c) quantos indivíduos tinham 34,5 anos de idade? d) quantos tinham idade entre 31,0 e 31,5 anos?

É possível respondermos às perguntas feitas acima, mas você verificou como isso seria trabalhoso e demorado? Essa dificuldade foi sentida há muito tempo pelos pioneiros da Estatística, que foram ao longo do tempo, em um esforço intelectual contínuo, desenvolvendo procedimentos que viabilizassem a descrição de um conjunto grande de dados.

Na planilha acima, os indivíduos estão listados por ordem crescente do número de identificação de cada um na pesquisa.

Experimente agora **organizar** melhor aquela planilha, colocando os dados em ordem crescente de idade, conforme apresentamos abaixo:

Número do indivíduo na pesquisa	Idade (em anos)	Número do indivíduo na pesquisa	Idade (em anos)	Número do indivíduo na pesquisa	Idade (em anos)
73	24,0	1	35,8	7	41,0
66	24,2	15	36,2	5	41,4
11	25,1	41	36,3	45	41,6
34	25,5	27	36,8	50	41,8
65	27,7	62	36,9	75	41,8
36	29,0	39	37,0	33	42,5
40	29,2	16	37,5	29	43,4
49	30,5	28	37,5	30	43,7
38	31,0	8	37,6	2	44,8
9	31,5	55	37,6	10	45,5
35	31,5	26	37,7	43	45,7
47	32,0	51	38,2	32	46,1
20	32,2	68	38,3	58	46,3
21	32,5	31	38,7	13	46,4
56	32,6	67	38,7	61	46,7
64	33,2	18	39,0	54	46,9
4	34,0	72	39,3	70	47,2
53	34,3	48	39,7	23	47,3
57	34,3	3	39,9	42	47,9
14	34,4	6	40,0	46	49,2
25	34,5	24	40,0	12	51,3
69	34,5	19	40,5	22	52,4
44	35,4	52	40,7	60	55,6
37	35,5	17	40,9	63	61,7
59	35,5	74	40,9	71	65,8

Observe que, nesta segunda planilha, os números de identificação dos indivíduos estão desordenados, começando por 73 e terminando por 71. A idade, porém, está listada do valor mais baixo ao mais alto. Tente novamente responder às perguntas formuladas anteriormente: a) qual o valor mínimo de idade? b) qual o valor máximo? c) quantos indivíduos tinham 34,5 anos de idade? d) quantos tinham idade entre 31,0 e 31,5 anos?

Viu como agora ficou muito mais fácil responder a essas perguntas? Observando rapidamente a segunda planilha, vemos que o valor mínimo observado de idade foi 24,0 anos; o máximo foi 65,8 anos; dois indivíduos tinham 34,5 anos de idade; e três tinham idade entre 31,0 e 31,5 anos.

Portanto, o primeiro passo na descrição dos dados obtidos por você deverá ser organizar aquela grande quantidade de números de um modo mais apropriado. Fique tranquilo, porque quando solicitar listagens de resultados ao computador através de programas estatísticos, você poderá, com uma simples apertada no rato (“mouse”)², obter o ordenamento ascendente ou descendente para cada uma das variáveis. Para alguns programas não será necessário nem mesmo apertar o rato, porque a opção-padrão já lhe fornecerá os dados em ordem crescente.

O segundo passo será o **cálculo de freqüências**. Elas podem ser simples, simples acumulada, relativa e relativa acumulada.

TIPOS DE FREQUÊNCIAS:
Simples
Simples acumulada
Relativa
Relativa acumulada

Para obter freqüências simples, como o nome indica, observaremos com que freqüência (quantas vezes) cada valor de uma determinada variável foi observado, ou seja, contaremos quantos indivíduos apresentaram cada valor daquela variável. Retomemos nosso exemplo da idade. Para simplificar, e garantir espaço para tantos números, vamos considerar apenas os valores de idade para os primeiros 25 indivíduos do banco de dados, ordenado pelo número de identificação de cada indivíduo na pesquisa.

— Mas, assim não iremos perder o ordenamento por idade?

— Sim, mas lembre-se de que quando solicitarmos a listagem de freqüências desta variável, o computador, automaticamente ou a seu critério, ordenará os valores de idade em ordem crescente ou decrescente.

Ao retomarmos o nosso exemplo, será mais conveniente também considerarmos a idade em anos completos.

— Por quê?

— Porque se continuarmos utilizando um decimal, é muito provável que, devido a essa maior precisão da medida da idade, para cada valor desta só exista um indivíduo entre os 25 estudados.

— E daí?

— Não haveria erro em usarmos um decimal, mas para o que estamos desejando que você aprenda, neste momento, será mais apropriado utilizarmos a idade em anos completos, porque assim teremos uma quantidade maior de indivíduos para cada valor de idade, já que, p. ex., indivíduos com 32,3 e 32,4 anos, seriam considerados como tendo a mesma idade em anos completos: 32 anos. Dessa forma poderemos

² “Camundongo” seria a melhor tradução para a palavra “mouse” do idioma inglês. Entretanto, preferimos a palavra “rato” por ser mais curta e por ter praticamente o mesmo significado pretendido.

encontrar diferentes freqüências de indivíduos para diferentes idades. Trabalhando com decimais e com apenas 25 indivíduos, seria muito provável que cada um desses indivíduos apresentassem valores diferentes de idade (por diferenças de décimos, centésimos ou milésimos), o que comprometeria o nosso exemplo, já que queremos mostrar como utilizar freqüências simples para descrevermos dados quantitativos. Em resumo, se utilizássemos decimais seria muito provável que a freqüência simples (número de indivíduos) obtida para cada idade fosse sempre igual a 1. Utilizando a idade em anos completos, nossa nova planilha de dados seria então:

Número do indivíduo na pesquisa	Valores de idade (em anos completos)
1	36
2	45
3	40
4	34
5	41
6	40
7	41
8	38
9	31
10	46
11	25
12	51
13	46
14	34
15	36
16	38
17	41
18	39
19	41
20	32
21	32
22	52
23	47
24	40
25	34

Solicitando a listagem de freqüências ao computador obtemos:

Valores de idade (em anos completos)	Freqüência simples
25	1
31	1
32	2
34	3
36	2
38	2
39	1
40	3
41	4
45	1
46	2
47	1
51	1
52	1

O computador verificou que apenas um indivíduo tinha idade de 25 anos completos, o mesmo

acontecendo para as idades 31, 39, 45, 47, 51 e 52 anos. Dois indivíduos tinham 32 anos completos, sendo esta mesma frequência observada para as idades 36, 38, e 46 anos. Três indivíduos tinham 34 anos de idade, o mesmo ocorrendo para a idade 40 anos. Quatro pessoas apresentavam 41 anos de idade. Poderemos utilizar esses resultados na descrição dos nossos dados. Com base nessas frequências simples, você poderá dizer em seu trabalho quantos indivíduos tinham essa ou aquela idade, com destaque para as idades com as maiores frequências ou as menores, a depender do que seja mais importante ou interessante para o seu estudo.

Solicitando agora ao computador as frequências simples acumuladas obtemos:

Valores de idade (em anos completos)	Frequência simples	Frequência simples acumulada
25	1	1
31	1	2
32	2	4
34	3	7
36	2	9
38	2	11
39	1	12
40	3	15
41	4	19
45	1	20
46	2	22
47	1	23
51	1	24
52	1	25

Observe que cada valor na coluna intitulada “frequência simples acumulada” representa o número de indivíduos acumulados até aquela idade, naquela série de observações. Por exemplo, a frequência acumulada até a idade 46 anos foi 22, porque até a idade 45 anos já havia um acúmulo de 20 indivíduos, aos quais nesta etapa foram acrescentados dois indivíduos com 46 anos de idade, resultando na soma $20 + 2 = 22$. Esse resultado poderá ser utilizado na descrição dos nossos dados, pois a frequência encontrada revelou que, dos 25 indivíduos estudados, 22 tinham idade até 46 anos. Contudo, esse tipo de frequência não é o mais utilizado, sendo mais comum usar-se a frequência relativa acumulada, que será discutida logo adiante.

As frequências relativas são apresentadas abaixo:

Valores de idade (em anos completos)	Frequência simples	Frequência simples acumulada	Frequência relativa (%)
25	1	1	4,0
31	1	2	4,0
32	2	4	8,0
34	3	7	12,0
36	2	9	8,0
38	2	11	8,0
39	1	12	4,0
40	3	15	12,0
41	4	19	16,0
45	1	20	4,0
46	2	22	8,0
47	1	23	4,0
51	1	24	4,0
52	1	25	4,0

Conforme sua denominação, a frequência relativa indica que proporção da frequência total é representada pela frequência simples de cada idade. Ou seja, o quanto a frequência simples de cada valor de idade representa em relação à (relativo à) frequência total. No nosso exemplo atual, o número de indivíduos estudado, denotado por n , é igual a 25. Assim, para ilustrar um dos cálculos, temos que a frequência relativa para a idade 41 anos foi obtida dividindo-se 4 (frequência simples para aquela idade) por 25 (frequência total), e multiplicando-se o resultado por 100. Poderíamos multiplicar por outro múltiplo de dez, mas todos nós estamos mais familiarizados com a expressão de resultados em percentuais. Dessa maneira, podemos dizer que 16 por cento dos indivíduos tinham 41 anos de idade. O cálculo feito acima é uma simples regra-de-três: se 25 representam 100 por cento, ou seja, representam o total dos indivíduos, 4 indivíduos (número de indivíduos com 41 anos de idade) representam quantos por cento? Como você aprendeu no segundo grau de sua formação pré-universitária, isso pode ser expresso por

$$\frac{25}{4} = \frac{100}{x},$$

que se lê: “vinte e cinco está para quatro, assim como cem está para xis”.

Essa equação é resolvida facilmente como mostrado abaixo:

$$\frac{25}{4} = \frac{100}{x}, \text{ donde}$$

$$25x = (100)(4)$$

$$25x = 400$$

$$x = \frac{400}{25} = 16,0 \%$$

Assim, a frequência relativa da idade 41 anos foi 16,0%, como já havíamos mencionado. Essas frequências relativas são muito utilizadas na descrição dos seus dados.

Uma frequência relativa informará aos(às) leitores(as) do seu trabalho que proporção dos indivíduos estudados possuem essa ou aquela característica de interesse.

As frequências relativas acumuladas estão apresentadas em tabela, logo no início da próxima página:

Valores de idade (em anos completos)	Frequência simples	Frequência simples acumulada	Frequência relativa (%)	Frequência relativa acumulada (%)
25	1	1	4,0	4,0
31	1	2	4,0	8,0
32	2	4	8,0	16,0
34	3	7	12,0	28,0
36	2	9	8,0	36,0
38	2	11	8,0	44,0
39	1	12	4,0	48,0
40	3	15	12,0	60,0
41	4	19	16,0	76,0
45	1	20	4,0	80,0
46	2	22	8,0	88,0
47	1	23	4,0	92,0
51	1	24	4,0	96,0
52	1	25	4,0	100,0

Observe que a frequência relativa acumulada de cada idade é obtida somando-se as frequências relativas existentes até aquela idade. Esse tipo de frequência também pode ser utilizado por você na descrição dos dados. No caso, verificamos que um pouco mais de três quartos (76,0%) dos indivíduos tinham idade até 41 anos, o que está nos revelando que a população ou amostra estudada era relativamente jovem.

Resumindo:

TIPOS DE FREQUÊNCIAS:	
Simples	Resultado da contagem de indivíduos em cada categoria de uma determinada variável
Simples acumulada	Resultado da contagem de indivíduos acumulada até cada categoria de uma determinada variável
Relativa	Proporção de cada contagem obtida para cada categoria de uma determinada variável em relação à contagem total de indivíduos
Relativa acumulada	Proporção de cada contagem obtida para cada categoria de uma determinada variável em relação à contagem total, acumulada até cada categoria desta variável

— Bem! Já vimos os dois primeiros passos mais comumente utilizados para descrição de dados quantitativos. Estou curioso para saber quais são os próximos.

— Para não nos alongarmos muito neste capítulo, vamos encerrá-lo aqui, sob a ressalva de que ainda não concluímos os procedimentos mais utilizados na Estatística Descritiva. Outra ressalva é a de que demos ênfase ao uso do cálculo de frequências na descrição de dados quantitativos, mas a comparação de frequências é fundamental para a análise de dados quantitativos. Dois dos procedimentos para compararmos frequências serão discutidos nos capítulos 16 e 17.

Os outros procedimentos mais utilizados na Estatística Descritiva ocuparão os próximos cinco capítulos, nos quais discutiremos com você as famosas medidas de tendência central (moda, média e mediana); as medidas de dispersão (amplitude, desvio médio, desvio-padrão, variância, coeficiente de variação); as medidas de posição (média, mediana, percentis); os diagramas; e as distribuições de frequências (que são um tipo de diagrama). Em breve, também explicaremos porque estas últimas são também aplicadas na Estatística Inferencial.

CAPÍTULO 5

-
- O que são medidas de tendência central e quais as suas aplicações?
 - Quais são as medidas de tendência central mais utilizadas, o que indicam e como calculá-las?
 - Em quais circunstâncias deveremos usar a moda, a média ou a mediana?
-



— **O que são medidas de tendência central e quais as suas aplicações?**

— No capítulo anterior, vimos que devemos iniciar a descrição de dados quantitativos organizando-os de modo adequado e contando as freqüências de cada valor, de cada variável estudada. Tais procedimentos representaram um grande progresso, mas são limitados, deixando a descrição dos dados ainda bastante incompleta. Outros procedimentos foram também desenvolvidos e melhoraram substancialmente nossa capacidade descritiva. O **cálculo de medidas de tendência central** é um desses procedimentos, e foi criado com o objetivo de fornecer medidas quantitativas que resumissem o conjunto de dados investigados. Ou seja, uma medida de tendência central é uma única medida que pode ser usada para representar toda uma série de observações. Esta medida única indica qual a tendência central daquele conjunto de valores (daí sua denominação), isto é, para qual valor de uma determinada variável os demais valores convergem, tendem.

— **Quais são as medidas de tendência central mais utilizadas e como calculá-las?**

— As medidas de tendência central mais utilizadas são: **a moda, a média e a mediana.**

MEDIDAS DE TENDÊNCIA CENTRAL
Moda
Média
Mediana

A **moda**, como sua denominação indica, é o valor de uma série de observações que aparece com a maior freqüência. Dizer que uma vestimenta está na moda é o mesmo que dizer que essa é usada por muitas pessoas, ou seja, muito freqüentemente.

Moda de uma variável é o valor daquela variável que ocorre mais freqüentemente na população ou amostra estudada.

Considere novamente o nosso exemplo das idades de 25 indivíduos:

Número do indivíduo na pesquisa	Valores de idade (em anos completos)
1	36
2	45
3	40
4	34
5	41
6	40
7	41
8	38
9	31
10	46
11	25
12	51
13	46
14	34
15	36
16	38
17	41
18	39
19	41
20	32
21	32
22	52
23	47
24	40
25	34

Para encontrarmos a moda temos de solicitar ao computador as freqüências simples de cada idade, uma vez que a moda é o valor de idade que mais ocorreu naquela série de observações. Solicitando as freqüências simples obtemos:

Valores de idade (em anos completos)	Freqüência simples
25	1
31	1
32	2
34	3
36	2
38	2
39	1
40	3
41	4
45	1
46	2
47	1
51	1
52	1

Vemos que o valor de idade com a maior freqüência é 41 anos, sendo esta a moda da série investigada. Note que não dissemos que a moda é 41. Dissemos que a moda é 41 anos.

— **Por que temos de colocar a palavra anos?**

— Porque a idade é uma variável dimensional, devendo sua moda ser também expressa na mesma dimensão, ou seja, em quantidade de anos completos vividos.

Como você observou, a moda é muito simples de ser encontrada, mas como uma mesma série pode apresentar mais de uma moda (série bimodal, trimodal, etc.), essa medida de tendência central não é a mais apropriada, pois lembre-se que nosso interesse é encontrar uma medida única que possa resumir toda aquela série. A média e a mediana não têm esse problema, porque para cada série de observações pode-se calcular apenas uma única média e uma única mediana.

A média é uma velha conhecida sua, e a todo o momento você a utiliza nas suas atividades pessoais e/ou profissionais.

Existem quatro tipos principais de médias: aritmética, ponderada, geométrica e harmônica.

TIPOS DE MÉDIAS	
Aritmética	Geométrica ¹
Ponderada	Harmônica ¹

1. Não será abordado neste livro.

A média aritmética é calculada somando-se todos os valores de uma série de observações e dividindo-se esta soma pelo número de valores da série. Não se esqueça de que o número de valores geralmente coincide com o número de indivíduos estudados, pois, p. ex., cada indivíduo só possui um valor de idade. Assim, o número de valores de idade será igual ao número de indivíduos estudados.

Média aritmética de uma variável é a soma de todos os valores da variável, obtidos em uma população ou amostra, dividida pelo número de indivíduos na população ou amostra.

Vamos considerar o exemplo visto anteriormente dos 25 valores de idade:

Número do indivíduo na pesquisa	Valores de idade (em anos completos)
1	36
2	45
3	40
4	34
5	41
6	40
7	41
8	38
9	31
10	46
11	25
12	51
13	46
14	34
15	36
16	38
17	41
18	39
19	41
20	32
21	32
22	52
23	47
24	40
25	34

A média aritmética da idade desses indivíduos, \bar{x} (lê-se “xis barra”), será dada por

$$\bar{x} = \frac{36 + 45 + 40 + 34 + 41 + 40 + 41 + 38 + 31 + \dots + 52 + 47 + 40 + 34}{25} = \frac{980}{25} = 39,2 \text{ anos}.$$

Pense um pouco sobre o que foi realizado: somamos todas as idades e depois dividimos pelo número de indivíduos, ou seja, a quantidade total de anos vividos foi dividida pelo número de indivíduos estudados. Você deve concordar, então, que o resultado vai nos indicar qual a quantidade de idade por indivíduo, isto é, qual o valor de idade de cada indivíduo caso todos eles tivessem a mesma idade, ou, o que é o mesmo, caso a idade não variasse entre eles. No nosso exemplo, é isso que chamamos de idade média ou média de idade dos 25 indivíduos estudados. Utilizaremos essa média como uma única medida que resume o conjunto de idades obtidas para aqueles 25 indivíduos.

Se quisermos expressar a média aritmética através de notações matemáticas teremos duas opções:

a) para populações:

$$\mu = \frac{\sum_{i=1}^N x_i}{N};$$

Observe que no lugar de \bar{x} usamos a letra grega μ (lê-se “mi”) para indicar a média aritmética, e a notação N para o número de indivíduos estudados, em letra maiúscula, justamente para indicar que estamos representando toda a população-alvo. A letra grega \sum , que corresponde ao “S” maiúsculo do nosso alfabeto (letra adequada para Soma ou Somatório), como você já sabe, foi utilizada para indicar a soma dos valores de idade, sendo esses valores representados por x_i , com o subscrito i indicando, de uma só vez, cada um dos valores de idade, pois, na notação para somatório vemos que i varia de 1 a N . Ou seja, para trabalharmos com uma notação resumida, em vez de colocarmos todos os valores de idade $(x_1, x_2, x_3, x_4, \dots, x_N)$, nós representamos todos eles de uma só vez através da notação x_i e indicamos na notação do somatório que os valores de i variam de 1 a N . Repetindo de uma outra forma, x_i indica cada valor de idade, por isso, varia de 1 a N já que N indivíduos (toda a população-alvo) foram estudados, ou seja, quando $i = 1$, $x_i = x_1$ indica a idade do indivíduo 1; quando $i = 2$, $x_i = x_2$ indica a idade do indivíduo 2, e assim por diante, até a idade do indivíduo N . Isto é, x_i representa desde a idade do primeiro indivíduo até a idade do último indivíduo da série, já que i varia de 1 a N .

A fórmula acima é lida da seguinte maneira: “Mi” é igual ao somatório de “xis i”, com “i” variando de um a “N”, dividido pelo número de indivíduos da população.

b) para amostras:

$$\bar{x} = \frac{\sum_{i=1}^n x_i}{n};$$

Nesta fórmula a média aritmética é indicada por \bar{x} e o valor de i varia de 1 a n , em letra minúscula, para indicar um número menor de indivíduos em uma amostra.

A fórmula acima é lida como: “xis barra” é igual ao somatório de “xis i”, com “i” variando de um a “n”, dividido pelo número de indivíduos na amostra.

Quando calculamos a média de idade no nosso exemplo obtivemos o resultado 39,2 anos. Observe então que a média de uma variável dimensional como é a idade, tal como a moda, deve ser expressa na mesma dimensão. Sua dimensão é a mesma da moda, ou seja, o número de anos. Por isso, dizemos que a média de idade obtida foi 39,2 anos. Não se esqueça de que podemos expressar a idade em outras unidades de tempo, tais como décadas ou meses, a depender de qual seja a mais adequada ao nosso estudo. As medidas de tendência central irão assumir as mesmas dimensões utilizadas na medição da variável para a qual estão sendo calculadas.

Conseguiu acompanhar bem? Podemos prosseguir? Se ainda está com aquela sensação de que não entendeu tudo que poderia ou deveria, releia esta parte do livro apenas mais uma vez e depois prossiga, porque geralmente quando vamos adiante, vários pontos que não tínhamos entendido muito bem vão ficando mais claros.

Para obtermos uma média ponderada damos pesos aos valores considerados no cálculo da média. Ou seja, admitimos que os valores considerados no cálculo têm influência quantitativamente diferente sobre o resultado e, portanto, devem receber pesos diferentes nesse cálculo.

Considere, por exemplo, o modo como é calculada a nota final nos cursos de graduação de várias universidades brasileiras. Suponha que você foi aluno da disciplina bioestatística no último semestre e obteve as seguintes notas, nas três avaliações parciais realizadas ao longo do semestre, cada uma valendo de 0 a 10 pontos: 8,1; 4,3 e 7,5. Você começou bem com uma nota 8,1, depois relaxou, fez muita farra e foi muito à praia (o que não é ruim, apenas você exagerou), obtendo uma nota baixa, 4,3, recuperando-se depois do susto com um 7,5. Sua média aritmética nessas três notas foi:

$$\frac{(8,1 + 4,3 + 7,5)}{3} = \frac{19,9}{3} = 6,63 \cong 6,6 \text{ pontos}.$$

Como a média que obteve, 6,6 pontos, não atingiu 7,0 pontos, você teve que fazer uma prova final, para revisar o assunto e aprender mais. Suponha que obteve nota 8,0 pontos na prova. Para saber sua nota final na disciplina você calcula uma média ponderada, considerando todas as notas obtidas durante o curso, dando peso seis à média obtida nas avaliações parciais e peso quatro à nota obtida na prova final. O objetivo é fazer com que as avaliações parciais tenham um peso maior na sua nota final do que a nota da prova final, porque aquelas resultaram de um esforço maior e mais prolongado. Assim, sua nota final seria:

$$\begin{aligned} \text{Nota final} &= \frac{6 (\text{média nas verificações parciais}) + 4 (\text{nota na prova final})}{\text{soma dos pesos}} = \\ &= \frac{6(6,6) + 4(8,0)}{6 + 4} = \frac{39,6 + 32,0}{10} = \frac{71,6}{10} = 7,16 \cong 7,1 \text{ pontos}. \end{aligned}$$

Note que 7,16 foi aproximado para 7,1 e não para 7,2, como seria o correto utilizando-se as regras de aproximação vigentes. Contudo, a orientação em algumas universidades é a de que a aproximação seja feita simplesmente desconsiderando-se as casas decimais após o primeiro decimal. Nessa etapa da avaliação, exige-se que o aluno alcance no mínimo a nota 5,0 pontos. Ao obter 7,1 pontos, você seria aprovado na disciplina. Nós professores, ficamos geralmente muito tristes quando um aluno é aprovado, porque, afinal de contas, nosso principal objetivo é reprovar o aluno, prejudicá-lo, arrasá-lo. Brincadeira!

Entendeu o que é uma média ponderada?

Considerando w_i o peso dado a cada observação, a média ponderada pode ser expressa por

$$\text{Média ponderada populacional} = \frac{\sum_{i=1}^N w_i x_i}{\sum_{i=1}^N w_i} \quad \text{ou por} \quad \text{Média ponderada amostral} = \frac{\sum_{i=1}^n w_i x_i}{\sum_{i=1}^n w_i}.$$

As médias geométrica e harmônica não serão abordadas. A primeira nada mais é do que a média aritmética dos valores de uma variável quando esses são expressos em escala logarítmica. Uma técnica que utiliza a segunda é a famosa análise de variância de uma via, em etapa denominada de pós-teste, ou seja, após a realização da análise. Neste livro, veremos apenas como comparar duas médias aritméticas. A análise de variância é usada quando precisamos comparar mais de duas médias aritméticas, mas tal técnica não será apresentada.

Entre as medidas de tendência central já vimos até aqui a moda, a média aritmética e a média ponderada. Agora vamos apresentar a mediana.

Vamos começar com um exemplo, porque isso lhe ajudará a entender o cálculo, o significado e a aplicação da mediana. Depois, então, ficará mais fácil entender sua definição. Para facilitar os nossos cálculos, considere apenas os sete primeiros valores de idade do exemplo que temos utilizado, apresentados a seguir:

Número do indivíduo na pesquisa	Valores de idade (em anos completos)
1	36
2	45
3	40
4	34
5	41
6	40
7	41

Não se esqueça de que a mediana é mais uma das medidas de tendência central. Assim, ao se calcular uma mediana busca-se um único valor de idade para o qual converge o conjunto de valores de idade. Até o momento, já utilizamos dois modos diferentes de obter este valor: contando qual o valor que apareceu com a maior frequência (moda), somando todos os valores e dividindo esta soma pelo número de valores (média aritmética), ou calculando uma média na qual os valores apresentam pesos diferentes (média ponderada). Para calcularmos a mediana uma outra maneira é usada. A mediana é aquele valor da variável "idade" que ocupa a posição central da série.

— Já sei, então! Na planilha acima vejo que o valor de idade que ocupa a posição central é 34 anos. Então a mediana é 34 anos.

— Não. A mediana é o valor central da série, quando os valores estão organizados em ordem crescente ou decrescente.

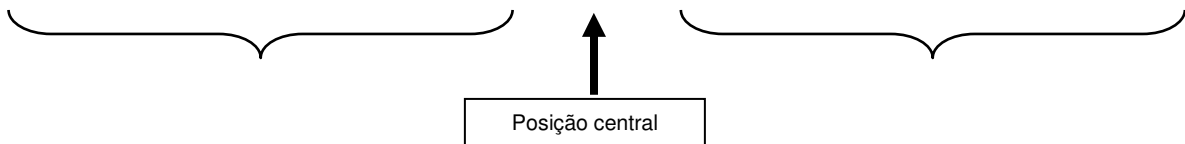
— Por que temos de ordenar os valores?

— Observe novamente a série de observações da planilha acima: 36; 45; 40; 34; 41; 40 e 41 anos. Como essa série não está ordenada, se dissermos que 34 anos é a mediana, não atingiremos nosso objetivo de obter o valor central da série de idades, pois, a série começa com 36 anos, que não é o menor valor de idade, sobe para 45 anos, desce para 40, depois para 34, voltando a subir, descer e subir em seguida. Para obtermos o valor central, o primeiro passo será organizarmos os valores de idade em ordem crescente ou decrescente. Sugerimos que você organize esses valores em ordem crescente, porque estamos mais acostumados a proceder assim. Já vimos que podemos solicitar ao computador para fazer o ordenamento que desejarmos. Abaixo apresentamos a planilha de dados com a idade ordenada crescentemente:

Número do indivíduo na pesquisa	Valores de idade (em anos completos)
4	34
1	36
3	40
6	40
5	41
7	41
2	45

É evidente que a posição central nessa série é a quarta posição, e sendo esta posição ocupada pelo valor de idade 40 anos, essa é a mediana dessa série de apenas sete valores de idade.

34	36	40	40	41	41	45
1ª posição	2ª posição	3ª posição	4ª posição	5ª posição	6ª posição	7ª posição



Posição central

— Ah! Mas isso foi evidente para uma série com um número ímpar de observações. E se a série tiver um número par de observações?

— Não há problema. Se incluíssemos mais um indivíduo no estudo cuja idade fosse 42 anos, a série de idades seria então: 36; 45; 40; 34; 41; 40; 41 e 42 anos. Note que teríamos oito observações, um número par, portanto. Organizando a série em ordem crescente obteríamos: 34; 36; 40; 40; 41; 41; 42 e 45 anos. Poderíamos olhar essa série e verificar que a posição central estaria entre os dois valores centrais da série que, neste exemplo, seriam as idades que ocupariam as quarta e quinta posições, com valores 40 e 41 anos, respectivamente.

34	36	40	40	41	41	42	45
----	----	----	----	----	----	----	----

1ª posição	2ª posição	3ª posição	4ª posição	5ª posição	6ª posição	7ª posição	8ª posição
------------	------------	------------	------------	------------	------------	------------	------------

Posição central

Em seguida, por convenção, calculamos a média aritmética desses dois valores centrais, sendo o resultado desse cálculo a mediana que estamos querendo encontrar. Assim, teríamos:

$$\text{Mediana} = \text{média aritmética dos valores centrais} = \frac{40 + 41}{2} = \frac{81}{2} = 40,5 \text{ anos}.$$

Lembre-se de que, no cálculo acima, o numerador $40 + 41$ representa a soma dos valores centrais da série de idades, e o denominador o número de valores centrais, que são dois: 40 e 41 anos. Uma média aritmética é a soma dos valores da variável dividida pelo número de valores, não é?

— **Está bem! Mas, se o número de observações for muito grande, não vai ser tão fácil encontrar a mediana.**

— É verdade. Vamos retomar nosso exemplo com 25 indivíduos, para mostrarmos como obter a mediana seja qual for o número de indivíduos estudados. Os dados são apresentados novamente a seguir:

Número do indivíduo na pesquisa	Valores de idade (em anos completos)
1	36
2	45
3	40
4	34
5	41
6	40
7	41
8	38
9	31
10	46
11	25
12	51
13	46
14	34
15	36
16	38
17	41
18	39
19	41
20	32
21	32
22	52
23	47
24	40
25	34

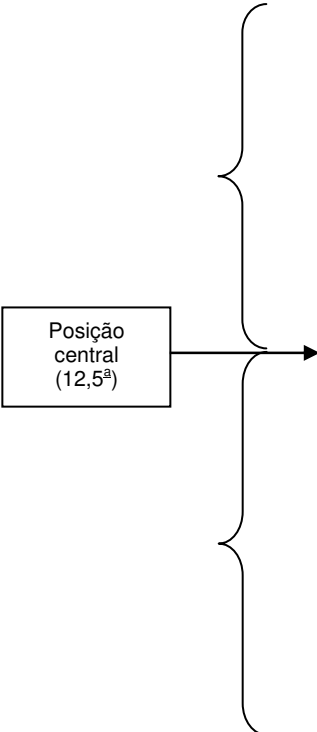
Como já vimos, o primeiro passo será a organização dos valores de idade em ordem crescente, como apresentado abaixo:

Número do indivíduo na pesquisa	Valores de idade (em anos completos)
11	25
9	31
20	32
21	32
4	34
14	34
25	34
1	36
15	36
8	38
16	38
18	39
3	40
6	40
24	40
5	41
7	41
17	41
19	41
2	45
10	46
13	46
23	47
12	51
22	52

O segundo passo será calcularmos em que posição nessa série já ordenada de valores de idade se encontra o valor central da série. O valor de idade que estiver ocupando essa posição é a mediana da série. Intuitivamente pensaríamos em obter a posição central da série dividindo o total de valores de idade ($n = 25$), ou seja, o número total de posições na série, por 2. Ao fazermos isso, estaríamos dividindo o número total de posições em duas partes iguais. Vejamos o que obteríamos:

$$\text{Posição central da série} = \frac{n}{2} = \frac{25}{2} = 12,5^{\text{a}} \text{ posição.}$$

Observe agora na planilha da próxima página os valores de idade e suas respectivas posições na série:



Posição na série	Valores de idade (em anos completos)
1ª	25
2ª	31
3ª	32
4ª	32
5ª	34
6ª	34
7ª	34
8ª	36
9ª	36
10ª	38
11ª	38
12ª	39
13ª	40
14ª	40
15ª	40
16ª	41
17ª	41
18ª	41
19ª	41
20ª	45
21ª	46
22ª	46
23ª	47
24ª	51
25ª	52

Embora tenhamos aumentado o nosso n para 25, ainda é possível encontrarmos a posição central visualmente como fizemos logo acima com um n de 7 ou 8. Note porém, que, na verdade, a posição central dessa série é a 13ª, e não a 12,5ª. Há doze posições acima da 12,5ª posição e treze abaixo. Separando visualmente as doze posições iniciais das doze últimas, verificamos que é o valor que ocupa a 13ª posição (40 anos), que divide a série em duas partes iguais, e não o valor que ocupa a 12,5ª posição, $(39 + 40) / 2 = 39,5$ anos. Assim, a mediana dessa série de idades é 40 anos e não 39,5 anos.

Observe que obteremos o resultado correto se utilizarmos a seguinte fórmula ao invés da anterior, que intuitivamente pensaríamos em usar:

$$\text{Posição central da série} = \frac{n+1}{2} = \frac{25+1}{2} = \frac{26}{2} = 13^{\text{ª}} \text{ posição.}$$

Dessa maneira, concluiremos corretamente que a mediana é o valor de idade que ocupa a 13ª posição na série ordenada, ou seja, 40 anos. Vimos então que o espectro total de posições de uma série de observações é dado por $n+1$ e não simplesmente por n como pensaríamos intuitivamente.

A fórmula $(n+1)/2$ pode ser deduzida da seguinte maneira: se dividirmos as posições da série de

valores de idade em quatro partes iguais, o total de posições corresponderia a quatro quartos ($4/4$) das posições, certo? Então, a posição que dividiria esta série em duas partes iguais equivaleria a dois quartos ($2/4$) do total de posições, ou seja, dois quartos de $(n + 1)$. Assim,

$$\text{Posição central da série} = \frac{2}{4}(n + 1) = \frac{1}{2}(n + 1) = \frac{1(n + 1)}{2} = \frac{(n + 1)}{2}^{\text{a}} \text{ posição.}$$

Se o número de observações for par, utilizaremos a mesma fórmula acima. Se o n no nosso exemplo for 26 teremos:

$$\text{Posição central da série} = \frac{n + 1}{2} = \frac{26 + 1}{2} = \frac{27}{2} = 13,5^{\text{a}} \text{ posição.}$$

O valor da mediana seria obtido então calculando-se a média aritmética dos valores de idade que ocupam as posições 13^{a} e 14^{a} : $(40 + 40)/2 = 40$ anos, já que $13,5^{\text{a}}$ nos indica uma posição intermediária entre essas posições.

E aí? Já sabe calcular a mediana de uma série de observações seja qual for o número de indivíduos estudados?

Agora já podemos definir mediana:

Mediana de um conjunto finito de valores é aquele valor que divide o conjunto em duas partes iguais, de modo que o número de valores menores ou iguais à mediana é igual ao número de valores maiores ou iguais à mediana.

Outra definição pode ser:

Mediana de uma série de observações é o número que fica exatamente no meio da série, quando os dados estão ordenados e o número de observações for ímpar, ou é a média aritmética dos dois valores do meio da série, quando o número de observações for par.

— Em quais circunstâncias deveremos usar a moda, a média ou a mediana?

— Vamos discutir as vantagens e desvantagens dessas medidas.

Já vimos que a **moda** é muito simples de ser calculada. Além disso, essa medida não é influenciada por valores extremos (valores anômalos muito altos ou muito baixos de uma determinada variável), porque estes tendem a aparecer em poucas pessoas, sendo muito improvável que um desses valores anômalos seja o mais freqüente da série. Considere, por exemplo, que você obteve os seguintes valores de idade, em anos completos, ao estudar uma amostra de nove indivíduos: 13; 14; 14; 15; 15; 15; 16; 16 e 17 anos. Nessa série

a moda seria 15 anos. Se a série apresentasse um valor extremo, p. ex., uma idade de 27 anos no lugar de 17 anos, a moda continuaria sendo 15 anos. Outra vantagem é que pode ser utilizada para resumir variáveis nominais. Por exemplo podemos verificar qual a raça mais freqüente em uma amostra ou população. Contudo, uma mesma série de observações pode apresentar mais de uma moda e já vimos que o nosso interesse é obter uma única medida que possa resumir aquele conjunto de dados. Por isso, a moda é uma medida pouco utilizada.

CARACTERÍSTICAS DA MODA	
Vantagens	Simplicidade de cálculo
	Não é afetada por valores extremos
	Pode resumir variável nominal
Desvantagem	Pode não ser única

Quanto à **média**, seu cálculo também é muito simples e apenas um único (singular) valor pode ser obtido para uma mesma série, o que já vimos ser uma característica desejável. Além disso, existem muitas técnicas estatísticas fundamentadas no cálculo da média, já bastante testadas e bem desenvolvidas. Entretanto, a média é muito vulnerável a valores extremos. Considere novamente o exemplo com nove valores de idade. A média aritmética seria dada pelo somatório das idades ($13 + 14 + 14 + 15 + 15 + 15 + 16 + 16 + 17 = 135$), dividido pelo número de indivíduos ($n = 9$), sendo então seu valor igual a 15 anos ($135 / 9 = 15$). Se a série tivesse o valor 27 no lugar de 17 (um valor mais extremo, portanto), a média aritmética seria ($13 + 14 + 14 + 15 + 15 + 15 + 16 + 16 + 27 = 145$) dividido por 9, que seria igual a 16,1 anos, valor diferente daquele obtido na série original, evidenciando a influência do valor anômalo no resultado da média.

CARACTERÍSTICAS DA MÉDIA	
Vantagens	Simplicidade de cálculo
	Singularidade
	Existem muitas técnicas disponíveis para o seu uso
Desvantagens	É muito influenciada por valores extremos
	Não pode resumir variável nominal

Quanto à **mediana**, é obtida também de maneira simples; é uma medida única para cada série de observações; há várias técnicas estatísticas baseadas em seu cálculo (embora em menor número do que aquelas baseadas na média); mas, não é uma medida vulnerável a valores extremos. Para você verificar esta última característica, vamos retomar o exemplo dos valores de idade observados para uma amostra de nove indivíduos: 13; 14; 14; 15; 15; 15; 16; 16 e 17 anos. Já vimos que a moda dessa série é 15 anos e a média também. A mediana é o valor de idade que ocupa a posição $(n + 1) / 2$, após o ordenamento de todos os valores da série. Note que os valores já estão ordenados e que $(n + 1) / 2 = (9 + 1) / 2 = 10 / 2 = 5$, indicando que a mediana é o valor que ocupa a quinta posição, nesse caso o valor 15 anos. Você deve ter notado que, nesse exemplo, os valores da moda, média e mediana, coincidem. Nem sempre os valores das medidas de

tendência central são iguais.

Agora, vamos novamente substituir a idade 17 anos por 27 anos. A série seria então: 13; 14; 14; 15; 15; 15; 16; 16 e 27 anos. Como o n continua o mesmo ($n = 9$), a mediana desta série ocupa a mesma quinta posição ($((n + 1) / 2 = (9 + 1) / 2 = 10 / 2 = 5)$) e continuará a ser 15 anos, sem influência, portanto, do valor mais extremo 27 anos. Lembre-se de que quando calculamos a média sem e com o valor extremo, essa mudou de 15 para 16,1 anos.

CARACTERÍSTICAS DA MEDIANA	
Vantagens	Simplicidade de cálculo
	Singularidade
	Não é influenciada por valores extremos
Desvantagens	Existem menos técnicas disponíveis para o seu uso
	Não pode resumir variável nominal

Considerando as vantagens e desvantagens apresentadas acima podemos dizer que, ao descrevermos ou analisarmos nossos dados, inicialmente tentamos aplicar técnicas que utilizem a média, devido à exuberância de técnicas disponíveis. Mas, se observarmos a existência de valores extremos no nosso banco de dados, revisaremos esses resultados anômalos, olhando qual a informação original constante do formulário, questionário ou instrumento de registro de respostas de entrevistas, no qual foi coletada a informação. O intuito será esclarecer se o valor na planilha eletrônica é aquele mesmo ou se houve erro de digitação do dado. Se o valor foi digitado corretamente, mas for muito extremo, fazendo-nos duvidar da sua veracidade, deveremos suspeitar de que o erro ocorreu na própria coleta do dado. Diante disso, poderemos retornar à nossa fonte de informações para obter a informação correta ou, se isso não for possível, devemos substituir aquele valor anômalo para uma determinada variável e indivíduo, por um código que indique ao computador que aquela informação foi perdida. Uma consequência indesejada deste último procedimento é a diminuição do número de indivíduos disponível para as análises envolvendo a variável para a qual perdemos a informação, o que pode comprometer a precisão estatística dos nossos resultados.

Se o valor considerado extremo for muito alto ou baixo, mas verossímil, não conseguiremos justificar a exclusão daquele valor da análise. Nesse caso, deveremos aplicar procedimentos que utilizem a mediana, já que essa não é vulnerável aos valores extremos.

Os valores anômalos podem ainda ter ocorrido porque os indivíduos que os apresentavam estavam expostos a variáveis que interagem, ou seja, o valor estava mais alto do que o esperado porque o indivíduo estava exposto a duas variáveis sendo que uma potencializava o efeito da outra sobre a doença estudada. Se essa for a situação, o resultado anômalo deve ser mantido na análise, pois reflete um aspecto importante da realidade estudada, não se constituindo em erro na coleta ou digitação dos dados.

A natureza (o tipo) da variável é outro critério que devemos considerar ao decidir utilizar a moda, a média ou a mediana. Se, por exemplo, uma das variáveis que estivermos investigando for o número de filhos do indivíduo (0, 1, 2, 3, 4,..., até k filhos), ao calcularmos a média aritmética dessa variável, poderíamos obter números fracionários que seriam incompatíveis com a realidade biológica estudada. Seria possível obter uma média de, p. ex., 3,5 filhos (três filhos e meio), o que, evidentemente, seria um absurdo em termos biológicos, pois, teríamos de admitir a existência de metade de um indivíduo. Sabemos que, muitas vezes, ao avaliarmos certos seres humanos quanto à sua responsabilidade, seriedade, caráter, generosidade, solidariedade, coerência ou ética, achamos que cada um deles é metade de um ser humano ou mesmo um terço ou um

quarto, mas geralmente não levamos isso em conta quando estamos utilizando técnicas estatísticas, embora devêssemos fazê-lo em outras atividades da nossa existência. A variável “número de filhos”, tal como categorizada no exemplo acima, é uma variável de razão, porque além de suas categorias poderem ser ordenadas, seus intervalos são regulares e “zero” filhos indica ausência de filhos. Contudo, como essa variável, pelo exposto mais acima, não pode ser expressa em uma escala contínua de valores (números fracionários), podemos classificá-la também como variável discreta (não-contínua ou descontínua). Assim, em termos estritamente estatísticos é inadequado calcularmos a média aritmética de variáveis discretas como o número de filhos, pelo motivo mencionado acima. Para variáveis desse tipo, e quando o número de observações for ímpar, é mais apropriado utilizarmos o cálculo da mediana.

— Por que apenas quando o número de observações for ímpar?

— Porque quando n é ímpar o valor da mediana será sempre discreto, não-contínuo, não-fracionário, uma vez que há uma única posição central ocupada pelo valor da mediana, que é expresso na escala original da variável. Se o número de observações for par, o valor da mediana poderá ser fracionário, já que este será obtido pelo cálculo da média aritmética dos dois valores centrais da série. Nesse caso, então, a mediana seria também imprópria, pois apresentaria um valor fracionário para uma variável originalmente discreta. Restaria para nós, nesse último caso, a possibilidade de utilizarmos a moda, que seria sempre um ou mais valores não-fracionários.

Na prática, porém, a maioria dos pesquisadores utiliza a mediana e a média mesmo quando seus resultados fracionários indicam situações biologicamente absurdas.

QUANDO UTILIZAR MODA, MÉDIA OU MEDIANA	
Moda	Série é unimodal
Média	Variável é contínua
	Série não contém valores anômalos
	Série continha valores extremos incorretos que foram corrigidos
Mediana	Série continha valores extremos que não puderam ser corrigidos e que foram retirados do banco de dados
	Variável é discreta e n é ímpar
	Série continha valores extremos que não foram retirados do banco de dados porque estavam corretos e/ou havia interação entre variáveis

— Qual o próximo assunto?

— No próximo capítulo apresentaremos as medidas de dispersão ou de variabilidade. Antes de continuar pegue um “cineminha”, uma praia, um teatro ou realize outra atividade de lazer; ou estude outras disciplinas, ou faça outras leituras importantes como os escritos de Darwin, Marx, Engels, Adam Smith, Keynes, Freud e tantos outros. Visite também seus amigos, irmãos ou seus pais. Não se esqueça de que o trabalho e o estudo são importantes, mas existem outras dimensões da existência humana que não devem ser suprimidas pela escravidão ao trabalho ou ao estudo.

CAPÍTULO 6

- O que são medidas de dispersão e quais as suas aplicações?
 - Quais as principais medidas de dispersão e como calculá-las?
 - O que são graus de liberdade?
 - O que nos faz perder graus de liberdade?
-



— O que são medidas de dispersão e quais as suas aplicações?

— Há muito tempo, os estatísticos perceberam que precisavam de outros procedimentos além daqueles apresentados por nós até o momento. Observe as séries de observações da variável “altura” (para descansarmos do exemplo das idades) apresentadas abaixo, e você logo entenderá o porquê.

Valor da altura (em metros, dois decimais)
1,45
1,62
1,42
1,80
1,55
1,81
1,82
1,83
Média = 1,66
Mediana = 1,71

Valor da altura (em metros, dois decimais)
1,44
1,72
1,40
1,79
1,48
1,83
1,70
1,94
Média = 1,66
Mediana = 1,71

Valor da altura (em metros, dois decimais)
1,34
1,82
1,30
1,89
1,38
1,93
1,60
2,04
Média = 1,66
Mediana = 1,71

Você diria que essas três séries são iguais? Evidentemente que não. O número de indivíduos (observações) em cada série é o mesmo ($n = 8$), mas veja que várias alturas em cada série são diferentes das alturas das outras séries. Se você utilizar apenas as principais medidas de tendência central, vistas no capítulo anterior, a média aritmética e a mediana, verá que as três distribuições têm a mesma média e mediana, cujos valores são 1,66 m e 1,71 m, respectivamente. Se estiver disposto, aproveite este exemplo para treinar, e calcule a média e a mediana de pelo menos uma das séries acima.

Fica claro, com esse exemplo, que precisamos de outros procedimentos que nos permitam comparar diferentes conjuntos de observações. Por isso, os estatísticos se esforçaram para desenvolver procedimentos adicionais para descrição de dados quantitativos, que indicassem outras características desses dados, além das medidas de tendência central. Alguns desses procedimentos foram elaborados com o objetivo de medir a variabilidade dos valores de uma série de observações. Vimos que as três séries de alturas acima apresentam a mesma média e mediana, mas é possível que não tenham a mesma variabilidade dos seus valores.

— Quais as principais medidas de variabilidade e como calculá-las?

— As **medidas de dispersão** (ou de variabilidade) mais utilizadas na pesquisa em saúde são: amplitude de variação, amplitude interquartil, desvio médio, variância, desvio-padrão e coeficiente de variação.

MEDIDAS DE DISPERSÃO
Amplitude
Amplitude interquartil
Desvio médio
Variância
Desvio-padrão
Coefficiente de variação

A **amplitude de variação** ou simplesmente **amplitude**, como sua denominação indica, mede o quão amplo é o intervalo entre o valor mínimo e o valor máximo de uma série de observações. Quanto maior esse intervalo, maior deve ser a dispersão dos valores, concorda? A amplitude é uma das medidas mais simples de variabilidade e pode ser usada por você para descrever ainda mais os dados que obteve, mostrando ao(à) leitor(a) o quanto esses variaram.

— **Como a amplitude é calculada?**

— Se quisermos medir a magnitude (amplitude) do intervalo entre os valores mínimo e máximo, vamos utilizar a operação matemática da subtração, diminuindo o valor máximo pelo mínimo, obtendo a distância entre esses dois valores.

Amplitude de uma variável é a magnitude do intervalo entre o valor mínimo e o valor máximo obtido para essa variável em uma série de observações.

$$\text{Amplitude} = \text{valor máximo} - \text{valor mínimo}.$$

Considere a série de observações de altura abaixo, obtida em uma pesquisa sobre intoxicação por chumbo em crianças com nove anos de idade ou menos:

Número da criança na pesquisa	Valores de altura (em metros, dois decimais)
1	1,14
2	0,86
3	1,24
4	1,17
5	0,94

Estamos considerando apenas cinco crianças para facilitar os cálculos que faremos ao longo deste capítulo. Em pesquisas concretas, utilizaremos amostras de tamanho maior. Você verá neste livro que, com números (n_s) suficientemente grandes, um pesquisador evita quase todos os possíveis problemas estatísticos em uma pesquisa quantitativa.

Organizando os valores em ordem crescente de altura temos:

Número da criança na pesquisa	Valores de altura (em metros, dois decimais)
2	0,86
5	0,94
1	1,14
4	1,17
3	1,24

Verificamos que os valores mínimo e máximo de altura são 0,86 m e 1,24 m, respectivamente. Podemos então calcular a amplitude dessa série de alturas:

$$\text{Amplitude} = \text{valor máximo} - \text{valor mínimo} = 1,24 - 0,86 = 0,38 \text{ m}.$$

Não há uma notação universalmente utilizada para a amplitude. Quando isso acontece, escrevemos o nome da medida por extenso, como fizemos acima.

Podemos apresentar o resultado obtido de duas formas: ou escrevemos que a amplitude foi 0,38 m, ou que foi de 0,86 m a 1,24 m. Na nossa opinião, a segunda maneira é mais informativa, porque o(a) leitor(a) do trabalho terá acesso aos valores mínimo e máximo, sendo que na primeira ele saberá apenas o valor da amplitude. Note que a amplitude, nesse exemplo, é uma medida dimensional (expressa em metros), porque a variável altura é dimensional.

A **amplitude interquartil** será vista no próximo capítulo (páginas 80 a 83) porque, até este ponto do livro, ainda não abordamos os conteúdos necessários para o seu cálculo.

As próximas três medidas de dispersão, **desvio médio**, **variância** e **desvio-padrão**, medem o quanto, em média, os valores da série se afastam (desviam) da média aritmética dos valores. Para obtermos essas medidas temos de primeiramente calcular a média aritmética dos valores da série, para em seguida medir o quanto cada valor se afasta (desvia) dessa média, somar todos esses desvios, e dividir o total de variabilidade pelo número de valores (indivíduos). Obtemos dessa maneira a quantidade de variação para cada um dos indivíduos estudados, ou seja, a variação média por indivíduo.

O que distingue o desvio médio do desvio-padrão é a forma utilizada para calcular essa variação média, mas o significado de ambos é o mesmo. A variância também tem o mesmo significado e nada mais é do que o quadrado do desvio-padrão, ou seja, é o desvio-padrão expresso em escala quadrática. Detalharemos este aspecto logo adiante.

Desvio médio, variância e desvio-padrão: medem o quanto, em média, os valores da série afastam-se (desviam-se, distanciam-se) da média aritmética dos valores.

A mais simples dessas três medidas é o **desvio médio**. A primeira etapa para o seu cálculo é medir o quanto cada valor da série desviou da média da série. Em seguida, soma-se o módulo desses desvios. Esta soma é finalmente dividida pelo número de valores (indivíduos) da série.

— **Por que temos de considerar o módulo dos desvios? Por que simplesmente não somamos os desvios obtidos?**

— Você já sabe que a forma mais freqüente de medirmos um desvio entre dois valores quaisquer é

através de uma operação de subtração, que vai quantificar o quanto esses dois valores se distanciam um do outro. Quanto maior a diferença entre os valores, maior o resultado da subtração de um pelo outro, e maior a distância entre os mesmos. Ao subtrair cada valor da série pelo valor médio da mesma série que, como já vimos, é a primeira etapa para o cálculo do desvio médio, o resultado será positivo (quando o valor para um determinado indivíduo for maior do que a média do grupo) ou negativo (quando o valor para um determinado indivíduo for menor do que a média do grupo). Você também já sabe que a segunda etapa será somarmos todos esses desvios. Se somarmos considerando os sinais, não obteremos o que desejamos que é o total de variação dos valores da série em relação à média da série, independentemente de se o valor variou para cima ou para baixo da média, porque os desvios positivos ao serem somados aos negativos, resultarão sempre em um total igual a zero. Para verificarmos isso, vamos retomar o exemplo da série de cinco observações de altura.

Número da criança na pesquisa	Valores de altura (em metros, dois decimais)
1	1,14
2	0,86
3	1,24
4	1,17
5	0,94

A média de altura dessa série é 1,07 m. Agora calcularemos os desvios de cada valor de altura em relação a essa média e, em seguida, somaremos esses desvios, para ilustrarmos o que dissemos acima:

$$1,14 - 1,07 = +0,07;$$

$$0,86 - 1,07 = -0,21;$$

$$1,24 - 1,07 = +0,17;$$

$$1,17 - 1,07 = +0,10; \text{ e}$$

$$0,94 - 1,07 = -0,13.$$

A soma dos desvios considerando seus sinais seria:

$$(+0,07) + (-0,21) + (+0,17) + (+0,10) + (-0,13) = 0,07 - 0,21 + 0,17 + 0,10 - 0,13 = 0,00.$$

O resultado obtido revela que a magnitude dos desvios para acima da média foi semelhante àquela dos desvios abaixo da média, de modo que, considerando os sinais, os desvios positivos anularam os negativos. Como desejamos obter o total de variação, independentemente de se essa variação ocorreu para cima ou para baixo da média, a soma correta não deveria considerar os sinais dos desvios porque, como acabamos de ver, os desvios de sinais diferentes se anulam. Uma das maneiras de neutralizarmos a influência dos sinais nas operações necessárias para calcularmos o total de variação, é considerarmos o módulo dos desvios, antes de proceder à soma dos mesmos. Assim teríamos:

$$(|1,14 - 1,07|) + (|0,86 - 1,07|) + (|1,24 - 1,07|) + (|1,17 - 1,07|) + (|0,94 - 1,07|) =$$

$$\begin{aligned}
 &= |+0,07| + |-0,21| + |+0,17| + |+0,10| + |-0,13| = \\
 &= 0,07 + 0,21 + 0,17 + 0,10 + 0,13 = 0,68 \text{ metro}.
 \end{aligned}$$

Esse resultado é aquele que realmente nos interessa, pelos motivos já expostos. Por isso, consideramos os módulos dos desvios, e não os desvios originais com seu respectivo sinal.

A próxima etapa para o cálculo do desvio médio é a divisão da soma dos desvios pelo número de observações, n , que é igual a 5:

$$\begin{aligned}
 \text{Desvio médio} &= \frac{|+0,07| + |-0,21| + |+0,17| + |+0,10| + |-0,13|}{5} = \\
 &= \frac{0,68}{5} = 0,136 \cong 0,14 \text{ m}.
 \end{aligned}$$

O resultado indica que os valores de altura variam 0,14 m, em média, em relação à média das alturas.

— E a variância?

— A **variância** utiliza outro modo de neutralização da influência dos sinais no cálculo, elevando cada desvio ao quadrado, antes de somá-los. Com isso garantimos que todos os desvios ao quadrado sejam positivos, já que mesmo os números negativos quando elevados ao quadrado dão resultados positivos. Lembre-se de que, por exemplo, $(-3)^2 = +9$. Uma outra vantagem de elevarmos os desvios ao quadrado é que os desvios maiores influenciarão mais o resultado final a ser obtido. Lembre-se que $(2)^2 = 4$ e $(3)^2 = 9$. Se 2 e 3 tivessem sido os desvios obtidos por nós, ao serem elevados ao quadrado, o desvio de valor 3 contribuirá com um valor 9 para a soma dos desvios, enquanto o valor 2 contribuirá com apenas 4, quando originalmente a diferença entre esses desvios era de apenas uma unidade ($3 - 2 = 1$, enquanto que depois da elevação ao quadrado essa diferença passa a ser $9 - 4 = 5$). Assim, para o cálculo da variância, obtemos inicialmente, com uma operação de subtração, o desvio de cada valor de altura em relação à média das alturas (tal como fizemos para o cálculo do desvio médio), mas, na próxima etapa, ao invés de tomarmos o módulo de cada desvio, elevamos cada um ao quadrado. A etapa seguinte será somarmos todos esses “desvios ao quadrado” e dividirmos essa soma por $n - 1$.

— $n - 1$? Por que não n apenas? Por que esse menos um?

— Como nosso objetivo é a obtenção do quanto há de variação por indivíduo, temos de dividir o total de variação das alturas em torno da média pelo número de indivíduos da série, concorda? Então, é natural você pensar que cada indivíduo cujo valor de altura teve a oportunidade de variar deva ser considerado no cálculo. A questão é que você está achando que todos os valores de altura estudados tiveram a liberdade de variar. Contudo, você verá que um dos valores de altura não teve essa liberdade e, portanto, não poderemos considerá-lo no denominador, sendo este então igual a $n - 1$. O “menos um” ($- 1$) corresponde justamente a

esse valor de idade que não teve a liberdade de variar.

— **Mas, o que é que faz com que uma das alturas não possa variar?**

— Uma das alturas não tem a liberdade de variar porque, para calcular a variância das alturas, precisamos antes calcular a média dos valores de altura, pois desejamos saber o quanto cada altura desviou em relação a essa média.

— **E daí?**

— O que ocorre é que, inicialmente, ao coletarmos a altura para os cinco indivíduos estudados, o valor de altura para cada um deles teve liberdade total para assumir qualquer valor de altura possível para um ser humano. Não havia nenhum critério prévio que obrigasse qualquer um dos valores a ter um determinado valor pré-fixado. Você concorda com isso?

Posteriormente, contudo, se quisermos calcular a variância das alturas, teremos de calcular previamente sua média e, ao calcularmos a média, criamos uma condição limitante para os valores de altura.

— **Condição limitante?**

— Sim. Ao calcularmos a média das alturas, \bar{x} , os valores de altura passam a ser obrigados a atender à condição de que a soma de seus valores dividida pelo número de indivíduos estudados tenha que ser igual a um determinado valor. No nosso exemplo temos:

$$\begin{aligned}\bar{x} &= \frac{1,14 + 0,86 + 1,24 + 1,17 + 0,94}{5} = \\ &= \frac{5,35}{5} = 1,07 \text{ m}.\end{aligned}$$

Ou seja, ao calcularmos a média, criamos a condição de que a soma dos valores de altura dividida por 5 seja igual a 1,07 m, e essa condição é limitante, porque ao calcularmos a média, os quatro primeiros valores a serem considerados nesse cálculo continuam tendo a liberdade de variar (de assumir qualquer dos valores possíveis de altura), mas o último valor de altura a entrar no cálculo da média é um valor já fixado pelos quatro valores de altura já colocados no cálculo, porque a soma dos cinco valores de altura tem que dar 5,35, pois esse é o valor obrigatório para a soma das alturas, já que quando dividirmos a soma por 5, o resultado terá que ser 1,07 m. Se não respeitarmos tal condição, estaremos aceitando alterar o resultado que realmente obtivemos para \bar{x} que foi 1,07 m, e será um absurdo estarmos dispostos a modificar os resultados do nosso estudo, não é?

Considere novamente o cálculo da média das alturas:

$$\begin{aligned}\bar{x} &= \frac{1,14 + 0,86 + 1,24 + 1,17 + 0,94}{5} = \\ &= \frac{5,35}{5} = 1,07 \text{ m}.\end{aligned}$$

O último valor colocado no cálculo teve que ser obrigatoriamente 0,94 para que a média seja 1,07 m. Se alterássemos a ordem de entrada das alturas no cálculo, e nada nos impede de fazer isso, porque o resultado final terá que ser o mesmo, encontraríamos:

$$\begin{aligned}\bar{x} &= \frac{0,86 + 1,24 + 1,17 + 0,94 + 1,14}{5} = \\ &= \frac{5,35}{5} = 1,07 \text{ m}.\end{aligned}$$

O último valor colocado no cálculo teria que ser obrigatoriamente 1,14 para que a média fosse 1,07 m. Os quatro primeiros valores teriam a liberdade de assumir qualquer valor possível de altura, mas o último não, porque seu valor teria que ser exatamente aquele que satisfizesse a condição limitante.

Entendeu? Se não entendeu essa perda de **graus de liberdade**, leia o APÊNDICE 1 no final do livro, no qual esse tópico é explicado mais detalhadamente.

Pelo que vimos, a variância, em notações amostrais, pode ser expressa por:

$$s^2 = \frac{\sum_{i=1}^n (x_i - \bar{x})^2}{n - 1};$$

e, em notações populacionais, por

$$\sigma^2 = \frac{\sum_{i=1}^N (x_i - \mu)^2}{N}.$$

— Mas, no denominador da segunda fórmula, não deveríamos colocar $N - 1$, pelo que acabamos de ver?

— Teoricamente sim, você está absolutamente certo(a). Deveríamos colocar $N - 1$ no denominador, mas você verá que, na prática, isso só nos interessará quando escolhermos uma amostra e quisermos utilizar a variância dessa amostra, s^2 , como um estimador da variância populacional, σ^2 . A variância amostral só será um estimador não-viciado da variância populacional, se dividirmos seu numerador por $n - 1$. Além disso, toda a teoria estatística que nos orienta nesse assunto foi desenvolvida admitindo-se uma população infinita e, evidentemente, nessa situação, tanto μ quanto σ^2 não são calculáveis. Entretanto, podemos calcular média e variância de populações finitas e, nesse caso, como o número de indivíduos, N , é muito grande, retirar uma unidade de um número muito grande é matematicamente irrelevante. Assim, podemos considerar que, como N é muito grande, $N - 1 \cong N$. Então, ao estudarmos uma população grande, não é necessário considerarmos a perda de um grau de liberdade, porque o resultado será praticamente o mesmo.

Calculando a variância para o nosso exemplo, onde $n = 5$, temos que

$$s^2 = \frac{\sum_{i=1}^n (x_i - \bar{x})^2}{n-1}$$

$$= \frac{(1,14 - 1,07)^2 + (0,86 - 1,07)^2 + (1,24 - 1,07)^2 + (1,17 - 1,07)^2 + (0,94 - 1,07)^2}{5-1}$$

$$= \frac{0,0049 + 0,0441 + 0,0289 + 0,01 + 0,0169}{4} = \frac{0,1048}{4} = 0,0262 \cong 0,03 \text{ m}^2.$$

Observe que, como os termos contidos no numerador foram elevados ao quadrado, a dimensão da variância também é elevada ao quadrado, sendo expressa em m^2 .

— E por que no cálculo do desvio médio não dividimos por $n-1$?

— Boa pergunta! Concordamos com você. O denominador do desvio médio deve ser $n-1$ para seu cálculo em amostras, pelo mesmo motivo apresentado para a variância. Acontece que, enquanto a variância é muito utilizada para inferência estatística, o desvio médio não é, devido a dificuldades matemáticas e, então, na prática, torna-se irrelevante, no caso do desvio médio, essa questão de n ou $n-1$ no denominador.

Finalmente, para obtermos o **desvio-padrão** extraímos a raiz quadrada da variância. Assim, em notações amostrais, o desvio-padrão é expresso por:

$$s = \sqrt{s^2} = \sqrt{\frac{\sum_{i=1}^n (x_i - \bar{x})^2}{n-1}};$$

e, em notações populacionais, por

$$\sigma = \sqrt{\sigma^2} = \sqrt{\frac{\sum_{i=1}^N (x_i - \mu)^2}{N-1}} \quad \text{ou} \quad \sigma = \sqrt{\sigma^2} = \sqrt{\frac{\sum_{i=1}^N (x_i - \mu)^2}{N}}.$$

Isso é o mesmo que dizermos que a variância é o desvio-padrão ao quadrado, ou seja, é o desvio-padrão expresso em escala quadrática. Ao calcular a variância você viu que elevamos cada desvio ao quadrado, por isso essa medida é expressa em escala quadrática (m^2). É, portanto, necessário extrairmos a raiz quadrada da variância para obtermos uma medida de dispersão na escala original, não-quadrática.

Calculando o desvio-padrão no nosso exemplo, obtemos:

$$s = \sqrt{s^2} = \sqrt{0,03} \cong 0,17 \text{ m}.$$

Observe que a dimensão do desvio-padrão das alturas é metros e não metros ao quadrado, que é a

dimensão da variância das alturas, como já tínhamos visto.

Então, concluindo esta parte, não se esqueça: desvio médio, variância e desvio-padrão são essencialmente a mesma medida de dispersão, isto é, medem o quanto, em média, os valores de uma determinada variável afastaram-se, distanciaram-se, desviaram-se, da média dos valores dessa variável. A diferença entre essas medidas está na maneira como são calculadas ou na sua expressão em escala quadrática ou não. Na prática, as mais utilizadas são a variância e o desvio-padrão.

Outra medida de variabilidade é o **coeficiente de variação**, denotado por CV . Suponha que, além das alturas, tenhamos coletado os pesos das cinco crianças do nosso exemplo, e que os valores obtidos tenham sido os seguintes:

Número da criança na pesquisa	Valores de altura (em metros, dois decimais)	Valores de peso (em quilogramas, dois decimais)
1	1,14	20,70
2	0,86	15,40
3	1,24	21,40
4	1,17	21,10
5	0,94	17,45

Podemos descrever esses resultados utilizando o cálculo de freqüências, moda, média, mediana, amplitude, desvio médio, variância e desvio-padrão, mas se quisermos enriquecer nossa descrição verificando qual dessas variáveis variou mais, teremos que lançar mão do cálculo do coeficiente de variação.

— Mas, por que precisamos de uma medida diferente das que já vimos? Por que simplesmente não calculamos a amplitude, o desvio médio, a variância ou o desvio-padrão para cada uma das variáveis, e comparamos os resultados? Aquela variável que apresentar os maiores valores dessas medidas de dispersão será aquela que variou mais na população ou amostra estudada!

— Não. Se você proceder assim poderá chegar a conclusões incorretas. Considere o nosso exemplo. Já vimos que o desvio-padrão e a média das alturas são 0,17 m e 1,07 m, respectivamente. Calculando o desvio-padrão e a média dos pesos, veremos que esses valores são 2,66 kg e 19,21 kg, respectivamente. Com base nesses valores, diríamos que a variável que variou mais foi o peso, mas não podemos afirmar isso, porque as alturas e os pesos possuem dimensões diferentes. Assim, o desvio-padrão dos pesos pode ser maior que o das alturas simplesmente porque a magnitude da escala de valores em que os pesos são medidos é maior do que aquela em que as alturas são medidas, e não porque os primeiros variem mais que os últimos. A altura é medida em metros e o peso em kg e o espectro de valores possíveis para cada uma dessas variáveis é diferente: a altura de seres humanos pode variar de zero até cerca de 2 m, em pessoas consideradas normais; já o peso de zero a, digamos, no máximo, um pouco mais de 100 kg. Do ponto de vista numérico, a variabilidade dos pesos tenderá a ser quantitativamente maior do que a da altura, porque a escala de variação dos primeiros é maior e, não necessariamente, porque variam mais. Por isso, deve-se utilizar o coeficiente de variação, que é uma maneira de padronizarmos os desvios-padrão dos pesos e das alturas, de modo a podermos comparar a variabilidade dessas variáveis.

O coeficiente de variação é dado por:

$$CV = \left(\frac{\text{desvio-padrão}}{\text{média}} \right) 100 = \left(\frac{s}{\bar{x}} \right) 100.$$

Então, o coeficiente de variação indica quantos por cento o valor do desvio-padrão de uma variável é, em relação ao valor da média daquela mesma variável. Devemos calcular o coeficiente de variação para cada uma das variáveis, e depois podemos comparar esses coeficientes para avaliar qual das variáveis variou mais, porque dessa maneira estamos utilizando uma medida de variabilidade padronizada, na qual cada desvio-padrão é expresso como porcentual da média da própria variável.

No nosso exemplo temos que

$$CV_{\text{altura}} = \left(\frac{\text{desvio-padrão da altura}}{\text{média da altura}} \right) 100 = \left(\frac{s_{\text{altura}}}{\bar{x}_{\text{altura}}} \right) 100 = \left(\frac{0,17}{1,07} \right) 100 \cong 15,89 \% ;$$

e que

$$CV_{\text{peso}} = \left(\frac{\text{desvio-padrão do peso}}{\text{média do peso}} \right) 100 = \left(\frac{s_{\text{peso}}}{\bar{x}_{\text{peso}}} \right) 100 = \left(\frac{2,66}{19,21} \right) 100 \cong 13,85 \% .$$

Utilizando o procedimento correto, concluímos que a altura variou mais que o peso. Veja que chegamos a uma conclusão diferente daquela à qual chegaríamos se, incorretamente, tivéssemos comparado diretamente os valores dos desvios-padrão dessas variáveis.

O coeficiente de variação deve ser usado também quando queremos comparar a variabilidade de uma mesma variável em trechos distintos do espectro de variação dessa variável. Se desejássemos saber, por exemplo, se a variável “idade” variou mais em indivíduos menores de 19 anos do que naqueles com 20 anos ou mais de idade, teríamos que calcular o coeficiente de variação para os dois subgrupos de idades, porque até os 19 anos haveria um intervalo de valores, dentro dos quais a idade poderia variar, muito menor do que o intervalo acima de 20 anos, no qual as idades poderiam atingir valores de até mais de 100 anos. Então, para chegarmos a uma conclusão teríamos de calcular o CV da idade nos menores de 19 anos e o CV naqueles com 20 anos ou mais, e depois compararíamos esses coeficientes para sabermos em que grupo a variação foi maior.

Outra aplicação importante do CV é na avaliação da precisão (reprodutibilidade) de testes diagnósticos, p. ex., do hemograma e da dosagem da glicemia. Se quisermos comparar a precisão de dois ou mais testes, calculamos o CV de cada um, sendo que aquele que apresentar o menor CV será o mais preciso, por ser o que variou menos seus resultados, quando repetido por diferentes observadores ou em diferentes momentos. Lembre-se de que um teste diagnóstico muito impreciso pode levar a diagnósticos incorretos, com consequências relevantes para seus pacientes.

— **Qual das medidas de dispersão devemos utilizar?**

— Como acabamos de ver acima, o CV tem aplicações específicas nas situações já mencionadas.

Quanto à amplitude, vimos que é uma medida muito incompleta, porque não leva em conta todas as observações da série, mas apenas os valores mínimo e máximo. Sugerimos usá-la como uma informação complementar às outras medidas de dispersão.

O desvio médio, o desvio-padrão e a variância são mais completos, pois levam em conta todos os valores da série de observações. Tais medidas indicam a mesma coisa, mas os dois primeiros são calculados de maneira diferente, e a última nada mais é do que o desvio-padrão expresso em escala quadrática. Há autores (*Guedes MLS, Guedes JS. Bioestatística para profissionais de saúde. Rio de Janeiro (RJ): Ao livro técnico; 1988*) que consideram uma vantagem do desvio-padrão e da variância o fato desses elevarem ao quadrado cada desvio da média, pois com isso os desvios maiores influenciam mais seu resultado do que os menores.

Os quadros abaixo mostram um resumo das vantagens e desvantagens das medidas de dispersão:

CARACTERÍSTICAS DA AMPLITUDE	
Vantagem	Simplicidade
Desvantagens	Não leva em conta todos os valores da série, mas apenas os valores mínimo e máximo
	Existem menos técnicas estatísticas que a utilizam
	É influenciada por valores extremos

CARACTERÍSTICAS DO DESVIO MÉDIO	
Vantagens	Leva em conta todos os valores da série
	Ao somar os módulos dos desvios, expressa o total de variabilidade em torno da média
Desvantagens	Cálculo é menos simples do que o da amplitude
	Os desvios maiores não influenciam bem mais seu resultado do que os menores
	Existem menos técnicas estatísticas que o utilizam quando comparado ao desvio-padrão e a variância
	É influenciado por valores extremos

CARACTERÍSTICAS DO DESVIO-PADRÃO E VARIÂNCIA	
Vantagens	Levam em conta todos os valores da série
	Ao somar os quadrados dos desvios, expressam o total de variabilidade em torno da média
	Os desvios maiores influenciam bem mais seus resultados do que os menores
	Existem muitas técnicas estatísticas que os utilizam
Desvantagens	Cálculos são menos simples do que o da amplitude
	São influenciados por valores extremos
	A variância é expressa em escala quadrática, à qual estamos menos acostumados

Já vimos até agora como utilizarmos vários procedimentos da Estatística Descritiva (cálculo de frequências, de medidas de tendência central e de dispersão). No próximo capítulo abordaremos mais um desses procedimentos: as medidas de posição.

CAPÍTULO 7

-
- Quais as principais medidas de posição?
 - Como os quartis são calculados?
 - Quais as principais aplicações dos percentis?
-



– Quais as principais medidas de posição?

– As principais medidas de posição são a **média**, a **mediana** e o **porcentil**.

MEDIDAS DE POSIÇÃO
Média
Mediana
Porcentil

– **Média e mediana? Mas, essas não são medidas de tendência central?**

– A **média** e a **mediana** além de serem medidas de tendência central são também medidas de posição. Você já entendeu porque a média e a mediana são medidas de tendência central, certo? Agora você ficará sabendo que essas medidas também podem ser usadas pelos estatísticos para avaliarem em que posição, no espectro de valores possíveis para uma determinada variável, se encontra o valor dessa variável para um certo indivíduo. Assim, poderemos verificar se o valor obtido para uma pessoa está posicionado acima ou abaixo da média ou da mediana. Provavelmente você mesmo(a) já utilizou a média como uma referência para avaliação do quanto alto ou baixo era um determinado valor: “Minha nota naquela prova foi boa, ficou acima da média da turma”.

Considere a série de 75 observações de peso de crianças com menos de 9 anos de idade, obtida em uma amostra aleatória, apresentada abaixo:

Número da criança na pesquisa	Valores de peso (em kg, dois decimais)	Número da criança na pesquisa	Valores de peso (em kg, dois decimais)	Número da criança na pesquisa	Valores de peso (em kg, dois decimais)
1	17,90	26	22,75	51	13,30
2	17,75	27	21,65	52	11,75
3	15,25	28	10,45	53	10,95
4	14,95	29	18,05	54	23,90
5	14,25	30	15,10	55	20,10
6	12,55	31	17,70	56	16,35
7	22,00	32	11,70	57	13,30
8	15,65	33	10,25	58	10,20
9	14,20	34	99,99	59	20,70
10	15,15	35	99,99	60	17,60
11	13,55	36	19,20	61	15,60
12	10,08	37	15,90	62	25,20
13	99,99	38	12,20	63	20,30
14	18,80	39	24,10	64	19,95
15	12,10	40	16,90	65	15,40
16	9,90	41	18,30	66	25,90
17	19,85	42	14,20	67	19,40
18	14,30	43	12,10	68	16,55
19	14,35	44	9,50	69	13,30
20	9,50	45	17,35	70	12,90
21	15,55	46	16,90	71	6,30
22	11,40	47	12,25	72	8,35
23	17,55	48	8,65	73	24,25
24	13,25	49	21,00	74	6,45
25	12,55	50	11,85	75	13,45

— Mas, tem alguma coisa estranha nesses dados! Três crianças têm quase 100 kg (99,99 kg) de peso!

— O valor 99,99 kg foi especificado pelos pesquisadores, como o valor de peso para crianças cujo peso não foi obtido (valor perdido ou valor não obtido ou valor ignorado). O programa de computador que você irá utilizar para processar seus dados, automaticamente excluirá esses valores quando você solicitar as estatísticas para descrição, análise ou inferência. Assim, na verdade, no exemplo acima $n = 72$ e não 75, porque há três crianças cujos pesos não foram obtidos. Se os valores 99,99 fossem utilizados nas estatísticas elas estariam incorretas. A média, por exemplo, ficaria superestimada, pois os três valores 99,99, ao serem somados aos outros no numerador, aumentariam artificialmente o numerador, alterando o resultado final. Não se esqueça de que o valor 99,99 foi um valor definido arbitrariamente pelos pesquisadores que, ao fazerem isso, devem ter o cuidado de não escolher um valor que possa ser encontrado em pelo menos um dos indivíduos estudados. Eles escolheram 99,99 justamente porque é um valor tão alto, que fica garantido que nenhuma criança apresentará esse peso, de modo que, corretamente, o computador não o utilizará nos cálculos dos percentuais válidos, a serem utilizados na redação do trabalho. Se uma das crianças estudadas no nosso exemplo pudesse apresentar um peso tão alto como 99,99 kg, esse valor não poderia ser utilizado para expressar um valor perdido, porque seria um valor válido que não poderia ser descartado nos cálculos. Neste caso, os pesquisadores teriam que utilizar, p. ex., o valor 999,99 kg, e assim estariam seguros de que nenhuma criança estudada apresentaria tal peso. No nosso exemplo, utilizamos o valor 99,99 kg, esperando que nenhuma criança até 9 anos de idade apresente um peso tão alto.

Você já sabe como calcular a média aritmética de uma variável e pode verificar facilmente que essa média para os pesos apresentados acima é 15,44 kg. Podemos utilizar esse valor como uma referência para avaliarmos o quanto alto ou baixo é o peso de um indivíduo qualquer dessa amostra. Por exemplo, se desejarmos julgar se o peso 25,10 kg é alto ou baixo, sem uma referência para nos orientar, nossa avaliação ficaria muito difícil. Mas, se usarmos a média dos pesos como uma referência, poderemos dizer que o peso 25,10 kg encontra-se bem acima da média, sendo, portanto, um valor relativamente alto. O peso 10,43 kg seria considerado relativamente baixo por estar abaixo da média. Assim, a média aritmética pode ser utilizada para verificarmos em que posição, alta ou baixa, se encontra determinado peso no espectro de valores possíveis para os pesos de crianças com até 9 anos de idade.

Podemos também utilizar a mediana como medida de posição, do mesmo modo como acabamos de fazer com a média, só que nesse caso diremos que o valor de certo indivíduo está abaixo ou acima da mediana. Você já sabe calcular a mediana e poderá confirmar que seu valor para os pesos do nosso exemplo é aproximadamente 15,12 kg. Valores acima desse serão considerados mais altos, e vice-versa.

— E o que são percentis?

— Existe uma medida tipicamente de posição que é o percentil. Conforme sua denominação indica, os percentis são valores que separam os valores de uma série de observações em duas ou mais partes, delimitando um certo percentual de valores abaixo, acima ou entre eles.

Porcentis são valores que separam os valores de uma série de observações em duas ou mais partes, delimitando um certo percentual de valores abaixo, acima ou entre eles.

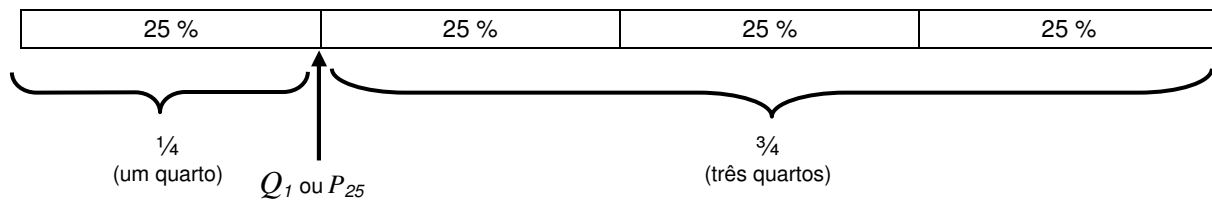
Existem vários tipos de percentis. Um dos tipos mais utilizados são os **quartis**. Esses são compostos pelo primeiro, segundo e terceiro quartis.

Depois de ordenarmos crescentemente a série de pesos do nosso exemplo, o valor que separa os valores correspondentes ao quarto (25%) de valores mais baixos dos três quartos (75%) de valores mais altos, é o **primeiro quartil**, também denominado de **quartil 1** ou **percentil 25**, denotado por Q_1 ou P_{25} .

— **Por que esse percentil é chamado de primeiro quartil?**

— Porque esse percentil separa o primeiro quarto de valores mais baixos dos três quartos subseqüentes de valores mais altos.

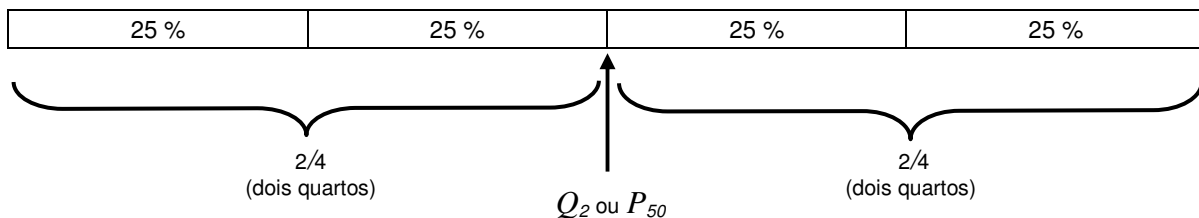
A figura abaixo representa o percentil 25 ou primeiro quartil da série de pesos:



O **segundo quartil**, também denominado de **quartil 2**, ou **percentil 50**, denotado por Q_2 ou P_{50} , é aquele valor de peso que separa os valores correspondentes aos dois quartos (50%) de valores mais baixos dos dois quartos (50%) de valores mais altos, na série ordenada de pesos.

— **Mas, isso é a mediana!**

— Exatamente! O segundo quartil é igual à mediana! Por isso, o segundo quartil é também chamado de **quartil mediano**. Não devemos estranhar, portanto, que a mediana seja também uma medida de posição, uma vez que é um tipo de percentil. O quartil mediano é apresentado na figura abaixo:



O **terceiro quartil**, também denominado de **quartil 3**, ou **percentil 75**, denotado por Q_3 ou P_{75} , é aquele valor de peso que separa os valores correspondentes aos três quartos (75%) de valores mais baixos do quarto (25%) de valores mais altos, na série ordenada de pesos. Veja na figura abaixo:



— **Existem outros percentis, além dos quartis?**

— Sim. Podemos calcular qualquer percentil que seja útil à descrição, análise ou inferência estatística, tais como os P_{10} , P_{90} , P_{95} , etc. Voltaremos a esse tópico mais adiante.

— **Como os quartis são calculados?**

— O primeiro passo será colocarmos os valores de peso em ordem crescente ou decrescente. Se desejarmos encontrar o valor de peso que separa o quarto dos valores mais baixos dos três quartos mais altos, será necessário colocarmos os valores em ordem. Propomos a você ordenar os valores em ordem crescente, como fizemos para o cálculo da mediana, no capítulo 5 (páginas 47 a 52). Tudo bem?

Na planilha a seguir apresentamos os valores de peso em ordem crescente:

Número da criança na pesquisa	Valores de peso (em kg, dois decimais)	Número da criança na pesquisa	Valores de peso (em kg, dois decimais)	Número da criança na pesquisa	Valores de peso (em kg, dois decimais)
71	6,30	57	13,30	31	17,70
74	6,45	69	13,30	2	17,75
72	8,35	75	13,45	1	17,90
48	8,65	11	13,55	29	18,05
20	9,50	9	14,20	41	18,30
44	9,50	42	14,20	14	18,80
16	9,90	5	14,25	36	19,20
12	10,08	18	14,30	67	19,40
58	10,20	19	14,35	17	19,85
33	10,25	4	14,95	64	19,95
28	10,45	30	15,10	55	20,10
53	10,95	10	15,15	63	20,30
22	11,40	3	15,25	59	20,70
32	11,70	65	15,40	49	21,00
52	11,75	21	15,55	27	21,65
50	11,85	61	15,60	7	22,00
15	12,10	8	15,65	26	22,75
43	12,10	37	15,90	54	23,90
38	12,20	56	16,35	39	24,10
47	12,25	68	16,55	73	24,25
6	12,55	40	16,90	62	25,20
25	12,55	46	16,90	66	25,90
70	12,90	45	17,35	13	99,99
24	13,25	23	17,55	34	99,99
51	13,30	60	17,60	35	99,99

Não se esqueça de que os três últimos valores de peso (99,99 kg) não serão considerados no procedimento para encontrarmos o valor do primeiro quartil.

Pronto! Agora só falta encontrarmos, nessa série ordenada de pesos, o valor que separa o quarto dos valores mais baixos dos três quartos mais altos. Para isso, pegamos o total de valores possíveis de peso nessa série, n (que, como já vimos, deverá ser representado por $n + 1$), e achamos qual a posição que está a um quarto ($1/4$) da posição inicial da série ordenada:

$$\frac{1}{4}(n+1) = \frac{n+1}{4}$$

O resultado desse cálculo não nos dá o valor do primeiro quartil, mas sua posição na série ordenada de pesos. O primeiro quartil é aquele valor de peso que está ocupando a $(n+1)/4$ ésima posição, na série de pesos já ordenada de modo crescente.

No nosso exemplo, como $n = 72$, temos que o Q_1 é aquele valor que ocupa a

$$\frac{n+1}{4} = \frac{72+1}{4} = \frac{73}{4} = 18,25^{\text{a}} \text{ posição.}$$

O resultado nos indica que o primeiro quartil é o valor de peso que ocupa a 18,25ª posição na série ordenada de pesos. Note que não podemos escrever “kg” após o 18,25, porque esse ainda não é o valor do Q_1 . Resta-nos então contar da primeira até a 18,25 ésima posição para acharmos Q_1 .

A planilha abaixo apresenta os valores ordenados de peso e suas respectivas posições na série ordenada:

Posição do peso da criança na série ordenada	Valores de peso (em kg, dois decimais)	Posição do peso da criança na série ordenada	Valores de peso (em kg, dois decimais)	Posição do peso da criança na série ordenada	Valores de peso (em kg, dois decimais)
1ª	6,30	26ª	13,30	51ª	17,70
2ª	6,45	27ª	13,30	52ª	17,75
3ª	8,35	28ª	13,45	53ª	17,90
4ª	8,65	29ª	13,55	54ª	18,05
5ª	9,50	30ª	14,20	55ª	18,30
6ª	9,50	31ª	14,20	56ª	18,80
7ª	9,90	32ª	14,25	57ª	19,20
8ª	10,08	33ª	14,30	58ª	19,40
9ª	10,20	34ª	14,35	59ª	19,85
10ª	10,25	35ª	14,95	60ª	19,95
11ª	10,45	36ª	15,10	61ª	20,10
12ª	10,95	37ª	15,15	62ª	20,30
13ª	11,40	38ª	15,25	63ª	20,70
14ª	11,70	39ª	15,40	64ª	21,00
15ª	11,75	40ª	15,55	65ª	21,65
16ª	11,85	41ª	15,60	66ª	22,00
17ª	12,10	42ª	15,65	67ª	22,75
18ª	12,10	43ª	15,90	68ª	23,90
19ª	12,20	44ª	16,35	69ª	24,10
20ª	12,25	45ª	16,55	70ª	24,25
21ª	12,55	46ª	16,90	71ª	25,20
22ª	12,55	47ª	16,90	72ª	25,90
23ª	12,90	48ª	17,35		
24ª	13,25	49ª	17,55		
25ª	13,30	50ª	17,60		

Vemos que a 18,25 ésima posição está entre as décima oitava e décima nona posições. Os valores

que ocupam as décima oitava e décima nona posições são 12,10 kg e 12,20 kg, respectivamente. Há um intervalo de 0,10 kg (12,20 kg – 12,10 kg) entre os valores que ocupam essas posições. O Q_1 está 0,25 (uma quarta parte) além do valor 12,10 kg (que é o valor que ocupa a 18ª posição) nesse intervalo (lembre-se de que o Q_1 ocupa a 18,25ª posição). Como 0,25 (um quarto) de 0,10 é $\frac{1}{4}(0,10) = \frac{1(0,10)}{4} = \frac{0,10}{4} = 0,025$, temos que

$$Q_1 = 12,10 + 0,025 = 12,125 \cong 12,12 \text{ kg}.$$

Concluindo, encontramos que o primeiro quartil é aproximadamente 12,12 kg, ou seja, esse é o valor abaixo do qual está o quarto de valores mais baixos de peso e acima do qual estão três quartos de valores mais altos da série observada. Entendeu?

Agora vamos mostrar como se calcula o **segundo quartil**, embora você já tenha aprendido a fazer isso, porque já vimos como se calcula a mediana. Vamos revisar esse cálculo?

As etapas são as mesmas: colocamos os pesos em ordem crescente; calculamos em que posição se encontra o Q_2 ; e contamos da primeira até a posição calculada para acharmos o valor do Q_2 . O ordenamento dos pesos já foi feito. Para calcularmos em que posição o Q_2 se encontra na série ordenada dos pesos, consideramos o total de valores possíveis de peso na série estudada, $n + 1$, e verificamos qual a posição que está a dois quartos ($2/4$) da posição inicial dessa série de valores ordenados. Dois quartos ($2/4$) não é igual a um meio ($1/2$)? E não é isto que queremos: a posição que separa os 50% (metade) correspondentes aos valores mais baixos dos 50% (metade) correspondentes aos valores mais altos, ou seja, que divide ao meio a série ordenada de valores de peso? Assim, temos que o Q_2 é aquele valor que ocupa a

$$\frac{2}{4}(n+1) = \frac{2(n+1)}{4} = \frac{1(n+1)}{2} = \frac{n+1}{2} \text{ éima posição}.$$

No nosso exemplo, o Q_2 ocupa a

$$\frac{2}{4}(72+1) = \frac{2(73)}{4} = \frac{1(73)}{2} = \frac{73}{2} = 36,5^a \text{ posição}.$$

Na planilha a seguir vemos que, como o Q_2 está na trigésima sexta e meia posição da série ordenada de pesos, está entre as 36ª e 37ª posições, ou seja, entre os valores 15,10 kg e 15,15 kg:

Posição do peso da criança na série ordenada	Valores de peso (em kg, dois decimais)	Posição do peso da criança na série ordenada	Valores de peso (em kg, dois decimais)	Posição do peso da criança na série ordenada	Valores de peso (em kg, dois decimais)
1ª	6,30	26ª	13,30	51ª	17,70
2ª	6,45	27ª	13,30	52ª	17,75
3ª	8,35	28ª	13,45	53ª	17,90
4ª	8,65	29ª	13,55	54ª	18,05
5ª	9,50	30ª	14,20	55ª	18,30
6ª	9,50	31ª	14,20	56ª	18,80
7ª	9,90	32ª	14,25	57ª	19,20
8ª	10,08	33ª	14,30	58ª	19,40
9ª	10,20	34ª	14,35	59ª	19,85
10ª	10,25	35ª	14,95	60ª	19,95
11ª	10,45	36ª	15,10	61ª	20,10
12ª	10,95	37ª	15,15	62ª	20,30
13ª	11,40	38ª	15,25	63ª	20,70
14ª	11,70	39ª	15,40	64ª	21,00
15ª	11,75	40ª	15,55	65ª	21,65
16ª	11,85	41ª	15,60	66ª	22,00
17ª	12,10	42ª	15,65	67ª	22,75
18ª	12,10	43ª	15,90	68ª	23,90
19ª	12,20	44ª	16,35	69ª	24,10
20ª	12,25	45ª	16,55	70ª	24,25
21ª	12,55	46ª	16,90	71ª	25,20
22ª	12,55	47ª	16,90	72ª	25,90
23ª	12,90	48ª	17,35		
24ª	13,25	49ª	17,55		
25ª	13,30	50ª	17,60		

Para encontrarmos o valor entre 15,10 kg e 15,15 kg calculamos a média aritmética desses dois valores, obtendo:

$$Q_2 = \frac{15,10 + 15,15}{2} = 15,125 \cong 15,12 \text{ kg}.$$

O mesmo resultado será obtido se considerarmos que o segundo quartil está na metade (1/2) do intervalo entre os valores 15,10 kg e 15,15 kg:

$$\text{Um meio de } 0,05 = \frac{1}{2}(0,05) = \frac{1(0,05)}{2} = \frac{0,05}{2} = 0,025 \text{ kg}.$$

Dessa maneira,

$$Q_2 = 15,10 + 0,025 = 15,125 \cong 15,12 \text{ kg}$$

Valor que ocupa a 36ª posição Dois quartos (um meio, metade) do intervalo entre as 36ª e 37ª posições Valor que ocupa a 36,5ª posição

Como acabamos de ver, o segundo quartil da série de 72 observações de peso do nosso exemplo é

aproximadamente 15,12 kg.

O terceiro quartil é calculado também de modo semelhante.

Como as observações já estão em ordem crescente, calculamos em que posição se encontra o Q_3 . Para isso, calculamos qual posição está a três quartos (3/4) do primeiro valor da série. Essa posição será dada por

$$\frac{3}{4}(n+1) = \frac{3}{4}(n+1) \text{ ésim}a \text{ posição}.$$

No exemplo, temos que o terceiro quartil é o peso que ocupa a

$$\frac{3}{4}(72+1) = \frac{3}{4}(73) = \frac{219}{4} = 54,75^a \text{ posição}.$$

Como a 54,75ª posição encontra-se entre as 54ª e 55ª posições, vemos na planilha abaixo que o terceiro quartil está entre os valores 18,05 kg e 18,30 kg, ou, mais especificamente, está 0,75 unidade entre 18,05 kg e 18,30 kg.

Posição do peso da criança na série ordenada	Valores de peso (em kg, dois decimais)	Posição do peso da criança na série ordenada	Valores de peso (em kg, dois decimais)	Posição do peso da criança na série ordenada	Valores de peso (em kg, dois decimais)
1ª	6,30	26ª	13,30	51ª	17,70
2ª	6,45	27ª	13,30	52ª	17,75
3ª	8,35	28ª	13,45	53ª	17,90
4ª	8,65	29ª	13,55	54ª	18,05
5ª	9,50	30ª	14,20	55ª	18,30
6ª	9,50	31ª	14,20	56ª	18,80
7ª	9,90	32ª	14,25	57ª	19,20
8ª	10,08	33ª	14,30	58ª	19,40
9ª	10,20	34ª	14,35	59ª	19,85
10ª	10,25	35ª	14,95	60ª	19,95
11ª	10,45	36ª	15,10	61ª	20,10
12ª	10,95	37ª	15,15	62ª	20,30
13ª	11,40	38ª	15,25	63ª	20,70
14ª	11,70	39ª	15,40	64ª	21,00
15ª	11,75	40ª	15,55	65ª	21,65
16ª	11,85	41ª	15,60	66ª	22,00
17ª	12,10	42ª	15,65	67ª	22,75
18ª	12,10	43ª	15,90	68ª	23,90
19ª	12,20	44ª	16,35	69ª	24,10
20ª	12,25	45ª	16,55	70ª	24,25
21ª	12,55	46ª	16,90	71ª	25,20
22ª	12,55	47ª	16,90	72ª	25,90
23ª	12,90	48ª	17,35		
24ª	13,25	49ª	17,55		
25ª	13,30	50ª	17,60		

Como há um intervalo de 0,25 kg (18,30 kg – 18,05 kg) entre as 54ª e 55ª posições, 0,75 unidade entre 18,05 kg e 18,30 kg será dada por

$$(0,75)(0,25) = 0,1875 \text{ kg}.$$

O mesmo resultado será obtido se considerarmos que 0,75 corresponde a três quartos $(0,75 = 0,25 + 0,25 + 0,25 = \frac{1}{4} + \frac{1}{4} + \frac{1}{4} = \frac{3}{4})$ do intervalo entre 18,05 kg e 18,30 kg:

$$\text{Três quartos de } 0,25 \text{ kg} = \frac{3}{4}(0,25) = \frac{3(0,25)}{4} = \frac{0,75}{4} = 0,1875 \text{ kg}.$$

Agora só falta somarmos o valor da 54ª posição (18,05 kg) ao valor correspondente a 0,75 unidade entre as 54ª e 55ª posições (0,1875 kg) para, finalmente, obtermos o Q_3 :

$$Q_3 = 18,05 + 0,1875 = 18,2375 \cong 18,24 \text{ kg}.$$

O valor do Q_3 é aproximadamente 18,24 kg.

— Quais as principais aplicações dos percentis?

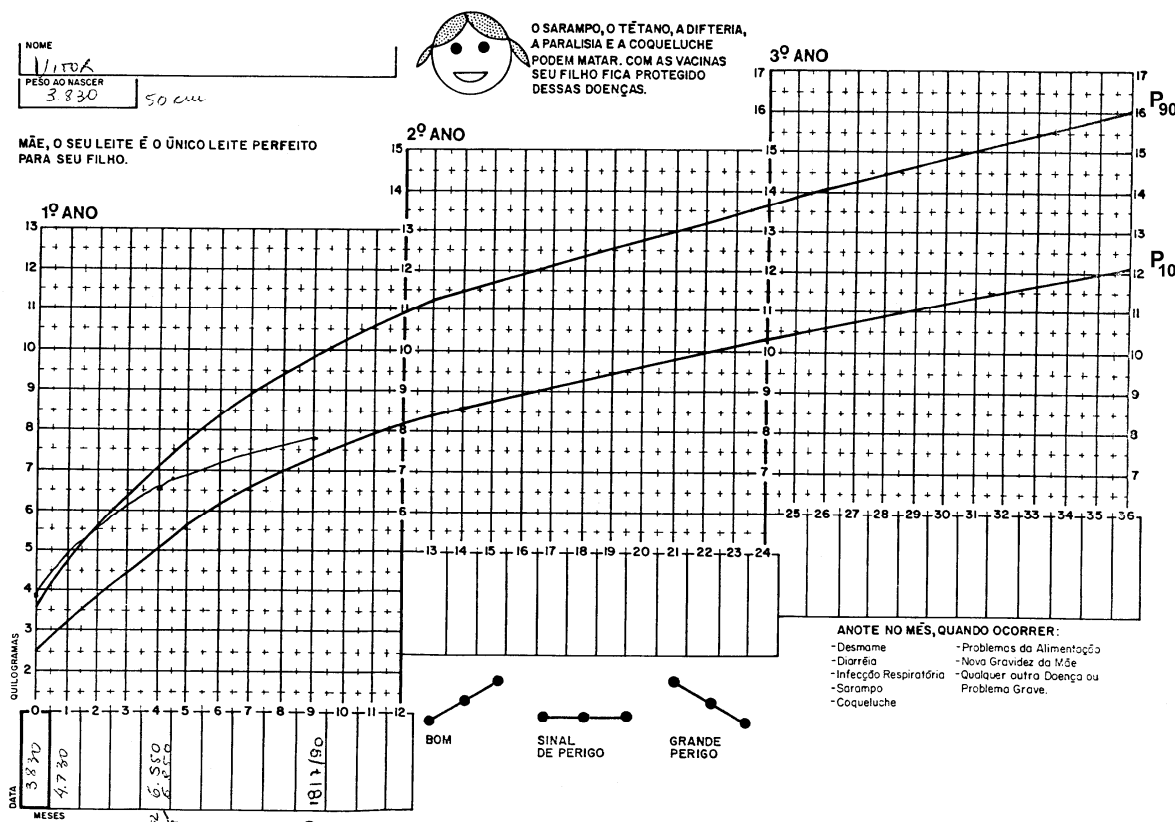
— Até agora só escrevemos que os percentis são medidas de posição e mostramos como são calculados.

Vamos apresentar uma situação freqüentemente utilizada por profissionais de saúde e que envolve a aplicação de percentis.

É muito importante fazer-se o acompanhamento da evolução dos pesos dos indivíduos a partir do seu nascimento. Com esse objetivo foi elaborada uma ficha (apresentada na página seguinte), na qual a cada mês de vida é registrado o peso da criança. Além da anotação do peso, marca-se um ponto (correspondente ao peso medido naquela criança a cada mês) em um diagrama (gráfico) contido na ficha. Este contém uma linha inferior, delimitando pesos muito baixos, e uma linha superior, delimitando pesos muito altos.

A linha inferior do diagrama foi obtida unindo-se com um traço os valores do P_{10} (percentil 10), obtidos para os pesos de crianças consideradas normais na população, para cada idade considerada. A linha superior do diagrama foi obtida unindo-se com um traço os valores do P_{90} (percentil 90), obtidos também para os pesos de crianças consideradas normais na população. Os pesos dentro da faixa central delimitada pelas linhas superior e inferior, são considerados aceitáveis para crianças nas idades correspondentes. Observe que há, na parte esquerda do diagrama, uma linha entre as linhas superior e inferior. Esta linha foi obtida marcando-se a cada mês no diagrama o ponto correspondente ao valor de peso medido na criança, e unindo-se também esses pontos. Note que nesse exemplo, a criança apresentava inicialmente peso próximo ao valor do P_{90} , mas a partir do terceiro mês de vida aquela criança começou a apresentar pesos cada vez mais próximos do P_{10} . Isso indica que a criança inicialmente estava com peso próximo ao de crianças com os 10%

de valores mais altos de peso, e posteriormente passou a perder peso aproximando-se da área correspondente a crianças com os 10% de valores mais baixos de peso. Com base nessas evidências, um médico poderá tentar descobrir, com maior antecedência, que fatores estão provocando aquela queda de peso na criança. O diagrama será útil também se o peso da criança em determinado mês aproximar-se da linha superior (indicando peso quase excessivo para a idade). Veja então que os percentis são aqui utilizados para nos ajudar a verificar se o valor de certa variável, apresentado por um determinado indivíduo, está muito ou pouco elevado, e avaliar, portanto, a **posição** daquele valor no espectro de valores possíveis para aquela variável.



Outra aplicação freqüente dos percentis é na aglutinação de categorias de uma variável. Considere as 72 observações de peso do nosso exemplo. Se desejarmos expressar essa variável em intervalos de classe (faixas), poderemos encontrar os quartis dessa série e utilizá-los como “escores de corte” (ou “pontos de corte”), que são os valores de peso que servirão como limites entre os intervalos de classe. Se quisermos classificar a variável peso em apenas dois intervalos de classe, poderemos utilizar o P_{50} (o quartil mediano) como “escore de corte”. Desse modo, passarão a existir duas categorias de peso, uma delas incluirá crianças com pesos menores ou iguais a 15,12 kg, e a outra, crianças com pesos maiores do que este valor.

Se desejarmos agrupar a série de pesos em quatro intervalos de classe, poderemos utilizar os outros quartis, além do quartil mediano. Assim, teremos uma faixa de pesos menores ou iguais ao Q_1 (12,12 kg), outra de 12,13 kg ao Q_2 (15,12 kg), outra de 15,13 kg ao Q_3 (18,24 kg) e uma última de pesos maiores do

que 18,24 kg.

Mais uma aplicação dos percentis é na avaliação da variabilidade dos valores de uma variável. Para isso, calculamos a **amplitude interquartil** (AIQ).

Como sua denominação indica, a AIQ mede a distância entre o primeiro e o terceiro quartis. É calculada diminuindo-se o valor do terceiro quartil pelo valor do primeiro quartil.

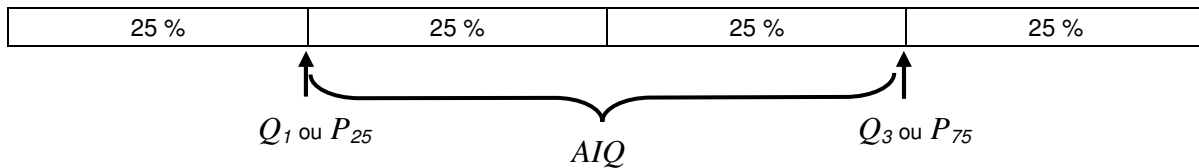
Amplitude interquartil é a distância entre o primeiro e o terceiro quartis.

Assim,

$$AIQ = Q_3 - Q_1.$$

— E para quê isso serve?

— Observe que a AIQ , sendo a amplitude entre Q_1 e Q_3 , indica o quanto ampla é a variação dos valores correspondentes aos 50% valores mais centrais, da série ordenada de valores de uma determinada variável (veja desenho abaixo).



Veja então que podemos usar percentis, que são medidas de posição, para calcularmos uma medida de dispersão.

Podemos calcular a AIQ de uma determinada variável para os homens e, separadamente, para as mulheres, e compará-las, para avaliar em que sexo aquela variável apresenta valores centrais com maior variabilidade. Lembre-se de que você já aprendeu a calcular a amplitude de variação (capítulo 6, páginas 59 e 60), que indica o quanto ampla é a distância entre o menor e o maior valor de uma determinada variável. Note que a amplitude interquartil não mede a mesma coisa. A AIQ não mede a amplitude entre o menor e o maior valor, mas entre o primeiro e o terceiro quartis, indicando-nos a amplitude de variação dos valores mais centrais da série. Além disso, a amplitude de variação é muito influenciada por valores extremos (muito altos ou baixos), podendo nos induzir a uma avaliação distorcida do quanto ampla é a variação, enquanto a AIQ não sofre essa influência. Lembra-se de que vimos como a mediana não sofre a influência de valores extremos? A mediana não é um dos quartis (segundo quartil ou quartil mediano)? Pois, os outros quartis (Q_1 e Q_3) têm essa mesma propriedade da mediana, já que são calculados do mesmo modo. Assim, a AIQ é uma medida de variabilidade menos vulnerável (mais robusta) aos valores extremos.

No nosso exemplo temos que

$$AIQ = Q_3 - Q_1 = 18,24 \text{ kg} - 12,12 \text{ kg} = 6,12 \text{ kg}.$$

Isso nos indica que os valores correspondentes aos 50% valores mais centrais da série ordenada de pesos, variam entre si não mais do que 6,12 kg.

As planilhas abaixo apresentam, separadamente, os pesos das crianças do sexo masculino ou feminino:

Dados para crianças do sexo masculino ($n = 43 - 2 = 41$) :

Posição do peso da criança na série ordenada	Valores de peso (em kg, dois decimais)	Posição do peso da criança na série ordenada	Valores de peso (em kg, dois decimais)	Posição do peso da criança na série ordenada	Valores de peso (em kg, dois decimais)
1 ^a	8,35	16 ^a	14,20	31 ^a	18,30
2 ^a	9,50	17 ^a	14,25	32 ^a	19,20
3 ^a	9,50	18 ^a	14,35	33 ^a	19,40
4 ^a	10,08	19 ^a	15,10	34 ^a	19,95
5 ^a	10,20	20 ^a	15,40	35 ^a	20,10
6 ^a	11,85	21 ^a	15,55	36 ^a	20,30
7 ^a	12,10	22 ^a	15,60	37 ^a	20,70
8 ^a	12,20	23 ^a	15,65	38 ^a	21,00
9 ^a	12,25	24 ^a	16,55	39 ^a	21,65
10 ^a	12,55	25 ^a	16,90	40 ^a	22,75
11 ^a	12,55	26 ^a	17,35	41 ^a	23,90
12 ^a	12,90	27 ^a	17,60	42 ^a	99,99
13 ^a	13,25	28 ^a	17,70	43 ^a	99,99
14 ^a	13,30	29 ^a	17,90		
15 ^a	13,55	30 ^a	18,05		

Dados para crianças do sexo feminino ($n = 32 - 1 = 31$) :

Posição do peso da criança na série ordenada	Valores de peso (em kg, dois decimais)	Posição do peso da criança na série ordenada	Valores de peso (em kg, dois decimais)	Posição do peso da criança na série ordenada	Valores de peso (em kg, dois decimais)
1 ^a	6,30	12 ^a	13,30	23 ^a	17,55
2 ^a	6,45	13 ^a	13,30	24 ^a	17,75
3 ^a	8,65	14 ^a	13,45	25 ^a	18,80
4 ^a	9,90	15 ^a	14,20	26 ^a	19,85
5 ^a	10,25	16 ^a	14,30	27 ^a	22,00
6 ^a	10,45	17 ^a	14,95	28 ^a	24,10
7 ^a	10,95	18 ^a	15,15	29 ^a	24,25
8 ^a	11,40	19 ^a	15,25	30 ^a	25,20
9 ^a	11,70	20 ^a	15,90	31 ^a	25,90
10 ^a	11,75	21 ^a	16,35	32 ^a	99,99
11 ^a	12,10	22 ^a	16,90		

Calculando a AIQ , separadamente, para crianças do sexo masculino ou feminino, obtemos:

$$AIQ_{\text{sexo masculino}} = Q_3 - Q_1 = 18,75 \text{ kg} - 12,55 \text{ kg} = 6,20 \text{ kg}$$

e

$$AIQ_{\text{sexo feminino}} = Q_3 - Q_1 = 17,75 \text{ kg} - 11,40 \text{ kg} = 6,35 \text{ kg}.$$

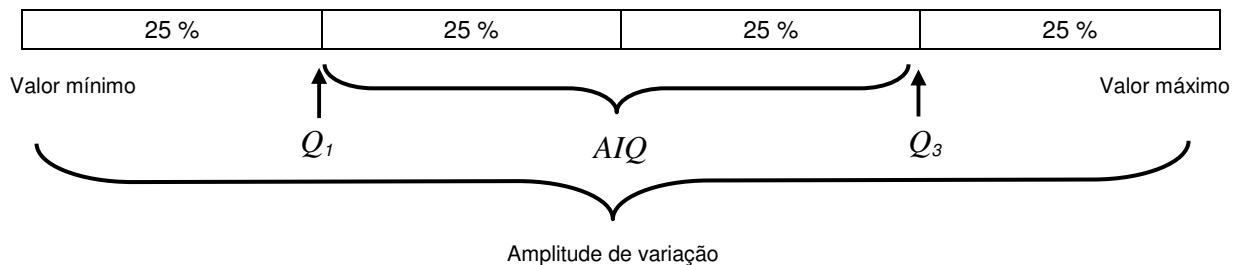
Como podemos ver, a variabilidade dos valores centrais da série é maior nas crianças do sexo feminino do que naquelas do sexo masculino.

Aproveite e, quando estiver disposto, calcule os primeiro e terceiro quartis de peso para cada sexo, como um exercício. Compare seus resultados com aqueles apresentados acima.

Podemos também expressar a AIQ como percentual da amplitude de variação:

$$AIQ \% = \frac{\text{amplitude interquartil}}{\text{amplitude de variação}} (100).$$

O resultado acima nos indicará quantos por cento da amplitude de variação a amplitude interquartil representa (veja desenho abaixo).



No nosso exemplo (sem considerar a variável sexo) temos que

$$AIQ \% = \frac{\text{amplitude interquartil}}{\text{amplitude de variação}} (100) = \frac{6,12}{19,60} (100) = \frac{(6,12)(100)}{19,60} = \frac{612}{19,60} = 31,22 \%$$

Veja esse resultado no capítulo anterior

O resultado obtido indica que o espectro de variação dos valores centrais da série de pesos representa 31,22% do espectro de variação de todos os valores de peso. Isso nos ajuda a avaliar se há uma aglutinação maior ou menor de valores no centro da série. E será mais útil ainda quando tivermos um outro valor de $AIQ\%$ para comparar. Podemos, por exemplo, comparar a $AIQ\%$ dos homens com a das mulheres.

Para o sexo masculino obtemos:

$$AIQ \% = \frac{\text{amplitude interquartil}}{\text{amplitude de variação}} (100) = \frac{6,20}{15,55} (100) = \frac{(6,20)(100)}{15,55} = \frac{620}{15,55} = 39,87 \%$$

E para o sexo feminino:

$$AIQ \% = \frac{\text{amplitude interquartil}}{\text{amplitude de variação}} (100) = \frac{6,35}{19,60} (100) = \frac{(6,35)(100)}{19,60} = \frac{635}{19,60} = 32,40 \%$$

Vemos que o sexo masculino, embora tenha sido o que apresentou menor variação interquartil, foi aquele no qual a variação interquartil apresentou o maior percentual quando comparado ao seu espectro total de variação. Isto é, os valores centrais para os homens variaram menos, mas esses valores eram mais esparsos em relação à amplitude de variação total, do que os do sexo feminino.

Os percentis **são também aplicados** freqüentemente nos procedimentos de inferência estatística. Nos próximos capítulos, você verá que o P_{95} e o $P_{97,5}$ são muito utilizados como valores de referência para concluirmos se um resultado é ou não estatisticamente significativo. Se você se considera pronto e motivado para aprender isso, inicie a leitura dos próximos capítulos. Antes, porém, não se esqueça de adiantar suas leituras fora da sua área profissional, para não ficar um ser humano alienado e fácil de ser manipulado pela propaganda (na maioria das vezes enganosa), financiada quase exclusivamente pelos seres humanos economicamente mais poderosos, porque essa propaganda é caríssima. Pouquíssimos têm condições de utilizar os meios de comunicação existentes para defender suas opiniões. Sendo assim, concentra-se em poucas pessoas o poder de influenciar decisivamente outras pessoas, eleger políticos do Poder Legislativo (que vão elaborar as leis), administradores do Poder Executivo (que vão executar essas leis e as que eles mesmos criam, pasmem, através das chamadas medidas provisórias), e o poder também de influenciar, direta ou indiretamente, na escolha dos indivíduos que irão ocupar os principais cargos do Poder Judiciário (que julgam sobre as questões levantadas na aplicação dessas leis). A essa sociedade humana damos o nome de “democracia”, ou seja, “poder do povo”. Bom! Vamos dar uma olhada em um dicionário para conferirmos o que significa a palavra “povo”.

O argumento mais utilizado em defesa da “democracia” e do capitalismo é que esses têm falhas, mas são os sistemas político e econômico menos piores. Ora! Como vamos saber se esses sistemas são os menos piores, se as sociedades humanas dominantes a cada época (romanos, ingleses, franceses, estadunidenses, e seus aliados, etc.) não permitem que os seres humanos (usando sua fantástica criatividade) experimentem outras formas de organizar as sociedades?

Como podemos aceitar que o tipo de organização atual das sociedades humanas seja inevitável e inexorável, como ouvimos quase todos os dias? E a apregoada liberdade da qual os mandantes nessas sociedades se dizem paladinos? Se somos livres, não devemos, em princípio, considerar nada inexorável. Qual a justificativa para substituímos a tirania comunista pela tirania das leis de mercado? Existe algum tipo de tirania aceitável? Devemos tolerar que um ser humano possa explorar o trabalho de outro, e as consequências resultantes disso: desigualdades econômicas e sociais, violência, fome, miséria, infelicidade, etc.?

CAPÍTULO 8

-
- Qual a diferença entre quadro e tabela?
 - Quais as partes de uma tabela?
 - Quais os tipos de gráficos que existem?
 - Como escolher o gráfico adequado?
 - Quando utilizar tabela ou gráfico?
-



– **Ainda existem outros procedimentos da Estatística Descritiva?**

– Existem. No capítulo 1 (página 4) mencionamos também a elaboração de **tabelas** e **gráficos**. Esses assuntos serão explicados agora.

– **E os quadros também não são utilizados na descrição de dados quantitativos?**

– Não. Os **quadros** são utilizados para apresentação de informações não-quantitativas. Veja o exemplo abaixo, transcrito do capítulo 1:

A Estatística pode ser dividida em três partes:		
Estatística Descritiva	Descreve	Caracterização dos indivíduos estudados
Estatística Analítica	Analisa	Investigação das relações entre as características estudadas
Estatística Inferencial	Inferre	Verificação da possibilidade de generalização

Note também que um quadro é todo contornado por linhas, o que justifica sua denominação.

No capítulo 4 (páginas 36 a 40) abordamos os diversos tipos de freqüências utilizados na descrição de dados quantitativos. Para começarmos a avaliar como essas freqüências se distribuem em uma determinada população ou amostra, podemos organizá-las em **tabelas**, como a apresentada abaixo (exemplo transposto do capítulo 4):

Valores de idade (em anos completos)	Freqüência simples	Freqüência simples acumulada	Freqüência relativa (%)	Freqüência relativa acumulada (%)
25	1	1	4,0	4,0
31	1	2	4,0	8,0
32	2	4	8,0	16,0
34	3	7	12,0	28,0
36	2	9	8,0	36,0
38	2	11	8,0	44,0
39	1	12	4,0	48,0
40	3	15	12,0	60,0
41	4	19	16,0	76,0
45	1	20	4,0	80,0
46	2	22	8,0	88,0
47	1	23	4,0	92,0
51	1	24	4,0	96,0
52	1	25	4,0	100,0

Observe que, diferentemente do quadro, uma tabela não é completamente contornada por linhas, pois não possui as linhas laterais, e seu conteúdo é quantitativo.

Atualmente, recomenda-se que as tabelas tenham o menor número possível de linhas, como mostramos abaixo:

Valores de idade (em anos completos)	Freqüência simples	Freqüência simples acumulada	Freqüência relativa (%)	Freqüência relativa acumulada (%)
25	1	1	4,0	4,0
31	1	2	4,0	8,0
32	2	4	8,0	16,0
34	3	7	12,0	28,0
36	2	9	8,0	36,0
38	2	11	8,0	44,0
39	1	12	4,0	48,0
40	3	15	12,0	60,0
41	4	19	16,0	76,0
45	1	20	4,0	80,0
46	2	22	8,0	88,0
47	1	23	4,0	92,0
51	1	24	4,0	96,0
52	1	25	4,0	100,0

— Quais as partes de uma tabela?

— As partes constantes de uma tabela são: **título**, **colunas**, **linhas**, **cabeçalho**, **coluna indicadora**, **corpo**, **células**, **fonte** e **nota**. Cada uma dessas partes pode ser vista na seqüência abaixo:

Título → **Distribuição da freqüência simples da idade de residentes¹ na cidade X, Bahia, em 2.005.**

Valores de idade (em anos completos)	Freqüência simples
25	1
31	1
32	2
34	3
36	2
38	2
39	1
40	3
41	4
45	1
46	2
47	1
51	1
52	1

Fonte: Secretaria da Saúde do Estado da Bahia.

1. Apenas para residentes com idade entre 25 e 52 anos.

Distribuição da frequência simples da idade de residentes¹ na cidade X, Bahia, em 2.005.

Coluna →

Valores de idade (em anos completos)	Frequência simples
25	1
31	1
32	2
34	3
36	2
38	2
39	1
40	3
41	4
45	1
46	2
47	1
51	1
52	1

Fonte: Secretaria da Saúde do Estado da Bahia.

1. Apenas para residentes com idade entre 25 e 52 anos.

Distribuição da frequência simples da idade de residentes¹ na cidade X, Bahia, em 2.005.

Valores de idade (em anos completos)	Frequência simples	
25	1	
31	1	
32	2	
34	3	
36	2	
38	2	
39	1	
40	3	
41	4	
45	1	
46	2	
47	1	
51	1	
52	1	

← Outra coluna

Fonte: Secretaria da Saúde do Estado da Bahia.

1. Apenas para residentes com idade entre 25 e 52 anos.

Distribuição da frequência simples da idade de residentes¹ na cidade X, Bahia, em 2.005.

Valores de idade (em anos completos)	Frequência simples	
25	1	Outras linhas
31	1	
32	2	
34	3	
36	2	
38	2	
39	1	Linha
40	3	
41	4	Outras linhas
45	1	
46	2	
47	1	
51	1	
52	1	

Fonte: Secretaria da Saúde do Estado da Bahia.

1. Apenas para residentes com idade entre 25 e 52 anos.

Distribuição da frequência simples da idade de residentes¹ na cidade X, Bahia, em 2.005.

Valores de idade (em anos completos)	Frequência simples	
25	1	Cabeçalho
31	1	
32	2	
34	3	
36	2	
38	2	
39	1	
40	3	
41	4	
45	1	
46	2	
47	1	
51	1	
52	1	

Fonte: Secretaria da Saúde do Estado da Bahia.

1. Apenas para residentes com idade entre 25 e 52 anos.

Distribuição da frequência simples da idade de residentes¹ na cidade X, Bahia, em 2.005.

Coluna indicadora	Valores de idade (em anos completos)	Frequência simples
	25	1
	31	1
	32	2
	34	3
	36	2
	38	2
	39	1
	40	3
	41	4
	45	1
	46	2
	47	1
	51	1
	52	1

Fonte: Secretaria da Saúde do Estado da Bahia.

1. Apenas para residentes com idade entre 25 e 52 anos.

Distribuição da frequência simples da idade de residentes¹ na cidade X, Bahia, em 2.005.

Valores de idade (em anos completos)	Frequência simples	Corpo
25	1	
31	1	
32	2	
34	3	
36	2	
38	2	
39	1	
40	3	
41	4	
45	1	
46	2	
47	1	
51	1	
52	1	

Fonte: Secretaria da Saúde do Estado da Bahia.

1. Apenas para residentes com idade entre 25 e 52 anos.

Distribuição da frequência simples da idade de residentes¹ na cidade X, Bahia, em 2.005.

Valores de idade (em anos completos)	Frequência simples	
25	1	Outras células
31	1	
32	2	
34	3	
36	2	
38	2	
39	1	
40	3	Célula
41	4	
45	1	Outras células
46	2	
47	1	
51	1	
52	1	

Fonte: Secretaria da Saúde do Estado da Bahia.

1. Apenas para residentes com idade entre 25 e 52 anos.

Distribuição da frequência simples da idade de residentes¹ na cidade X, Bahia, em 2.005.

Valores de idade (em anos completos)	Frequência simples
25	1
31	1
32	2
34	3
36	2
38	2
39	1
40	3
41	4
45	1
46	2
47	1
51	1
52	1

Fonte: Secretaria da Saúde do Estado da Bahia. ← Fonte

1. Apenas para residentes com idade entre 25 e 52 anos.

Distribuição da frequência simples da idade de residentes¹ na cidade X, Bahia, em 2.005.

Valores de idade (em anos completos)	Frequência simples
25	1
31	1
32	2
34	3
36	2
38	2
39	1
40	3
41	4
45	1
46	2
47	1
51	1
52	1

Fonte: Secretaria da Saúde do Estado da Bahia.

Nota → 1. Apenas para residentes com idade entre 25 e 52 anos.

O título descreve sucintamente o conteúdo da tabela; os cruzamentos de **colunas** e **linhas** formam a tabela; o **cabeçalho** especifica o conteúdo de cada coluna; a **coluna indicadora** informa o conteúdo de cada linha; o **corpo** contém as frequências observadas; as **células** compõem o corpo e cada uma mostra a frequência observada para cada combinação possível para as variáveis representadas na linha e na coluna; a **fonte** indica de onde os dados foram obtidos; e a **nota** informa algum aspecto importante para compreensão dos dados colocados na tabela.

— As tabelas só apresentam frequências?

— Não. Podemos também utilizá-las para apresentar medidas de tendência central ou de dispersão, e no capítulo 16 utilizaremos tabelas com mais de uma variável.

Para ilustrar a primeira situação vamos considerar como exemplo a variável “idade”, coletada em um estudo sobre condições de trabalho e saúde de uma amostra aleatória de 311 agentes penitenciários da Região Metropolitana de Salvador. Os dados foram digitados em um computador e, facilmente, foi possível obtermos a média e o desvio-padrão dessa variável.

Veja, abaixo, uma comparação entre as médias e os desvios-padrão das idades de homens e mulheres da amostra estudada, apresentadas em uma tabela:

Médias de idade, segundo o sexo, de agentes penitenciários da Região Metropolitana de Salvador, Bahia, 2.000.

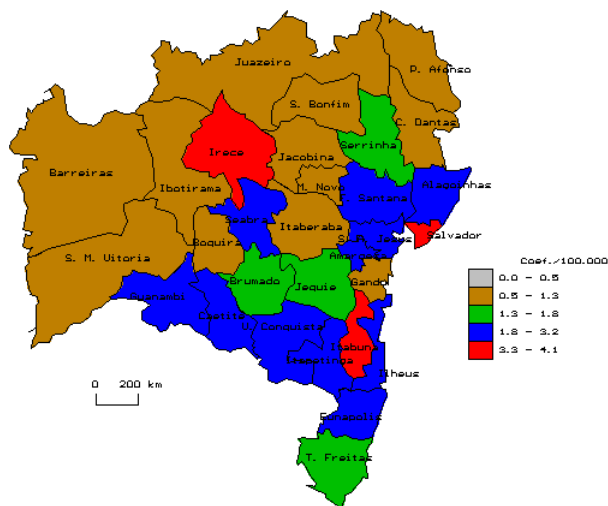
Variável	<i>n</i>	Média	Desvio-padrão
Idade			
Masculino	216	40,20	7,61
Feminino	57	40,25	7,16
Total	303	40,21	7,51

— E os gráficos? Quais os seus tipos e aplicações?

— Os resultados desse estudo podem também ser apresentados através de **gráficos**, que podem ser de dois tipos: cartogramas ou diagramas.

Os **cartogramas** são mapas utilizados para apresentar a distribuição de eventos de saúde de interesse, em diferentes áreas geográficas. Veja um exemplo abaixo (*gentilmente cedido pelo Professor Marco Antônio Vasconcelos Rêgo e elaborado pelo Geógrafo Davi Félix Martins Júnior*):

Mortalidade por Câncer de Estômago, Estado da Bahia - 2000



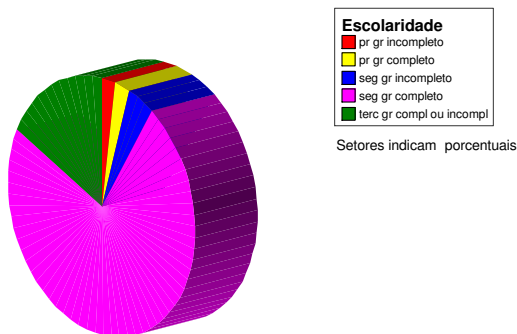
Os **diagramas** utilizam figuras geométricas (perfeitas ou imperfeitas), linhas ou pontos, para representar a magnitude de eventos de saúde em determinada população e local.

Os diagramas mais frequentemente utilizados são: o de setores, o de barras, o histograma, o polígono de freqüências, o diagrama de talo e folha, o de pontos, o de linha, o de dispersão, de caixa, e o de linhas de afastamento. Vamos elaborar cada um desses, utilizando os resultados da pesquisa com os agentes penitenciários.

O **diagrama de setores** pode conter setores simples, compostos ou diagramados.

Diagrama de setores simples:

Distribuição dos agentes, segundo a escolaridade, Região Metropolitana de Salvador, 2000.

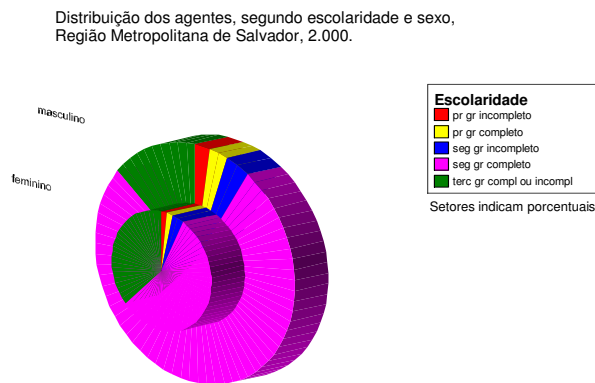


Neste diagrama, cada setor representa a freqüência relativa em percentuais (a freqüência absoluta pode ser também utilizada) com que cada categoria de uma variável, nominal ou ordinal, apareceu na amostra ou população estudada.

— Quer dizer que não podemos utilizar este tipo de diagrama para apresentar resultados de variáveis intervalares ou de razão?

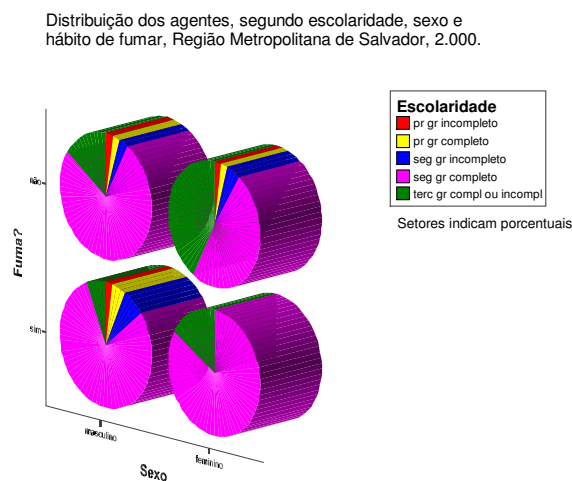
— Isso mesmo. Note que, como variáveis intervalares ou de razão possuem muitas categorias, o número de setores será tão grande, que o diagrama de setor não conseguirá apresentar todas essas com clareza.

Diagrama de setores compostos:



No diagrama acima são apresentadas duas variáveis, “escolaridade” e “sexo”, sendo possível compararmos visualmente o tamanho dos setores correspondentes a cada categoria da variável “escolaridade” obtidos para o sexo masculino e, separadamente, para o feminino. Se desejarmos, podemos solicitar ao computador que coloque os percentuais obtidos para cada setor em cada sexo, permitindo-nos uma comparação numérica, além da visual.

Diagrama de setores diagramados:

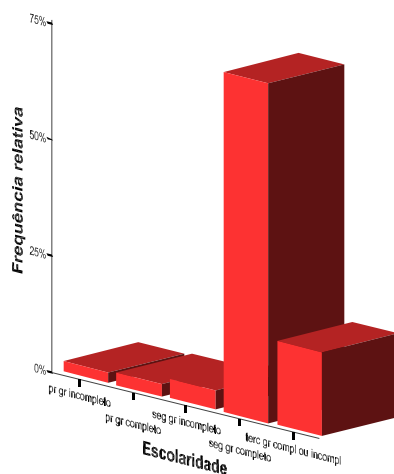


Observe no diagrama anterior que foram apresentadas três variáveis: “escolaridade”, “sexo” e “hábito de fumar”. São mostrados os percentuais (indicados pelo tamanho dos setores) das categorias da variável “escolaridade”, para homens ou mulheres, fumantes ou não-fumantes.

O **diagrama de barras** utiliza retângulos (denominados barras ou colunas nesse contexto) para indicar frequências. Pode ser de barras simples, de barras compostas ou de barras múltiplas.

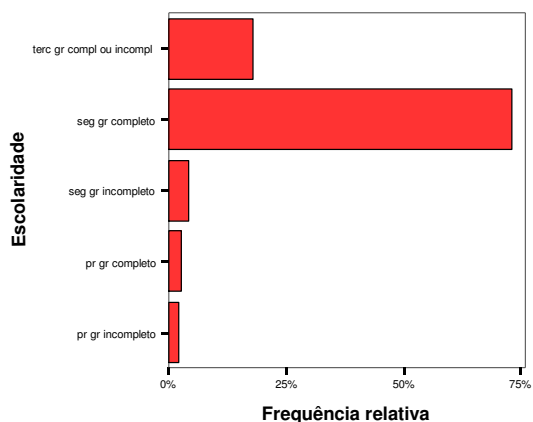
Diagrama de barras simples:

Distribuição dos agentes, segundo a escolaridade,
Região Metropolitana de Salvador, 2.000.



No diagrama acima, retângulos são usados para representar o percentual de agentes em cada categoria da variável “escolaridade”. Note que as denominações dessas categorias são muito extensas e não há espaço adequado para sua colocação lado a lado na abscissa. Uma solução para isso é elaborarmos esse diagrama colocando as barras horizontalmente, como mostrado a seguir:

Distribuição dos agentes, segundo a escolaridade,
Região Metropolitana de Salvador, 2.000.

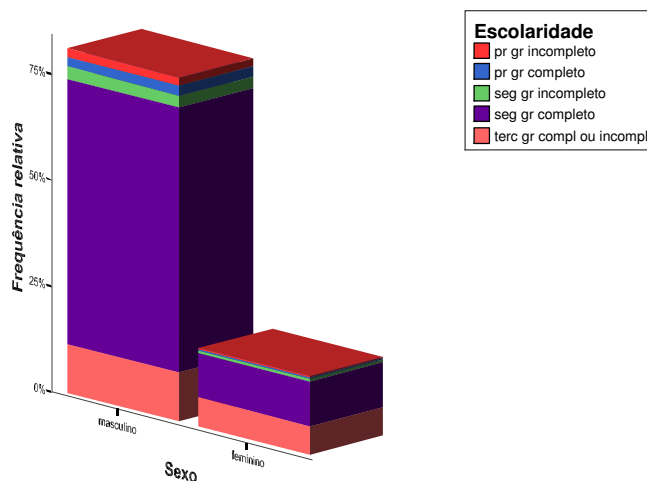


Observe, contudo, que, ao optarmos por barras horizontais, não foi possível manter o diagrama em

três dimensões, por uma limitação do programa estatístico que estamos utilizando.

Diagrama de barras compostas:

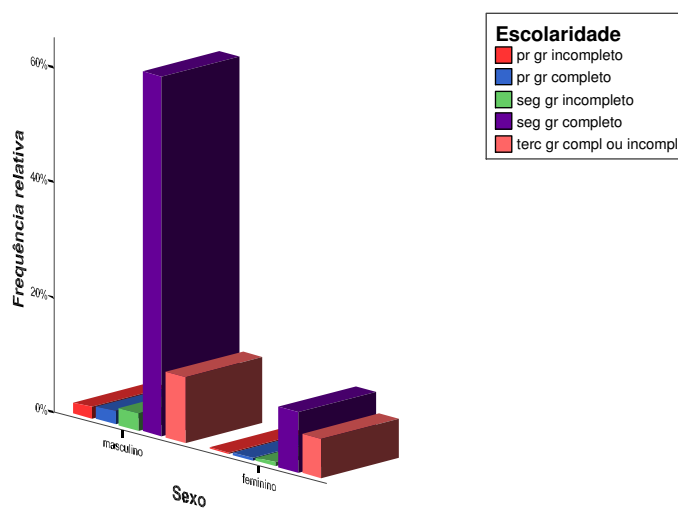
Distribuição dos agentes, segundo escolaridade e sexo,
Região Metropolitana de Salvador, 2.000.



Neste diagrama comparamos os percentuais das diversas categorias de escolaridade nos sexos masculino e feminino. Tais percentuais aparecem como subdivisões dentro de cada barra, sendo que cada barra representa uma categoria da variável “sexo”.

Diagrama de barras múltiplas:

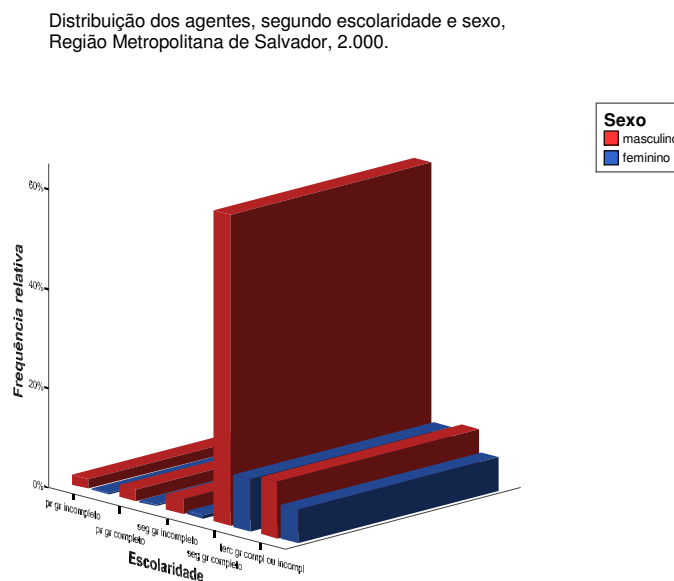
Distribuição dos agentes, segundo escolaridade e sexo,
Região Metropolitana de Salvador, 2.000.



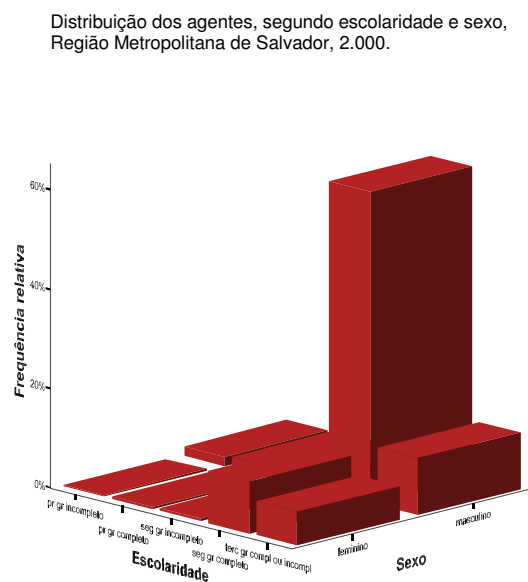
O diagrama acima tem o mesmo conteúdo do anterior. A única diferença é que agora, em vez dos

porcentuais de escolaridade para a variável “sexo” serem apresentados em subdivisões das barras, aparecem em barras separadas, daí a denominação “de barras múltiplas” desse tipo de diagrama.

Uma opção melhor ainda para compararmos os percentuais de escolaridade dos homens e mulheres, é fazermos o mesmo tipo de diagrama acima, invertendo as posições das variáveis “escolaridade” e “sexo”. Veja como fica o diagrama:



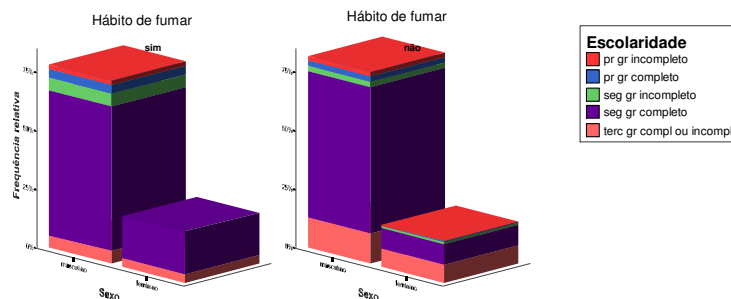
Outra opção para apresentarmos duas variáveis simultaneamente, seria acrescentarmos mais um eixo no diagrama, como mostramos abaixo:



Se precisarmos acrescentar uma terceira variável no diagrama, procedemos das seguintes maneiras:

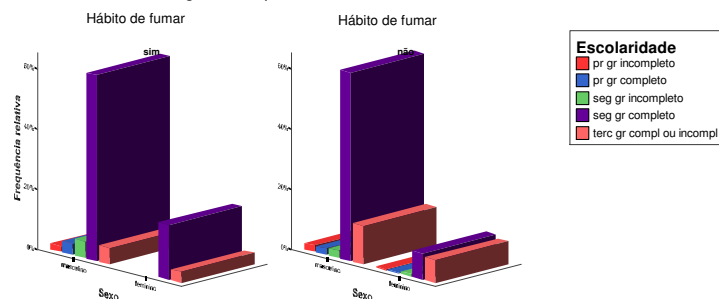
a)

Distribuição dos agentes, segundo escolaridade, sexo e hábito de fumar, Região Metropolitana de Salvador, 2.000.



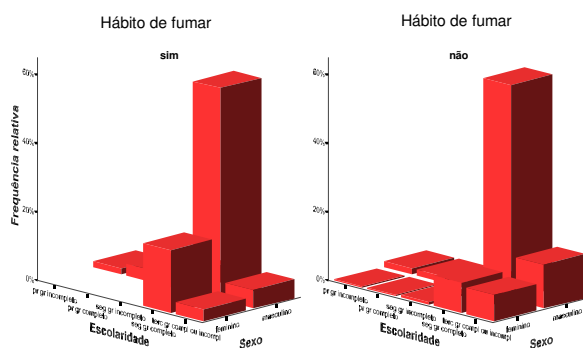
b)

Distribuição dos agentes, segundo escolaridade, sexo e hábito de fumar, Região Metropolitana de Salvador, 2.000.



c)

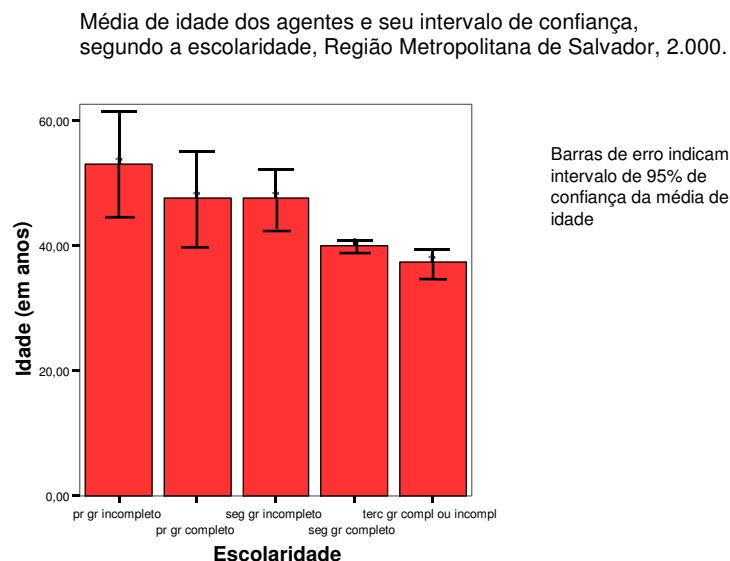
Distribuição dos agentes, segundo escolaridade, sexo e hábito de fumar, Região Metropolitana de Salvador, 2.000.



Os diagramas de barras podem ser utilizados também para a realização de inferência estatística. No caso, recebem a denominação especial de **diagrama de barras de erro**. Vamos dar um exemplo e explicar como interpretá-lo, mas achamos que você ainda não o entenderá completamente, porque nosso livro não

abordou ainda os procedimentos utilizados para inferência estatística. Se encontrar muita dificuldade, pule o exemplo e a interpretação seguintes, deixando para rever esse trecho após o estudo dos capítulos 9 e 10.

Exemplo:



Interpretação: como as linhas verticais na parte superior das barras representam os intervalos de confiança das médias de idade para as diferentes categorias de escolaridade, a avaliação de se essas médias são estatisticamente diferentes (processo chamado de inferência estatística) consistirá da avaliação da existência de superposição entre essas linhas verticais. Você verá nos próximos capítulos que um intervalo de confiança expressa o quanto a média (ou outra medida de interesse) variaria caso tivéssemos estudado numerosas amostras, e não uma apenas. Então, se os intervalos de confiança se superpuserem, isso indicará que, caso estudássemos numerosas amostras, poderíamos ter obtido resultados semelhantes para os grupos que estivessem sendo comparados, embora na única amostra estudada os resultados tivessem sido matematicamente diferentes. Por outro lado, se os intervalos não se superpuserem, concluiremos que as médias dos grupos comparados são estatisticamente diferentes.

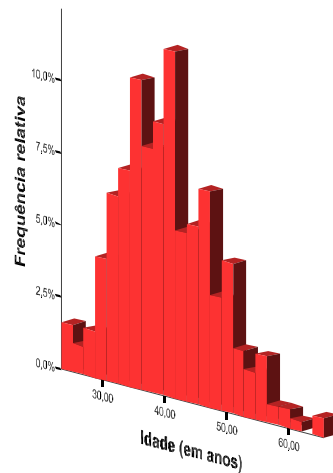
Note, no diagrama acima, que os intervalos para as médias de idade dos agentes com primeiro grau incompleto, primeiro grau completo ou segundo grau incompleto se superpõem, indicando que as médias de idade desses indivíduos não diferem estatisticamente. O mesmo ocorre quando comparamos os intervalos das médias de idade dos trabalhadores com segundo grau completo com aqueles com terceiro grau completo ou incompleto. Também não há diferença estatisticamente significativa entre aqueles com segundo grau completo, ou terceiro grau completo ou incompleto, e aqueles com primeiro grau completo. As médias de idade estatisticamente diferentes foram aquelas dos agentes com segundo grau completo, ou terceiro grau incompleto ou completo, e aquelas dos indivíduos com primeiro grau incompleto ou segundo incompleto.

Outro aspecto que pode ser observado no diagrama acima, é que podemos também utilizar medidas de tendência central (como a média aritmética) de variáveis de razão (como a idade) na ordenada, em vez das freqüências simples ou relativas de variáveis nominais ou ordinais (como sexo ou escolaridade).

Histograma:

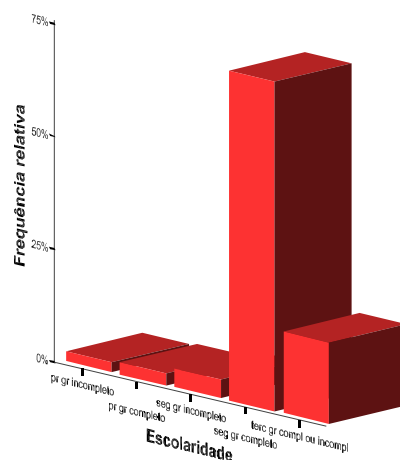
Solicitando ao computador a elaboração de um histograma para a variável idade, obtemos:

Distribuição dos agentes, segundo a idade,
Região Metropolitana de Salvador, 2.000.



Observe que na abscissa é apresentada a variável “idade”, que foi automaticamente organizada pelo computador em intervalos de classe (faixas de idade). Na ordenada são apresentadas as frequências relativas de indivíduos em cada intervalo de classe de idade. Note que o efeito tridimensional deve ser utilizado com cautela, pois dificulta um pouco a avaliação da simetria da distribuição. Veja também que diferentemente do diagrama de barras, o histograma contém barras contíguas. Nosso exemplo de diagrama de barras simples foi com a variável “escolaridade”. Apresentamos novamente a seguir esse diagrama, para ajudá-lo a verificar que nele as barras são separadas umas das outras:

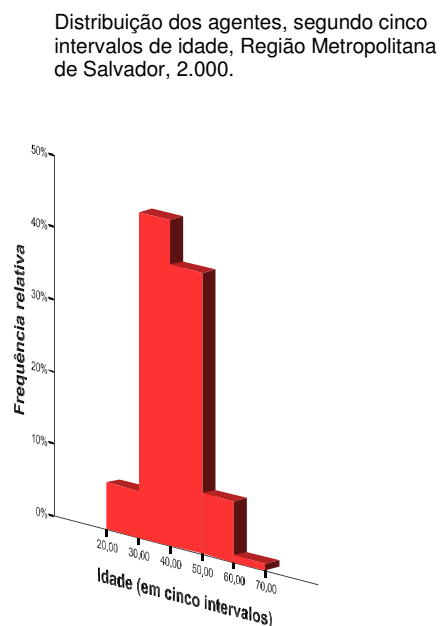
Distribuição dos agentes, segundo a escolaridade,
Região Metropolitana de Salvador, 2.000.



— **Existe uma justificativa para essa diferença?**

— Existe. É que as categorias da variável “escolaridade” representam níveis de instrução, que podem ser ordenados do menor ao maior ou vice-versa, mas não podem ser considerados como contíguos. Um indivíduo pode completar o primeiro grau e não iniciar imediatamente o nível seguinte de escolaridade, por razão econômica ou de outra natureza. Por isso devemos colocar espaços entre as barras. Se em vez da escolaridade considerarmos a variável “sexo”, fica claro também que existe uma separação entre ser masculino ou feminino, que resulta das diferenças marcantes existentes entre os dois sexos, biologicamente falando. Já os intervalos nos quais organizamos a idade são contíguos, porque ao findar-se um desses, imediatamente inicia-se outro. Assim, no caso, não são colocados espaços entre as barras, e o diagrama recebe o nome especial de histograma, como já vimos.

Se quisermos estabelecer intervalos diferentes daqueles definidos pelo computador, podemos obter isso alterando o padrão de intervalo utilizado pelo computador ao elaborar o diagrama ou recodificando previamente a variável “idade”, organizando-a conforme desejarmos. Veja a seguir um histograma no qual a variável está organizada em cinco intervalos de classe (21 a 30 anos; 31 a 40 anos; 41 a 50 anos; 51 a 60 anos; e 61 a 70 anos):

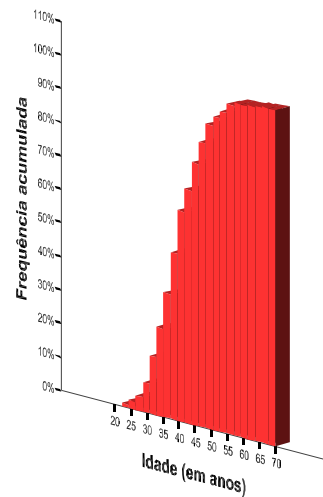


Do mesmo modo como fizemos para o diagrama de barras, o histograma pode também considerar, simultaneamente, duas ou três variáveis. O formato geral desses histogramas será semelhante àquele apresentado para os diagramas de barras, tendo como única diferença a contigüidade entre as barras.

Note, no diagrama acima, que os intervalos de classe utilizados têm o mesmo comprimento (10 em 10 anos). Isto é importante porque se os intervalos forem desiguais, obteremos barras com áreas diferentes, o que pode nos conduzir a erros na interpretação dos resultados. Mas não precisamos nos preocupar com isso, porque os programas estatísticos utilizam intervalos de mesmo comprimento, como procedimento padrão.

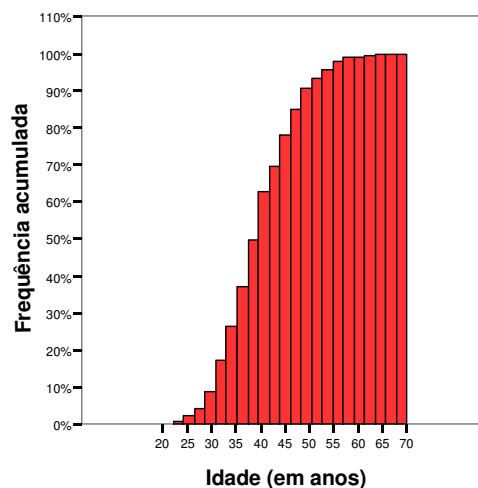
Existe também um histograma que apresenta frequências acumuladas:

Distribuição dos agentes, segundo a idade,
Região Metropolitana de Salvador, 2.000.



No diagrama acima, podemos verificar que cerca de 50% dos agentes penitenciários estudados tinham idade até cerca de 40 anos. Nesse caso, talvez seja visualmente mais fácil de olhar os resultados, se abrirmos mão do efeito tridimensional, como mostrado abaixo:

Distribuição dos agentes, segundo a idade,
Região Metropolitana de Salvador, 2.000.



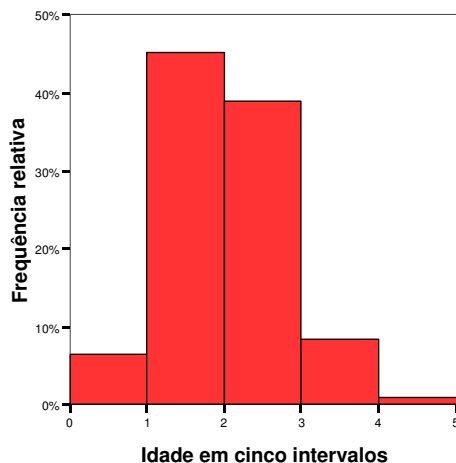
O histograma de freqüências acumuladas também pode apresentar duas ou três variáveis simultaneamente.

Os histogramas são muito utilizados na descrição dos nossos dados, para verificarmos a forma com que se distribuíram as freqüências de variáveis de razão, como a idade. Você verá isso em breve, e também a utilização dessas distribuições de freqüências para a realização de inferência estatística.

Com base em um histograma pode ser elaborado um outro diagrama denominado “**polígono de freqüências**”.

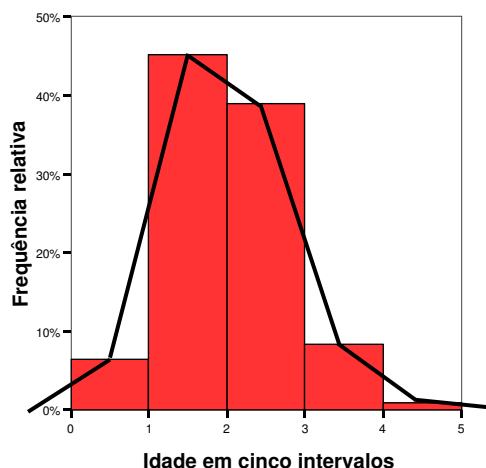
Considere o histograma a seguir:

Distribuição dos agentes, segundo a idade,
Região Metropolitana de Salvador, 2.000.



Se traçarmos retas interligando os pontos médios dos segmentos superiores de cada barra deste histograma, obteremos o polígono de freqüências.

Distribuição dos agentes, segundo a idade,
Região Metropolitana de Salvador, 2.000.

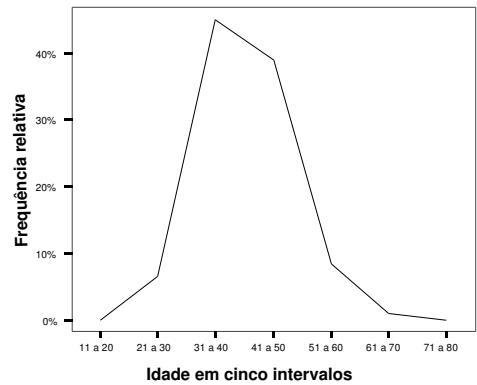


Observe que tivemos que considerar dois outros intervalos de idade, um imediatamente antes da primeira barra e outro imediatamente depois da última barra. Ao ser elaborado dessa maneira, a área sob o polígono de freqüências será igual àquela resultante da soma das áreas das barras do histograma.

— **Como essas áreas podem ser iguais se há partes das barras que não são englobadas pelo polígono, e partes do polígono que ultrapassam as barras?**

— O que ocorre é que essas partes que você mencionou são equivalentes. Verifique que cada parte do histograma que o polígono não engloba (em vermelho), é compensada por uma de mesmo tamanho (em branco), correspondente a uma parte em que o polígono ultrapassa o histograma. Por isso, a afirmativa que fizemos acima é correta. Veja a seguir, isoladamente, o polígono de freqüências do nosso exemplo:

Distribuição dos agentes, segundo a idade,
Região Metropolitana de Salvador, 2.000.



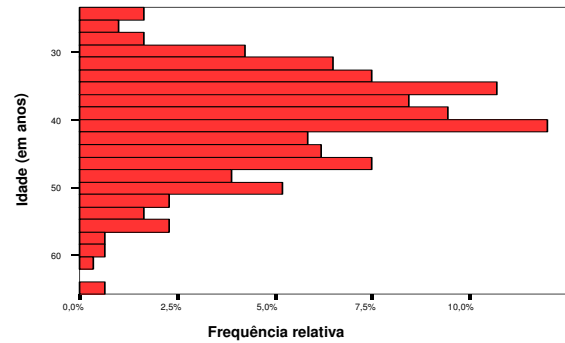
Há um diagrama que pode ser utilizado em substituição ao histograma, que é o **diagrama de talo e folha**. Observe-o a seguir:

Distribuição dos agentes, segundo a idade, Região Metropolitana de Salvador, 2.000.		
Frequência	Talo ,	Folha
1	2 ,	3
6	2 ,	444555
3	2 ,	677
10	2 ,	8888999999
20	3 ,	000000111111111111
21	3 ,	2222222222223333333
35	3 ,	44444444444444445555555555555555
30	3 ,	666666666666666677777777777777
33	3 ,	88888888888888888888889999999999
37	4 ,	000000000000000000000011111111111111
18	4 ,	222222223333333333
26	4 ,	4444444444445555555555555555
18	4 ,	66666666666677777777
21	4 ,	8888888899999999999999
6	5 ,	001111
8	5 ,	22223333
7	5 ,	455555
3	5 ,	667
2	5 ,	88
3 Extremos	(>=62)	

Cada folha representa um caso.

Compare-o ao histograma abaixo:

Distribuição dos agentes, segundo a idade,
Região Metropolitana de Salvador, 2.000.



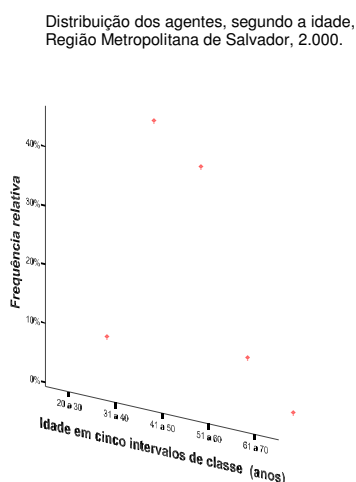
Não são iguais, mas são muito semelhantes, não são? É possível tornar os comprimentos das barras do histograma praticamente iguais à extensão das linhas horizontais numéricas do diagrama de talo e folha, modificando adequadamente os comprimentos dos intervalos de classe do histograma.

— **Há alguma razão especial para escolhermos entre um ou outro?**

— Visualmente, o histograma é mais agradável para o(a) leitor(a) do nosso trabalho ou para quem estiver assistindo uma apresentação do mesmo. Mas, evidentemente, o diagrama de talo e folha é mais informativo, na medida em que informa o valor aproximado de cada idade observada. Para entender isso, vamos considerar os primeiros sete valores mais baixos de idade, que são os primeiros a serem apresentados: 23,37; 24,02; 24,20; 24,71; 25,15; 25,54 e 25,99 anos. Veja que cada talo representa o primeiro dígito da idade e cada folha o segundo. Então, o primeiro conjunto de talo e folha que aparece no diagrama apresenta a idade mais baixa observada na série, 23,37, que foi aproximada para 23, resultando no talo 2 e na folha 3. O segundo, terceiro e quarto conjuntos apresentam a segunda, terceira e quarta idades mais baixas observadas na série, respectivamente, 24,02; 24,20 e 24,71, aproximadas para 24, resultando no segundo talo 2 e nas três folhas 444, que aparecem na segunda linha do diagrama. O quinto, sexto e sétimo conjuntos apresentam a quinta, sexta e sétima idades mais baixas observadas na série, respectivamente, 25,15; 25,54 e 25,99, aproximadas para 25, resultando no segundo talo 2 e nas três folhas 555, que aparecem na segunda linha do diagrama, após as folhas 444, já mencionadas; e assim por diante. Na primeira coluna do diagrama são apresentadas as frequências de cada talo e suas folhas.

Na prática, usamos muito mais o histograma do que os diagramas de talo e folha, apesar deste ser mais informativo.

Embora não seja muito utilizado, o **diagrama de pontos** é uma outra opção para apresentação dos nossos dados. Neste tipo de gráfico, utilizamos pontos, cuja altura em relação à ordenada indica a magnitude do percentual de indivíduos para cada intervalo da variável expressa na abscissa. Veja um exemplo abaixo:

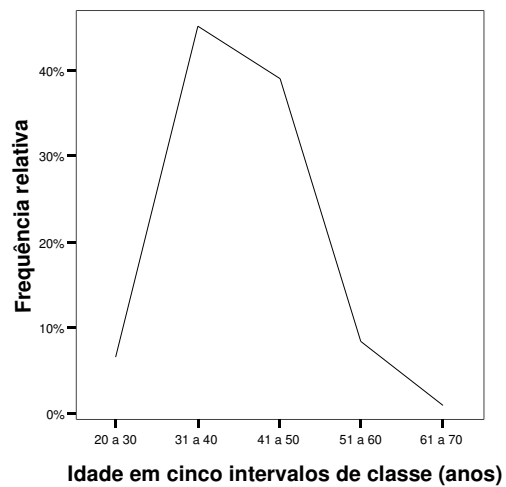


O diagrama de pontos pode também apresentar duas ou três variáveis simultaneamente.

Na elaboração do **diagrama de linhas**, como sua denominação aponta, utilizamos linhas, cuja localização mais alta ou mais baixa nos indica valores mais altos ou mais baixos da variável representada na ordenada, ou cuja tendência descendente ou ascendente nos ajuda a identificar o aumento ou diminuição

desses valores. Veja um exemplo abaixo:

Distribuição dos agentes, segundo a idade,
Região Metropolitana de Salvador, 2.000.

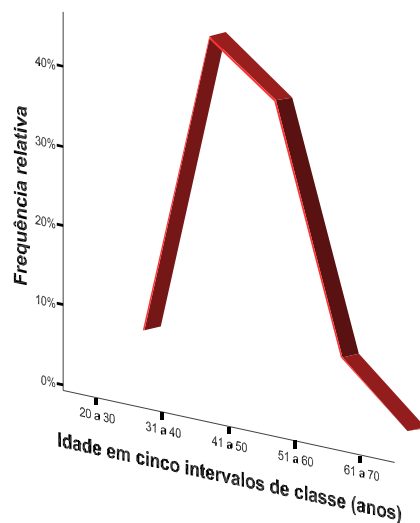


O diagrama acima resultou da ligação, através de linhas, dos pontos representativos dos percentuais de agentes em cada intervalo de idade, vistos no diagrama de pontos.

Note que o diagrama acima se assemelha muito ao polígono de freqüências, faltando apenas as categorias imediatamente anterior à primeira e imediatamente posterior à última, que são necessárias para a elaboração do polígono. Assim, este deve ser considerado como um tipo de diagrama de linhas.

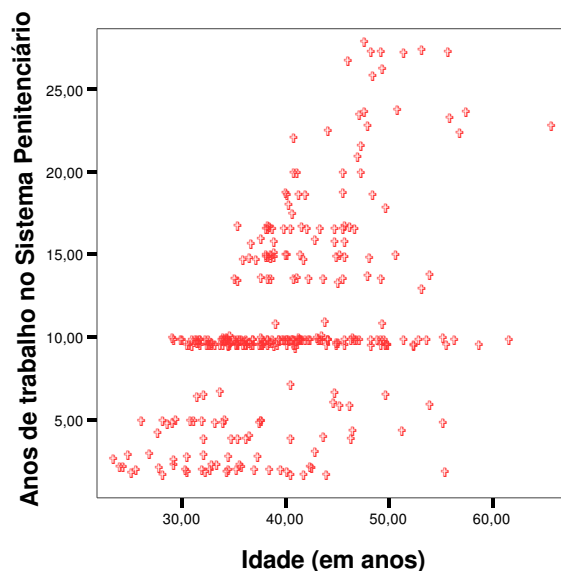
O diagrama de linhas pode ser elaborado com efeito tridimensional, como mostrado abaixo, e permite também considerarmos duas ou três variáveis simultaneamente.

Distribuição dos agentes, segundo a idade,
Região Metropolitana de Salvador, 2.000.



O **diagrama de dispersão** expressa a associação entre duas, três ou quatro variáveis contínuas, através da dispersão de pontos em um espaço bi ou tridimensional. Veja a seguir um exemplo com duas variáveis:

Diagrama de dispersão entre idade e anos de trabalho no Sistema Penitenciário, de agentes penitenciários da Região Metropolitana de Salvador, 2.000.

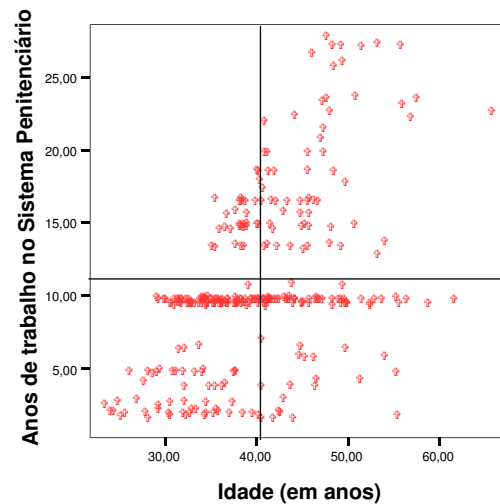


A forma com que os pontos se distribuem pode indicar o tipo de relação matemática entre as duas variáveis representadas. Os pontos podem, por exemplo, distribuir-se de modo a seguirem aproximadamente uma linha reta, ascendente ou descendente. Quando isso ocorrer, podemos prosseguir a análise utilizando a equação de uma linha reta para analisar a associação entre as duas variáveis ou, se o diagrama assim sugerir, usaremos outra função matemática. A técnica que utiliza a equação de uma linha reta para avaliar a associação entre duas variáveis é chamada de análise de regressão linear, que não será abordada neste livro.

Podemos também dividir esse espaço bi-dimensional em quatro quadrantes para facilitar nossa avaliação sobre se há uma concentração maior de pontos em algum ou alguns dos quadrantes. Você verá em outros livros que existe uma análise chamada de correlação, que nos permite quantificar a concentração de pontos nos quadrantes.

Veja na próxima página a divisão do diagrama de dispersão em quadrantes.

Diagrama de dispersão entre idade e anos de trabalho no Sistema Penitenciário, de agentes penitenciários da Região Metropolitana de Salvador, 2.000.



A linha horizontal para definição dos quadrantes corta a ordenada no ponto correspondente à média dos anos trabalhados, e a vertical corta a abscissa no valor da média da idade. Existe um indicador quantitativo para a concentração dos pontos nos quadrantes, chamado coeficiente de correlação de Pearson.

Observe, em seguida, como ficam os diagramas de dispersão com três ou quatro variáveis:

Diagrama de dispersão do tempo de trabalho no Sistema Penitenciário, idade e total de horas semanais de trabalho, de agentes penitenciários, Região Metropolitana de Salvador, 2000.

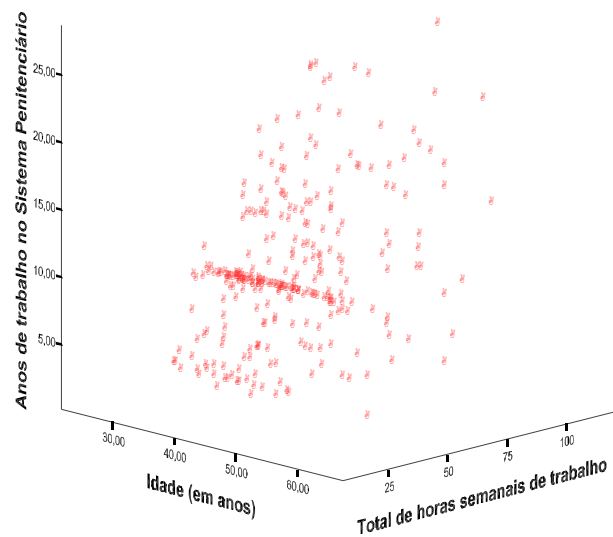
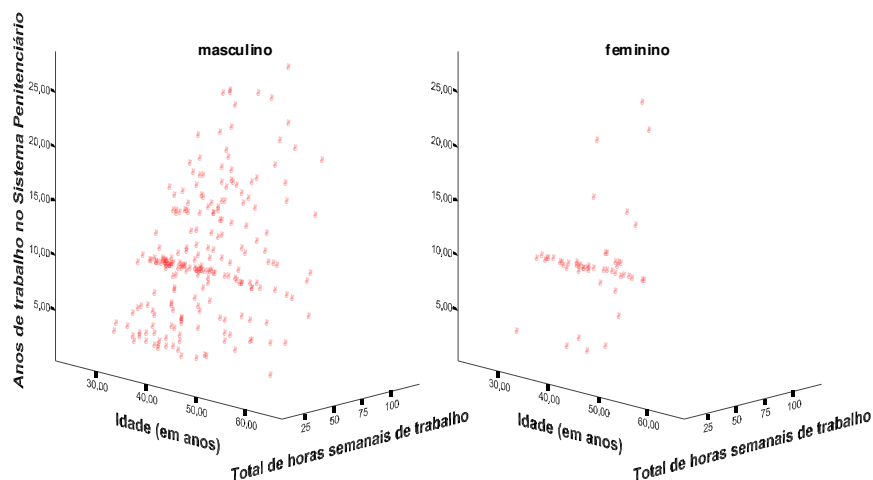
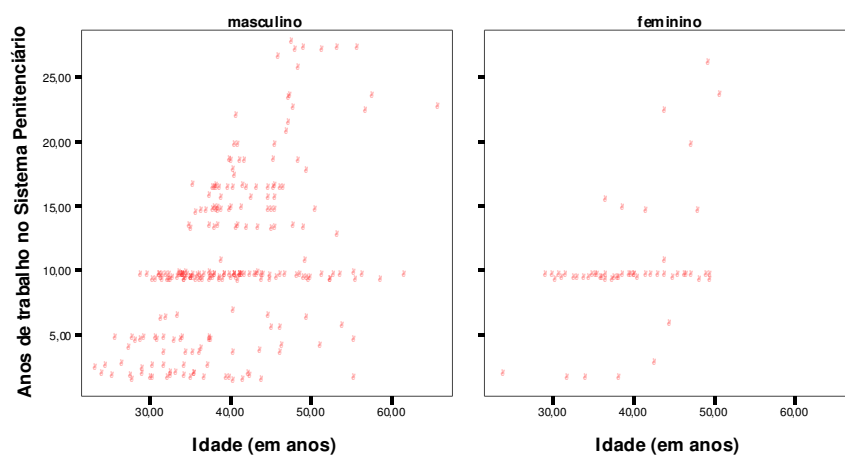


Diagrama de dispersão do tempo de trabalho no Sistema Penitenciário, idade e total de horas semanais de trabalho, segundo o sexo, de agentes penitenciários, Região Metropolitana de Salvador, 2.000.



Uma opção com três variáveis, sendo uma dessas nominal, é mais utilizada do que as apresentadas acima. Veja essa opção a seguir:

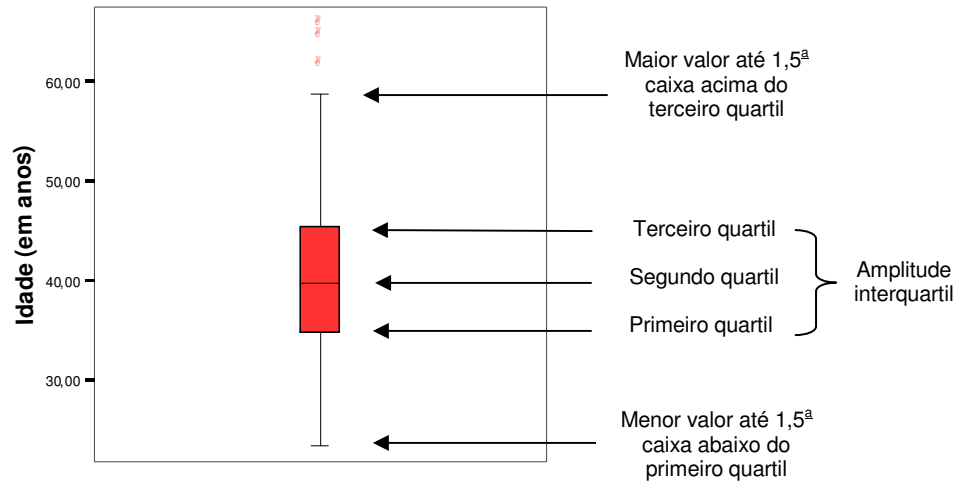
Diagrama de dispersão do tempo de trabalho no Sistema Penitenciário e idade, segundo o sexo, de agentes penitenciários, Região Metropolitana de Salvador, 2.000.



Existe um diagrama, chamado de **diagrama de caixa**, que pode ser utilizado para apresentarmos quartis e amplitude interquartil. Lembra-se dessas medidas?

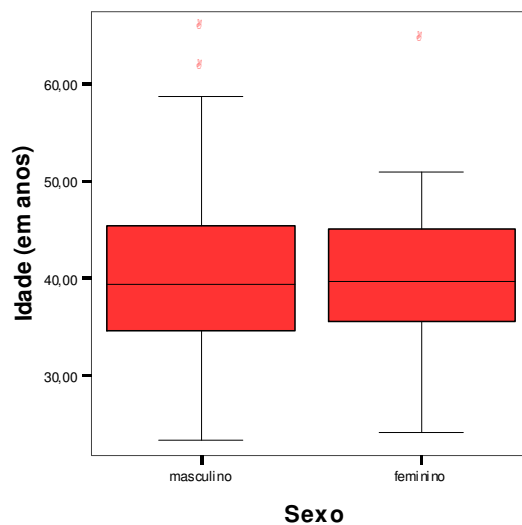
Veja a seguir o diagrama de caixa elaborado para o nosso exemplo, para a variável "idade":

Diagrama de caixa da idade, de agentes penitenciários, Região Metropolitana de Salvador, 2.000.



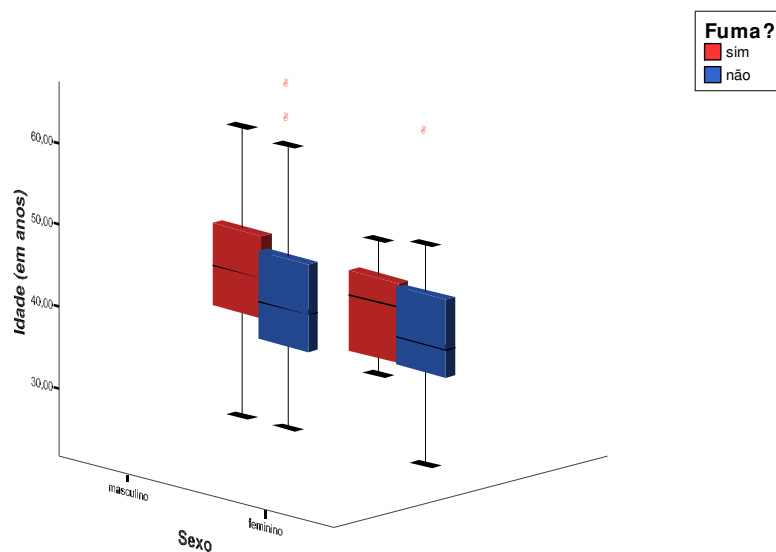
Ao considerarmos mais uma variável (sexo, p. ex.) em um diagrama de caixa, obtemos:

Diagrama de caixa da idade, segundo o sexo, de agentes penitenciários, Região Metropolitana de Salvador, 2.000.



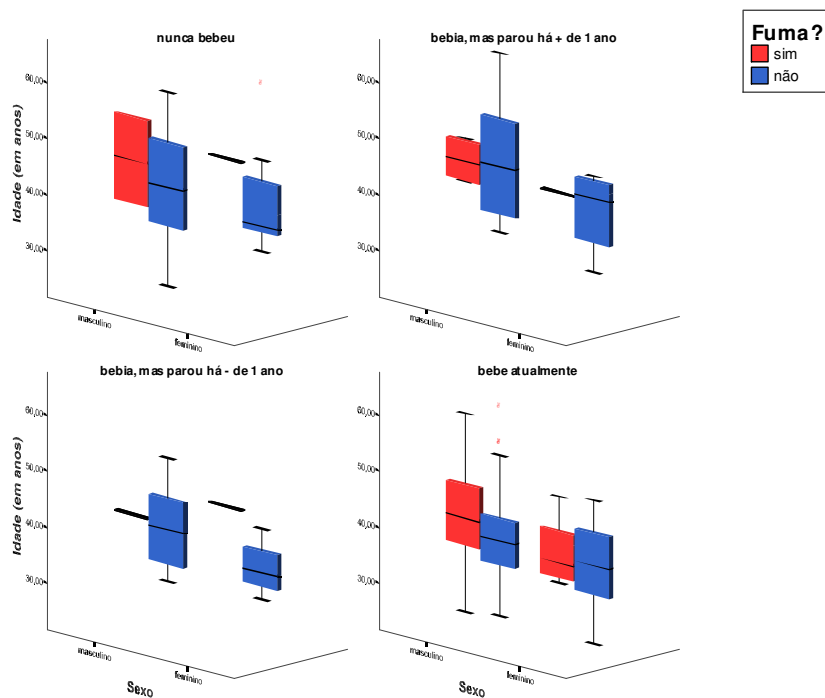
E mais uma, o hábito de fumar:

Diagrama de caixa da idade, segundo o sexo e o hábito de fumar, de agentes penitenciários, Região Metropolitana de Salvador, 2.000.



Uma quarta variável, etilismo, pode ainda ser considerada, do seguinte modo:

Diagrama de caixa da idade, segundo sexo, hábito de fumar e etilismo, de agentes penitenciários, Região Metropolitana de Salvador, 2.000.

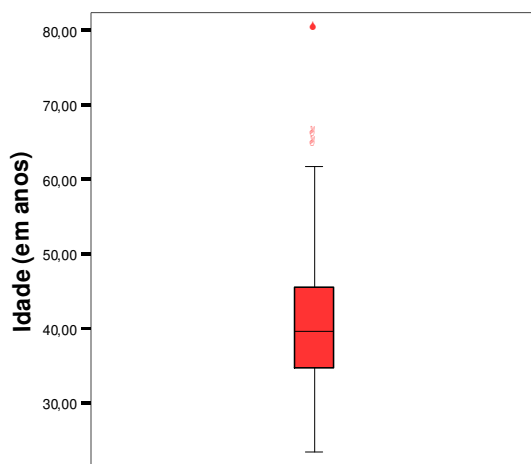


Note que algumas caixas não aparecem no diagrama acima, porque não havia agentes em alguns dos subgrupos definidos pelas diversas combinações das quatro variáveis consideradas. Observe também que é possível utilizarmos o efeito tridimensional.

— **E o que significam os três pequenos círculos vermelhos acima do diagrama?**

— Esses círculos representam agentes com idades tão altas (valores anômalos de idade), a ponto desses círculos estarem localizadas no gráfico em um local que dista entre 1,5 a 3,0 vezes o comprimento da caixa (retângulo em vermelho), a partir do limite superior desta (terceiro quartil). Poderemos também obter círculos abaixo do último traço inferior do diagrama, que indicarão valores muito baixos de idade, situados entre 1,5 e 3,0 comprimentos da caixa obtida, contando para baixo, a partir do primeiro quartil. Valores de idade acima ou abaixo de 3,0 comprimentos da caixa são também apresentados, sendo chamados de valores extremos. Se mudarmos um dos valores anômalos de idade no nosso banco de dados para 80 anos, obteríamos o seguinte diagrama:

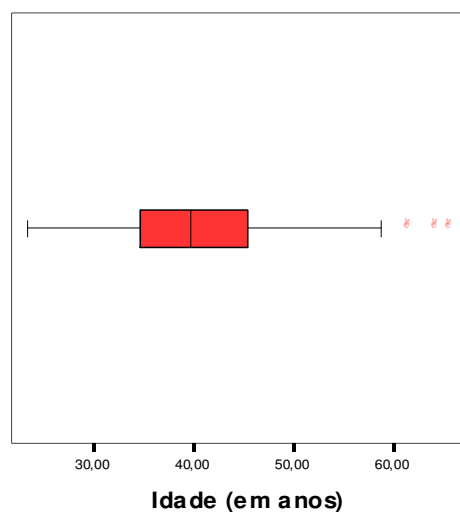
Diagrama de caixa da idade, agentes penitenciários, Região Metropolitana de Salvador, 2.000.



O valor extremo, correspondente à idade 80 anos, aparece representado por um asterisco, na parte de cima do diagrama. Seu afastamento da caixa mostra o quanto esse valor de idade se distanciou do conjunto das outras idades.

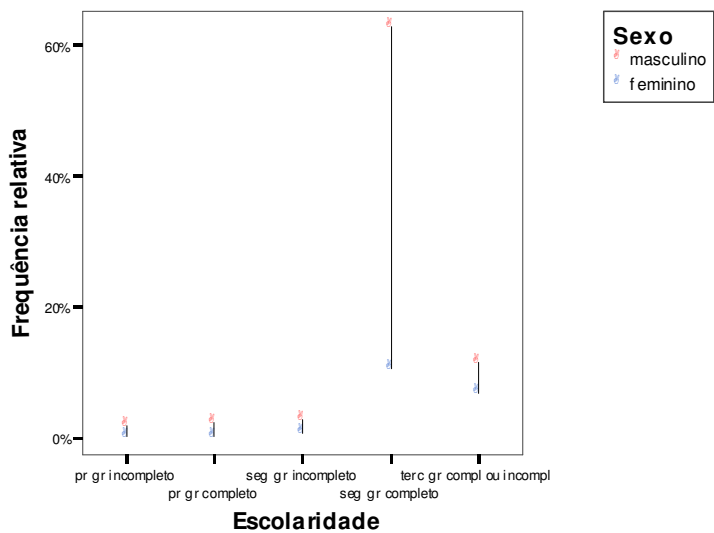
O diagrama de caixa pode ser elaborado horizontalmente, como mostrado a seguir:

Diagrama de caixa da idade, agentes penitenciários,
Região Metropolitana de Salvador, 2.000.



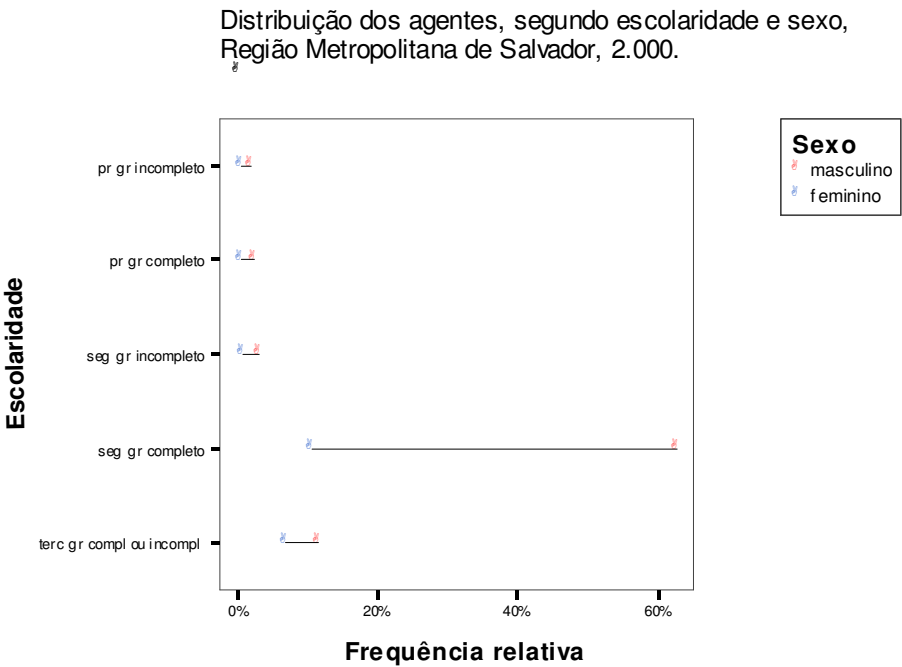
Outra opção para apresentarmos nossos resultados é o **diagrama de linhas de afastamento**. Veja o exemplo abaixo:

Distribuição dos agentes, segundo escolaridade e sexo,
Região Metropolitana de Salvador, 2.000.

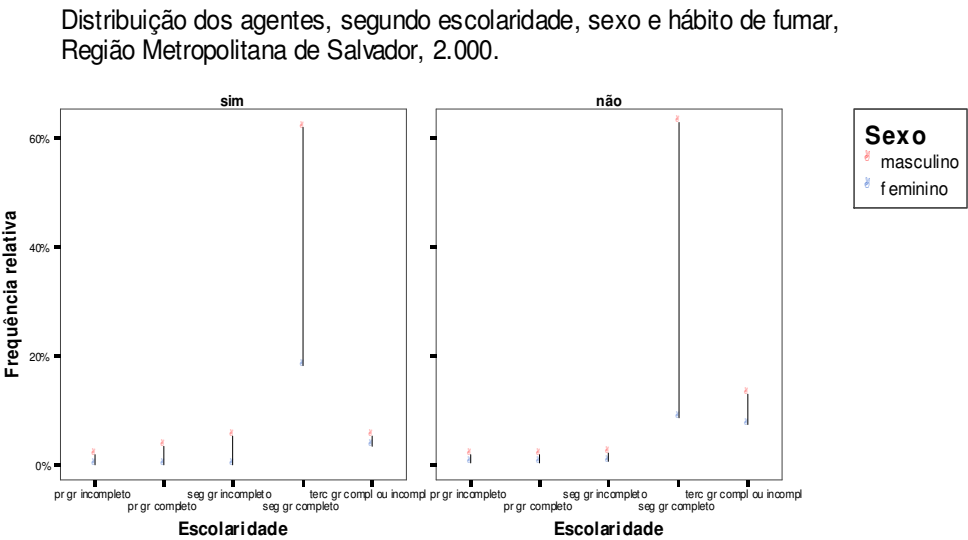


Neste diagrama, a altura das linhas verticais representam a magnitude das diferenças (do afastamento) entre os percentuais de homens e mulheres, em cada nível de escolaridade. Vemos que essa diferença foi acentuadamente maior nos agentes com segundo grau completo. As variáveis incluídas são nominais, podendo também ser ordinais. Esse gráfico não é adequado para apresentarmos variáveis intervalares ou de razão, porque a quantidade de valores (categorias) desses tipos de variáveis é tão grande que ficariam sobrepostos, impossibilitando sua visualização.

Este diagrama pode ser apresentado também no sentido horizontal, como mostramos abaixo:



Veja a seguir que podemos acrescentar mais uma variável (hábito de fumar):



O programa que estamos utilizando não nos permite utilizar o efeito tridimensional no diagrama de linhas de afastamento.

Observe que em todos os gráficos que apresentamos neste capítulo, não levamos em conta os

indivíduos cujo valor para a variável em foco não era conhecido. Se desejarmos, podemos solicitar ao computador que represente também esses indivíduos, embora isso não seja o mais adequado.

— Na apresentação dos resultados devo preferir as tabelas ou os gráficos?

— Em dissertações, teses e artigos científicos, até o momento, há um predomínio do uso de tabelas, sob a justificativa de que elas contêm mais informações. Os gráficos são mais utilizados em apresentações orais ou em cartazes expostos em painéis de seminários e congressos, nas quais se busca uma maior comunicação visual. É possível que a curto ou médio prazo haja uma maior utilização dos gráficos porque, atualmente, com o aperfeiçoamento dos programas de computador, essa forma de apresentação tem ficado cada vez mais fácil de realizar, sendo que os gráficos já conseguem ser quase tão informativos quanto uma tabela.

— Pronto? Esgotamos esse assunto?

— Não. Ainda existem outros diagramas úteis para a análise quantitativa de dados. O **diagrama de seqüência** é utilizado para apresentar e analisar dados relativos a séries temporais (esse diagrama, na verdade, é expresso como um diagrama de pontos, de linhas ou de barras); o **diagrama de controle** é também um tipo especial de diagrama de linha, usado por epidemiologistas para avaliar se está ocorrendo uma epidemia de determinada doença em populações, ou por médicos para verificar a ocorrência de doença em um indivíduo (um exemplo é o diagrama de acompanhamento do peso de uma criança, apresentado na página 79); os **diagramas PP** e **QQ**, aplicados quando desejamos verificar se uma determinada série de observações tem uma distribuição normal; a **curva ROC** que nos ajuda a encontrar o melhor escore de corte para um determinado teste diagnóstico, buscando equilibrar o melhor possível os valores da sensibilidade e especificidade; o **diagrama de metanálise** (também chamado de diagrama de floresta), que nos permite analisar conjuntamente os resultados de vários estudos sobre uma mesma associação entre um determinado fator de risco, de proteção ou de prognóstico e uma determinada doença ou agravo à saúde. Esses diagramas não serão abordados neste livro.

No próximo capítulo, explicaremos as famosas distribuições de frequências.

CAPÍTULO 9

- O que são distribuições de frequências e distribuições probabilísticas e quais as suas aplicações?
 - O que são distribuições reais?
 - O que são distribuições normais?
 - Qual a definição estatística de normalidade?
 - Quais as limitações da definição estatística de normalidade?
 - O que é distribuição normal padrão?
 - Qual a expressão matemática da distribuição normal?
 - Quais as propriedades da distribuição normal?
 - Como calculamos áreas sob as distribuições probabilísticas?
-



— O que são distribuições de freqüências e distribuições probabilísticas e quais as suas aplicações?

— No capítulo 4 (páginas 36 a 40) abordamos os diversos tipos de freqüências utilizados na descrição de dados quantitativos, e no capítulo 8 mostramos que essas freqüências podem ser apresentadas em tabelas ou gráficos. Ao serem assim organizadas, é possível avaliarmos a forma com que as freqüências de uma determinada variável se distribuem em certa amostra ou população finita. Daí que, resultados sobre freqüências de eventos em tabelas ou gráficos, são chamados de distribuições de freqüências. O formato dessa distribuição é mais um recurso disponível para descrevermos os nossos dados.

— **Sim, mas como uma tabela me auxilia a avaliar uma distribuição de freqüências?**

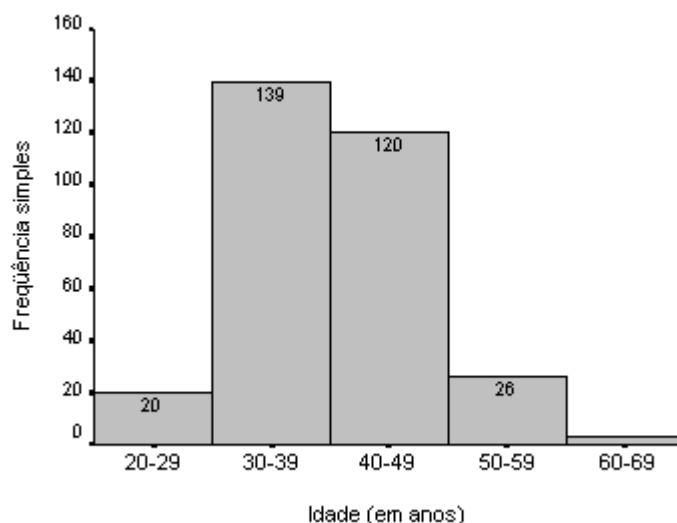
— Vamos voltar a utilizar, como exemplo, a variável “idade”, coletada no estudo sobre condições de trabalho e saúde em uma amostra de 311 agentes penitenciários da Região Metropolitana de Salvador. Os dados foram digitados em um computador e facilmente foi possível obtermos o número de trabalhadores (a freqüência simples) para cada valor de idade. Vamos considerar inicialmente a idade classificada em cinco faixas. Cada uma dessas faixas, como já vimos, é chamada de um “intervalo de classe”. A tabela abaixo apresenta as freqüências obtidas para esses intervalos:

Faixas etárias	Freqüência simples
20,00 a 29,99 anos	20
30,00 a 39,99 anos	139
40,00 a 49,99 anos	120
50,00 a 59,99 anos	26
60,00 a 69,99 anos	3
Total válido	308
Ignorada	3
Total	311

Observe a tabela e note que os valores mais baixos ou mais altos de idade apresentam freqüências mais baixas, enquanto os valores mais centrais apresentam freqüências mais elevadas. Isso nos informa que as freqüências das idades são inicialmente pequenas para valores mais baixos de idade, tendem a ser maiores nos valores mais centrais, voltando a cair para os valores mais altos.

Solicitando ao computador que elabore um histograma para a variável “idade”, no qual a ordenada expresse a freqüência de indivíduos em cada intervalo de classe, e a abscissa as diversas faixas de idade consideradas, obtemos o seguinte:

Distribuição das freqüências de idade nos cinco intervalos de classe.



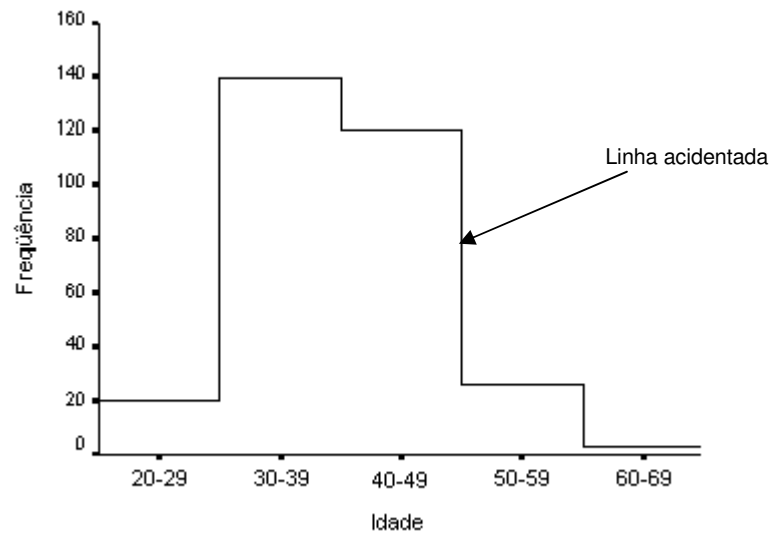
Observe que, como sempre, não levamos em conta os valores não obtidos. Assim, não há nenhuma coluna representando os três indivíduos com idade ignorada.

Lembre-se também de que este diagrama surgiu a partir da colocação dos intervalos de idade na abscissa, e do número (frequência) de trabalhadores em cada um dos intervalos, na ordenada. Cada barra do diagrama atingiu a altura correspondente à frequência de trabalhadores em cada intervalo de idade. Assim, para o primeiro intervalo a barra subiu na ordenada até o valor 20, porque 20 trabalhadores apresentaram idade dentro deste intervalo; a barra seguinte subiu até uma altura correspondente a 139, porque 139 trabalhadores apresentaram idade entre 30 e 39 anos; e assim por diante. Portanto, fica claro que este diagrama é uma maneira bastante eficiente de apresentarmos a distribuição das frequências de uma determinada variável, porque visualmente fica mais fácil verificarmos a subida e a descida das colunas. Um histograma pode ser utilizado para avaliarmos a forma com que as frequências de uma variável se distribuíram. No histograma acima, podemos verificar que há uma frequência pequena no intervalo que contém os trabalhadores mais jovens. Vemos que nos dois intervalos seguintes aparecem as maiores frequências e nos dois últimos, frequências novamente muito baixas, tal como já tínhamos verificado na tabela.

Observe também que a linha que faz o contorno das barras é acidentada, apresentando mudanças de direção de 90 graus, formando como se fossem degraus de uma escada, subindo até a frequência máxima e descendo em seguida até a frequência mínima observada.

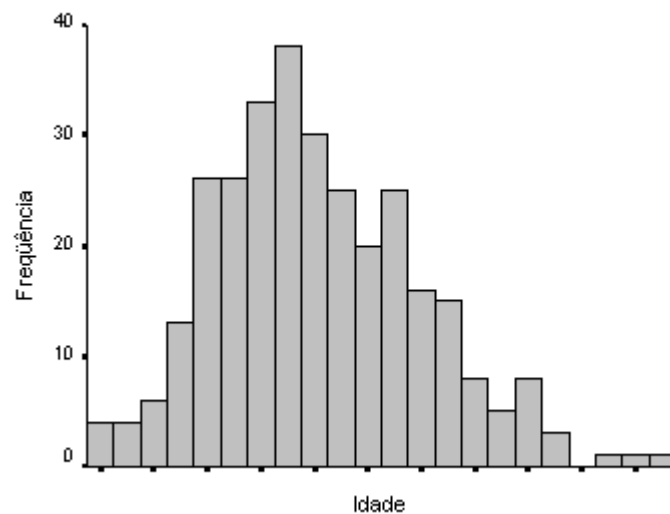
Em breve, abordaremos diversos tipos de distribuições teóricas, que nos ajudarão a avaliar o formato com que as frequências das variáveis estudadas se distribuíram e também a fazer inferência estatística.

Distribuição das frequências de idade em cinco intervalos de classe.

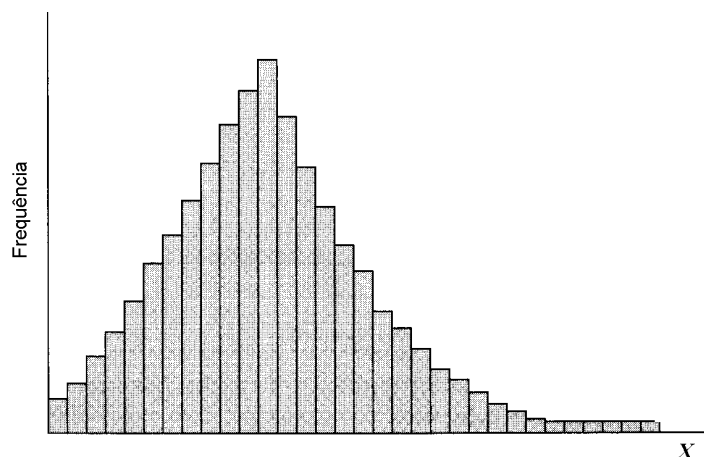


Vamos agora aumentar o número de indivíduos (observações) estudados, n , e o número de intervalos de classe da variável “idade”, para ver o que acontece. Perceba que ao aumentarmos o número de intervalos diminuimos a amplitude (o comprimento) de cada intervalo. O resultado, em termos da distribuição de frequências, pode ser visto abaixo:

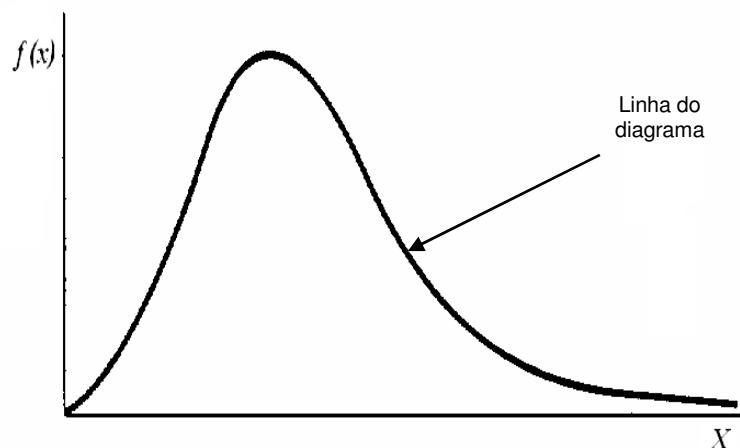
Distribuição das frequências de idade em vinte e dois intervalos de classe.



À medida que o número de observações aumenta e a amplitude dos intervalos de classe diminui, as subidas e descidas da linha vão se tornando menos acentuadas, como mostrado a seguir, sendo que X representa a variável estudada:



À medida que o número de observações tende a infinito e a amplitude de cada intervalo tende a zero (isto equivale a não organizar a variável em intervalos de classe, sendo cada valor de idade representado na abscissa e a freqüência com que cada um desses valores ocorreu representado na ordenada, concorda?), a linha que contorna as barras vai se tornando cada vez mais contínua, até adquirir a forma apresentada a seguir:



Na teoria estatística, quando n tende a infinito, a ordenada de uma distribuição de freqüências torna-se uma função de X , denotada por $f(x)$, indicando, portanto, os resultados obtidos quando infinitos valores possíveis de X são substituídos em uma equação matemática que expressa a linha do diagrama. Prossiga, pois isso ficará mais claro nas próximas páginas.

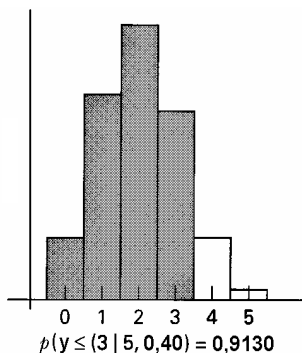
Outra característica importante dessas distribuições, nesse contexto teórico de população infinita, é que a área entre a curva (a linha) e a abscissa, equivale à probabilidade de ocorrência dos valores de X . Por isso, como já afirmamos na página 8, nessa situação, essas distribuições são chamadas de **distribuições de**

probabilidades ou probabilísticas.

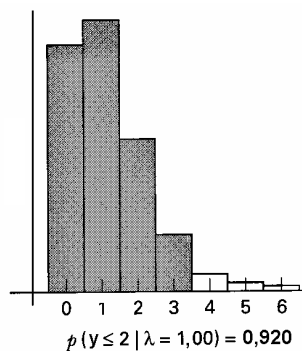
As distribuições probabilísticas de variáveis contínuas como a idade, apresentarão uma forma semelhante à do diagrama acima.

Já as distribuições probabilísticas de variáveis discretas, assemelham-se àquela vista inicialmente quando consideramos a idade em intervalos de classe. Apenas para lhe informar, mostramos rapidamente, duas das mais importantes distribuições probabilísticas de variáveis discretas: a distribuição binomial e a de Poisson.

Distribuição binomial



Distribuição de Poisson

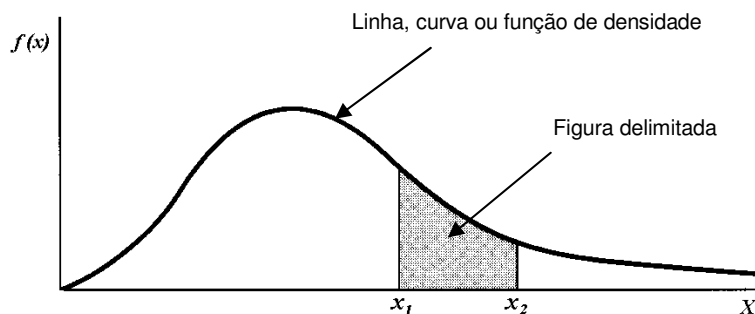


Tanto a distribuição binomial quanto a de Poisson constituem uma família de distribuições, ou seja, não existe uma única distribuição, mas várias, a depender dos valores dos parâmetros que as definem. As distribuições acima representam apenas uma das possíveis distribuições binomiais ou de Poisson. Mas o formato geral dessas distribuições é mais ou menos esse visto acima, sendo nossa intenção aqui apenas mostrar a “cara” das mesmas, pois não as utilizaremos neste livro.

Vários estatísticos, ao longo do tempo, têm demonstrado que as distribuições probabilísticas têm certas propriedades que as tornam muito úteis. Uma dessas propriedades, já mencionada, é que a área sob a curva corresponde à probabilidade de ocorrência de valores de X .

— O quê?

— Vamos explicar melhor. Observe o diagrama abaixo:



No diagrama acima e ao longo de todo o livro, utilizaremos letra maiúscula para representar todos os valores de uma variável e letra minúscula para indicar um valor específico dessa variável, exceto para os valores de F , como veremos mais adiante. Por isso no diagrama acima a variável está representada em letra maiúscula, X , ao final da abscissa (poderia ser no centro ou no início), e dois valores específicos desta são indicados por letras minúsculas, x_1 e x_2 .

Pode ser demonstrado que, ao rebatermos os pontos x_1 e x_2 na linha (é comum os estatísticos dizerem na “curva” ou na “função de densidade”), a área da figura que fica delimitada sob a curva (área sombreada) corresponde à probabilidade de obtermos valores de X entre os valores x_1 e x_2 . Se considerarmos que X representa a idade de agentes penitenciários, e que x_1 e x_2 sejam, respectivamente, 35 e 45 anos, a área sombreada (dizemos também “a área sob a curva”) corresponderá à probabilidade de obtermos agentes com idade entre 35 e 45 anos.

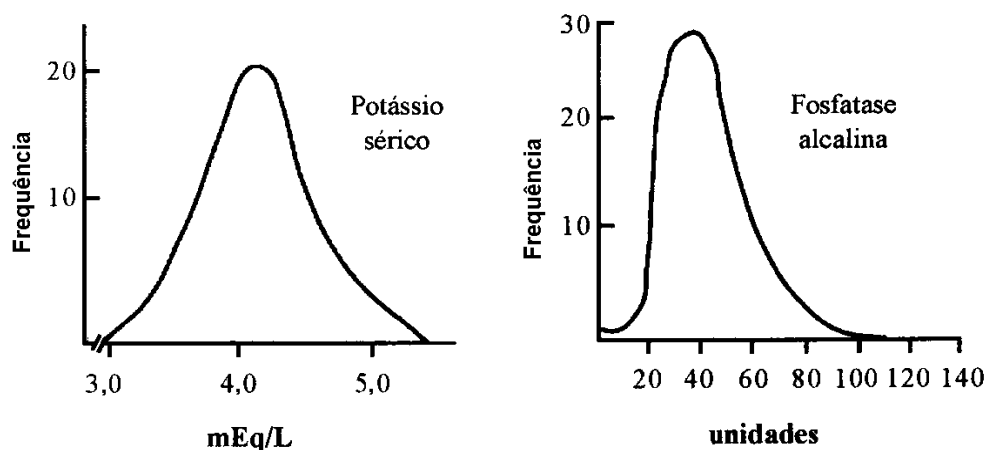
— **E como calculamos essa ou outra área qualquer sob a curva?**

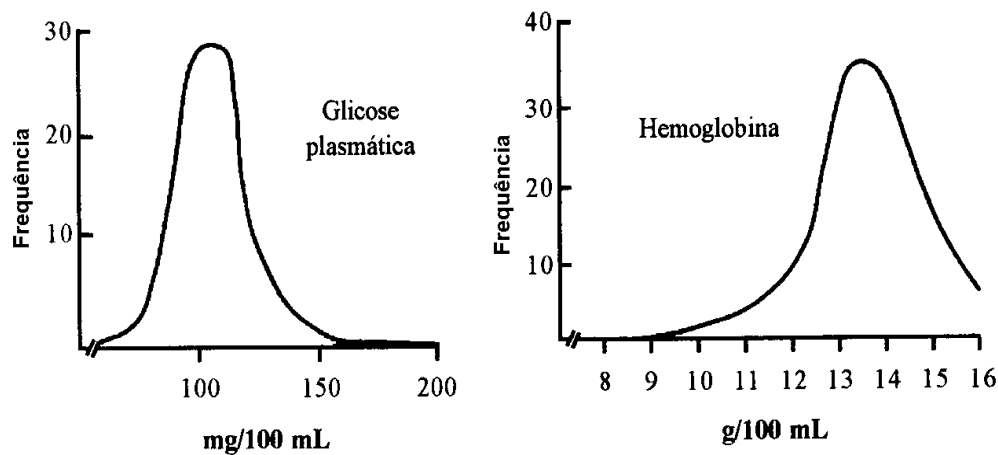
— Note que a figura delimitada não é uma figura geométrica perfeita. É bom lembrarmos aqui algumas das figuras geométricas perfeitas, caso você não esteja se lembrando. O círculo, o triângulo, o retângulo e o quadrado são figuras geométricas perfeitas. Veja então que a figura sombreada no diagrama acima não é nenhuma destas, sendo, portanto, uma figura geométrica imperfeita. Você aprendeu a calcular a área de figuras geométricas perfeitas no segundo grau de educação. A área do círculo, por exemplo, é igual à constante π vezes o seu raio elevado ao quadrado, πr^2 , sendo seu raio a distância entre o centro do círculo e qualquer ponto do seu perímetro. Os seres humanos levaram muito tempo até poderem calcular diretamente a área de figuras geométricas imperfeitas. Um passo fundamental foi o desenvolvimento do cálculo integral feito por vários matemáticos, embora mereçam destaque as contribuições de Isaac Newton (1643-1727) e Gottfried Leibniz (1646-1716). Foi necessário ainda que outros indivíduos, entre estes Georg Riemann (1826-1866), adequassem o cálculo integral ao objetivo de calcular a área de figuras geométricas imperfeitas. Essa é a forma atual de calcularmos áreas sob uma determinada curva. Resumindo, para o cálculo da área sob a curva em foco, a função de densidade é integrada de x_1 a x_2 .

— **Então não vou poder continuar porque não sei nada sobre cálculo integral.**

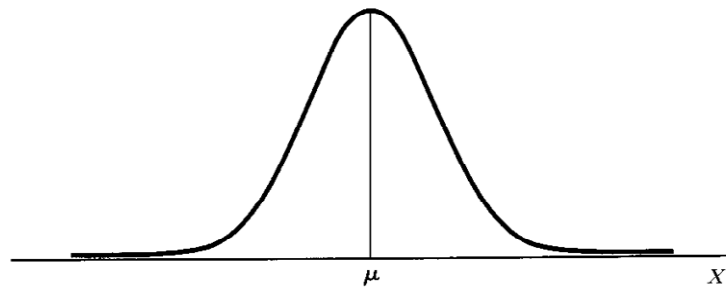
— Não desanime, pois não será necessário integrarmos funções de densidade. Alguns matemáticos e estatísticos fizeram isto para nós, existindo tabelas prontas com todas as áreas possíveis já calculadas para as distribuições probabilísticas mais conhecidas e utilizadas, facilmente encontradas nos livros-texto e incorporadas aos programas estatísticos de computador. Você aprenderá neste livro como usar essas tabelas.

Considere agora que você mediu alguns parâmetros de interesse clínico em certo número de indivíduos, e elaborou diagramas de linha para cada um desses parâmetros. Se o número de indivíduos estudados tiver sido suficientemente grande, os diagramas devem ter apresentado os seguintes formatos (dados de Martin HF, Gudzinowicz BJ, Fanger H. *Normal values in clinical Chemistry*. New York, Marcel Dekker, 1975):





As distribuições de freqüências acima são chamadas de **distribuições reais**, porque foram elaboradas com dados de determinados indicadores clínicos obtidos em populações concretas (reais) de tamanho finito. Observe-as atentamente. Note que todas apresentam freqüências mais baixas para os valores mais baixos e mais altos, e freqüências maiores para os valores mais centrais da distribuição, formando uma curva que começa bem próxima à abscissa, vai se elevando até atingir seu ápice, e depois declina até aproximar-se novamente da abscissa. Muitas distribuições reais, portanto, apresentam uma forma semelhante à apresentada a seguir:



Nesta distribuição, a média, μ (você verá mais adiante que a moda e a mediana também) ocupa a posição mais central.

Ao observar tal comportamento em várias distribuições de fenômenos da natureza, incluindo indicadores relacionados ao processo saúde-doença tais como os mencionados acima, alguns estatísticos desenvolveram uma curva teórica para população infinita, com propriedades semelhantes, e que pode ser utilizada como modelo para representar algumas distribuições reais. Em 12 de novembro de 1733, Abrahan de Moivre publicou um artigo apresentando a fórmula matemática de uma distribuição com o formato acima. Mostramos essa fórmula a seguir:

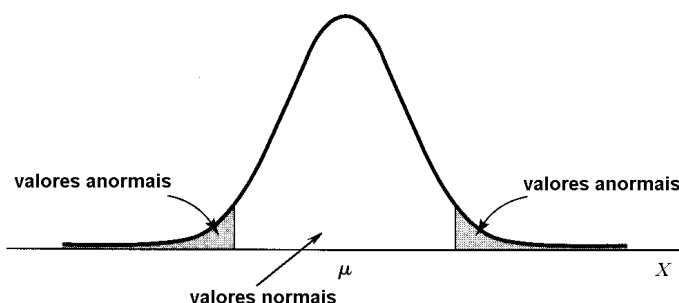
$$f(x) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(x-\mu)^2}{2\sigma^2}}, -\infty < X < \infty.$$

Veja então, que uma distribuição probabilística pode também ser expressa por uma fórmula matemática.

A distribuição desenhada acima é denominada por **distribuição normal**. É possível que essa curva tenha sido chamada de normal, por representar o comportamento natural de certos fenômenos. Posteriormente, tal denominação deve ter sido reforçada pelo uso dessa distribuição para definir normalidade, ou seja, para estabelecer um limite entre o que deve ser considerado normal ou patológico (anormal). Indivíduos com valores muito baixos ou muito altos de determinado indicador biológico devem, por este modelo, ser considerados anormais.

— **Sim, mas o que é um valor muito baixo ou alto?**

— A distribuição normal é utilizada justamente para nos ajudar a avaliar isso. Valores situados nos extremos da curva (valores menos freqüentes) são considerados anormais, enquanto valores mais centrais (mais freqüentes) são considerados normais. Veja o diagrama abaixo:



Esse é o famoso **critério estatístico para definição de normalidade**. Há muitas limitações nesse critério (Fletcher RH, Fletcher SW e Wagner EH. *Epidemiologia Clínica: Elementos Essenciais*. 3ª ed., Artes Médicas, Porto Alegre, 1996.):

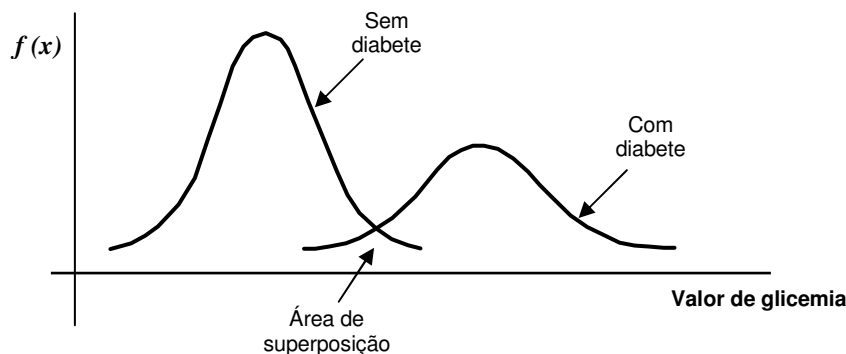
- a) Nem todos os parâmetros biológicos indicativos de doença se distribuem de modo normal; nesses casos, não teria cabimento utilizarmos a distribuição normal como modelo para definição de normalidade;
- b) Se todos os valores abaixo ou acima de um mesmo limite estatístico arbitrário fossem considerados anormais, a prevalência de todas as doenças seria a mesma, o que evidentemente não ocorre na realidade: não há um limite estatístico único para todas as doenças, além do qual se possa garantir que há doença clínica; para algumas doenças este limite é 5% dos valores mais altos ou mais baixos; para outras somente a partir do limite que define 1% dos valores mais altos ou mais baixos, se consegue identificar claramente a presença de doença; não há, portanto, uma relação constante entre raridade estatística e presença de doença;
- c) Para muitas doenças não há um limite definido a partir do qual a doença apareça; por exemplo, à medida que o nível de colesterol sérico aumenta, a probabilidade de o indivíduo ter doença coronariana também aumenta, não existindo um nível específico a partir do qual a doença ocorra;
- d) Nem sempre um valor muito alto ou muito baixo significa anormalidade; por exemplo, uma pressão arterial sistólica (PAS) de 100 mmHg é um valor situado na cauda inferior da distribuição de PAS, mas resulta em um risco menor de adoecer;
- e) Nem sempre um valor mais comum (localizado na região central da distribuição) significa

normalidade: há casos de hidrocefalia com pressão intracraniana baixa, glaucoma com pressão ocular normal e hiperparatireoidismo com cálcio normal.

Um aperfeiçoamento desse critério estatístico de normalidade é a utilização de duas distribuições, uma para indivíduos que com certeza têm a doença e outra para aqueles que com certeza não a têm.

— **Ora! Mas como vou ter certeza de que o indivíduo está realmente doente? Não estamos vendo justamente as dificuldades de fazermos isso?**

— Considerando a diabetes como exemplo, utilizaremos um conjunto de procedimentos diagnósticos (anamnese, que inclui a coleta de informações epidemiológicas, exame físico, testes laboratoriais, biópsia, ou autópsia e exames histopatológicos em pacientes que foram a óbito) que nos permitam ter certeza ou quase certeza, sobre se um indivíduo tem ou tinha a doença. Em seguida, mediremos a concentração de glicose no sangue (ou verificaremos qual foi a glicemia registrada nos prontuários médicos dos pacientes já falecidos). Poderemos representar as distribuições das freqüências dos valores encontrados, pelas distribuições probabilísticas abaixo, sendo a primeira usada para modelar os indivíduos que, pelo conjunto de procedimentos diagnósticos de certeza não tinham diabetes, e a outra para aqueles que, pelos mesmos procedimentos, tinham essa doença:



Note que, como esperado, os níveis glicêmicos dos indivíduos sem diabetes encontram-se em um espectro de valores mais baixos, e os daqueles com diabetes em uma faixa de valores mais altos. Observe também, que o ápice da curva dos sadios é mais alto do que o da curva para os diabéticos, indicando freqüências maiores para os primeiros, já que existem mais indivíduos sadios do que doentes. Mas, além disso, verifique que as duas curvas se superpõem.

— **E agora? Se as duas curvas se superpõem, como poderemos definir a partir de que valor de glicemia um indivíduo será considerado diabético?**

— Os clínicos e epidemiologistas consideram como limite de normalidade, separadamente, cada um dos valores de glicemia dentro da área de superposição das curvas, e calculam indicadores de validade (sensibilidade, especificidade, etc.). Aquele valor de glicemia que apresentar as maiores sensibilidade e especificidade será considerado como o melhor limite (escore de corte) para separar os sadios dos doentes. Os resultados deste procedimento são apresentados em diagrama denominado “Curva ROC”, mas esse tópico não será desenvolvido neste livro. A depender das características da doença em foco e do seu tratamento, podemos também escolher um escore de corte que aumente um desses indicadores de validade mais do que o outro.

Outros critérios para definição de normalidade, além do critério estatístico visto acima, são: o **critério clínico**, pelo qual só consideraremos como doente aqueles indivíduos que apresentarem alterações clinicamente significativas em termos de perda do bom estado de saúde, e o **critério terapêutico**, segundo o qual o indivíduo só será considerado doente se existir tratamento eficaz para sua doença.

Voltando ao nosso tema principal neste capítulo, é importante lembrarmos que após a contribuição inicial de Abraham de Moivre, outros matemáticos e estatísticos colaboraram para o desenvolvimento teórico da distribuição normal. Entre eles se destacou Carl Friedrich Gauss (1777-1855). Suas contribuições foram tão importantes que a distribuição normal é denominada também por **distribuição de Gauss** ou **distribuição gaussiana**.

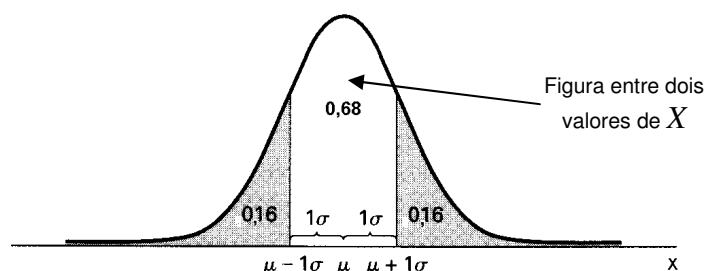
A distribuição normal tem algumas **propriedades matemáticas** muito importantes:

1ª propriedade:

A área delimitada entre dois valores quaisquer de X corresponde à probabilidade de obtermos valores de X entre esses dois valores.

Já havíamos mencionado essa propriedade, que é comum a todas as distribuições probabilísticas.

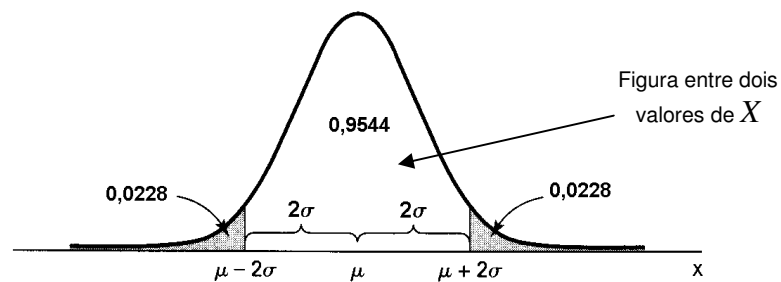
Veja a distribuição normal abaixo:



Para simplificar, não estamos desenhando a ordenada, que representa os valores de $f(x)$, indicando até aonde a distribuição sobe ou desce, ou seja, probabilidades de ocorrência maiores ou menores.

Se considerarmos os valores de X equivalentes à média mais ou menos um desvio-padrão, e rebatermos esses pontos na curva, obteremos uma figura cuja área corresponderá à probabilidade de obtermos valores de X entre $x = \mu - 1\sigma$ e $x = \mu + 1\sigma$. Esta área foi calculada através de cálculo integral como sendo 0,68. Isso significa que a probabilidade de obtermos valores de X dentro de um intervalo de mais ou de menos um desvio-padrão em torno da média é de 68% (0,68 é igual a 68/100 ou 68 por cento, não é?). Observe que as áreas das duas caudas da curva (uma à esquerda, outra à direita) correspondem cada uma a 16%, que é a probabilidade de obtermos valores muito altos ou muito baixos de X , evidenciando que a área total sob a curva equivale a uma probabilidade de 100% de obtenção dos valores de X .

Do mesmo modo, se escolhermos os dois valores de X correspondentes à média, mais ou menos dois desvios-padrão, obteremos a figura apresentada na próxima página. A área desta figura é 0,9544, o que significa que a probabilidade de obtermos valores de X entre a média e mais ou menos dois desvios-padrão é de 95,44%.



Você verá em breve que uma área muito utilizada em Estatística Inferencial é aquela que representa exatamente uma probabilidade de 95% dos valores de X ocorrerem. Essa área é delimitada pelos valores $x = \mu - 1,96\sigma$ e $x = \mu + 1,96\sigma$. Veja que esta situação é muito parecida com a que acabamos de mostrar acima, na qual a área delimitada era 95,44%. Observe que ali os valores de X que delimitavam a área eram $x = \mu - 2\sigma$ e $x = \mu + 2\sigma$, e agora são $x = \mu - 1,96\sigma$ e $x = \mu + 1,96\sigma$. A única diferença é que agora consideramos 1,96 desvios-padrão acima ou abaixo da média, e não 2. Mas, como 1,96 é aproximadamente igual a 2, é muito comum os bioestatísticos considerarem 1,96 como sendo 2, mesmo ao se referirem aos valores de X que delimitariam uma área exatamente equivalente a 95%. Não devemos porém esquecer que o valor exato é 1,96.

Considerando o nosso exemplo atual, e sabendo que a média aritmética dessa série de idades é 40,27 anos e seu desvio-padrão 7,60 anos, se pudermos assumir que a distribuição de freqüências das idades seja normal, podemos afirmar que a probabilidade dos agentes terem idade entre 25,374 e 55,166 anos é de 95%. Obtivemos o primeiro valor, calculando:

$$40,27 - [1,96(7,60)] = 40,27 - 14,896 = 25,374 ;$$

e o segundo, calculando:

$$40,27 + [1,96(7,60)] = 40,27 + 14,896 = 55,166 .$$

Aproximando os valores, já que estamos expressando a idade em anos completos, podemos dizer que a probabilidade dos agentes apresentarem idade entre 25 e 55 anos é de 95%. Isto é o mesmo que afirmar que, se pudermos assumir que os resultados obtidos na amostra estudada sejam aplicáveis a todos os agentes penitenciários, ao escolhermos aleatoriamente um desses agentes, a probabilidade da idade do agente sorteado estar entre 25 e 55 anos é de 95%.

Vimos, então, que as áreas sob as distribuições probabilísticas correspondem às probabilidades de ocorrência de valores de uma determinada variável. Essas distribuições são teóricas e pressupõem que estejamos considerando populações infinitas.

As distribuições de freqüências, por sua vez, são reais e elaboradas para amostras ou populações finitas. As áreas sob suas curvas representam as proporções com que os valores de determinada variável ocorreram na amostra ou população finita estudada.

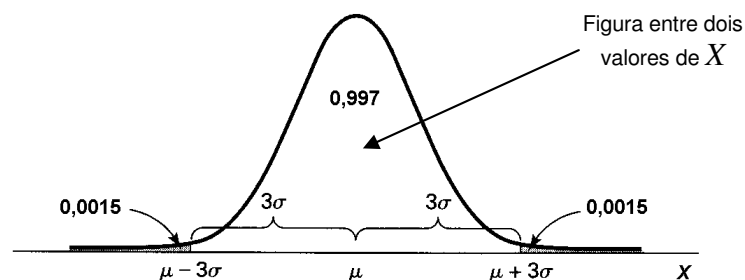
Você verá nas próximas páginas, que os estatísticos utilizam distribuições probabilísticas como modelo gráfico e matemático para as distribuições de freqüências. A finalidade é lançar mão das propriedades teóricas das primeiras como “ferramentas” para inferir os resultados obtidos em uma amostra para a população mais ampla de onde esta foi retirada.

Resumindo:

Podemos definir as distribuições de freqüências como formas de sumarização da relação entre os valores de uma variável e a proporção com que ocorreram em uma amostra ou população finita.

Podemos definir as distribuições probabilísticas como formas de sumarização da relação entre os valores de uma variável e sua probabilidade de ocorrer em uma população infinita.

Continuando nosso raciocínio, se considerarmos a área da figura central delimitada pelos valores $x = \mu - 3\sigma$ e $x = \mu + 3\sigma$, veremos que essa área compreenderá 99,7% dos valores de X , como nos mostra o diagrama abaixo:



A distribuição normal tem outras propriedades matemáticas:

2ª propriedade:

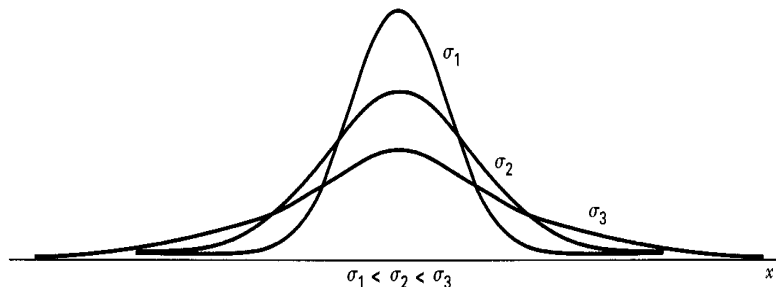
É completamente determinada por sua média e seu desvio-padrão.

Olhe novamente com atenção a fórmula matemática que expressa a distribuição normal:

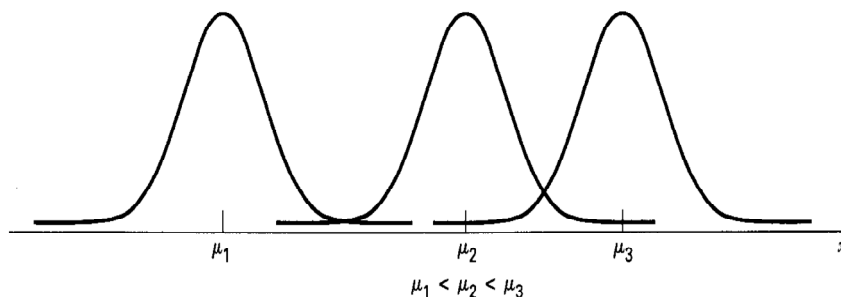
$$f(x) = \frac{1}{\sqrt{2\pi}\sigma} e^{- (x-\mu)^2 / 2\sigma^2}, -\infty < x < \infty.$$

Verifique que π e e são constantes cujos valores são aproximadamente 3,1416 e 2,7183, respectivamente. Os valores de X não são constantes (variam de menos infinito a mais infinito) e é justamente sua variação que vai fazer com que os valores de $f(x)$ também variem. A notação $f(x)$ é lida “função de xis” e indica justamente os valores que vão resultando da fórmula acima, à medida que os valores de X variam, isto é, indicam valores que variam em função de X . Assim, ao alocarmos no diagrama cada ponto gerado por um determinado valor de X , e seu valor correspondente de acordo com a fórmula acima (valor de $f(x)$), veremos que irá surgindo em um espaço bi-dimensional uma distribuição normal. Resta saber

qual a distribuição normal a ser formada. Isso dependerá dos outros parâmetros da fórmula: a média, μ , e o desvio-padrão, σ , da variável a ser representada na distribuição de probabilidades. O que irá distinguir uma distribuição normal de outra serão os valores de suas médias e desvios-padrão. Podem existir distribuições normais com a mesma média e diferentes desvios-padrão, como mostramos a seguir:

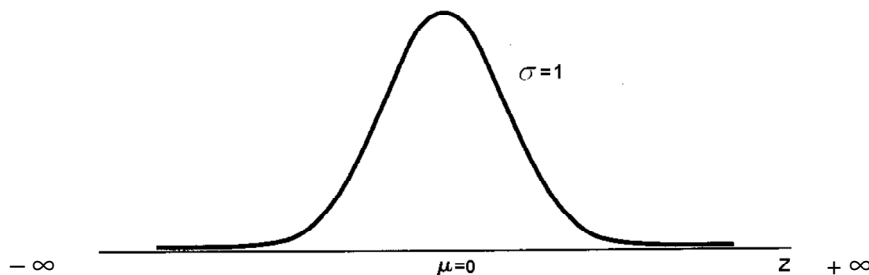


Ou com médias diferentes e desvios-padrão iguais:



Mas, se duas distribuições normais apresentarem a mesma média e desvio-padrão, essas distribuições serão absolutamente iguais, concorda? É justamente por isso que se diz que a distribuição normal é completamente determinada pelos parâmetros μ e σ .

Pelo exposto, existem várias distribuições normais, a depender dos diversos valores de μ e σ . Entretanto, há uma distribuição normal mais importante que é a **distribuição normal padrão**. Você verá no próximo capítulo, que sua média e desvio-padrão são 0 e 1, respectivamente. Assim, ao substituirmos os valores de π , e , μ e σ , na equação da distribuição normal, à medida que formos variando os valores de X de menos infinito a mais infinito, iremos obtendo os valores de $f(x)$ que, ao serem considerados como coordenadas juntamente com os valores de X , darão origem graficamente à distribuição normal padrão, tal como você pode ver a seguir:



Observe que na distribuição normal padrão os valores de X são expressos em valores de Z . Estes últimos são expressos em unidades de desvios-padrão, ou seja, se $z = 2,50$, isso significaria que z estaria representando um valor de X que estaria localizado dois desvios-padrão e meio acima da média. Do mesmo modo, se $z = -1,50$, z estaria representando um valor de X que estaria um desvio-padrão e meio abaixo da média.

— Qual a vantagem de transformarmos valores de X em valores de Z , se o que estamos estudando é a variável X ?

— É que, como já foi mencionado, alguns estatísticos calcularam todas as áreas sob a curva normal padrão, isto é, considerando valores de Z , e colocaram esses resultados em uma tabela chamada de **tabela Z** , que pode ser encontrada em quase todos os livros de estatística e que é usada nos programas estatísticos de computador. A vantagem então é que, ao utilizarmos valores de Z , poderemos facilmente encontrar nessa tabela a área sob a curva que seja do nosso interesse, sem precisar fazer cálculo integral, que já foi feito para nós previamente. Assim, para encontrarmos a área sob a curva entre dois valores de Z , digamos, entre z_0 e z_1 , temos duas alternativas: avaliarmos a integral

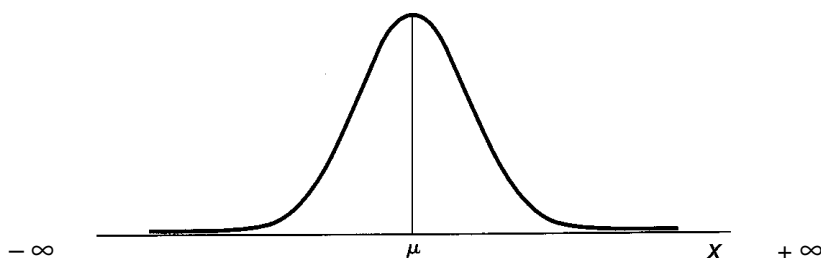
$$\int_{z_0}^{z_1} \frac{1}{\sqrt{2\pi}} e^{-\frac{z^2}{2}} dz,$$

ou olharmos a área já calculada na tabela. Qual delas você prefere?

3ª propriedade:

É simétrica em relação à sua média.

Outra propriedade da distribuição normal é que sua forma é simétrica, como pode ser visto no diagrama a seguir:

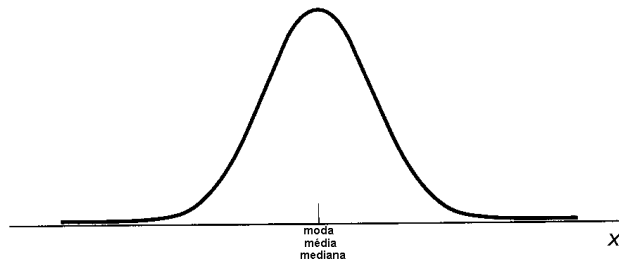


As áreas compreendidas entre a linha central passando por μ e menos infinito, e entre aquela e mais infinito, são exatamente iguais, formando uma distribuição simétrica.

4ª propriedade:

Sua média, mediana e moda são iguais.

Obviamente, como consequência de sua simetria, a curva normal apresenta média, mediana e moda iguais, já que todas essas são medidas de tendência central dos valores de X :

**5ª propriedade:**

A distribuição normal é assintótica, isto é, seus extremos à esquerda e à direita nunca tocam a abscissa.

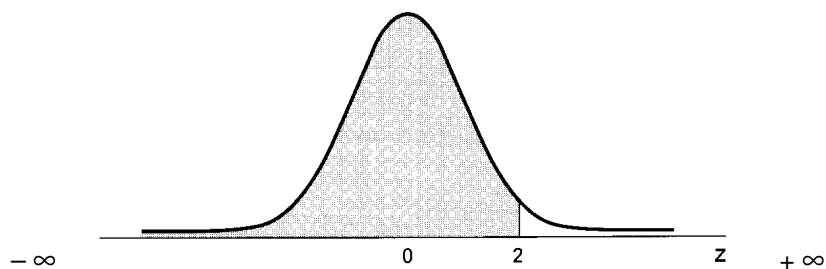
Se considerarmos as distribuições reais (aquelas realmente encontradas na natureza), fica evidente que essa e outras propriedades da distribuição normal não se aplicam. Ou você conhece alguém com idade, peso ou altura, iguais a menos infinito ou a mais infinito?

— **E aí! Como vamos utilizar a curva normal como modelo para representar distribuições reais, se estas não são assintóticas, podem ter média, mediana e moda diferentes e podem não ser simétricas?**

— No próximo capítulo, você verá que o cálculo de áreas sob a curva Z é essencial para fazermos inferência estatística. Acontece que as áreas sob a curva, nos extremos desta, são muito pequenas, não influenciando significativamente os resultados, nos quais basearemos nossa conclusão. Além disso, as técnicas estatísticas chamadas de “paramétricas” são razoavelmente robustas, permitindo um certo grau de violação dos seus pressupostos. Ainda assim, um bom estatístico deve utilizar os vários procedimentos existentes para verificação de normalidade da distribuição, quando isso for necessário. Tais procedimentos não serão vistos neste livro, pois estão fora dos nossos objetivos.

Para terminarmos o capítulo, vamos treinar juntos a obtenção de áreas sob a curva normal padrão.

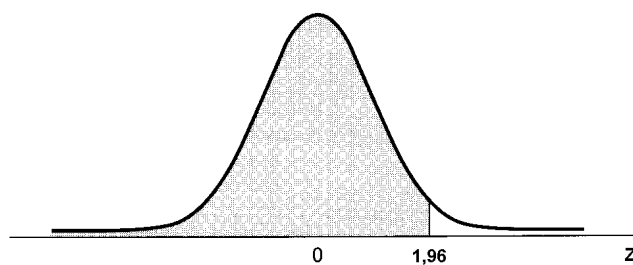
Qual a área sob a curva correspondente a valores de Z menores do que 2,00? O diagrama a seguir mostra a área a ser calculada (área sombreada):



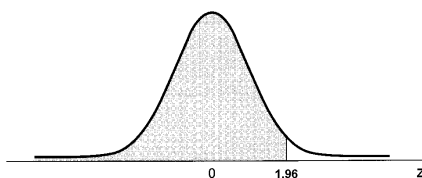
Olhando na tabela Z , apresentada na próxima página, verificamos que a área sob a curva entre $-\infty$ e $z = 2,00$ é 0,9772 (veja sombreamento do diagrama acima da tabela indicando qual o conteúdo das suas células).

No topo da tabela aparece escrito que esta é uma continuação. É que omitimos a primeira parte da tabela com os valores negativos de Z (veja essa parte na página 176), que não foi apresentada agora porque no nosso exemplo atual o valor de z é positivo. Além disso, como a distribuição Z é simétrica, na prática é possível encontrarmos as áreas de interesse apenas com uma das partes da tabela.

É importante também que você olhe o diagrama que é sempre apresentado no topo da tabela, e que, neste caso, é o seguinte:



O diagrama nos indica que, em cada célula da tabela Z , consta o valor da área entre $-\infty$ e um determinado valor de Z que seja do nosso interesse. Para sabermos isto, nos baseamos na área sombreada do diagrama. Em outros livros, você poderá encontrar uma tabela Z que apresente as áreas entre 0 e z , ou entre z e mais infinito, e não entre menos infinito e z .

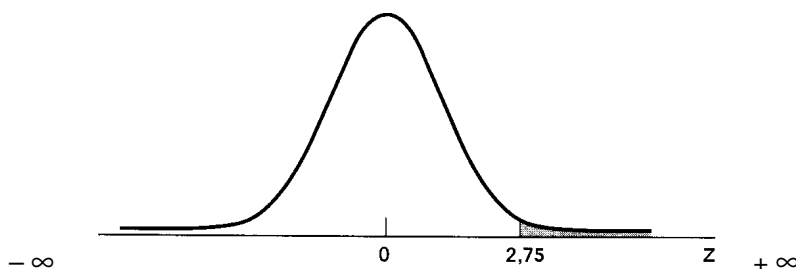
TABELA COM VALORES POSITIVOS DE Z (continuação)

z	0,00	0,01	0,02	0,03	0,04	0,05	0,06	0,07	0,08	0,09
0,00	0,5000	0,5040	0,5080	0,5120	0,5160	0,5199	0,5239	0,5279	0,5319	0,5359
0,10	0,5398	0,5438	0,5478	0,5517	0,5557	0,5596	0,5636	0,5675	0,5714	0,5753
0,20	0,5793	0,5832	0,5871	0,5910	0,5948	0,5987	0,6026	0,6064	0,6103	0,6141
0,30	0,6179	0,6217	0,6255	0,6293	0,6331	0,6368	0,6406	0,6443	0,6480	0,6517
0,40	0,6554	0,6591	0,6628	0,6664	0,6700	0,6736	0,6772	0,6808	0,6844	0,6879
0,50	0,6915	0,6950	0,6985	0,7019	0,7054	0,7088	0,7123	0,7157	0,7190	0,7224
0,60	0,7257	0,7291	0,7324	0,7357	0,7389	0,7422	0,7454	0,7486	0,7517	0,7549
0,70	0,7580	0,7611	0,7642	0,7673	0,7704	0,7734	0,7764	0,7794	0,7823	0,7852
0,80	0,7881	0,7910	0,7939	0,7967	0,7995	0,8023	0,8051	0,8078	0,8106	0,8133
0,90	0,8159	0,8186	0,8212	0,8238	0,8264	0,8289	0,8315	0,8340	0,8365	0,8389
1,00	0,8413	0,8438	0,8461	0,8485	0,8508	0,8531	0,8554	0,8577	0,8599	0,8621
1,10	0,8643	0,8665	0,8686	0,8708	0,8729	0,8749	0,8770	0,8790	0,8810	0,8830
1,20	0,8849	0,8869	0,8888	0,8907	0,8925	0,8944	0,8962	0,8980	0,8997	0,9015
1,30	0,9032	0,9049	0,9066	0,9082	0,9099	0,9115	0,9131	0,9147	0,9162	0,9177
1,40	0,9192	0,9207	0,9222	0,9236	0,9251	0,9265	0,9279	0,9292	0,9306	0,9319
1,50	0,9332	0,9345	0,9357	0,9370	0,9382	0,9394	0,9406	0,9418	0,9429	0,9441
1,60	0,9452	0,9463	0,9474	0,9484	0,9495	0,9505	0,9515	0,9525	0,9535	0,9545
1,70	0,9554	0,9564	0,9573	0,9582	0,9591	0,9599	0,9608	0,9616	0,9625	0,9633
1,80	0,9641	0,9649	0,9656	0,9664	0,9671	0,9678	0,9686	0,9693	0,9699	0,9706
1,90	0,9713	0,9719	0,9726	0,9732	0,9738	0,9744	0,9750	0,9756	0,9761	0,9767
2,00	0,9772	0,9778	0,9783	0,9788	0,9793	0,9798	0,9803	0,9808	0,9812	0,9817
2,10	0,9821	0,9826	0,9830	0,9834	0,9838	0,9842	0,9846	0,9850	0,9854	0,9857
2,20	0,9861	0,9864	0,9868	0,9871	0,9875	0,9878	0,9881	0,9884	0,9887	0,9890
2,30	0,9893	0,9896	0,9898	0,9901	0,9904	0,9906	0,9909	0,9911	0,9913	0,9916
2,40	0,9918	0,9920	0,9922	0,9925	0,9927	0,9929	0,9931	0,9932	0,9934	0,9936
2,50	0,9938	0,9940	0,9941	0,9943	0,9945	0,9946	0,9948	0,9949	0,9951	0,9952
2,60	0,9953	0,9955	0,9956	0,9957	0,9959	0,9960	0,9961	0,9962	0,9963	0,9964
2,70	0,9965	0,9966	0,9967	0,9968	0,9969	0,9970	0,9971	0,9972	0,9973	0,9974
2,80	0,9974	0,9975	0,9976	0,9977	0,9977	0,9978	0,9979	0,9979	0,9980	0,9981
2,90	0,9981	0,9982	0,9982	0,9983	0,9984	0,9984	0,9985	0,9985	0,9986	0,9986
3,00	0,9987	0,9987	0,9987	0,9988	0,9988	0,9989	0,9989	0,9989	0,9990	0,9990
3,10	0,9990	0,9991	0,9991	0,9991	0,9992	0,9992	0,9992	0,9992	0,9993	0,9993
3,20	0,9993	0,9993	0,9994	0,9994	0,9994	0,9994	0,9994	0,9995	0,9995	0,9995
3,30	0,9995	0,9995	0,9995	0,9996	0,9996	0,9996	0,9996	0,9996	0,9996	0,9997
3,40	0,9997	0,9997	0,9997	0,9997	0,9997	0,9997	0,9997	0,9997	0,9997	0,9998
3,50	0,9998	0,9998	0,9998	0,9998	0,9998	0,9998	0,9998	0,9998	0,9998	0,9998
3,60	0,9998	0,9998	0,9999	0,9999	0,9999	0,9999	0,9999	0,9999	0,9999	0,9999
3,70	0,9998	0,9998	0,9999	0,9999	0,9999	0,9999	0,9999	0,9999	0,9999	0,9999
3,80	0,9998	0,9998	0,9999	0,9999	0,9999	0,9999	0,9999	0,9999	0,9999	0,9999

Continuando nosso exemplo, verificamos que a área sob a curva entre $-\infty$ e $z = 2,00$ é 0,9772. Podemos então afirmar que há uma probabilidade de 97,72%, de um valor qualquer de Z selecionado aleatoriamente, estar entre $-\infty$ e 2,00.

Agora vamos encontrar a área sob a curva entre $z = 2,75$ e mais infinito. A área do nosso interesse

está apresentada no diagrama abaixo (área sombreada):



Olhando na tabela Z verificamos que a área sob a curva entre menos infinito e $z = 2,75$ é 0,9970. Este valor foi encontrado na tabela procurando-se pelo valor 2,70 na coluna correspondente ao valor de z , e pelo valor 0,05 (ou 5 centésimos) nas colunas, que indicam os centésimos do valor de z , completando o valor de z de interesse que é 2,75. A área entre menos infinito e $z = 2,75$ foi então encontrada na interseção entre a linha correspondente a $z = 2,70$ e a coluna correspondente a 0,05. O valor encontrado foi, portanto, 0,9970, o que significa que há uma probabilidade de 99,7% de um valor qualquer de Z selecionado aleatoriamente, estar entre menos infinito e 2,75. Mas, a área do nosso interesse não é essa e sim aquela entre $z = 2,75$ e mais infinito. Para encontrarmos esta área diminuimos 0,9970 ou 99,70% (que é a área entre $-\infty$ e $z = 2,75$) de 1 ou 100% (que corresponde à área total). A diferença obtida é a área entre $z = 2,75$ e $+\infty$. Assim, temos que

$$1 - 0,9970 = 0,003 = 0,3 \%,$$

que é nossa área de interesse. Podemos então dizer que se escolhermos aleatoriamente um valor qualquer de Z , a probabilidade desse valor estar entre 2,75 e $+\infty$ é de apenas 0,3%. Como a área mede também probabilidade, podemos expressar as operações acima como

$$P(Z > 2,75) = 1 - P(Z < 2,75) = 1 - 0,9970 = 0,003 = 0,3 \%,$$

onde $P(Z > 2,75)$ indica a probabilidade de Z ser maior do que 2,75 e $P(Z < 2,75)$ a probabilidade de Z ser menor do que 2,75. Observe que não consideramos a probabilidade de Z ser igual a 2,75, mas as probabilidades de Z ser maior ou menor do que esse valor. Procedemos assim porque, com base na teoria estatística, não é possível calcularmos a probabilidade de nenhum valor específico, mas apenas de um intervalo de valores. Não se preocupe, já que isso não nos trará dificuldade alguma.

Só nos falta agora usar os valores de Z em um exemplo concreto.

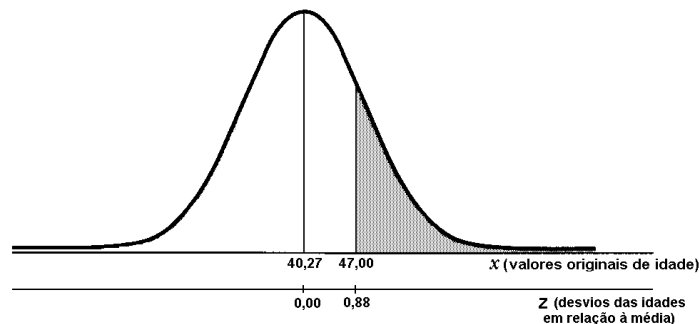
Considere as idades dos 311 agentes penitenciários e lembre-se de que a média de idade obtida foi $\bar{x} = 40,27$ anos e o desvio-padrão $s = 7,60$ anos. Você verá brevemente que inúmeras vezes desejaremos

calcular a probabilidade de obtermos idades maiores do que uma determinada idade. Suponha, por exemplo, que desejamos saber a probabilidade de um agente penitenciário qualquer selecionado aleatoriamente e exercendo sua atividade na Região Metropolitana de Salvador, ter idade maior do que 47 anos. Se pudermos assumir que os resultados obtidos na amostra estudada sejam aplicáveis a todos os agentes e que as frequências de idade apresentem uma distribuição normal na população de agentes, podemos usar esta distribuição probabilística como modelo gráfico e matemático para calcularmos a probabilidade desejada. Você já sabe como encontrar áreas sob a curva para valores de Z , não é? Então, poderemos facilmente obter essa probabilidade, transformando 47 anos em um valor de Z . Como sabemos que Z é expresso em número de desvios-padrão, teremos que calcular o quanto a idade $x = 47$ anos se afastou (desviou) da média das idades, 40,27 anos, e a quantos desvios-padrão esse desvio equivale. Matematicamente, para obtermos o valor de z correspondente a 47 anos fazemos a seguinte operação:

$$z = \frac{x - \mu}{\sigma} = \frac{x - \bar{x}}{s} = \frac{47,00 - 40,27}{7,60} = \frac{6,73}{7,60} \cong 0,88.$$

Note que, por serem desconhecidos por nós, os parâmetros μ e σ foram substituídos por seus estimadores \bar{x} e s .

Graficamente, temos:



A área sombreada é a que queremos calcular. A média de idade corresponde a $z = 0,00$, e 47 anos é um valor de idade que desvia um pouco menos de um desvio-padrão (0,88 desvio-padrão) acima da média. A média de idade corresponde a $z = 0,00$ porque, evidentemente, a média não varia nada (zero desvio-padrão) em relação a ela mesma (explicaremos isto mais detalhadamente no próximo capítulo).

Agora que encontramos o valor de z correspondente a 47 anos de idade, que é $z = 0,88$, podemos utilizar a tabela de áreas sob a curva Z para obter a área sob esta curva entre $-\infty$ e esse valor de z . Depois é só diminuirmos essa área de 1 (que equivale a 100% dos valores sob a curva), para encontrarmos a área que nos interessa. Na tabela Z verificamos que a área entre $-\infty$ e $z = 0,88$ é 0,8106. Como

$$1 - 0,8106 = 0,1894,$$

este é o valor da área sob a curva com valores de z maiores do que 0,88. Isso indica que, se pudermos

assumir que os resultados obtidos nesta amostra são aplicáveis a todos os agentes penitenciários e que a distribuição dessa variável na população de agentes seja normal, ao selecionarmos um desses agentes aleatoriamente, a probabilidade dele apresentar uma idade maior do que 47 anos é de 18,94%. Note que evitamos afirmar “a probabilidade de ele apresentar uma idade entre 47 anos e mais infinito”, porque sabemos que, biologicamente, isso não faria sentido.

— A distribuição normal é a única distribuição probabilística que existe?

— Boa pergunta! Não, não é a única. Pode ter passado despercebido, mas neste mesmo capítulo mencionamos duas outras distribuições: a binomial e a de Poisson. Lembra-se? E existem várias outras, tais como as distribuições T , F e a distribuição qui-quadrado. Estas ainda serão utilizadas no livro, nos próximos capítulos.

Consideramos que você agora está quase pronto para entender inferência estatística, tema que começaremos a abordar a partir do próximo capítulo. Quanto a alguns dos procedimentos da Estatística Analítica, esses serão abordados muito parcialmente, simultaneamente aos da Estatística Inferencial, pois uma apresentação mais completa daqueles nos afastaria dos nossos objetivos.

CAPÍTULO 10

PRIMEIRA PARTE: Preparação para inferência estatística:

- Por que precisamos fazer inferência estatística?
- O que é afinal inferência estatística?
- O que é variação amostral?
- O que é inferência não-estatística?
- Como se distribuem as frequências dos resultados de diferentes amostras?
- O que é erro-padrão e como calculá-lo?
- Qual é o teorema central do limite e qual a sua importância?

SEGUNDA PARTE: Teste z :

- Como a inferência estatística é feita?
- O que é nível de significância estatística e qual a sua utilidade?
- O que é um teste de hipóteses e como realizá-lo?
- Quando um teste é mono ou bicaudado?
- Para quê os valores originais da variável testada são transformados em valores de z ?
- O que é valor- p e qual a sua utilidade?
- Quais os erros que podemos cometer ao realizarmos inferência estatística?
- O que é poder do teste estatístico e como calculá-lo?

TERCEIRA PARTE: Intervalo de confiança:

- Como calcular e interpretar um intervalo de confiança?
-

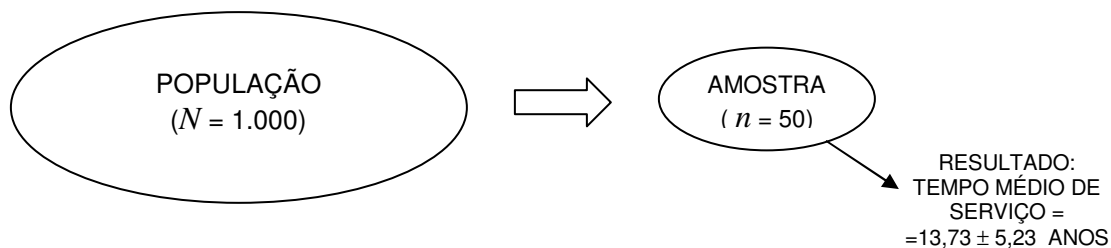


▣ PRIMEIRA PARTE ▣

— Por que precisamos fazer inferência estatística?

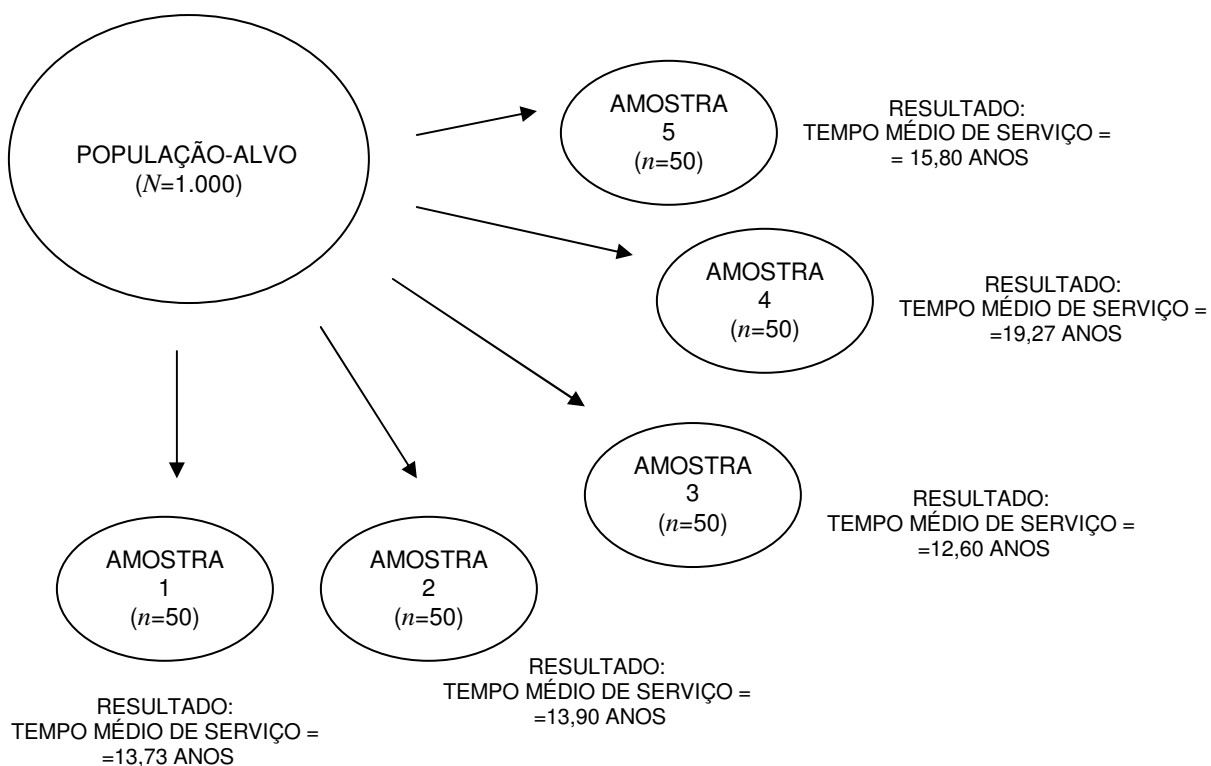
— Suponha que estejamos interessados em saber qual o tempo médio de serviço de uma população de 1.000 trabalhadores em certa refinaria de petróleo. A melhor maneira de sabermos isso seria coletarmos o número de anos de trabalho na Refinaria para cada um desses mil trabalhadores e depois calcularmos a média aritmética dessa série de observações. Mas suponha que não dispuséssemos do tempo, recursos humanos e/ou financeiros necessários para realizarmos tal levantamento de informações. Nesse caso, a alternativa viável seria selecionarmos aleatoriamente uma amostra de tamanho suficiente dos trabalhadores e estimarmos o tempo médio de trabalho populacional, ou seja, do conjunto dos mil trabalhadores, com base no tempo médio obtido nessa amostra.

Considere que tenhamos utilizado essa alternativa. Através de uma fórmula adequada (veja capítulo 15, páginas 262 a 269), calculamos o número mínimo (n mínimo) necessário para a amostra, tendo sido este de 50 trabalhadores. Foram obtidas informações sobre o número de anos que cada um dos 50 indivíduos tinha de trabalho na Refinaria. Suponha que tenhamos obtido nessa amostra um tempo médio de serviço de 13,73 anos com desvio-padrão de 5,23 anos.



Nossa pergunta seria então a seguinte: com base no resultado da amostra estudada, o que podemos afirmar sobre o verdadeiro tempo médio de serviço na população de trabalhadores?

Só poderemos respondê-la se utilizarmos procedimentos adequados. Lembre-se de que nós selecionamos e investigamos apenas uma amostra. Se outras equipes tivessem estudado o mesmo tema com a mesma metodologia, cada uma investigando uma amostra do mesmo tamanho, retirada da mesma população-alvo, você esperaria que os tempos médios de serviço obtidos fossem iguais àquele encontrado pela nossa equipe? Evidentemente que não, porque muito provavelmente cada uma das amostras conteria alguns ou mesmo todos os trabalhadores diferentes, concorda? Como os trabalhadores em cada amostra não seriam sempre os mesmos, seria esperado que cada amostra apresentasse um resultado diferente. A figura a seguir representa o que poderia ser encontrado se numerosas amostras do mesmo tamanho, retiradas da mesma população-alvo e submetidas à mesma metodologia, tivessem sido estudadas:



A amostra 1 representa a única amostra que realmente foi estudada e é isso o que geralmente ocorre: apenas uma amostra é estudada, já que na maioria das vezes é inviável investigarmos muitas amostras ou toda a população-alvo. Mas, para você nos entender, é importante considerarmos hipoteticamente que várias amostras de mesmo tamanho tenham sido estudadas. A figura acima mostra o que seria esperado: os resultados obtidos nas diversas amostras seriam diferentes entre si. Embora os resultados de uma ou outra amostra pudessem até ser iguais (isto não foi mostrado na figura), essas coincidências de valores seriam pouco freqüentes. Assim, caso tivéssemos estudado numerosas amostras, haveria uma **variação amostral** dos resultados. Ao escolhermos uma única amostra para estudarmos, qualquer uma das possíveis amostras do mesmo tamanho poderia ter sido aquela selecionada por nós, já que o processo de seleção foi aleatório. Na única amostra estudada (amostra 1) o tempo médio de serviço encontrado foi 13,73 anos, mas se em vez de termos selecionado essa amostra tivéssemos selecionado a amostra 4 a média seria 19,27 anos. Considerando as cinco amostras acima apresentadas, vemos que o resultado poderia variar entre 12,60 e 19,27 anos.

É por esse motivo que, ao analisarmos resultados quantitativos obtidos em estudo feito em amostra aleatória retirada de uma população, temos de nos perguntar se o(s) resultado(s) encontrado(s) é(são) estatisticamente significativo(s). O processo para respondermos a essa pergunta é chamado de **inferência estatística**.

— Mas o que é afinal inferência estatística?

— Podemos definir inferência estatística como:

o processo pelo qual tiramos conclusões sobre uma população a partir de resultados observados em uma amostra aleatória

ou,

o processo pelo qual avaliamos a probabilidade de resultados observados em uma amostra aleatória terem ocorrido por variação amostral.

Se isto foi bem entendido por você, temos certeza de que concordará com a afirmação de que em um estudo que investigou toda a população não tem sentido algum fazermos inferência estatística, porque se não houve amostragem, qual seria o motivo para avaliarmos se um determinado valor poderia ter ocorrido devido à variação amostral? Alguns pesquisadores argumentam que a população estudada pode ser considerada como uma amostra de uma “população ainda maior” e que isso justificaria a realização de inferência estatística. De modo algum isso pode ser argumentado! Quando esse estudo foi planejado e realizado, a população estudada não foi retirada aleatoriamente de uma “população ainda maior”. Os outros indivíduos dessa “população ainda maior” não tiveram nenhuma oportunidade de participar do estudo. Assim, não poderíamos afirmar que a população estudada foi uma amostra aleatória de uma população maior, e isso é obrigatório para fazermos inferência estatística.

Quando toda a população tiver sido estudada, os resultados encontrados devem ser considerados como fidedignos da realidade estudada naquela população, caso a investigação tenha sido realizada sem erros (vieses), não sendo necessário, portanto, avaliarmos o papel da variação amostral para tirarmos nossas conclusões.

Quando o estudo incluir apenas uma parte dos indivíduos de uma população-alvo, ou seja, uma amostra, mas a escolha deles não for feita aleatoriamente, também não fará sentido fazer-se inferência estatística, porque toda a teoria que a fundamenta assume que houve amostragem e que esta foi aleatória.

— Nesse caso, como poderei avaliar se os resultados obtidos nessa amostra não-aleatória podem ser generalizados para a população-alvo?

— Você fará uma inferência não-estatística, ou seja, avaliará com sua equipe ou com pessoas que possam lhe auxiliar nisso, se há razões para crerem que a amostra estudada seja representativa (se tem as mesmas ou características muito semelhantes) da população-alvo ou, ao contrário, se há razões para crerem que ela não seja representativa da população-alvo. Se chegarem à conclusão de que a amostra é representativa (simplesmente comparando suas características com aquelas da população-alvo), poderão generalizar os resultados obtidos na amostra para a população-alvo de onde esta foi retirada não-aleatoriamente. E não poderão fazer tal generalização se concluírem o contrário.

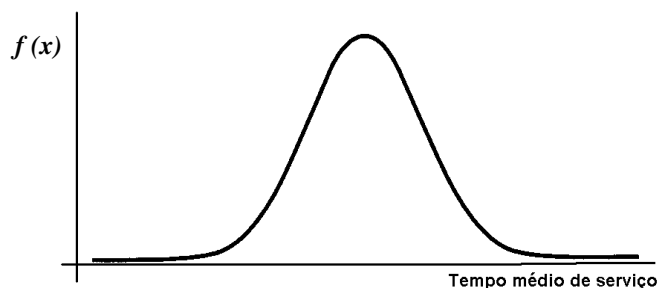
Esse mesmo tipo de inferência poderá ser feito quando estudarmos toda uma população e desejarmos generalizar os resultados obtidos para uma “população ainda maior”, como discutimos mais acima. O que não podemos fazer nesse caso é inferência estatística.

Há uma outra situação que merece ser considerada, que é a de um ensaio clínico randomizado, para o qual os indivíduos não foram selecionados de modo aleatório. Neste caso, os indivíduos foram incluídos no estudo por voluntarismo ou conveniência dos pesquisadores, e após obtenção do consentimento, os que aceitaram participar do estudo foram aleatoriamente (por randomização) alocados em um dos grupos a serem comparados. Os estatísticos admitem que seja feita inferência estatística neste estudo, mesmo que a seleção inicial dos indivíduos não tenha sido aleatória, porque houve sorteio na alocação dos indivíduos, o que torna apropriada a avaliação sobre se os resultados obtidos ocorreram por acaso. No entanto, a inferência neste caso, só é válida dentro dos grupos, isto é, na comparação dos grupos, ao avaliar a influência do acaso nas diferenças encontradas entre eles, e não no sentido de generalização para a população de onde os grupos comparados foram não-aleatoriamente retirados.

— **Como se distribuem as freqüências dos resultados de diferentes amostras?**

— Antes de lhe mostrarmos como a inferência estatística é feita, será necessário discutirmos com você qual a forma com que se distribuem as freqüências de resultados obtidos em diferentes amostras.

Vamos supor novamente que tenhamos estudado várias amostras aleatórias do mesmo tamanho retiradas da mesma população-alvo. Como já vimos, os resultados obtidos nessas amostras devem ser diferentes entre si. Considerando o nosso exemplo, para cada uma das numerosas amostras que porventura tivéssemos estudado, obteríamos 50 valores de tempos de serviço na Refinaria, um para cada um dos trabalhadores estudados, certo? Sendo assim, para cada amostra poderíamos somar todos os 50 tempos de serviço e, se dividíssemos esta soma por 50, obteríamos a média aritmética desses tempos de serviço, ou seja, o tempo médio de serviço em cada amostra. Se tivéssemos estudado 60 amostras teríamos 60 tempos médios, um para cada amostra. Se, em seguida, elaborássemos um diagrama de freqüências desses tempos médios de serviço das amostras, verificaríamos que sua forma se assemelharia àquela de uma distribuição normal, como mostrado abaixo:



Pelo que vimos em distribuições probabilísticas apresentadas anteriormente, o que era indicado na ordenada eram as probabilidades de obtermos indivíduos (e não amostras) com determinados valores de idade, ou outra variável, como o tempo de serviço em determinada empresa. Acontece que antes estávamos interessados em verificar como se distribuíam as probabilidades de medidas feitas para indivíduos, e agora em como essas medidas se distribuem para amostras de indivíduos, ou seja, para grupos de indivíduos.

— **E por que agora estamos interessados em amostras e não em indivíduos?**

— Porque agora desejamos saber se o resultado obtido em uma única amostra representa ou não o verdadeiro resultado que seria obtido caso tivéssemos estudado toda a população. Como sabemos que os resultados variam de uma amostra para outra, a suposição de que estudamos numerosas amostras, e o uso de uma distribuição probabilística compatível com os resultados que se esperaria obter nessas diferentes amostras, vai nos permitir avaliar o quanto esperaríamos que esses resultados amostrais variassem entre si, simplesmente como consequência natural das diferenças entre as amostras. Com base nisso, poderemos avaliar quais os resultados mais prováveis de serem o verdadeiro tempo médio de serviço da população de trabalhadores da Refinaria, que é o que realmente nos interessa no momento. Os tempos médios de serviço que estiverem dentro de certo intervalo de variação, serão considerados como possíveis valores para o verdadeiro tempo médio de serviço populacional, por serem valores esperados por simples variação amostral, qualquer um deles podendo, portanto, ser o verdadeiro tempo médio de serviço na população.

— Quer dizer então que o máximo que conseguiremos nessa situação será uma série de valores possíveis para a verdadeira média populacional?

— Isso mesmo. Não espere um procedimento mágico que lhe forneça com total exatidão a verdadeira média populacional a partir da investigação de apenas uma amostra dessa população. Haverá sempre uma imprecisão no processo de inferência estatística.

Contudo, apesar dessas limitações, há alguns fatos estatísticos que nos ajudam no processo de inferência. É possível demonstrarmos empiricamente (e, teoricamente também, quando n tende a infinito) que se o n de cada amostra for suficientemente grande, ao retirarmos numerosas amostras de mesmo tamanho da mesma população-alvo, a distribuição de freqüências das médias amostrais obtidas tenderá a ser normal, mesmo que a distribuição da variável estudada não seja normal na população-alvo de onde as amostras forem retiradas. Isto é importante porque nos permitirá utilizar a distribuição normal na realização da inferência estatística. Uma demonstração empírica, ou seja, com base em dados reais, do que acabamos de afirmar é feita abaixo:

Considere a população-alvo constituída pelos trabalhadores da Refinaria do nosso exemplo atual e uma série de 1.000 valores de tempo de serviço, um para cada trabalhador. Não apresentaremos esta série porque isto ocuparia muito espaço desnecessariamente, mas imagine um banco de dados contendo todos esses valores. Agora vamos utilizar um computador para retirar aleatoriamente daquela população-alvo 100 amostras do mesmo tamanho ($n = 50$). Para cada amostra obteremos então 50 valores da variável “tempo de serviço”, um para cada trabalhador selecionado para aquela amostra. Para cada amostra poderemos somar os 50 valores de “tempo de serviço” e dividir por 50, obtendo a média aritmética para essa variável. Isto resultará em 100 médias aritméticas, uma para cada uma das 100 amostras, certo?

As médias são apresentadas na planilha a seguir:

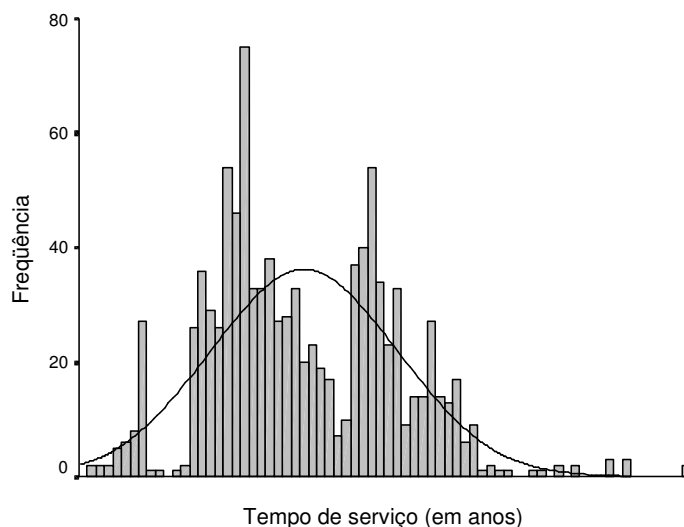
NÚMERO DA AMOSTRA	MÉDIA ARITMÉTICA DO TEMPO DE SERVIÇO (em anos) (\bar{x})	NÚMERO DA AMOSTRA	MÉDIA ARITMÉTICA DO TEMPO DE SERVIÇO (em anos) (\bar{x})
1	13,73	51	15,40
2	13,90	52	13,90
3	12,57	53	12,57
4	14,59	54	14,59
5	14,00	55	14,00
6	16,24	56	16,24
7	14,38	57	14,39
8	14,82	58	14,82
9	14,17	59	14,17
10	14,38	60	14,38
11	15,44	61	14,65
12	14,21	62	15,44
13	14,23	63	14,21
14	13,64	64	14,23
15	14,19	65	13,65
16	13,96	66	14,19
17	13,26	67	13,96
18	13,87	68	13,26
19	13,88	69	13,88
20	13,01	70	13,88
21	15,44	71	13,01
22	12,72	72	15,44
23	14,55	73	12,72
24	14,65	74	14,55
25	15,07	75	14,65
26	13,84	76	15,07
27	14,53	77	13,84
28	15,01	78	14,53
29	14,80	79	15,01
30	13,37	80	14,80
31	13,38	81	13,37
32	14,34	82	13,38
33	14,68	83	14,34
34	14,47	84	14,68
35	13,99	85	14,47
36	13,99	86	13,99
37	14,28	87	13,99
38	14,91	88	14,28
39	13,47	89	14,91
40	14,87	90	13,47
41	15,61	91	14,87
42	13,35	92	15,61
43	14,34	93	13,35
44	13,54	94	14,34
45	14,68	95	13,54
46	14,92	96	14,68
47	13,74	97	14,92
48	13,62	98	13,74
49	14,36	99	13,62
50	15,68	100	14,36

Já vimos que, geralmente, quando realizamos uma pesquisa, retiramos e estudamos apenas uma dessas possíveis amostras do mesmo tamanho. Na demonstração empírica que estamos fazendo, utilizando um banco de dados real e a ajuda do computador, analisando numerosas amostras retiradas da mesma população, você pode verificar que há uma variação dos resultados das diferentes amostras. Observe no quadro acima que, embora algumas amostras tenham apresentado tempos médios de serviço iguais, há uma variação nos valores destes tempos médios, entre 12,57 e 16,24 anos. Poderíamos afirmar, grosseiramente,

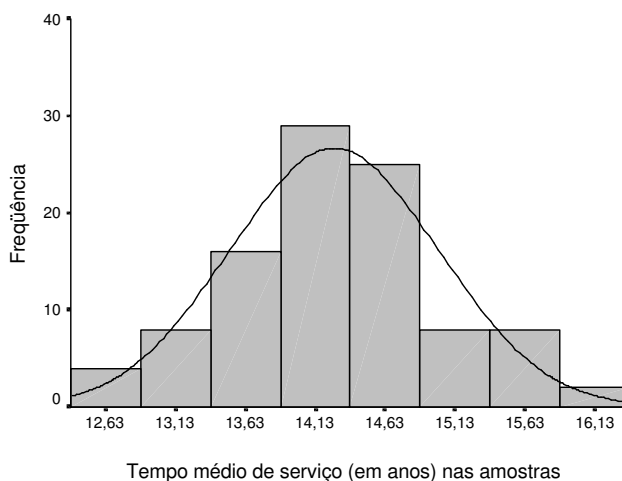
que a única amostra de tamanho $n = 50$ a ser efetivamente estudada poderia apresentar tempo médio de serviço tão baixo quanto 12,57 anos ou tão alto quanto 16,24 anos. E é por isso que, como já vimos, temos de avaliar em toda pesquisa que estudou uma amostra aleatória, até que ponto o tempo médio de serviço obtido nessa única amostra representa o verdadeiro tempo médio de serviço na população-alvo dos 1.000 trabalhadores da Refinaria.

Vamos agora elaborar dois diagramas de freqüências, o primeiro para os valores obtidos na população-alvo (lembre-se de que esses valores não foram apresentados), e o segundo para as médias amostrais que acabamos de apresentar na página anterior:

Distribuição de freqüências do tempo de serviço dos 1.000 trabalhadores da Refinaria, com a curva normal sobreposta.



Distribuição de freqüências do tempo médio de serviço para as 100 amostras selecionadas, com a curva normal sobreposta.



Veja que a distribuição das freqüências dos tempos de serviço para os 1.000 trabalhadores não apresenta um formato que possa ser modelado pela distribuição normal, enquanto aquela dos tempos médios

para as 100 amostras tende à normalidade, evidenciando o que queríamos demonstrar empiricamente.

Se somarmos todos os tempos médios de serviço obtidos nas 100 amostras estudadas e dividirmos o resultado por 100 (que é o número de amostras), obteremos a média aritmética dos tempos médios de serviço, concorda? E se calcularmos o quanto o tempo médio de cada amostra se desviou dessa média dos tempos médios, elevarmos ao quadrado cada um desses desvios, somarmos todos esses desvios ao quadrado, e dividirmos por $k - 1$, sendo k o número de amostras estudadas, obteremos a variância dos tempos médios de serviço naquelas amostras, não é?

Observe que as operações feitas acima são idênticas às aquelas que utilizamos no capítulo 6 (páginas 62 a 65) para calcular a variância de resultados obtidos dentro de uma mesma amostra (e não entre amostras, como acabamos de fazer). Então, se extrairmos a raiz quadrada da variância dos tempos médios de serviço, obteremos o seu desvio-padrão, concorda? Só que este desvio-padrão de resultados amostrais é tão importante para a inferência estatística que recebeu o nome especial de **erro-padrão**.

O desvio-padrão de resultados amostrais é chamado de erro-padrão.

Vamos calcular a média dessas médias amostrais e seu erro-padrão?

Média das médias amostrais =

= média dos tempos médios de serviço obtidos nas amostras = $\bar{\bar{x}}$ =

$$= \frac{\text{soma dos tempos médios de serviço das amostras}}{\text{número de amostras}} = \frac{\sum_{i=1}^k \bar{x}_i}{k} = \frac{1.426,01}{100} \cong 14,26 \text{ anos}.$$

Desvio-padrão das médias amostrais = erro-padrão = EP =

$$= \sqrt{\frac{\text{soma dos desvios ao quadrado de cada tempo médio de serviço em relação à média dos tempos médios}}{\text{número de amostras} - 1}}$$

$$= \sqrt{\frac{\sum_{i=1}^k (\bar{x}_i - \bar{\bar{x}})^2}{k - 1}} = \sqrt{\frac{\sum_{i=1}^k (\bar{x}_i - 14,26)^2}{100 - 1}} = \sqrt{\frac{55,26}{99}} = \sqrt{0,56} \cong 0,75 \text{ anos}.$$

Agora vamos considerar toda a população-alvo constituída pelos 1.000 trabalhadores da Refinaria e, tirando vantagem do fato de, neste caso, termos os tempos de serviço de toda essa população, vamos calcular os verdadeiros valores populacionais para a média e desvio-padrão (lembre-se de que, geralmente, não podemos fazer isso porque não dispomos dos dados para toda uma população):

Média dos tempos de serviço na população = μ = $\frac{\text{soma dos tempos de serviço dos trabalhadores}}{\text{número de trabalhadores}}$ *=*

$$= \frac{\sum_{i=1}^N x_i}{N} = \frac{14.259,13}{1.000} \cong 14,26 \text{ anos}.$$

Desvio-padrão dos tempos de serviço dos trabalhadores = σ =

$$= \sqrt{\frac{\text{soma dos desvios ao quadrado de cada tempo de serviço em relação à média dos tempos de serviço}}{\text{número de trabalhadores}}} =$$

$$= \sqrt{\frac{\sum_{i=1}^N (x_i - \bar{x})^2}{N}} = \sqrt{\frac{\sum_{i=1}^N (x_i - 14,26)^2}{1.000}} = \sqrt{\frac{30.139,35}{1.000}} = \sqrt{30,14} \cong 5,49 \text{ anos}.$$

O que você observou nos resultados acima?

Note que $\bar{x}_{\bar{x}} = \mu \cong 14,26$ anos. Acabamos também, então, de demonstrar empiricamente que a média de médias amostrais é igual à média populacional. Isso pode ser entendido por você intuitivamente. Não lhe parece claro que se estudássemos numerosas amostras de tamanho suficientemente grande, a média das médias amostrais obtidas se aproximaria muito da verdadeira média que obteríamos se estudássemos toda a população?

Quanto ao desvio-padrão (erro-padrão) das médias amostrais, vemos que seu valor, 0,75 ano, é bem menor do que o desvio-padrão populacional, 5,49 anos, mas isto não nos trará dificuldade, pois é sabido que o erro-padrão é realmente menor do que o desvio-padrão populacional, e não precisaremos assumir que esses valores sejam iguais para realizarmos inferência estatística.

— **Mas, como vamos calcular o erro-padrão para numerosas amostras se na prática estudamos apenas uma amostra?**

— É verdade. Em nossas pesquisas estudamos apenas uma amostra e, por isso, será impossível computarmos o erro-padrão do modo como fizemos acima, pois não temos os resultados obtidos em várias amostras.

Felizmente, os estatísticos puderam demonstrar que, para populações infinitas, se conhecermos a variância populacional, σ^2 , e a dividirmos pelo número de indivíduos na única amostra estudada, n , obteremos σ^2/n , que é a variância das médias amostrais que esperaríamos obter caso tivéssemos estudado infinitas amostras. E, como já vimos, se extrairmos a raiz quadrada desta variância,

$$\sqrt{\frac{\sigma^2}{n}} = \frac{\sigma}{\sqrt{n}},$$

obteremos o erro-padrão das médias amostrais, denotado por EP . Anote então que

$$EP_{\text{médias amostrais}} = \frac{\sigma}{\sqrt{n}}.$$

Embora o σ calculado no início desta página seja de uma população finita e o que temos discutido tenha sido demonstrado, a rigor, para populações infinitas, veja que o erro-padrão obtido considerando os

resultados das 100 amostras estudadas, 0,75 ano, é aproximadamente o mesmo que obtemos calculando

$$EP_{\text{médias amostrais}} = \frac{\sigma}{\sqrt{n}} = \frac{5,49}{\sqrt{50}} = \frac{5,49}{7,07} \cong 0,78 \text{ ano}.$$

— Ora, mas raramente conhecemos o desvio-padrão na população, porque quase nunca estudamos toda a população, pois isto é muito caro e se torna geralmente inviável como já vimos!

— Você está novamente certo(a). Quando não conhecemos o desvio-padrão populacional, podemos substituí-lo pelo desvio-padrão obtido na única amostra estudada, tendo sido demonstrado que este é um estimador aceitável para o desvio-padrão populacional, se o n for suficientemente grande. Assim, mais freqüentemente, obtemos a variância das médias de diversas amostras, através da fórmula s^2/n , e, portanto, seu erro-padrão por

$$EP_{\text{médias amostrais}} = \sqrt{\frac{s^2}{n}} = \frac{s}{\sqrt{n}}.$$

Suponha que a única amostra que estudamos tenha sido a primeira das amostras relacionadas na planilha da página 145, cuja média dos tempos de serviço foi 13,73 anos. O desvio-padrão de cada amostra não foi apresentado na planilha, mas estamos lhe informando que o desvio-padrão obtido para a primeira amostra selecionada foi 5,23 anos. Se o desvio-padrão populacional fosse desconhecido, nós o substituiríamos por 5,23 anos, para calcularmos o erro-padrão das médias amostrais. Dessa maneira obteríamos:

$$EP_{\text{médias amostrais}} = \sqrt{\frac{s^2}{n}} = \frac{s}{\sqrt{n}} = \frac{5,23}{\sqrt{50}} = \frac{5,23}{7,07} = 0,74 \text{ ano}.$$

Observe que o erro-padrão calculado com o desvio-padrão populacional, 0,78 ano, não difere muito do valor obtido utilizando o desvio-padrão na única amostra estudada, 0,74 ano, e que esses valores, por sua vez, não diferem muito do erro-padrão calculado com as próprias médias obtidas em numerosas amostras, 0,75 ano.

Vimos então que:

Dada uma população qualquer de qualquer forma funcional não-normal com média μ e variância σ^2 , a distribuição de médias, \bar{x} , computadas de amostras de tamanho n , retiradas dessa população, será aproximadamente normal, com média μ e erro-padrão σ/\sqrt{n} , quando o tamanho da amostra for grande e o desvio-padrão populacional for conhecido.

Esta é uma versão simplificada do famoso **teorema central do limite**. É com base nele que

utilizaremos a distribuição normal como modelo para fazermos inferência estatística sobre médias, já que podemos assumir que, seja qual for o tipo de distribuição de frequências de uma variável em uma população, a distribuição de frequências dos resultados obtidos para as médias dessa variável em numerosas amostras retiradas dessa população será normal, se o tamanho da amostra for suficientemente grande. Vimos também que a forma de calcularmos o erro-padrão de resultados amostrais dependerá do fato de conhecermos ou não o desvio-padrão populacional. Assim, podemos também afirmar que:

Dada uma população qualquer de qualquer forma funcional não-normal com média μ e variância σ^2 , a distribuição de médias, \bar{x} , computadas de amostras de tamanho n , retiradas dessa população, será aproximadamente normal, com média μ e erro-padrão s/\sqrt{n} , quando o tamanho da amostra for grande e o desvio-padrão populacional não for conhecido. (Quando o erro-padrão for calculado assim e a distribuição da variável investigada for normal na população da qual a amostra tiver sido retirada, a distribuição de médias amostrais será, a rigor, uma distribuição T , que abordaremos no próximo capítulo, mas se o n for grande, esta se aproximará da normal).

▣ SEGUNDA PARTE ▣

— Como a inferência estatística é feita?

— Há duas maneiras de fazermos inferência estatística, sendo uma delas através de teste de hipóteses estatísticas e a outra pelo cálculo de intervalo de confiança.

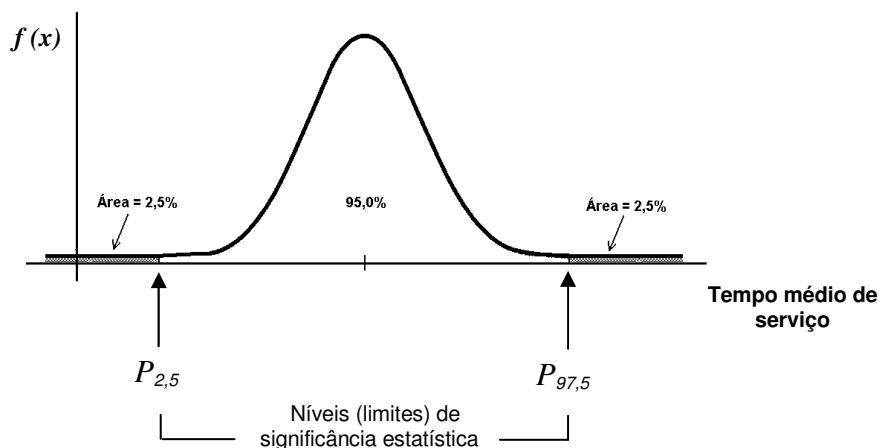
No teste de hipóteses nós formulamos duas hipóteses estatísticas e as testamos (uma em contraposição à outra) utilizando como referência uma das distribuições probabilísticas conhecidas, tal como a distribuição normal. A seguir, utilizando nosso exemplo dos trabalhadores de uma refinaria, explicaremos detalhadamente como isso é feito, e logo depois abordaremos o cálculo de intervalo de confiança.

A primeira etapa de um teste de hipóteses é a definição do nível de significância estatística, que é denotado pela letra grega α (alfa). O valor comumente escolhido para α é 0,05 ou 5,0%. Ao escolhermos um α de 5,0%, estamos definindo limites, com base nos quais um valor qualquer de tempo de serviço será considerado muito ou pouco provável de ser obtido, em amostras retiradas da população-alvo dos 1.000 trabalhadores da Refinaria. Estabelecer um $\alpha = 5,0\%$ implica em que vamos considerar como estatisticamente iguais ao tempo médio de serviço populacional, aqueles tempos médios de serviço que poderiam ser obtidos em numerosas amostras, e cuja probabilidade de ocorrer fosse maior do que 5,0%. Correspondentemente, vamos considerar como estatisticamente diferentes do tempo de serviço populacional, aqueles tempos médios de serviço que poderiam ser obtidos em numerosas amostras, e cuja probabilidade de ocorrer fosse menor ou igual a 5,0%. Vá adiante para entender melhor isso.

Já vimos que, se o tamanho da amostra for suficientemente grande (como uma regra prática, se $n \geq 30$), a distribuição dos tempos médios de serviço a serem obtidos se estudarmos diversas amostras tende a ser normal. Sendo assim, podemos utilizar a distribuição normal como modelo para realização do teste de

hipóteses estatísticas no nosso exemplo, já que a medida estatística que nos interessa é uma média (tempo médio de serviço) e nosso n é maior do que 30 ($n = 50 > 30$).

O gráfico a seguir mostra como o α determina nossos limites (níveis) de significância estatística:



Veja que o α , nessa explicação inicial, foi repartido em duas quantidades iguais de 2,5% nas duas caudas da distribuição normal. Na próxima etapa do teste de hipóteses, você verá que o modo de considerarmos o α dependerá das hipóteses que forem formuladas.

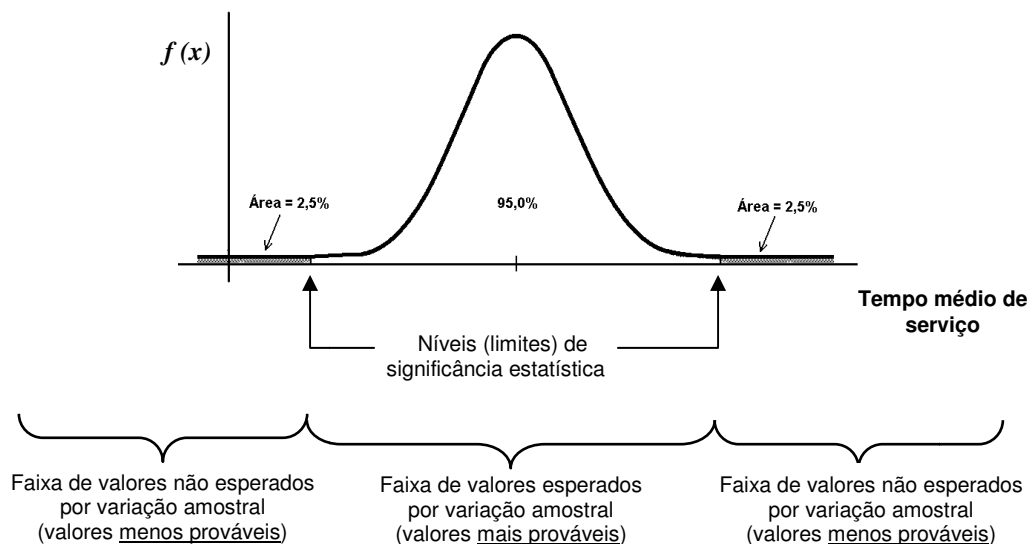
Os 2,5% de α em cada extremidade da curva nos permitirão estabelecer quais os valores de tempo médio de serviço, cuja probabilidade de ocorrer, caso tivéssemos estudado numerosas amostras, seria menor do que 2,5%.

Se o nosso alfa for 5,0% e distribuído nas duas caudas da curva, podemos também dizer que os valores de tempo médio de serviço equivalentes aos percentis 2,5 e 97,5, denotados respectivamente por $P_{2,5}$ e $P_{97,5}$, delimitam os nossos níveis de significância estatística.

Todos os valores localizados entre estes limites de significância seriam muito prováveis de serem obtidos e, portanto, seriam considerados por nós como estatisticamente iguais ao verdadeiro tempo médio de serviço populacional, porque assumiríamos que esses valores só teriam sido diferentes da verdadeira média populacional por simples variação amostral. Todos os valores igual ou menores, ou igual ou maiores do que aqueles limites de significância seriam muito improváveis de serem obtidos e, portanto, seriam considerados por nós como estatisticamente diferentes da verdadeira média populacional. Seriam valores não esperados por simples variação amostral. Concluiríamos que seria muito improvável que um desses valores fosse o verdadeiro resultado que obteríamos caso tivéssemos estudado toda a população-alvo. Como tais valores estariam em localizações muito extremas na distribuição, se verificássemos que o estudo realizado não tinha vieses de seleção, informação ou de confundimento, isso nos indicaria que seria mais provável (pelo menos 95,0% mais provável) que os valores muito extremos pertencessem a uma outra população, e não à população de trabalhadores da qual a amostra estudada foi retirada.

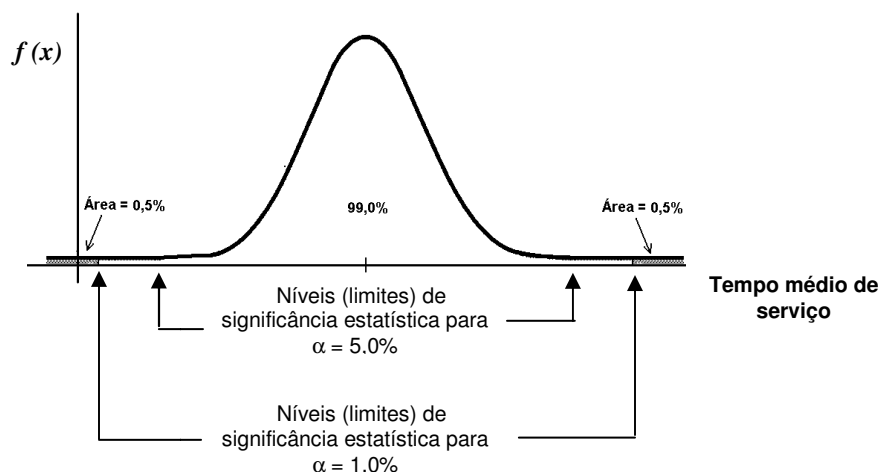
Assim, podemos afirmar que a definição do nível de significância delimita duas áreas distintas no

diagrama da distribuição probabilística utilizada como modelo (no exemplo, a distribuição normal): uma incluindo valores de tempo médio de serviço que seriam esperados por simples variação amostral de resultados (valores mais prováveis de serem obtidos em amostras retiradas aleatoriamente da população-alvo), e outra com valores não esperados por variação amostral (valores menos prováveis de serem obtidos em amostras retiradas aleatoriamente da população-alvo) e que, por serem muito extremos, ultrapassariam o nível de significância estatística, como pode ser visto no diagrama a seguir:



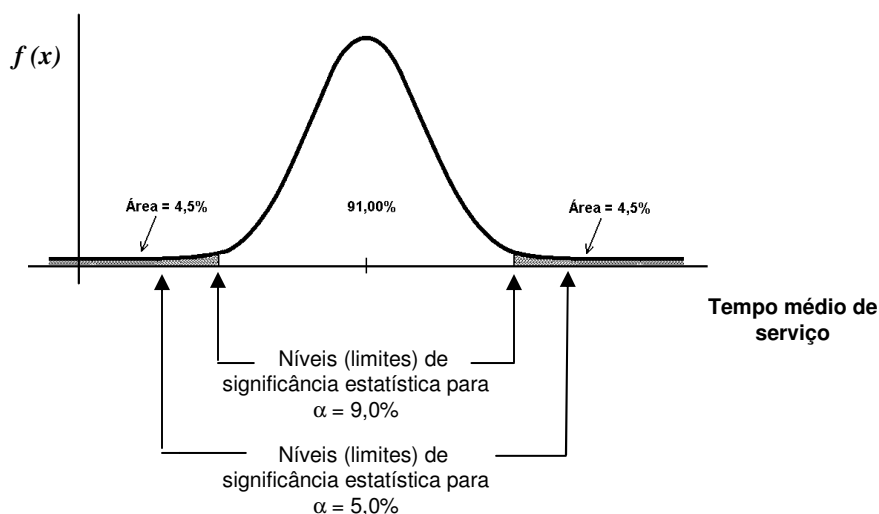
— E o alfa tem que ser sempre igual a 0,05?

— Não. Este é o valor mais utilizado, mas se você quiser um teste de hipóteses mais exigente no sentido de tornar mais difícil considerar um determinado valor estatisticamente diferente, você pode escolher um α menor, como 0,01 (1,0%), por exemplo. Observe que esta escolha fará com que os limites de significância fiquem mais extremos na distribuição, resultando em uma maior dificuldade para que certo valor os ultrapasse, como mostrado abaixo:



Haverá situações em que você escolherá um α menos conservador (menos exigente), de 10,0% ou mais, por exemplo. Vários autores (*Bendel RB e Afifi AA. Comparison of stopping rules in forward regression. Journal of the American Statistical Association 1977; 72:46-53; Costanza MC e Afifi AA. Comparison of stopping rules in forward stepwise discriminant analysis. Journal of the American Statistical Association 1979; 74:777-785; Hosmer DW e Lemeshow S. Applied logistic regression. 2ª ed. New York(NY): John Wiley; 2000*) recomendam fortemente a utilização de níveis de significância estatística tão altos quanto 0,15 (15,0%) ou 0,20 (20,0%), ou mesmo 0,25 (25,0%), em algumas técnicas estatísticas multivariáveis, como os diversos tipos de análise de regressão. Uma das justificativas para isso é que muitas das associações estudadas não são muito fortes, embora importantes, o que torna necessário o uso de um nível de significância menos exigente para que as técnicas estatísticas utilizadas sejam sensíveis o suficiente para detectá-las. Isso é especialmente aplicável quando se está investigando interações entre variáveis.

Tenha sempre em mente que quando mudamos o nosso α de 0,05 para 0,09, p. ex., uma das consequências desta decisão é que se torna mais fácil concluirmos que o resultado é estatisticamente significativo, já que esse aumento faz aumentar as áreas que estamos considerando nas extremidades da curva, o que resulta em limites de significância estatística mais baixos, mais fáceis de serem ultrapassados, não é? Veja isso no diagrama abaixo:



Com um alfa de 0,05, admitiríamos que, mesmo havendo uma probabilidade de 5,0% de um valor muito extremo pertencer à população de trabalhadores da Refinaria, concluiríamos que esse valor não pertenceria a esta distribuição, já que a probabilidade disso ocorrer seria muito pequena (de no máximo 5,0%). Com um alfa de 0,09, seria maior a probabilidade de concluirmos incorretamente que um valor extremo não pertenceria à população, pois nessa situação essa probabilidade seria de 9,0%. Então veja que, ao aumentarmos o nosso alfa, nossa probabilidade de errar aumentaria. No caso subiria de 5,0% para 9,0%, quatro pontos percentuais a mais. Mas, você verá mais adiante, quando explicarmos melhor os erros envolvidos na inferência estatística, que, mesmo aumentando o nosso erro para 9,0%, nossa probabilidade de acertar ainda seria muito grande, pois seria de 91,0%.

Contudo, se o pesquisador avaliasse que quatro pontos percentuais a mais de probabilidade de errar seriam expressivos clínica ou epidemiologicamente, ele deveria manter o seu nível de significância em 0,05 (5,0%) ou mesmo diminuí-lo. Ficará mais fácil para você entender isso um pouco mais adiante. Prossiga.

A segunda etapa é de definição das hipóteses. Estas são denominadas de hipóteses estatísticas. Existem várias hipóteses possíveis, mas para cada teste apenas duas são testadas, como escrevemos anteriormente. Uma delas é denominada de hipótese nula e a outra de hipótese alternativa.

Para continuarmos, e voltando ao nosso exemplo, será necessário especificarmos um determinado valor que esperamos ser o verdadeiro tempo médio de serviço na população de trabalhadores da Refinaria. Faremos isso com base em informações encontradas na literatura ou em nossas próprias observações prévias, se essas existirem. Ou ainda, poderemos testar vários possíveis valores para a média populacional, um de cada vez. Testaremos então, separadamente, se a verdadeira média pode ser 18,0 anos, 18,5 anos, 20,0 anos, ou 16,0 anos, por exemplo. Concordamos que este último processo seria extremamente trabalhoso e demorado. Mais adiante você verá que, na situação do nosso exemplo atual, será mais prático calcularmos o intervalo de confiança, em vez de fazermos vários testes de hipóteses, um para cada valor a ser avaliado como possível para a população. Aguarde um pouco. Isso ficará mais claro em breve.

Por enquanto, suponha que desejemos verificar se a média na população de onde retiramos a amostra estudada pode ser 16,5 anos, sendo esse valor obtido em um artigo encontrado na literatura, em investigação realizada em população de trabalhadores semelhante àquela para a qual queremos estimar a média do tempo de serviço. Suponha também, que o desvio-padrão relatado nesse mesmo artigo tenha sido 5,53 anos.

— Mas, por que verificarmos isso, se já sabemos, no nosso exemplo atual, que a verdadeira média populacional é 14,26 anos, como calculamos na página 147?

— É que, na situação atual, estamos supondo que μ seja desconhecido. É justamente para obter uma estimativa de μ que estamos realizando esse estudo em uma amostra. Assim, vamos considerar 16,5 anos como um valor possível para a verdadeira média da população que estamos estudando, porque esse valor foi encontrado em uma população semelhante. Mais adiante, você verá também que utilizaremos 5,53 anos como o valor esperado para o verdadeiro desvio-padrão populacional.

A hipótese nula é chamada assim porque afirma uma nulidade, isto é, que não há diferença entre a verdadeira média e aquela esperada pelo pesquisador para a população, sendo o valor desta estabelecida nesta hipótese. Na formulação das hipóteses utilizamos sempre notações populacionais, já que queremos tirar conclusões sobre a verdadeira média na população-alvo. Assim, podemos indicar a hipótese nula, ou de nulidade ou “H zero”, por

$$H_o : \mu = 16,5 \text{ anos},$$

onde μ representa a verdadeira média na população. Quando definimos que $H_o : \mu = 16,5$ anos, estamos estabelecendo que a verdadeira média populacional, μ , é 16,5 anos. Decidiremos por rejeitar ou não essa hipótese, baseando-nos na média obtida na única amostra estudada, que foi 13,73 anos. Nossa estratégia será utilizar a média defendida na hipótese nula, 16,5 anos, como referência, e avaliarmos o quanto esperaríamos que as médias amostrais variassem, caso tivéssemos estudado numerosas amostras, e não apenas uma, retiradas de uma população cuja média fosse 16,5 anos. Se o valor 13,73 anos não ultrapassar

os limites de significância estatística definidos na etapa anterior, quando escolhemos um α de 0,05 ou 5,0%, concluiremos que, estatisticamente, 13,73 anos é igual a 16,5 anos, porque seria alta a probabilidade de obtermos uma média de 13,73 anos em numerosas amostras retiradas de uma população cuja verdadeira média fosse 16,5 anos. Então, este último valor pode ser a verdadeira média populacional, pois a média obtida na única amostra estudada é compatível com isso. Na única amostra que estudamos, a média foi 13,73 anos, mas poderia também ter sido 16,5 anos, caso a amostra estudada (retirada da mesma população) tivesse sido outra e não aquela. A média 13,73 anos pode ter sido menor do que 16,5 anos devido apenas à variação nos resultados, variação esta que ocorreria se tivéssemos estudado numerosas amostras.

A hipótese alternativa à hipótese nula afirma o oposto ao colocado por esta última, e é expressa por

$$H_A : \mu \neq 16,5 \text{ anos} .$$

As duas hipóteses acima poderiam ser assumidas como nossas hipóteses estatísticas, mas outras também poderiam. Nossa hipótese alternativa poderia ser a de que o verdadeiro tempo médio de serviço na população era menor do que aquele esperado, ou seja,

$$H_A : \mu < 16,5 \text{ anos} .$$

Nesse caso, nossa hipótese nula seria:

$$H_O : \mu \geq 16,5 \text{ anos} .$$

— **Mas não é óbvio que μ é menor do que 16,5 anos, já que 13,73 anos é um tempo médio menor do que 16,5 anos?**

— Como 13,73 anos é uma quantidade de anos menor do que 16,5 anos, poderíamos pensar em concluir o estudo dizendo que a verdadeira média populacional é menor do que a esperada, já que na amostra retirada da população-alvo e supostamente representativa desta, encontramos o valor 13,73 anos. Contudo, embora matematicamente 13,73 anos seja menor do que 16,5 anos, a questão a ser esclarecida é se estatisticamente 13,73 anos é menor do que 16,5 anos. Já vimos anteriormente que essa comparação é estatística, porque só investigamos uma amostra aleatória, e haveria uma variação dos resultados amostrais caso tivéssemos estudado uma outra amostra e não justamente aquela. Isto nos deixa em dúvida acerca do resultado obtido na única amostra estudada..

Outra possibilidade seria assumirmos as seguintes hipóteses:

$$H_A : \mu > 16,5 \text{ anos} \text{ e } H_O : \mu \leq 16,5 \text{ anos} .$$

— **Por que testaríamos se μ é maior do que 16,5 anos, se o valor encontrado na única amostra retirada da população cuja média se supõe ser 16,5 anos foi menor do que este valor?**

— Você já sabe a resposta: esta desigualdade é definida estatisticamente e não matematicamente.

Na única amostra que estudamos obtivemos um tempo médio de serviço de 13,73 anos, mas este valor poderia ter sido maior ou menor, a depender da amostra que tivéssemos selecionado, podendo ser, portanto, maior ou menor do que 16,5 anos. Lembra-se de que, se tivéssemos estudado várias amostras, naquela de número 4 poderíamos ter obtido um tempo médio de 19,27 anos, que seria maior do que 16,5 anos? Suponha que tenha sido esta a amostra selecionada. Testarmos se 19,27 anos é estatisticamente maior do que 16,5 anos, equivaleria a testarmos se o simétrico de 13,73 anos na distribuição é maior do que 16,5 anos, porque 19,27 anos é o tempo médio que se afasta para cima de 16,5 anos a mesma quantidade que 13,73 anos se afasta para baixo.

— Certo. Entendi. Mas, diante dessas várias possibilidades para minhas hipóteses estatísticas, como saberei quais as duas mais adequadas ao meu estudo?

— Se o seu estudo for um dos primeiros sobre o tema, escolha as hipóteses $H_O : \mu = 16,5$ anos e $H_A : \mu \neq 16,5$ anos, porque nesse caso sua hipótese alternativa estará ao mesmo tempo testando se $\mu < 16,5$ anos ou se $\mu > 16,5$ anos. Se μ é diferente de 16,5 anos, então ou μ é menor do que 16,5 anos ou maior, não é? Assim, quando o estudo sobre o tema for muito inicial, não havendo ainda um conjunto razoável de evidências sobre o assunto, é recomendável que você seja mais cauteloso(a) e admita que o resultado encontrado possa ser maior ou menor que um determinado valor. Mais adiante você verá que escolhendo essas hipóteses estará também realizando um teste estatístico mais rigoroso, portanto, mais adequado a uma situação na qual você tem mais incertezas do que certezas, como ocorre em fases muito iniciais de investigação de um tema.

Entretanto, se vários estudos já foram feitos sobre o assunto, suas hipóteses estatísticas poderão ser: a) $H_O : \mu \geq 16,5$ anos e $H_A : \mu < 16,5$ anos, se as evidências pré-existentes sugerirem que a média populacional é menor do que 16,5 anos; ou b) $H_O : \mu \leq 16,5$ anos e $H_A : \mu > 16,5$ anos, se essas evidências apontarem para uma média populacional maior do que 16,5 anos;

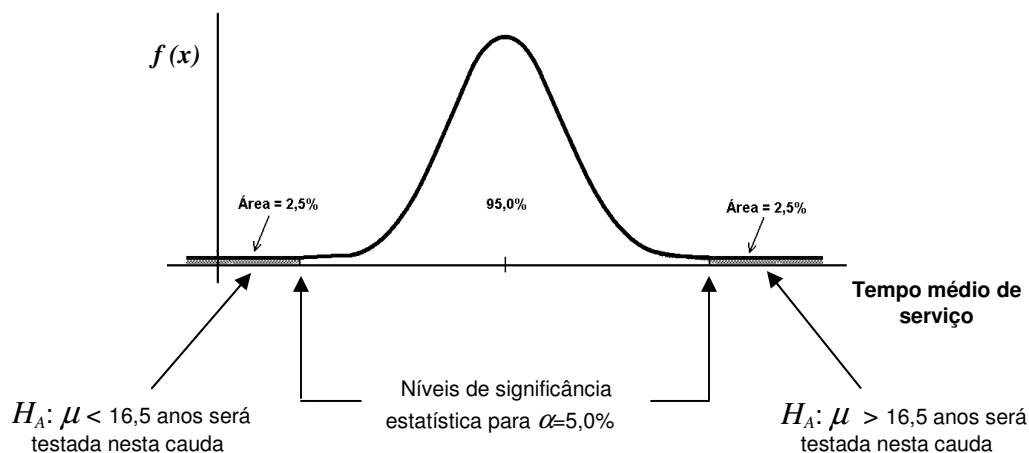
As hipóteses mais freqüentemente utilizadas são do tipo $H_O : \mu = 16,5$ anos e $H_A : \mu \neq 16,5$ anos, não porque os estudos sejam iniciais, mas por implicarem em testes mais conservadores (mais exigentes), como você verá em breve.

Para passarmos à próxima etapa de um teste de hipóteses sobre uma média, vamos assumir que as hipóteses definidas por nós no exemplo atual foram:

$$H_O : \mu = \mu_o \text{ e } H_A : \mu \neq \mu_o.$$

Observe que μ_o (lemos “mi zero”) denota o valor estabelecido como o verdadeiro pela hipótese nula, que no nosso exemplo é 16,5 anos, sendo seu subscrito $_o$ escolhido justamente para indicar nulidade.

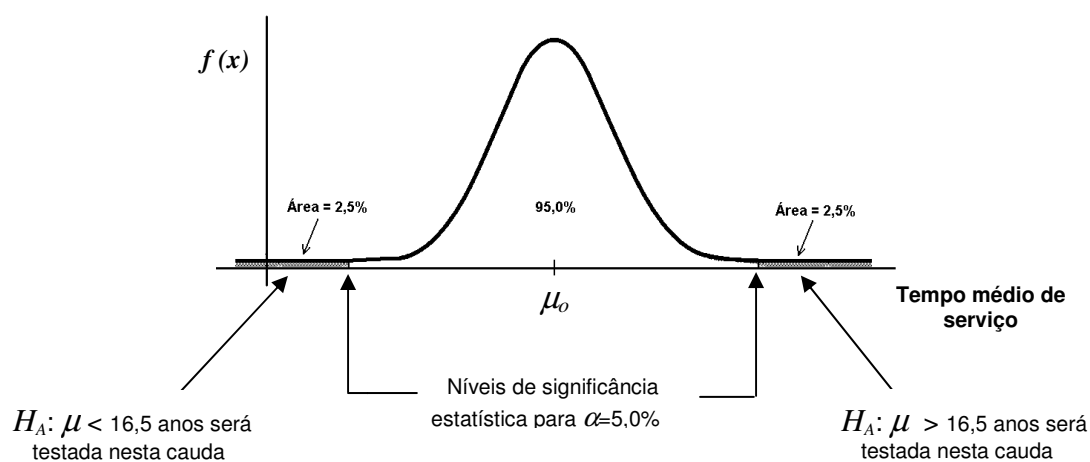
Já vimos que a hipótese $H_A : \mu \neq \mu_o$, engloba na verdade duas hipóteses, porque se $\mu \neq 16,5$ anos, ou $\mu < 16,5$ anos ou $\mu > 16,5$ anos. A implicação prática disso é que, como estamos testando duas hipóteses alternativas, uma dessas, $\mu < 16,5$ anos, é avaliada na cauda esquerda da distribuição normal e a outra, $\mu > 16,5$ anos, na direita, como mostrado no próximo diagrama. Nesse caso, o teste é denominado **bicaudado**.



Os 5,0% do nosso nível de significância tiveram que ser divididos em duas áreas iguais a 2,5%, uma em cada cauda da distribuição, porque há duas hipóteses alternativas a serem testadas. Por isso, dissemos anteriormente que a definição das hipóteses é importante para sabermos a maneira correta de considerarmos o nosso alfa.

Tanto a parte central da distribuição, que contém os valores com maior probabilidade de ocorrerem, como as áreas nas suas extremidades, correspondentes a valores com menor probabilidade de ocorrerem, serão utilizadas para testar as hipóteses formuladas.

Definidos nosso nível de significância e nossas hipóteses estatísticas, passaremos à terceira etapa do teste de hipóteses. Nesta etapa, calcularemos o quanto o tempo médio de serviço observado na amostra estudada, $\bar{x} = 13,73$ anos, desvia-se (afasta-se) do tempo médio de serviço esperado para a população, $\mu_o = 16,5$ anos. Nosso objetivo será avaliar se 13,73 anos afasta-se de 16,5 anos o suficiente para ultrapassar o limite de significância estatística. Como pode notar no diagrama a seguir, usamos a média estabelecida na hipótese nula, μ_o , como referência para a elaboração da distribuição normal do nosso teste:



Nossa estratégia consiste em, tomando 16,5 anos como referência, utilizarmos o erro-padrão que seria esperado caso tivéssemos estudado numerosas amostras, para quantificarmos o quanto de variação em

torno dessa média haveria, de modo a podermos avaliar se 13,73 anos estaria dentro ou fora dessa faixa de variação. Esta será considerada por nós como a faixa de variação esperada por simples variação de resultados amostrais. Se 13,73 anos estiver dentro dessa faixa, concluiremos que o desvio (afastamento, diferença) dessa média em relação à média esperada para a população, decorreu de variação de resultados amostrais, e não de uma diferença estatisticamente significativa. Isso nos faria concluir que 13,73 anos seria estatisticamente igual a 16,5 anos e que, portanto, este último valor poderia ser o verdadeiro valor populacional, por ser uma das médias que mais provavelmente poderiam ser encontradas caso investigássemos numerosas amostras retiradas da mesma população-alvo de onde a única amostra estudada foi retirada. Termos obtido uma amostra cuja média é 13,73 anos seria estatisticamente compatível com uma média populacional de 16,5 anos. Quando isso ocorrer, aceitaremos a hipótese nula e rejeitaremos a alternativa. Se 13,73 anos estiver fora da faixa, ou seja, nos extremos da distribuição, concluiremos o contrário.

— Por que μ_o é utilizada como referência e não a média (\bar{x}) obtida na amostra?

— Ainda não dá para respondermos a essa sua importante pergunta. Deixe-nos prosseguir mais um pouco para que possa entender isso.

Depois de calcular o quanto 13,73 anos desviou de 16,5 anos, dividiremos esse desvio pelo erro-padrão dos tempos médios de serviço, esperado em numerosas amostras. Essa operação pode ser expressa por

$$z = \frac{\bar{x} - \mu_o}{EP} = \frac{\bar{x} - \mu_o}{\sqrt{\frac{\sigma^2}{n}}} = \frac{\bar{x} - \mu_o}{\frac{\sigma}{\sqrt{n}}},$$

onde z é um valor da distribuição normal padrão, \bar{x} a média amostral, μ_o a média populacional estabelecida na hipótese nula, e σ/\sqrt{n} o erro-padrão das médias amostrais.

No numerador da fórmula acima está representado o desvio de \bar{x} em relação a μ_o , que foi calculado por uma operação de diminuição ($\bar{x} - \mu_o$). Quanto maior o resultado dessa operação, maior o desvio, o afastamento, de \bar{x} em relação a μ_o . Ao dividirmos esse desvio pelo erro-padrão das médias amostrais, obtemos tal desvio expresso em número de erros-padrão, concorda? Veja bem: se dividirmos, uma quantidade maior por outra menor, o resultado não representará quantas vezes uma é maior do que a outra? E isso não é o mesmo que calcular o número de quantidades menores que cabem dentro da quantidade maior? Então, ao fazermos a divisão das duas quantidades acima, $(\bar{x} - \mu_o)$ por EP , obtemos quantos erros-padrão cabem dentro do valor do desvio em questão, ou, correspondentemente, quantos erros-padrão equivalem ao desvio de \bar{x} em relação a μ_o . Esse número de erros-padrão é denotado por z .

Na terceira etapa do teste de hipóteses calculamos o valor de z , que expressa, em número de erros-padrão, o desvio da média amostral em relação à média esperada na população.

— Para que fazemos isso?

— Você já sabe a resposta. Ao expressarmos esse desvio em número de erros-padrão, ou seja, em

valor de z , poderemos utilizar a tabela de áreas sob a curva normal padrão para encontrarmos as probabilidades sob a curva normal, abaixo, acima ou entre dois valores quaisquer de X , sendo que X no nosso exemplo representa os tempos médios de serviço. Se não calculássemos o valor de z , teríamos que utilizar cálculo integral toda vez que efetuássemos um teste de hipóteses, para obtermos probabilidades sob a curva.

Ainda neste capítulo, você verá que o famoso **valor- p** é uma probabilidade sob uma distribuição probabilística (na situação atual, sob a distribuição normal), e verá como utilizaremos esse valor para verificarmos se 13,73 anos é estatisticamente diferente ou não de 16,5 anos. Calcular o valor de z nos permitirá saber qual seria a probabilidade de obtermos tempos médios de serviço menores do que 13,73 anos, caso tivéssemos estudado numerosas amostras. Se essa probabilidade, que é chamada de valor- p , for menor do que o valor que estabelecemos para o nosso α , consideraremos essa probabilidade como muito baixa, e concluiremos que 13,73 anos é um valor muito improvável de ser obtido em numerosas amostras que porventura tivéssemos investigado. Isso equivaleria a dizermos que 13,73 anos era estatisticamente diferente de 16,5 anos, e que, muito provavelmente (pelo menos com uma probabilidade de 95,0%, que é o complemento do nosso α), 16,5 anos não era a verdadeira média populacional.

No capítulo anterior (página 136), utilizamos um valor de Z para expressarmos o desvio de um valor de X obtido para um indivíduo de uma amostra ou população em relação à média de X obtida para todos os indivíduos dessa amostra ou população. Vimos, também, que o valor de z foi calculado através da expressão $z = (x - \bar{x})/s$, na qual, considerando agora o nosso exemplo, x indica um tempo de serviço específico ($X = x$) obtido para um trabalhador, \bar{x} o tempo médio de serviço dos trabalhadores da única amostra estudada (usado como estimador de μ , que é geralmente desconhecida), e s o desvio-padrão dessa variável nessa amostra (usado como estimador de σ , também geralmente desconhecido). O valor de z obtido desse modo expressará, portanto, em número de desvios-padrão, o desvio do tempo de serviço de um indivíduo em relação ao tempo médio de serviço de todos os indivíduos da amostra.

Na nossa situação atual, isto é, no contexto da inferência estatística, o valor de z expressará, em número de erros-padrão, o desvio de uma média de valores de X , \bar{x} , obtida para uma amostra, em relação à média das médias de X (média de \bar{x} ou $\bar{\bar{x}}$) obtidas em numerosas amostras que porventura tivéssemos estudado, sendo $\bar{\bar{x}}$ um estimador de μ , que geralmente é desconhecida. Este estimador é válido porque podemos assumir que $\bar{\bar{x}}$ converge para μ , se considerarmos um número cada vez maior de amostras. Acontece que, na nossa prática científica não dispomos também desse possível estimador de μ , já que não investigamos numerosas amostras. Faremos então o seguinte: utilizaremos $\mu_o = 16,5$ anos, que é a nossa expectativa para o valor da média populacional estabelecida na hipótese nula, em substituição a μ . Esta é a razão para utilizarmos μ_o como referência no nosso teste e para assumirmos que a hipótese nula é verdadeira durante a realização do teste.

O desvio que nos interessa, que teoricamente seria calculado por $z = (\bar{x} - \mu)/(\sigma/\sqrt{n})$, quando o desvio-padrão populacional fosse conhecido, ou por $t = (\bar{x} - \mu)/(s/\sqrt{n})$ (veja isto no próximo capítulo), quando o desvio-padrão populacional fosse desconhecido, será, portanto, calculado através das expressões $z = (\bar{x} - \mu_o)/(\sigma/\sqrt{n})$ e $t = (\bar{x} - \mu_o)/(s/\sqrt{n})$, respectivamente. Nas expressões acima, e considerando o nosso exemplo, \bar{x} indica o tempo médio de serviço obtido na única amostra investigada; μ_o , que representa a média populacional considerada como verdadeira pelos pesquisadores, substitui μ , que denota a verdadeira média populacional; e σ ou s indica o desvio-padrão populacional ou amostral, respectivamente. Lembre-se de que σ/\sqrt{n} e s/\sqrt{n} representam o erro-padrão de médias amostrais, ou

seja, o quanto, em média, as médias amostrais desviariam da média das médias amostrais, caso numerosas amostras tivessem sido estudadas. O valor de z obtido desse modo expressará, portanto, em número de erros-padrão, o desvio do tempo médio de serviço da única amostra estudada em relação ao tempo médio de serviço esperado para a população.

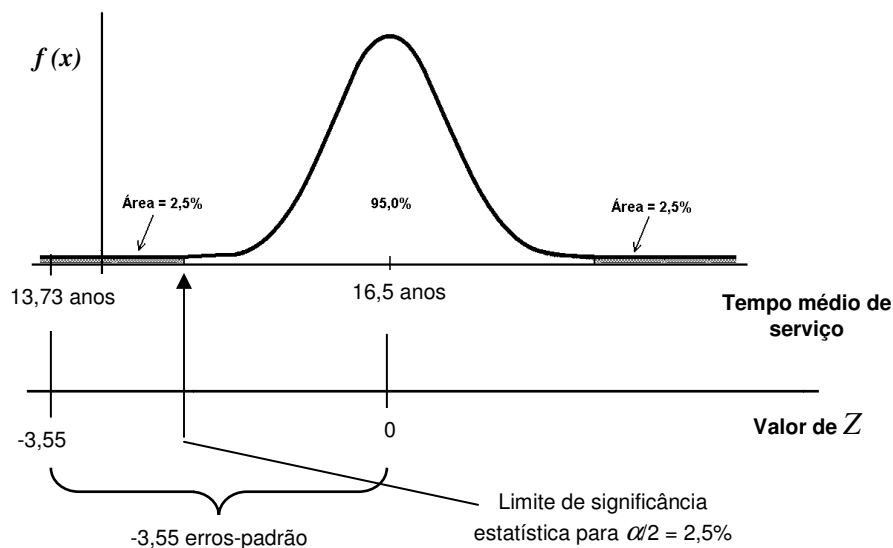
Feito isto, avaliaremos se a magnitude do desvio da média obtida ($\bar{x} = 13,73$ anos) em relação à média populacional esperada ($\mu_o = 16,5$ anos) é suficientemente grande para concluirmos pela provável existência de uma diferença estatisticamente significativa entre a verdadeira média e a média esperada, ou seja, para concluirmos que a verdadeira média pode ser 16,5 anos.

No nosso exemplo, para testarmos a hipótese de que $\mu < 16,5$ anos (uma das hipóteses contidas na hipótese alternativa $\mu \neq 16,5$ anos), como o desvio-padrão esperado para a população, σ , é conhecido, sendo igual a 5,53 anos, valor obtido em estudo já mencionado encontrado na literatura e realizado em população de trabalhadores semelhante à que estamos estudando, nosso z será

$$z = \frac{\bar{x} - \mu_o}{\frac{\sigma}{\sqrt{n}}} = \frac{13,73 - 16,50}{\frac{5,53}{\sqrt{50}}} = \frac{-2,77}{\frac{5,53}{7,07}} = \frac{-2,77}{0,78} \cong -3,55.$$

Assim, encontramos que a média amostral, 13,73 anos, desvia-se, afasta-se 3,55 erros-padrão para baixo do valor esperado, 16,5 anos.

Veja no diagrama a seguir a correspondência entre os valores originais de tempo médio de serviço e os valores de Z :



— Por que o valor de z correspondente à média dos tempos médios de serviço é zero?

— Para entender isso tente responder à seguinte questão: se a única amostra investigada tivesse apresentado um tempo médio de serviço de 16,5 anos, quantos erros-padrão essa média se desviaria da média populacional esperada, ou seja, quantos erros-padrão essa média se desviaria dela mesmo?

Obviamente, não desviaria nada, concorda? Para tirarmos qualquer dúvida sobre isso, podemos fazer o cálculo do valor de z para um tempo médio de serviço de 16,5 anos:

$$z = \frac{\bar{x} - \mu_o}{\frac{\sigma}{\sqrt{n}}} = \frac{16,5 - 16,5}{\frac{5,53}{\sqrt{50}}} = \frac{0}{\frac{5,53}{7,07}} = \frac{0}{0,78} = 0.$$

Logo, o valor de z correspondente à média populacional esperada é zero. Isso é o mesmo que dizer que a média da distribuição normal padrão é zero. Como a abscissa dessa distribuição é expressa em valores de Z , e estes são expressos em número de desvios-padrão (no caso de variabilidade em uma amostra ou população) ou de erros-padrão (no caso de variabilidade em diversas amostras), considerando que a média dos tempos médios não desvia nada dela mesmo, seu desvio em relação a si mesma é igual a zero erro-padrão.

A média da distribuição normal padrão é igual a zero.

O erro-padrão dessa distribuição é 1. Para entender isso lembre-se de que $z = (\bar{x} - \mu) / (\sigma / \sqrt{n})$ e $Var(\bar{x}) = (\sigma / \sqrt{n})^2 = \sigma^2 / n$, onde $Var(\bar{x})$ é a notação para a variância de médias amostrais, σ^2 representa a variância populacional, e n o tamanho da amostra estudada. É necessário também considerarmos a seguinte propriedade de uma média aritmética: $Var(k\bar{x}) = k^2 Var(\bar{x})$, isto é, se multiplicarmos cada média de uma série de médias por uma constante k , a variância dos novos valores obtidos será igual à constante ao quadrado vezes a variância dos valores originais da série. Experimente demonstrar isto empiricamente: escolha uma série de três médias quaisquer; multiplique cada uma delas por uma constante qualquer; calcule a variância de \bar{x} e depois a de $k\bar{x}$; você verá que $Var(k\bar{x}) = k^2 Var(\bar{x})$.

Agora você está pronto(a) para entender que $Var(z) = 1$, pois $Var(z) = Var\left(\frac{\bar{x} - \mu}{\sigma / \sqrt{n}}\right) = Var\left(\frac{\bar{x}}{\sigma / \sqrt{n}}\right) + Var\left(\frac{\mu}{\sigma / \sqrt{n}}\right)$ e, como $Var\left(\frac{\mu}{\sigma / \sqrt{n}}\right) = 0$, porque ao interior de um mesmo estudo e para cada variável do estudo, μ , σ , e n são todos constantes e a variância de uma constante é obviamente zero, a expressão pode ser simplificada para $Var(z) = Var\left(\frac{\bar{x}}{\sigma / \sqrt{n}}\right) + 0 = Var\left(\frac{\bar{x}}{\sigma / \sqrt{n}}\right) = Var\left(\frac{1}{\sigma / \sqrt{n}} \bar{x}\right) = \left[\left(\frac{1}{\sigma / \sqrt{n}}\right)^2 Var(\bar{x})\right] = \frac{(1/1)^2}{(\sigma / \sqrt{n})^2} Var(\bar{x}) = \left(\frac{1}{1}\right)\left(\frac{n}{\sigma^2}\right) Var(\bar{x}) = \left(\frac{n}{\sigma^2}\right) Var(\bar{x})$.

Substituindo $Var(\bar{x})$ por $\frac{\sigma^2}{n}$, temos que $Var(z) = \left(\frac{n}{\sigma^2}\right)\left(\frac{\sigma^2}{n}\right) = \frac{\sigma^2}{\sigma^2} = 1$. Como o erro-padrão de z é

igual a $\sqrt{\text{Var}(z)}$, e como acabamos de mostrar, a $\text{Var}(z) = 1$, o erro-padrão de z é igual a $\sqrt{1} = 1$.

O erro-padrão (o desvio-padrão também) da distribuição normal padrão é igual a 1.

Faltaria ainda nesta etapa calcularmos o valor de z correspondente à hipótese de que $\mu > 16,5$ anos (a outra hipótese contida na hipótese alternativa $\mu \neq 16,5$ anos). Os estatísticos consideram este cálculo desnecessário, pois a distribuição normal é simétrica e será obtido um resultado igual àquele encontrado na cauda esquerda da curva, com exceção do sinal, que será positivo. Vamos, contudo, fazer esse cálculo, porque talvez isso lhe ajude a entender melhor um teste de hipóteses.

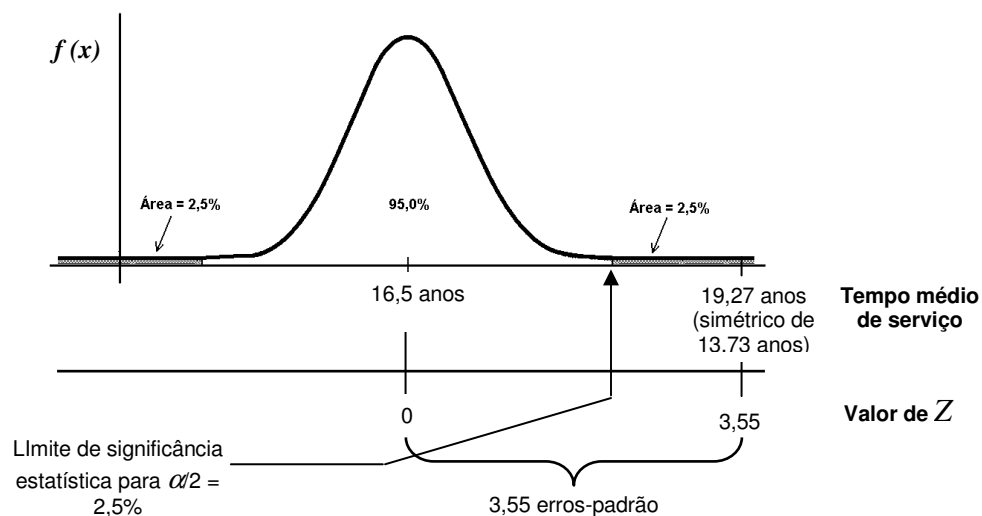
Como vimos, a questão que nos é posta aqui é a de que só estudamos uma amostra, mas se tivéssemos estudado várias, poderíamos ter obtido outros tempos médios de serviço, inclusive um tempo médio maior do que 16,5 anos. A média encontrada poderia desviar para cima deste valor, tanto quanto 13,73 anos desviou para baixo. O que desejaríamos saber seria qual o número de erros-padrão de afastamento, caso esse afastamento ocorresse para cima da média esperada, e fosse da mesma magnitude daquele ocorrido para baixo dessa média. Essa média desviada à direita poderia ter sido então 19,27 anos, obtido pela soma $16,5 + 2,77$. Nosso teste tem que considerar essa possibilidade e é isso que estamos fazendo ao trabalhar com a extremidade direita da distribuição.

Como σ é conhecido, sendo igual a 5,53 anos, nosso z seria

$$z = \frac{\bar{x} - \mu_o}{\frac{\sigma}{\sqrt{n}}} = \frac{19,27 - 16,50}{\frac{5,53}{\sqrt{50}}} = \frac{2,77}{\frac{5,53}{7,07}} = \frac{2,77}{0,78} \cong 3,55.$$

Observe que o valor de z é igual ao obtido para a cauda esquerda, exceto pelo sinal.

Veja no diagrama abaixo a correspondência entre os valores originais de tempo médio de serviço e os valores de Z :



A quarta etapa de um teste de hipóteses consiste em verificarmos na tabela de áreas sob a curva Z , qual a probabilidade de obtermos valores de Z maiores do que 3,55 (cauda direita de curva) ou menores do que -3,55 (cauda esquerda da curva).

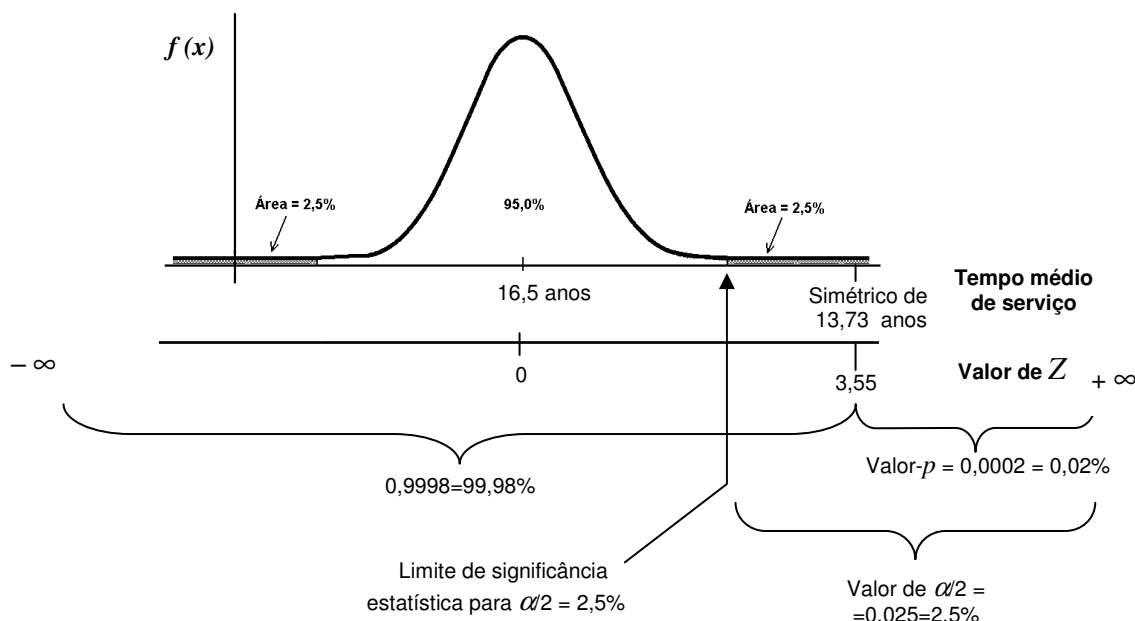
Na prática, só será necessário calcularmos a segunda dessas probabilidades, porque na amostra do nosso exemplo, encontramos uma média menor do que a esperada, mas, novamente, vamos considerar as duas extremidades da curva, com a finalidade de facilitar o seu entendimento. Começaremos com a cauda direita também por este último motivo. Fica claro ainda, que se em um estudo a média amostral for matematicamente maior do que a esperada, você já fica sabendo como proceder, pois bastará considerar a cauda direita da distribuição, sabendo que o mesmo resultado será obtido na outra extremidade.

Veja a seguir como essas probabilidades são obtidas:

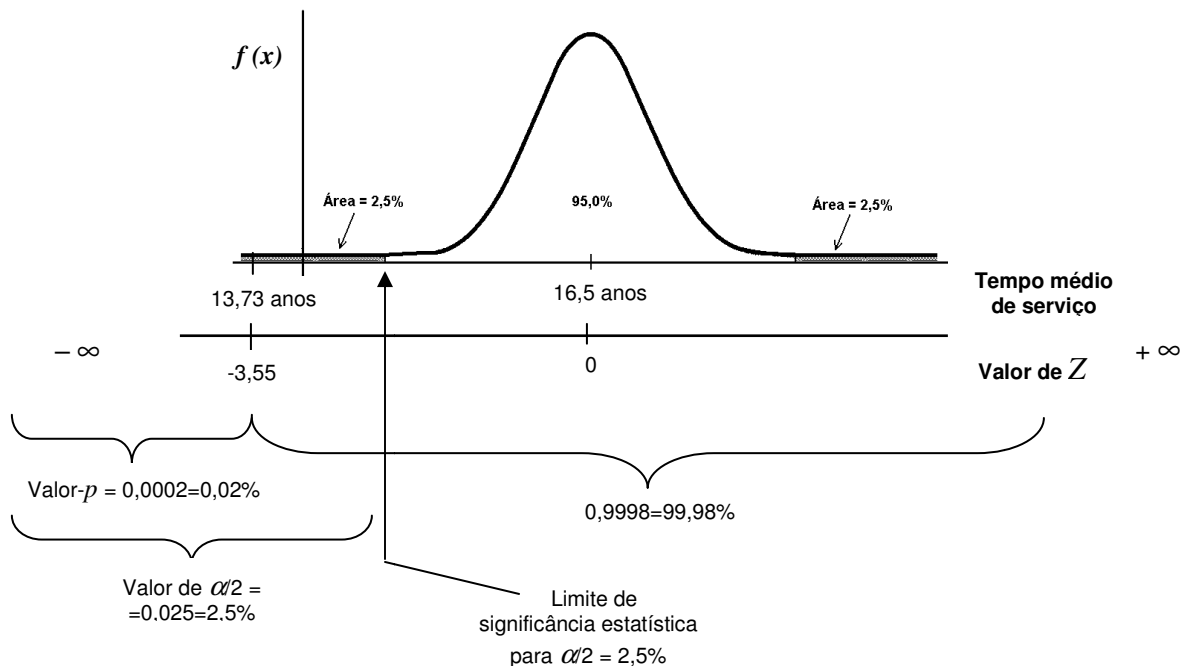
a) Para a cauda direita: olhe na tabela (página 134) e encontre a probabilidade de obtermos valores de z menores do que 3,55, isto é, $P(Z < 3,55)$. A probabilidade é de 0,9998 ou 99,98%. Logo, a probabilidade de obtermos valores de z maiores do que 3,55, que é a probabilidade que nos interessa, será calculada por

$$P(Z > 3,55) = 1 - P(Z < 3,55) = 1 - 0,9998 = 0,0002 \text{ ou } 0,02\%.$$

Esta probabilidade é denominada por **valor- p** , tendo sido possivelmente escolhida a notação p justamente para indicar uma probabilidade. Veja no diagrama abaixo o valor- p e outras probabilidades envolvidas no teste:



b) Para a cauda esquerda: como a distribuição normal é simétrica, a probabilidade de obtermos valores de z menores do que -3,55 é a mesma obtida acima, 0,02%. Veja isso no diagrama a seguir:



Não fique preocupado(a) porque, ainda neste capítulo, você verá como obter o valor- p diretamente na cauda esquerda da distribuição.

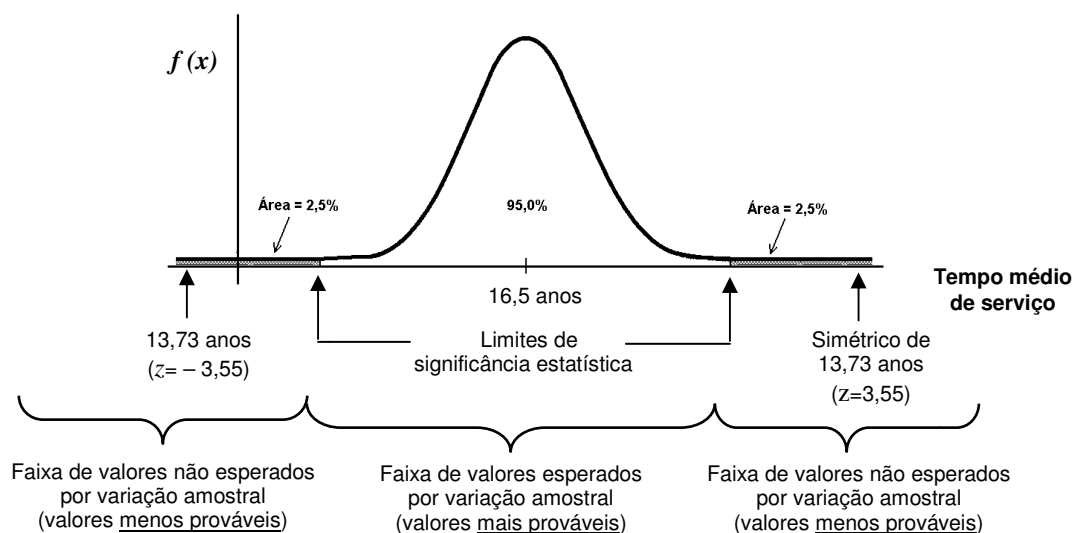
A probabilidade obtida em cada cauda da curva está testando uma hipótese específica ($\mu > 16,5$ anos ou $\mu < 16,5$ anos), mas essas duas hipóteses estão sendo verificadas ao mesmo tempo no teste de hipóteses, já que estamos testando a hipótese alternativa de que $\mu \neq 16,5$ anos. Então, temos de somar as probabilidades obtidas nas duas caudas ou multiplicar seu valor por dois, para chegarmos à probabilidade total de obtermos valores de Z maiores do que $3,55$ ou menores do que $-3,55$.

Considerando então as duas caudas, o valor- p final será

$$p = 0,02 + 0,02 = 2(0,02) = 0,04\%.$$

Na quinta etapa comparamos o valor- p ao valor de α , que é o nosso nível de significância, e concluimos o teste estatístico. Um valor- p menor ou igual ao valor de α , vai nos indicar que a probabilidade de obtermos valores de Z maiores do que $3,55$ ou menores do que $-3,55$ é muito pequena. O valor- p encontrado em nosso exemplo, $0,04\%$, é menor do que $5,0\%$. O valor- p é pequeno o suficiente para podermos afirmar que $-3,55$ e $3,55$, (que correspondem, respectivamente, a $13,73$ anos e seu simétrico na abscissa, não se esqueça disso), são valores que não estão dentro da faixa de valores esperados por variação amostral. Isso é o mesmo que dizermos que $3,55$ e $-3,55$ são valores grande ou pequeno demais, respectivamente, para serem obtidos por simples variação amostral, caso tivéssemos estudado numerosas amostras, retiradas de uma população cuja média fosse $16,5$ anos. São valores que ultrapassam os valores máximo e mínimo esperados por variação amostral. Podemos também dizer que seria muito improvável obtermos valores mais extremos do que $13,73$ anos ou seu simétrico, em amostras retiradas de uma população com média igual a $16,5$ anos. Então, a média populacional não deve ser esta. Rejeitaremos a

hipótese nula e aceitaremos a alternativa. Veja a situação atual no diagrama abaixo, que considera simultaneamente ambas as hipóteses contidas na nossa hipótese nula:



Uma opção equivalente seria não multiplicarmos a probabilidade encontrada por dois, mas nesse caso em vez de compararmos 0,0002 a 0,05 (ou 0,02% a 5,0%) teríamos que comparar 0,0002 a $\alpha/2 = 0,05/2 = 0,025$, ou 0,02% a $\alpha/2 = 5,0\%/2 = 2,5\%$. Observe que ao compararmos 0,02% a 2,5%, chegaríamos à mesma conclusão acima, já que $0,02\% < 2,5\%$.

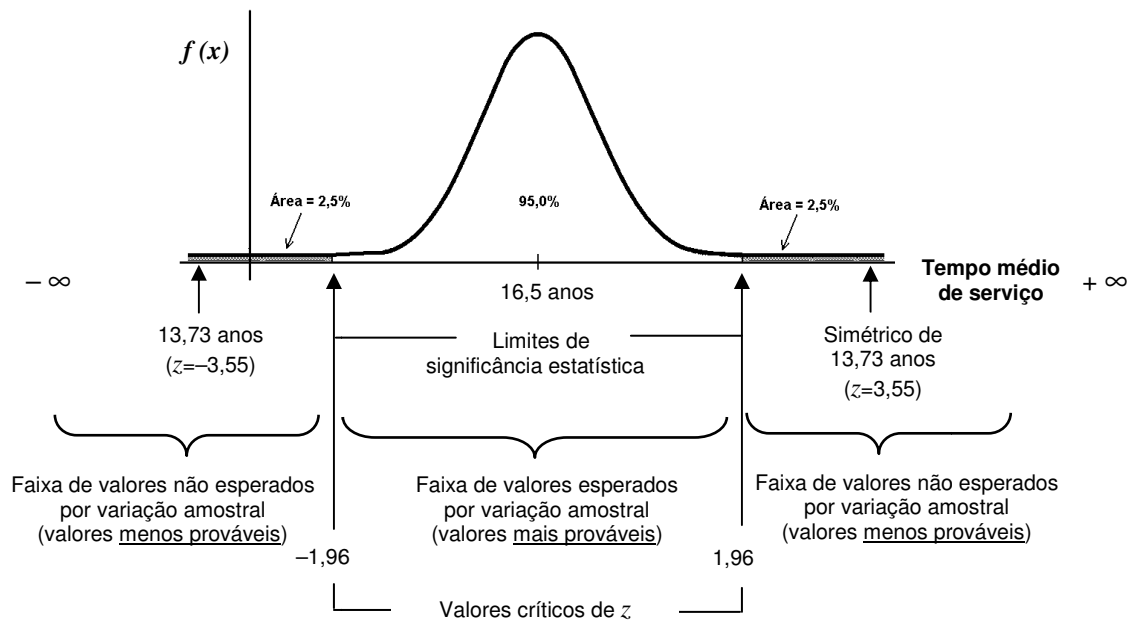
A comparação do valor- p ao valor de α é o modo mais utilizado por pesquisadores para conclusão do teste, mas há uma outra maneira que consiste em compararmos os valores de z calculados ($-3,55$ e $3,55$) aos valores de z correspondentes aos limites de significância estatística (denominados "valores críticos de z "), que no nosso exemplo são $-1,96$ e $1,96$, já que o α considerado é 0,05 e o teste é bicaudado. Se os valores de z calculados ultrapassarem os valores críticos de z , concluiremos que 13,73 anos e seu simétrico são valores suficientemente extremos para localizar-se na área de valores não esperados por variação amostral. Se não, concluiremos o contrário.

— E como encontramos esses valores críticos de z ?

— Procuramos na tabela Z (página 134) o valor de z que separa os 97,5% (0,975) valores mais baixos dos 2,5% valores mais altos da distribuição Z , já que nosso α é 5,0% em um teste bicaudado.

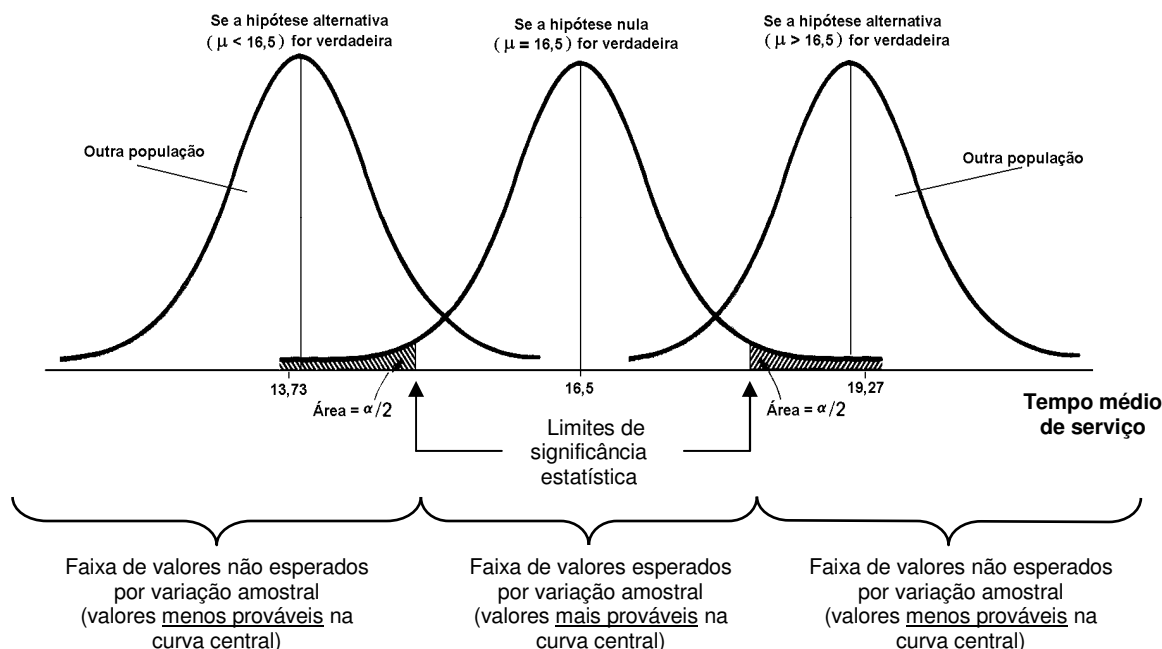
Olhou na tabela?

Vemos que o valor crítico de z , na situação atual e na cauda direita, é igual a 1,96. Para encontrar este valor na tabela Z , procure no corpo dela, aquela célula que contenha o valor 0,975, porque este valor indica a área sob a curva entre $-\infty$ e o valor de z que estamos procurando. Olhando os valores de Z na coluna indicadora e no cabeçalho da tabela Z , verifique que 0,975 é a área entre $-\infty$ e $z = 1,96$. Por simetria, o outro valor crítico é $-1,96$. No diagrama a seguir apresentamos os valores críticos de Z :



No nosso exemplo, vemos que os valores observados de z são menor e maior, respectivamente, do que os valores críticos de z ($-3,55 < -1,96$ e $3,55 > 1,96$). Concluiremos, então, que o tempo médio de serviço obtido na amostra, 13,73 anos, e seu simétrico à direita da média esperada, são tão extremos que ultrapassaram o limite da faixa de valores esperados por variação amostral, o que nos permitirá afirmar que, caso tivéssemos estudado numerosas amostras retiradas da população de trabalhadores da Refinaria, seria muito improvável obtermos tempos médios menores do que 13,73 anos ou maiores do que seu simétrico. É muito mais provável que esses tempos médios possam ser obtidos em amostras de uma outra população e não naquela da Refinaria. Portanto, como muito provavelmente 13,73 anos e 16,5 anos pertencem a populações diferentes, 16,5 anos não deve ser o verdadeiro tempo médio de serviço populacional. Se fosse, seria muito improvável obtermos a média encontrada, 13,73 anos, na única amostra que estudamos. Rejeitaremos a hipótese nula e aceitaremos a alternativa. Note que essa é a mesma conclusão à qual chegamos mais acima quando comparamos o valor- p ao valor de α .

Escrevemos "muito mais provável" porque estamos pelo menos 95,0% seguros de que um tempo médio menor do que 13,73 anos ou maior do que seu simétrico não seria obtido em numerosas amostras retiradas de uma população cuja média fosse 16,5 anos e com desvio-padrão igual a 5,53 anos. Conseqüentemente, admitimos uma probabilidade máxima de 5,0% de errarmos ao concluirmos o teste dessa maneira, porque 5,0% das amostras retiradas da população da Refinaria estudada e não de outra poderiam apresentar resultados menores do que 13,73 anos ou maiores do que seu simétrico, embora isso seja muito pouco provável. Mas o importante é que a probabilidade de acertarmos (95,0%) é muito maior do que a de errarmos (5,0%). Veja isso no diagrama a seguir:



Três populações estão representadas: uma dos tempos médios esperados em amostras retiradas da população de 1.000 trabalhadores da Refinaria (curva central), outra dos tempos médios esperados em amostras retiradas de uma população de trabalhadores com tempos de serviço maiores do que os da Refinaria (curva da direita) e outra dos tempos médios esperados em amostras retiradas de uma população de trabalhadores com tempos de serviço menores (curva da esquerda). Olhando inicialmente apenas a distribuição do centro, observe que um tempo médio de 13,73 anos poderia ser encontrado em amostras retiradas da população-alvo, mas isso seria muito improvável, porque 13,73 anos seria um valor situado em uma posição muito extrema nesta curva, delimitando uma área sob a curva muito pequena entre esse valor e $-\infty$. Isso indicaria justamente que a probabilidade (representada pelo valor- p) de encontrarmos aquele valor em amostras retiradas da população-alvo seria muito pequena. Olhando agora as outras distribuições, verifique que um tempo médio de 13,73 anos (ou seu simétrico) poderia estar situado em uma posição mais central destas curvas, indicando uma alta probabilidade desse valor ser obtido em amostras retiradas de populações com tempos médios diferentes de 16,5 anos.

— Quer dizer que podemos chegar a uma conclusão incorreta ao fazermos um teste de significância estatística?

— Sim, mas como você viu, a probabilidade de um desses erros ocorrer é de no máximo 5,0%, sendo que tal probabilidade de erro é definida pelo pesquisador. Se ele desejar correr um risco ainda menor de cometer esse tipo de erro, poderá estabelecer um nível de significância, α , menor do que 5,0%.

Existem dois tipos de erros possíveis envolvidos no processo de inferência estatística: um desses é a rejeição de H_0 quando H_0 é verdadeira, e o outro a aceitação de H_0 quando H_0 é falsa. O quadro a seguir apresenta os erros que podem ocorrer em um teste de hipóteses:

Conclusão do teste	Realidade sobre H_0	
	É verdadeira	É falsa
Aceitação de H_0	Conclusão correta	Erro tipo II
Rejeição de H_0	Erro tipo I	Conclusão correta

Suponha que exista um ser todo poderoso que saiba com certeza que 16,5 anos é a verdadeira média populacional, pois nós, simples mortais, quase nunca temos certeza absoluta acerca disso. Então, se H_0 é com certeza verdadeira, concluiremos corretamente nosso teste estatístico aceitando-a. Se, contudo, ao invés de aceitá-la nós a rejeitarmos quando for verdadeira, cometeremos o **erro tipo I**. Quando H_0 é com certeza falsa, a decisão correta obviamente será rejeitá-la. Se a aceitarmos estaremos cometendo o **erro tipo II**.

— Entendi, mas se sabemos da possibilidade desses erros ocorrerem, deve ser muito fácil saber se nós os cometemos, não?

— A situação não é tão simples assim. Em primeiro lugar, nunca sabemos com certeza se H_0 é verdadeira ou falsa. Estamos vendo neste capítulo que a única coisa que podemos fazer é verificar se 16,5 anos pode ser ou não o verdadeiro tempo médio populacional. Então, estamos sempre sujeitos ao risco de cometer os erros acima, quando investigamos uma amostra. A maneira de os evitarmos será estudarmos toda a população-alvo, em um estudo sem vieses importantes de seleção, informação ou de confundimento.

— Vocês escreveram que a probabilidade mais freqüentemente admitida para o erro tipo I é de 5,0% e que o pesquisador a define quando estabelece o nível de significância estatística (α) do teste que irá realizar. E qual é a probabilidade mais freqüentemente usada para o erro tipo II?

— A probabilidade é de 20,0%.

— Por que podemos admitir uma probabilidade maior para esse erro?

— Porque, freqüentemente, esse erro é menos grave do que o do tipo I.

— Como assim, menos grave?

— Suponha que estejamos comparando as eficácias de dois medicamentos, A (novo) e B (já utilizado há algum tempo), para o tratamento do câncer de mama. Nossa hipótese alternativa poderia estabelecer que a eficácia da droga A seria maior do que a da droga B, e a nula que a eficácia da primeira seria menor ou igual à segunda. Se cometêssemos o erro tipo I, concluiríamos que a droga A teria uma eficácia maior, quando, na verdade, sua eficácia poderia ser igual ou até mesmo menor. A implicação disso seria que os pacientes seriam tratados com uma droga supostamente mais eficaz, mas poderiam estar tendo um tratamento ineficaz, o que seria uma situação que poderia comprometer não somente o seu bem estar, mas a sua própria vida. Se cometêssemos o erro tipo II, concluiríamos que a eficácia da droga A não era maior, quando, na verdade, era. Como implicação disso, teríamos que os pacientes continuariam sendo tratados com a droga B, já testada e com eficácia já conhecida, não correndo o risco de ficarem sem tratamento eficaz, como na situação anterior. Haveria prejuízo para aqueles pacientes que melhorariam apenas com o novo tratamento, mas este seria um malefício menor do que o anterior, concorda? Por isso, afirmamos que o erro

tipo II é geralmente menos grave que o erro tipo I e, pela mesma razão, utilizamos um β de 20,0% e um α de apenas 5,0%.

Mas nem sempre o erro tipo I é o mais grave. Suponha que o Ministério do Trabalho esteja implementando um Programa de Redução da Rotatividade no Trabalho, e que irá considerar como tendo estabilidade aceitável, empresas com tempo médio de serviço dos seus trabalhadores igual ou maior do que 16,5 anos. As empresas a serem incluídas no Programa deveriam ter tempo médio menor do que esse. Assim, no nosso exemplo, com base no resultado de uma única amostra, iríamos avaliar se a Refinaria estudada deveria ser ou não incluída no Programa. Nossa hipótese nula seria: $\mu \geq 16,5$ anos; e a alternativa: $\mu < 16,5$.

No nosso exemplo, o que significaria cometermos o erro tipo I, ou seja, rejeitarmos uma H_0 verdadeira? Significaria concluirmos, no teste de significância estatística, que o tempo médio de serviço na população da Refinaria era menor do que 16,5 anos, quando na verdade era igual ou maior. A implicação dessa conclusão incorreta seria a inclusão da Refinaria no programa, o que, certamente não traria malefícios aos seus trabalhadores. Ao contrário, a participação no Programa poderia melhorar ainda mais a estabilidade. A consequência negativa seria a alocação dos recursos do Programa em empresa cuja situação não era tão ruim, enquanto outras em situação mais crítica não usufruiriam os possíveis benefícios do Programa.

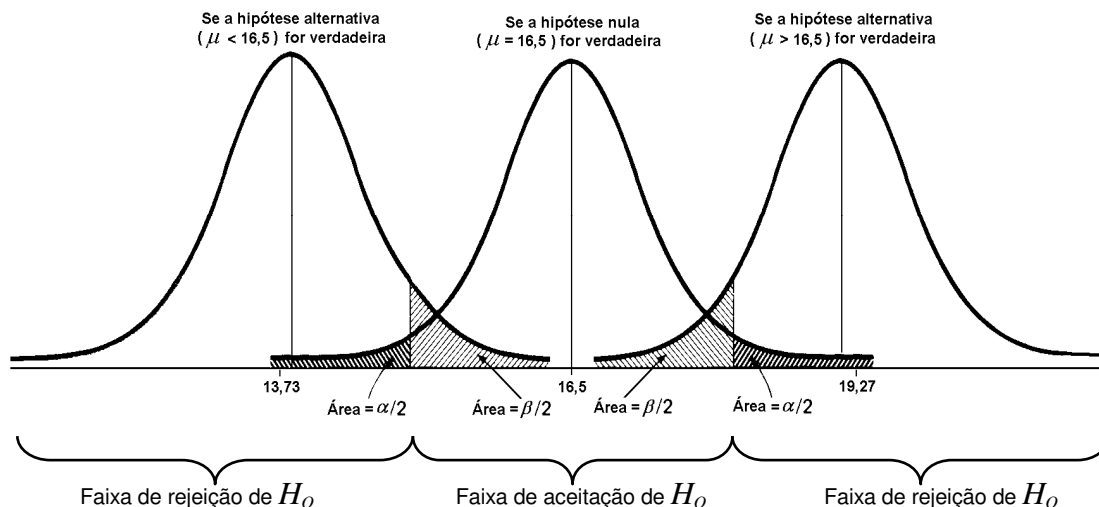
E o que significaria cometermos o erro tipo II, ou seja, aceitarmos uma H_0 falsa? Significaria concluirmos que a verdadeira média populacional era maior ou igual a 16,5 anos, quando na verdade era menor. A implicação desse erro seria a não inclusão da Refinaria no Programa. Isso prejudicaria muito os trabalhadores, porque perderiam a oportunidade de permanecer mais tempo naquele emprego, tornando-os mais expostos às agruras decorrentes do desemprego, consequências mais graves do que aquelas decorrentes do erro tipo I.

Pode-se argüir que se tivéssemos formulado as hipóteses $H_0 : \mu < 16,5$ e $H_A : \mu \geq 16,5$, o erro tipo I continuaria sendo o mais grave, mas seria forçado formularmos nossas hipóteses dessa maneira, porque, por definição, a hipótese nula deve estabelecer uma nulidade, neste caso a não inclusão no Programa, e a alternativa um efeito ou ação, a inclusão.

Nosso objetivo, portanto, é utilizar os menores valores possíveis de α e β . Um valor baixo de α implica em rejeitarmos uma hipótese nula verdadeira menos vezes. Um valor baixo de β resulta em aceitarmos uma hipótese nula falsa menos vezes. Essas ações são, contudo, antagônicas, ou seja, quando α aumenta, β diminui, e vice-versa, obrigando-nos a encontrar o melhor equilíbrio entre ambas.

Pense agora no seguinte: se β representa a probabilidade máxima admitida para o erro de aceitarmos uma H_0 falsa, o que o complemento de β , ou seja, $(1-\beta)$, representa? Se estabelecermos um β de 20,0% para o nosso teste, teremos que $1-\beta = 80,0\%$. Se β indica um erro, seu complemento representa um acerto, concorda? O valor $(1-\beta)$ indica a probabilidade do pesquisador corretamente não aceitar uma H_0 falsa. No nosso exemplo, se utilizarmos um β de 20,0%, haverá uma probabilidade de 80,0% de concluirmos corretamente que 13,73 anos e 16,5 anos são estatisticamente diferentes. A quantidade $(1-\beta)$ é denominada de poder do teste estatístico, porque expressa a capacidade (o poder) que o teste tem para detectar as diferenças estudadas.

Graficamente, podemos representar o erro tipo II da seguinte maneira:



Olhe inicialmente apenas para as distribuições da direita e esquerda. Todos os tempos médios englobados por essas curvas, incluindo os valores nas áreas sombreadas e rotuladas por “área= $\beta/2$ ”, poderiam ter sido obtidos em amostras retiradas de populações diferentes daquela da Refinaria, certo? Embora os valores nessas áreas pudessem pertencer a populações diferentes daquela estudada, se obtivéssemos tempos médios dentro dessas “áreas= $\beta/2$ ”, nós aceitaríamos H_0 (olhe agora também a distribuição do centro). Dessa maneira, poderíamos errar aceitando uma H_0 falsa, já que tempos médios de outras populações poderiam ser considerados por nós como tempos médios da população estudada. Assim, essas “áreas= $\beta/2$ ” representam o erro tipo II.

Resumindo:

Quando $\alpha = 5,0\%$, o pesquisador assume um risco de 5,0% de rejeitar uma H_0 verdadeira, mas a probabilidade dele não rejeitar uma H_0 verdadeira é de 95,0%.

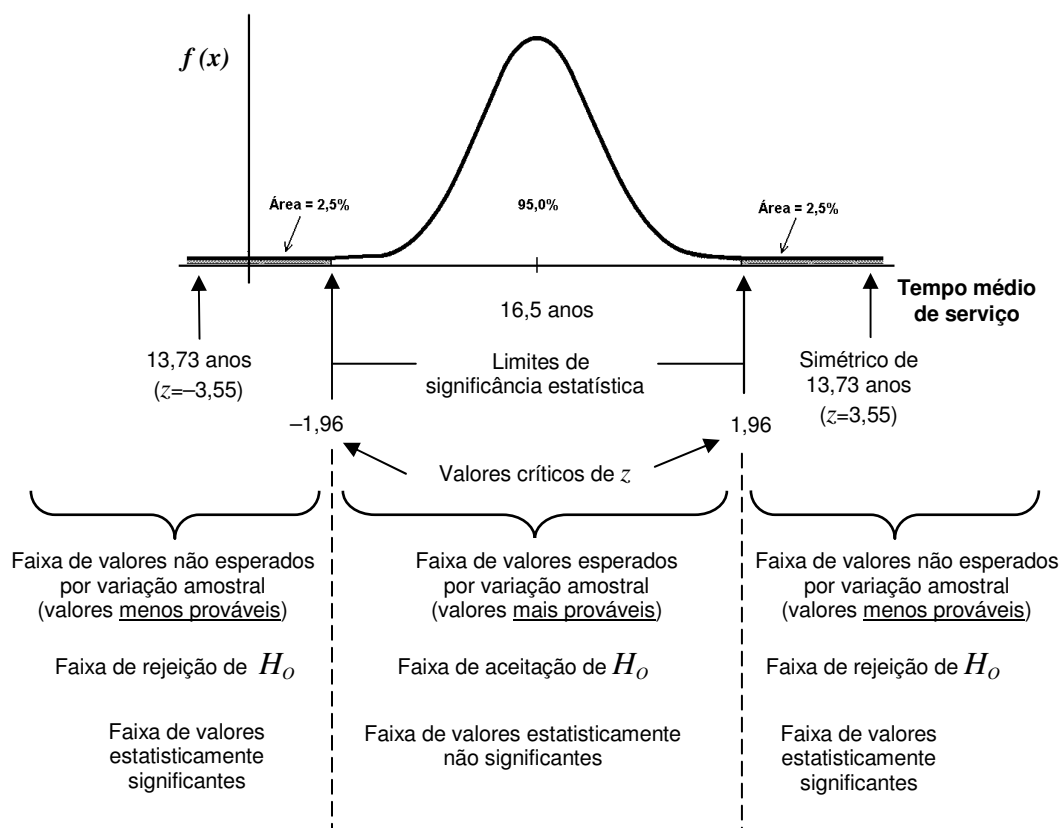
Quando $\beta = 20,0\%$, o pesquisador assume um risco de 20,0% de não rejeitar uma H_0 falsa, mas a probabilidade dele rejeitar uma H_0 falsa é de 80,0%.

Veja abaixo as conclusões equivalentes que podem ser feitas em um teste de hipóteses:

Estatisticamente significante = rejeição da hipótese nula = aceitação da hipótese alternativa = a verdadeira média pode ser considerada como estatisticamente diferente do valor populacional estabelecido na hipótese nula = variação amostral provavelmente não explica a diferença encontrada.

Estatisticamente não significativo = aceitação da hipótese nula = rejeição da hipótese alternativa = a verdadeira média não pode ser considerada como estatisticamente diferente do valor populacional estabelecido na hipótese nula = variação amostral provavelmente explica a diferença encontrada.

Essas equivalências podem ser vistas no diagrama da próxima página:



— Vimos que existem outras possibilidades de escolha para as hipóteses estatísticas a serem testadas. Como realizamos o teste com essas outras hipóteses?

— Até o momento, testamos as hipóteses $H_O : \mu = 16,5 \text{ anos}$ e $H_A : \mu \neq 16,5 \text{ anos}$. As outras hipóteses possíveis são:

- a) $H_O : \mu \leq 16,5 \text{ anos}$ e $H_A : \mu > 16,5 \text{ anos}$; e
- b) $H_O : \mu \geq 16,5 \text{ anos}$ e $H_A : \mu < 16,5 \text{ anos}$.

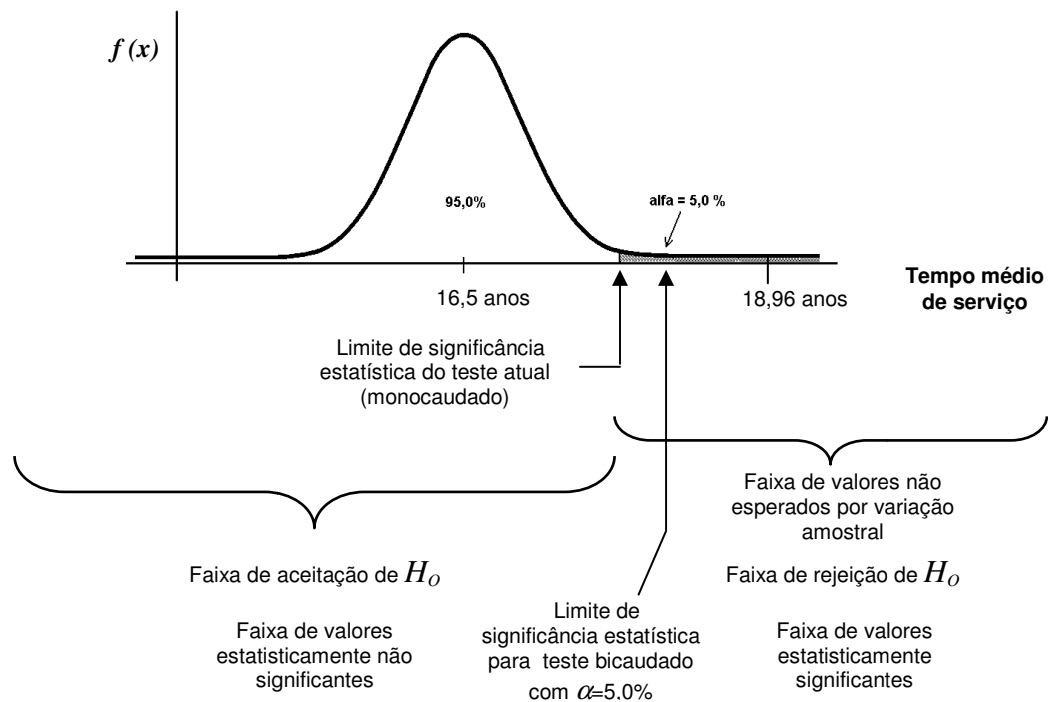
Escolheremos as hipóteses $H_O : \mu \leq 16,5 \text{ anos}$ e $H_A : \mu > 16,5 \text{ anos}$, quando já existirem evidências suficientes na literatura de que o tempo médio de serviço em Refinarias como a que estamos estudando é maior do que 16,5 anos. Podemos então, mantendo o nosso exemplo, estar interessados em testar se o verdadeiro tempo médio de serviço populacional é maior do que 16,5 anos.

Para o teste das hipóteses $H_O : \mu \leq 16,5 \text{ anos}$ e $H_A : \mu > 16,5 \text{ anos}$, continuam válidos todos os procedimentos utilizados no teste das hipóteses anteriormente testadas. A única diferença é que o teste é feito considerando-se apenas a cauda direita da distribuição normal, porque nossa suposição é a de que a média populacional é maior do que 16,5 anos. Isso faz com que esse teste seja chamado de **monocaudado**.

Nesta situação, ao calcularmos o tempo médio de serviço na única amostra estudada será muito

provável que encontremos uma média matematicamente maior do que 16,5 anos, já que com base na literatura, esperamos que a média populacional seja maior do que este valor. Suponha, então, que tenhamos obtido uma média de 18,96 anos na amostra investigada, e não 13,73 anos, como anteriormente. Nossa questão é sabermos se 18,96 anos é estatisticamente maior do que 16,5 anos.

Veja a situação atual no diagrama abaixo:



Trabalhamos nessa situação com apenas um limite de significância estatística, porque o nosso α será considerado integralmente na cauda direita da curva, já que só temos uma hipótese alternativa a ser testada, $H_A : \mu > 16,5$ anos. Com isso, a área sombreada na cauda direita da distribuição acima é maior do que a área sombreada na cauda direita da distribuição utilizada no teste bicaudado. Agora não precisamos mais dividir o nosso α nas duas pontas da curva e, por isso, temos de representar integralmente na cauda direita a área correspondente a 5,0%, e não apenas a 2,5%. E como a área agora é maior (5,0% em vez de 2,5%), o limite de significância estatística fica mais próximo da média populacional esperada, ficando, portanto, mais fácil de ser ultrapassado. É por isso que o teste monocaudado é considerado menos conservador (menos exigente) do que o bicaudado. No primeiro tipo de teste é mais fácil o valor observado de z ultrapassar o valor crítico de z , já que este é 1,65 e não 1,96.

As etapas do teste das hipóteses atuais são:

- 1ª) Definição do nível de significância: $\alpha = 0,05$;
- 2ª) Definição das hipóteses: $H_0 : \mu \leq 16,5$ anos e $H_A : \mu > 16,5$ anos;
- 3ª) Cálculo do valor de z :

$$z = \frac{\bar{x} - \mu_o}{\frac{\sigma}{\sqrt{n}}} = \frac{18,96 - 16,50}{\frac{5,53}{\sqrt{50}}} = \frac{2,46}{\frac{5,53}{7,07}} = \frac{2,46}{0,78} \cong 3,15.$$

4ª) Obtenção do valor- p :

Olhando na tabela Z , encontramos a probabilidade de obtermos um valor de z menor do que 3,15, isto é, $P(z < 3,15)$. Essa probabilidade é 0,9992 ou 99,92%. Logo, a probabilidade de obtermos um valor de z maior do que 3,15 (que é a probabilidade que nos interessa) será calculada por

$$P(z > 3,15) = 1 - P(z < 3,15) = 1 - 0,9992 = 0,0008 \text{ ou } 0,08\%;$$

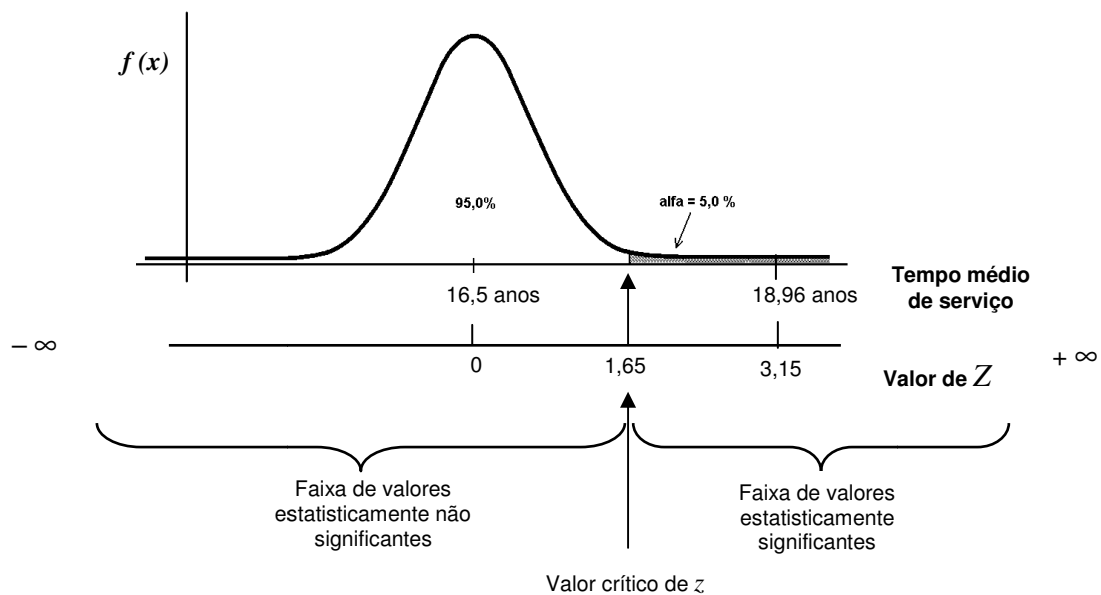
5ª) a) Comparação do valor- p ao valor de α e conclusão do teste:

Já que se trata de um teste monocaudado, não multiplicaremos o valor- p por dois, porque apenas uma cauda da curva está sendo considerada. Como $0,08\% < 5,0\%$, concluiremos que, embora valores maiores do que 18,96 anos pudessem ser tempos médios de serviço obtidos em amostras retiradas de uma população de trabalhadores cujo tempo médio de serviço fosse 16,5 anos com desvio-padrão de 5,53 anos, isso seria muito improvável (apenas 0,08% de probabilidade). Seria muito mais provável obtermos médias com essa magnitude em amostras retiradas de uma população cuja média fosse maior do que 16,5 anos. Assim, é muito provável que a verdadeira média populacional seja maior do que 16,5 anos. Outra interpretação é que 18,96 anos não é um dos valores esperados por variação amostral e, ainda outra, a de que esse tempo médio está localizado na área de rejeição da hipótese nula. Assim, assumindo que nossa pesquisa não apresenta vieses importantes, podemos rejeitar H_0 e aceitar H_A . Caso tivéssemos retirado numerosas amostras de mesmo tamanho $n = 50$, de uma população cujo tempo médio fosse 16,5 anos com desvio-padrão de 5,53 anos, seria muito improvável obtermos amostras com média entre 18,96 anos e valores mais altos, o que faria com que fosse muito provável obtermos esses valores em amostras retiradas de uma população cuja média fosse maior do que 16,5 anos, pois o resultado encontrado na única amostra estudada seria mais compatível com isto;

ou b) Comparação do valor observado de z ao valor crítico de z :

Quando o teste é bicaudado, vimos que os valores críticos de z são 1,96 e -1,96. Na situação atual (teste monocaudado na cauda direita da curva com α de 5,0%), o valor crítico de z é 1,645, ou aproximadamente 1,65. Para encontrar esse valor, procure no corpo na tabela Z (página 134) aquela célula que contém o valor 0,95, porque é esse valor que indica a área sob a curva entre $-\infty$ e z , que engloba 95% dos valores de Z . Na tabela Z utilizada neste livro não encontramos o valor exatamente igual a 0,95, mas 0,9505. Olhando os valores de Z na coluna indicadora e no cabeçalho da tabela, verifique que 0,9505 é a área entre $-\infty$ e $z = 1,65$. Se quiséssemos ser mais precisos, consultaríamos uma tabela mais completa e

veríamos que o valor mais preciso seria 1,645. Veja o limite crítico para a situação atual no diagrama abaixo:



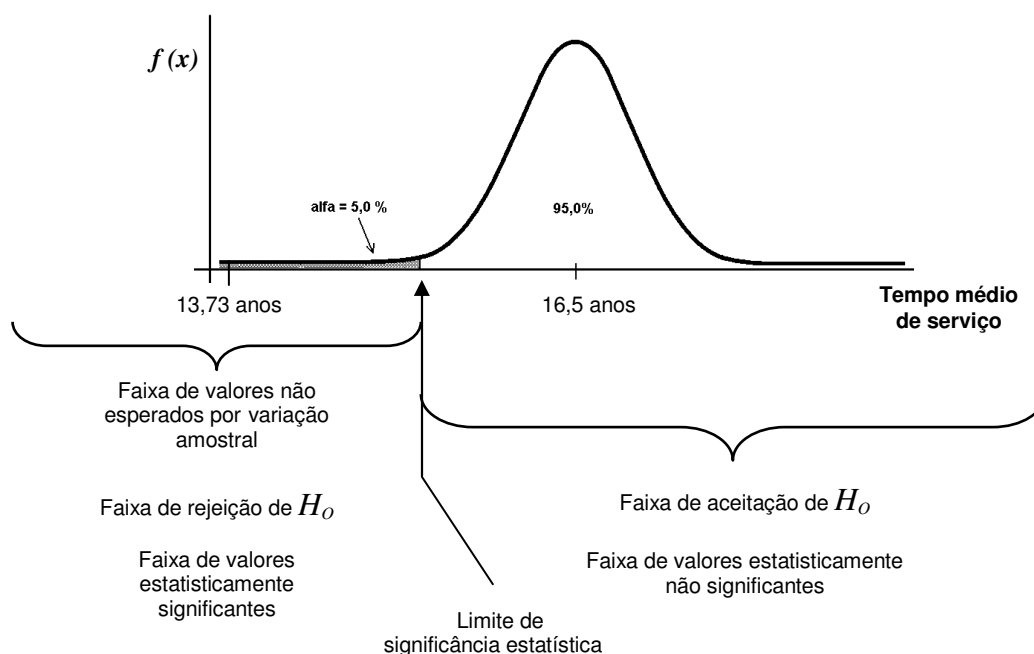
Como $3,15 > 1,65$, concluiremos que o valor de z correspondente a 18,96 anos está em uma localização muito extrema na cauda direita da curva, pois ultrapassa o valor crítico de z , situando-se bem à direita deste. Outra interpretação é a de que 18,96 anos é um valor não esperado por variação amostral. Então, assumindo que nossa pesquisa não apresente vieses importantes, concluiremos que o verdadeiro tempo médio de serviço na Refinaria deve ser maior do que 16,5 anos, porque 18,96 anos é um valor muito improvável de ser obtido em amostras retiradas de uma população cuja média fosse 16,5 anos. Isto é, o resultado obtido na única amostra retirada da população de trabalhadores da Refinaria é mais compatível com uma média populacional maior do que 16,5 anos. Rejeitamos H_0 e aceitamos H_A . Note que, como esperávamos, chegamos à mesma conclusão à qual chegamos na letra **a** da quinta etapa.

Escolheremos as hipóteses $H_0 : \mu \geq \mu_o$ e $H_A : \mu < \mu_o$ quando quisermos verificar se um determinado parâmetro de interesse é, do ponto de vista estatístico, significantemente menor do que determinado valor e já existirem evidências suficientes na literatura de que o esperado deve ser mesmo que um seja menor do que o outro.

Portanto, no nosso exemplo, poderíamos estar interessados em verificar se o tempo médio de serviço naquela Refinaria era menor do que 16,5 anos. Suponha, como já fizemos mais acima, que tenhamos obtido uma média de 13,73 anos na única amostra estudada.

Para o teste das hipóteses $H_0 : \mu \geq 16,5$ anos e $H_A : \mu < 16,5$ anos, continuam válidos todos os procedimentos utilizados nos testes das hipóteses anteriormente testadas. A única diferença é que o teste é feito considerando-se apenas a cauda esquerda da distribuição normal, porque nossa suposição é a de que a

verdadeira média é estatisticamente menor do que o valor esperado, 16,5 anos. Isso faz com que o teste seja também **monocaudado**. Veja essa situação no diagrama abaixo:

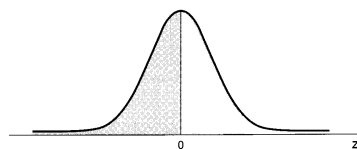


Vamos ter novamente apenas um limite de significância estatística, porque o nosso α será considerado integralmente na cauda esquerda da curva, já que só temos uma hipótese alternativa a ser testada, $H_A : \mu < 16,5$ anos. As etapas do teste das hipóteses atuais são:

- 1ª) Definição do nível de significância: $\alpha = 0,05$;
- 2ª) Definição das hipóteses: $H_O : \mu \geq 16,5$ anos e $H_A : \mu < 16,5$ anos;
- 3ª) Cálculo do valor de z :

$$z = \frac{\bar{x} - \mu_o}{\frac{\sigma}{\sqrt{n}}} = \frac{13,73 - 16,50}{\frac{5,53}{\sqrt{50}}} = \frac{-2,77}{\frac{5,53}{7,07}} = \frac{-2,77}{0,78} \cong -3,55;$$

4ª) Obtenção do valor- p : olhando na tabela Z encontramos a probabilidade de obtermos um valor de z entre menos infinito e $-3,55$, isto é, $P(Z < -3,55)$. Essa probabilidade é 0,0002 ou 0,02%. Observe que ao trabalharmos na cauda esquerda e com a parte de valores negativos da tabela Z , que vamos utilizar agora pela primeira vez (veja essa tabela a seguir), o valor encontrado já nos fornece diretamente a probabilidade de obtermos um valor de z entre menos infinito e $-3,55$, que é a probabilidade que nos interessa. Assim, como estamos trabalhando na cauda esquerda da curva e utilizando a tabela Z com valores negativos, você já sabe que não precisamos diminuir essa área de 1 ou 100,0%;

TABELA COM VALORES NEGATIVOS DE Z

z	-0,09	-0,08	-0,07	-0,06	-0,05	-0,04	-0,03	-0,02	-0,01	0,00
-3,80	0,0001	0,0001	0,0001	0,0001	0,0001	0,0001	0,0001	0,0001	0,0001	0,0001
-3,70	0,0001	0,0001	0,0001	0,0001	0,0001	0,0001	0,0001	0,0001	0,0001	0,0001
-3,60	0,0001	0,0001	0,0001	0,0001	0,0001	0,0001	0,0001	0,0001	0,0002	0,0002
-3,50	0,0002	0,0002	0,0002	0,0002	0,0002	0,0002	0,0002	0,0002	0,0002	0,0002
-3,40	0,0002	0,0003	0,0003	0,0003	0,0003	0,0003	0,0003	0,0003	0,0003	0,0003
-3,30	0,0003	0,0004	0,0004	0,0004	0,0004	0,0004	0,0004	0,0005	0,0005	0,0005
-3,20	0,0005	0,0005	0,0005	0,0006	0,0006	0,0006	0,0006	0,0006	0,0007	0,0007
-3,10	0,0007	0,0007	0,0008	0,0008	0,0008	0,0008	0,0009	0,0009	0,0009	0,0010
-3,00	0,0010	0,0010	0,0011	0,0011	0,0011	0,0012	0,0012	0,0013	0,0013	0,0013
-2,90	0,0014	0,0014	0,0015	0,0015	0,0016	0,0016	0,0017	0,0018	0,0018	0,0019
-2,80	0,0019	0,0020	0,0021	0,0021	0,0022	0,0023	0,0023	0,0024	0,0025	0,0026
-2,70	0,0026	0,0027	0,0028	0,0029	0,0030	0,0031	0,0032	0,0033	0,0034	0,0035
-2,60	0,0036	0,0037	0,0038	0,0039	0,0040	0,0041	0,0043	0,0044	0,0045	0,0047
-2,50	0,0048	0,0049	0,0051	0,0052	0,0054	0,0055	0,0057	0,0059	0,0060	0,0062
-2,40	0,0064	0,0066	0,0068	0,0069	0,0071	0,0073	0,0075	0,0078	0,0080	0,0082
-2,30	0,0084	0,0087	0,0089	0,0091	0,0094	0,0096	0,0099	0,0102	0,0104	0,0107
-2,20	0,0110	0,0113	0,0116	0,0119	0,0122	0,0125	0,0129	0,0132	0,0136	0,0139
-2,10	0,0143	0,0146	0,0150	0,0154	0,0158	0,0162	0,0166	0,0170	0,0174	0,0179
-2,00	0,0183	0,0188	0,0192	0,0197	0,0202	0,0207	0,0212	0,0217	0,0222	0,0228
-1,90	0,0233	0,0239	0,0244	0,0250	0,0256	0,0262	0,0268	0,0274	0,0281	0,0287
-1,80	0,0294	0,0301	0,0307	0,0314	0,0322	0,0329	0,0336	0,0344	0,0351	0,0359
-1,70	0,0367	0,0375	0,0384	0,0392	0,0401	0,0409	0,0418	0,0427	0,0436	0,0446
-1,60	0,0455	0,0465	0,0475	0,0485	0,0495	0,0505	0,0516	0,0526	0,0537	0,0548
-1,50	0,0559	0,0571	0,0582	0,0594	0,0606	0,0618	0,0630	0,0643	0,0655	0,0668
-1,40	0,0681	0,0694	0,0708	0,0721	0,0735	0,0749	0,0764	0,0778	0,0793	0,0808
-1,30	0,0823	0,0838	0,0853	0,0869	0,0885	0,0901	0,0918	0,0934	0,0951	0,0968
-1,20	0,0985	0,1003	0,1020	0,1038	0,1056	0,1075	0,1093	0,1112	0,1131	0,1151
-1,10	0,1170	0,1190	0,1210	0,1230	0,1251	0,1271	0,1292	0,1314	0,1335	0,1357
-1,00	0,1379	0,1401	0,1423	0,1446	0,1469	0,1492	0,1515	0,1539	0,1562	0,1587
-0,90	0,1611	0,1635	0,1660	0,1685	0,1711	0,1736	0,1762	0,1788	0,1814	0,1841
-0,80	0,1867	0,1894	0,1922	0,1949	0,1977	0,2005	0,2033	0,2061	0,2090	0,2119
-0,70	0,2148	0,2177	0,2206	0,2236	0,2266	0,2296	0,2327	0,2358	0,2389	0,2420
-0,60	0,2451	0,2483	0,2514	0,2546	0,2578	0,2611	0,2643	0,2676	0,2709	0,2743
-0,50	0,2776	0,2810	0,2843	0,2877	0,2911	0,2946	0,2981	0,3015	0,3050	0,3085
-0,40	0,3121	0,3156	0,3192	0,3228	0,3264	0,3300	0,3336	0,3372	0,3409	0,3446
-0,30	0,3483	0,3520	0,3557	0,3594	0,3632	0,3669	0,3707	0,3745	0,3783	0,3821
-0,20	0,3859	0,3897	0,3936	0,3974	0,4013	0,4052	0,4090	0,4129	0,4168	0,4207
-0,10	0,4247	0,4286	0,4325	0,4364	0,4404	0,4443	0,4483	0,4522	0,4562	0,4602
0,00	0,4641	0,4681	0,4721	0,4761	0,4801	0,4840	0,4880	0,4920	0,4960	0,5000

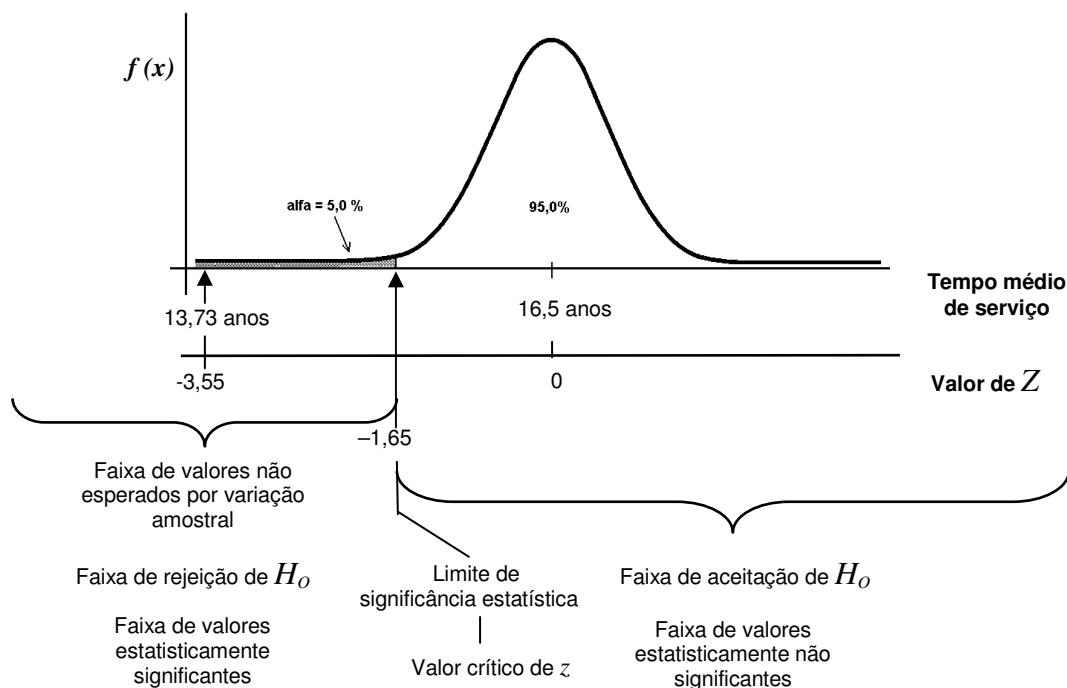
5ª) a) Comparação do valor- p ao valor de α e conclusão do teste:

Já que se trata de um teste monocaudado, não multiplicaremos o valor- p por dois, porque apenas uma cauda da curva está sendo considerada. Como $0,02\% < 5,0\%$, concluiremos que, muito provavelmente, a verdadeira média é menor do que 16,5 anos, ou seja, a verdadeira média é estatisticamente menor do que este valor. Outra interpretação é a de que 13,73 anos não é um dos valores que seriam esperados por variação amostral, e ainda outra a de que esse tempo médio estaria localizado na área de rejeição da hipótese nula. Assim, assumindo que a nossa pesquisa não apresente vieses importantes,

podemos rejeitar H_0 e aceitar H_A ;

ou b) Comparação do valor observado de z ao valor crítico de z :

Quando o teste é bicaudado com α de 5,0% vimos que os valores críticos de z são 1,96 e $-1,96$. Quando monocaudado à direita com α de 5,0%, o valor crítico utilizado é 1,65. Na situação atual (teste monocaudado à esquerda com α de 5,0%), o valor crítico de z é $-1,65$. Veja isso no diagrama abaixo:



Como $-3,55 < -1,65$, concluímos que o valor de z correspondente a 13,73 anos está em uma localização muito extrema na cauda esquerda da curva, pois ultrapassa o valor crítico de z , situando-se à esquerda deste. Outra interpretação é que 13,73 anos não é um valor esperado por variação amostral. Assim, assumindo que nossa pesquisa não apresenta vieses importantes, concluiremos que a verdadeira média populacional deve ser menor do que 16,5 anos, porque o valor obtido na única amostra estudada, retirada da população de trabalhadores da Refinaria, é mais compatível com isso. Rejeitaremos H_0 e aceitaremos H_A . Note que chegaremos à mesma conclusão à qual chegamos na letra **a** da quinta etapa deste teste.

▣ TERCEIRA PARTE ▣

— E o que é um intervalo de confiança?

— Bem lembrado! Como já escrevemos anteriormente neste capítulo, a inferência estatística também pode ser feita através de **cálculo de intervalo de confiança**. O intervalo de confiança é o intervalo que inclui os valores de tempo médio de serviço que poderiam ser aceitos como o verdadeiro tempo médio na população, denotado por μ . Como sua denominação indica, ao calcularmos esse intervalo teremos um certo grau de confiança sobre quais valores podem ou não ser o verdadeiro tempo médio na população, de onde a única amostra estudada foi retirada. Os tempos médios contidos dentro do intervalo são valores aceitos como

possíveis para μ , e aqueles fora do intervalo não são. Vamos explicar melhor isso.

Nosso objetivo é estimar μ , que teoricamente denota o tempo médio de serviço em uma população infinita de trabalhadores em refino de petróleo. Na prática, porém, queremos estimar a média de uma população finita (trabalhadores de determinada refinaria de petróleo), com base nos resultados obtidos em uma única amostra retirada dessa população. Vamos continuar utilizando a notação μ porque, embora toda a teoria subjacente aos nossos procedimentos tenha sido desenvolvida considerando-se populações infinitas, seus fundamentos são aplicáveis a populações finitas. Então, para estimarmos μ , coletamos o tempo de serviço de cada trabalhador em uma amostra aleatória retirada da população finita de trabalhadores da Refinaria pesquisada. O tempo médio obtido nessa amostra, $\bar{x} = 13,73$ anos, é chamada de **estimativa pontual** de μ . Mas, sabemos que outra amostra sorteada por nós poderia ser diferente daquela que realmente estudamos, devendo haver, portanto, em numerosas amostras que porventura tivéssemos investigado, uma variação dos resultados amostrais. Então, para utilizarmos \bar{x} como um estimador de μ , temos de levar em conta tal variabilidade. Como não podemos obter o valor exato de μ com base nos resultados de uma única amostra, o melhor que podemos fazer é calcular um intervalo que nos indique a provável magnitude de μ . Esse intervalo é chamado de **estimativa intervalar** de μ .

Aprendemos anteriormente que, se fizermos amostragem em uma população na qual a variável estudada esteja distribuída normalmente, a distribuição de médias amostrais em numerosas amostras que tenham sido estudadas será também distribuída normalmente, e terá média $\bar{x}_{\bar{x}}$, convergente para μ , e erro-padrão $EP_{\bar{x}}$, dado por σ/\sqrt{n} ou s/\sqrt{n} . Vimos também que, com base no teorema central do limite, mesmo que a distribuição na população não seja normal, se o tamanho da nossa amostra for suficientemente grande, a distribuição das médias amostrais em numerosas amostras será aproximadamente normal. Sendo normal, sabemos que 95% das possíveis médias amostrais dessa distribuição estariam localizadas na área sob a curva compreendida entre $\bar{x}_{\bar{x}}$, mais ou menos aproximadamente dois ($1,96 \cong 2,00$) erros-padrão.

Note, entretanto, que geralmente não conhecemos $\bar{x}_{\bar{x}}$, que é a média das médias amostrais. Só poderíamos obtê-la se estudássemos numerosas amostras retiradas de uma população finita, mas geralmente investigamos apenas uma amostra. Desse modo, não podemos utilizar a expressão $\bar{x}_{\bar{x}} \pm 1,96(EP_{\bar{x}})$ para calcular uma estimativa intervalar de μ . O valor do qual dispomos é a média $\bar{x} = 13,73$ anos, que é uma estimativa pontual de μ , obtida na única amostra estudada.

— Podemos usar a expressão $\bar{x} \pm 1,96(EP_{\bar{x}})$ para obter uma estimativa intervalar de μ ?

— Podemos. Para entender isso, suponha que desejemos calcular um intervalo que contenha μ com 95% de probabilidade. Isso nos torna interessados em uma área da curva normal delimitada por $-1,96$ e $1,96$, pois 95% dos valores de Z estão entre esses dois valores. Assim, essa probabilidade é expressa por

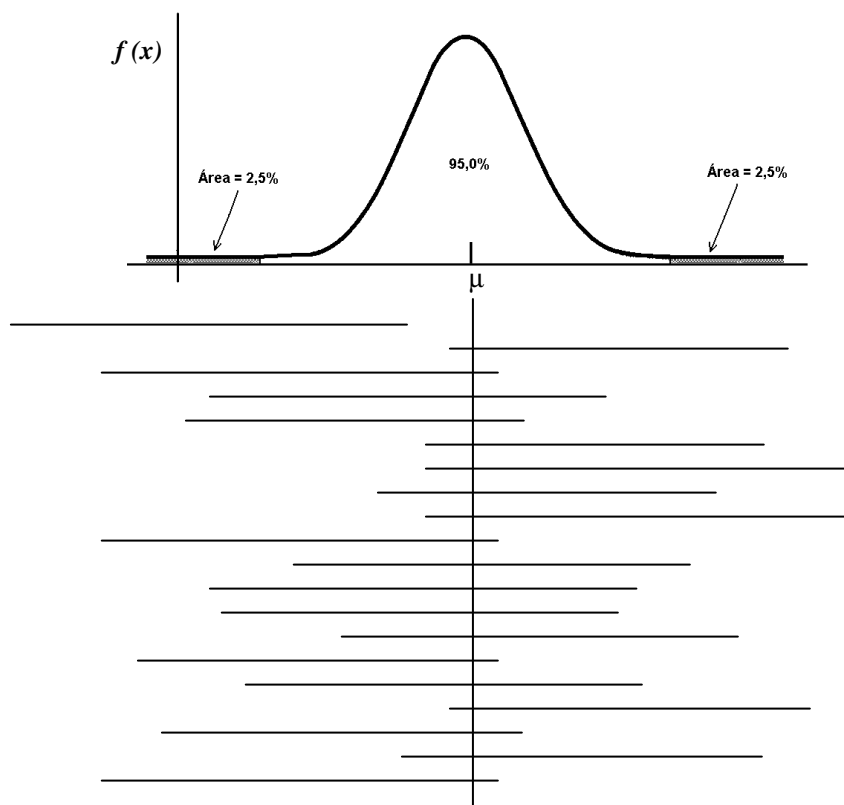
$P\{-1,96 < z < 1,96\} = 0,95$. Como $z = \frac{\bar{x} - \mu}{\sigma/\sqrt{n}}$, a expressão anterior pode também ser escrita como

$P\left\{-1,96 < \frac{\bar{x} - \mu}{\sigma/\sqrt{n}} < 1,96\right\} = 0,95$. No momento, estamos interessados em calcular um intervalo que

contenha μ com 95% de probabilidade. Logo, devemos resolver para μ a desigualdade contida na

expressão acima, o que resulta em $P\left\{-1,96\left(\sigma/\sqrt{n}\right) < \bar{x} - \mu < 1,96\left(\sigma/\sqrt{n}\right)\right\} = 0,95$, donde $P\left\{-\bar{x} - 1,96\left(\sigma/\sqrt{n}\right) < -\mu < -\bar{x} + 1,96\left(\sigma/\sqrt{n}\right)\right\} = 0,95$. Multiplicando todos os termos por -1 , obtemos $P\left\{\bar{x} + 1,96\left(\sigma/\sqrt{n}\right) > \mu > \bar{x} - 1,96\left(\sigma/\sqrt{n}\right)\right\} = 0,95$, que é o mesmo que $P\left\{\bar{x} - 1,96\left(\sigma/\sqrt{n}\right) < \mu < \bar{x} + 1,96\left(\sigma/\sqrt{n}\right)\right\} = 0,95$.

Observe que os limites inferior e superior do intervalo que contém μ com 95% de probabilidade são calculados com base em \bar{x} , que é a média obtida na única amostra estudada, mas sabemos que se estudássemos numerosas amostras as médias variariam entre si, de modo que, para cada amostra, obteríamos limites diferentes para esse intervalo. Assim, temos que considerar todos os possíveis intervalos, cujos limites seriam calculados por $\bar{x}_i \pm 1,96(EP_{\bar{x}_i})$, i variando de 1 a k , e k denotando o número de amostras. Pode ser demonstrado que, se numerosas amostras forem estudadas e todos esses intervalos forem realmente calculados, 95% deles conterão μ . Os diversos intervalos de valores mais prováveis de μ , cujos limites foram calculados com base nas diversas médias amostrais obtidas, estão representados graficamente a seguir:



Observe que as amplitudes dos intervalos não seriam iguais, porque raramente conhecemos σ , sendo este parâmetro substituído pelo desvio-padrão, s , obtido em cada amostra, no cálculo do erro-padrão, cujo valor, então, variaria de amostra para amostra.

Veja também que, propositadamente, na figura acima, fizemos com que 19 dos 20 (95%) intervalos incluíssem μ , enquanto apenas um (5%) não incluísse. Se 95% dos intervalos capturariam μ , ou seja, se 95 em cada 100 intervalos englobariam μ , podemos concluir que ao obtermos uma única amostra qualquer, o intervalo $\bar{x} \pm 1,96(EP_{\bar{x}})$, calculado com os resultados obtidos nesta amostra, terá uma probabilidade de 95% de englobar μ , concorda?

É essa propriedade desses intervalos, de muito provavelmente englobarem o valor de μ (que é o parâmetro que nos interessa estimar), que nos permite utilizar a expressão $\bar{x} \pm 1,96(EP_{\bar{x}})$ para obtenção de uma estimativa intervalar de μ .

— Mas, afinal de contas, o que esse intervalo indica?

— Como já vimos, se 95% dos intervalos com limites $\bar{x}_i \pm 1,96(EP_{\bar{x}_i})$, construídos a partir de numerosas médias amostrais, incluiriam a verdadeira média populacional μ , podemos concluir que qualquer um desses intervalos tem 95% de probabilidade de incluir μ . Isso equivale a dizer que estamos 95% confiantes de que o intervalo calculado na única amostra investigada, $\bar{x} \pm 1,96(EP_{\bar{x}})$, contém a média populacional. Por isso, esse intervalo é denominado “intervalo de confiança”. Qualquer valor dentro desse intervalo será aceito como possível valor da verdadeira média populacional.

O cálculo de todo e qualquer intervalo de confiança baseado no modelo normal contém os mesmos elementos básicos que são mostrados na fórmula abaixo:

$$IC(95\%) = \text{estimador} \pm z_{(1-\alpha/2)} (EP),$$

onde IC é a notação para intervalo de confiança; 95% é o quanto de confiança desejamos; *estimador* é o indicador estatístico utilizado para estimar o parâmetro de interesse; $z_{(1-\alpha/2)}$ é o valor de z correspondente ao grau de confiança desejado, sendo chamado por alguns estatísticos de coeficiente de confiança (note que $1 - \alpha/2 = 1 - 0,05/2 = 1 - 0,025 = 0,975 = 97,5\%$ é a área entre menos infinito e o valor $z = 1,96$ correspondente ao nível de confiança de 95%, pois o alfa é dividido nas duas caudas da curva); e EP é o erro-padrão do estimador.

No nosso exemplo, o estimador é uma média aritmética (tempo médio de serviço em uma amostra), utilizada para estimar a verdadeira média em uma população finita de trabalhadores de uma refinaria de petróleo.

Especificando a fórmula acima para o estimador de uma média, temos:

$$IC(95\%) = \bar{x} \pm 1,96(EP_{\bar{x}}),$$

que é lida como “o intervalo de 95% de confiança de valores aceitáveis para a média populacional é igual à média aritmética obtida na única amostra, mais ou menos 1,96 erro-padrão desta média aritmética”.

Observe atentamente o que é feito ao utilizarmos a expressão acima: consideramos o tempo médio obtido na única amostra estudada, \bar{x} , e somamos e diminuímos deste uma certa quantidade que corresponde a cerca de duas vezes o valor do seu erro-padrão, $EP_{\bar{x}}$. Ao somarmos obtemos o limite superior do intervalo de confiança, e ao diminuirmos encontramos seu limite inferior.

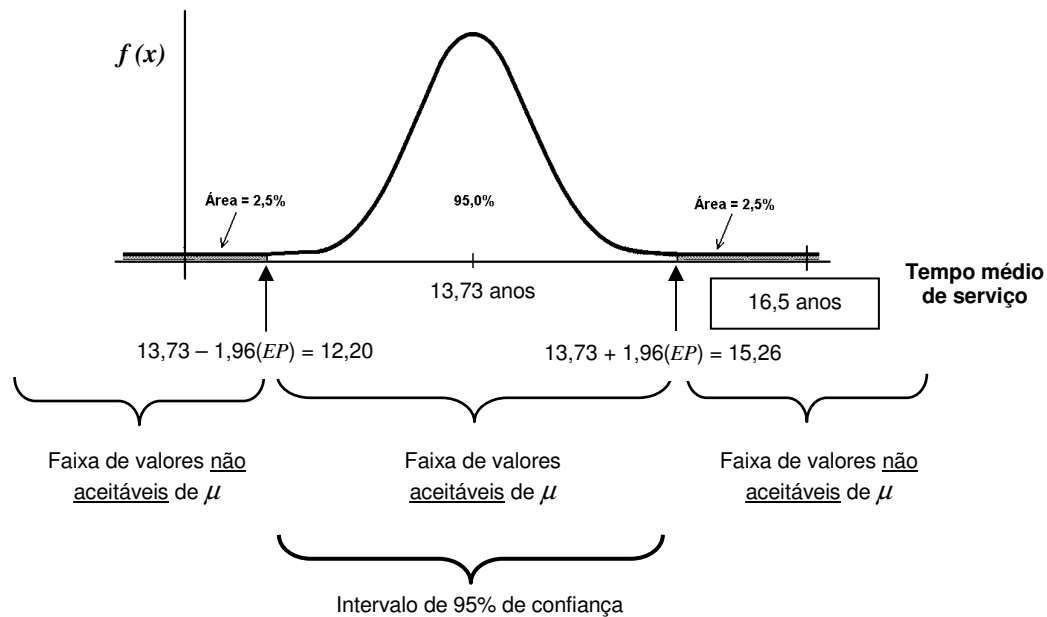
Vamos fazer esses cálculos para o nosso exemplo:

$$\begin{aligned}
 IC(95\%) &= \bar{x} \pm z_{(1-\alpha/2)} (EP_{\bar{x}}) = \bar{x} \pm z_{(1-\alpha/2)} \left(\sqrt{\frac{\sigma^2}{n}} \right) = \bar{x} \pm z_{(1-\alpha/2)} \left(\frac{\sigma}{\sqrt{n}} \right) = 13,73 \pm 1,96 \left(\frac{5,53}{\sqrt{50}} \right) = \\
 &= 13,73 \pm 1,96 \left(\frac{5,53}{7,07} \right) = 13,73 \pm (1,96)(0,78) = 13,73 \pm 1,53 \cong (12,20 \text{ a } 15,26).
 \end{aligned}$$

Obtivemos um intervalo de 95% de confiança que varia de 12,20 a 15,26 anos. Estamos, portanto, 95% confiantes de que o intervalo calculado com base na média obtida na única amostra investigada contenha a verdadeira média populacional, porque sabemos que tal intervalo tem uma probabilidade de 95% de englobar μ .

Podemos também verificar, com pequena probabilidade de erro, se um determinado tempo médio, 16,5 anos, p. ex., é o verdadeiro tempo médio populacional. Para isso, verificamos se o valor considerado, 16,5 anos, está localizado dentro ou fora do intervalo de confiança calculado. Se estiver dentro, concluiremos que 16,5 anos é um dos tempos médios que podem ser aceitos como a verdadeira média populacional e que, muito provavelmente, o valor avaliado foi diferente daquele obtido na única amostra estudada, apenas em consequência da simples variação de resultados amostrais. Se estiver fora, concluiremos que 16,5 anos não é um dos tempos médios que podem ser aceitos como a verdadeira média populacional e que, mais provavelmente, o valor avaliado foi diferente daquele obtido na única amostra estudada, porque esta deve pertencer a uma população com média diferente de 16,5 anos. Temos obtido uma média amostral de 13,73 anos não é estatisticamente compatível com uma média populacional de 16,5 anos.

Verifique no diagrama da próxima página que o valor 16,5 anos está localizado fora do intervalo de confiança, o que nos leva a concluir que esse valor não é um dos valores mais prováveis de μ , evidenciando que, caso não haja vieses importantes na pesquisa, 16,5 anos não deve ser o verdadeiro tempo médio de serviço na Refinaria. Foi a mesma conclusão à qual chegamos, quando realizamos o teste de hipóteses bicaudado, lembra-se?



Você deve ter notado que os fundamentos para se fazer inferência estatística utilizando-se o cálculo de intervalo de confiança são muito semelhantes aos de um teste de hipóteses. Então, lembre-se de que esses são métodos um pouco diferentes para se fazer a mesma coisa. As conclusões deles devem ser semelhantes, como já verificamos em nosso exemplo.

— **Por que no cálculo do intervalo de confiança utilizamos a média amostral como referência e não a média populacional esperada?**

— Se você percebeu isso, é porque está conseguindo acompanhar bem nossa apresentação. Ao colocarmos 13,73 anos no centro da distribuição normal utilizada como modelo para inferência, estamos considerando essa média amostral como referência, enquanto no teste de hipóteses usamos 16,5 anos, que era a média esperada para a população. Isto ocorre por dois motivos: a) lembre-se de que todo teste de hipóteses é feito sob o pressuposto de que a hipótese nula é verdadeira, sendo que, no nosso exemplo, isso nos levou a assumirmos que $\mu = 16,5$ anos. Como no cálculo do intervalo de confiança não há hipóteses a serem testadas, não faz sentido assumirmos que uma hipótese nula seja verdadeira; b) quase sempre não conhecemos a média e o desvio-padrão populacionais e, como no cálculo do intervalo de confiança não testamos $H_0: \mu = 16,5$ anos, teoricamente não temos nem mesmo um possível valor da média populacional sendo testado, o que nos leva a utilizar a média amostral como referência; c) o cálculo de intervalo de confiança é feito sob o fundamento estatístico de que seus limites são obtidos com base na média da única amostra estudada, porque tal intervalo tem 95% de probabilidade de capturar μ .

Em outros livros, você aprenderá ou revisará alguns indicadores de associação estatística entre eventos (razão de prevalências, risco relativo, razão de chances, etc.), e verá que há certas nuances para a interpretação correta dos seus intervalos de confiança, que não serão apresentadas aqui.

Observe que nossa apresentação sobre cálculo de intervalo de confiança utilizou as duas caudas da distribuição normal. Essa é a maneira mais utilizada, embora seja possível calcularmos intervalos de confiança monocaudados. Na situação equivalente ao teste na cauda esquerda da distribuição, o intervalo

variaria entre menos infinito e seu limite superior (que seria dado por $\bar{x} + 1,65(EP_{\bar{x}})$), se $\alpha = 5,0\%$. Na situação equivalente ao teste na cauda direita da distribuição, o intervalo variaria entre seu limite inferior (que seria dado por $\bar{x} - 1,65(EP_{\bar{x}})$) e mais infinito, se $\alpha = 5,0\%$. Nossas conclusões seriam feitas do mesmo modo visto acima, ou seja, verificando se determinado valor se encontraria dentro ou fora desses intervalos. No capítulo 13, nas páginas 239 e 240, calculamos um intervalo de confiança monocaudado.

— Tanto faz usarmos teste de hipóteses ou intervalo de confiança?

— Há muitos artigos na literatura discutindo esse aspecto, não havendo um consenso entre os estatísticos. Muitos consideram o cálculo de intervalo de confiança mais informativo, porque nos indica um espectro de valores dentro do qual deve estar o verdadeiro valor populacional, enquanto o valor- p nos indica apenas qual a probabilidade de obtermos valores maiores ou menores a determinado valor. Mas só o valor- p nos fornece a probabilidade específica de cometermos o erro tipo I naquele teste. Uma boa idéia é apresentarmos sempre em nossos artigos tanto o valor- p como o intervalo de confiança.

— E se o desvio-padrão populacional for desconhecido?

— Essa é a situação mais freqüente e diante da qual temos de substituir o desvio-padrão populacional, σ , pelo desvio-padrão obtido na única amostra estudada, s . Foi demonstrado que se o tamanho da amostra for suficientemente grande ($n \geq 30$), s será um estimador válido de σ .

No nosso exemplo, no qual $s = 5,23$ anos, teríamos:

$$IC(95\%) = \bar{x} \pm z_{(1-\alpha/2)} (EP_{\bar{x}}) = \bar{x} \pm z_{(1-\alpha/2)} \left(\sqrt{\frac{s^2}{n}} \right) = \bar{x} \pm z_{(1-\alpha/2)} \left(\frac{s}{\sqrt{n}} \right) = 13,73 \pm 1,96 \left(\frac{5,23}{\sqrt{50}} \right) =$$

$$= 13,73 \pm 1,96 \left(\frac{5,23}{7,07} \right) = 13,73 \pm 1,96 (0,74) = 13,73 \pm 1,45 = (12,28 \text{ a } 15,18).$$

Como 16,5 anos estaria fora desse intervalo, nossa conclusão seria a mesma à qual chegamos anteriormente.

Veja que o intervalo obtido não diferiria muito daquele calculado utilizando o desvio-padrão populacional, $\sigma = 5,53$ anos, embora essa pequena diferença pudesse nos levar a uma conclusão diferente, se o valor avaliado se localizasse muito próximo a um dos limites do intervalo. Isso ocorreria, por exemplo, se estivéssemos verificando se $\mu = 15,22$ anos, porque com base nesse segundo intervalo concluiríamos que 15,22 anos não seria um dos valores prováveis de μ , pois estaria fora do intervalo; mas se considerássemos o primeiro intervalo concluiríamos o contrário, pois 15,22 anos estaria dentro do intervalo. Essa é uma das razões pelas quais os estatísticos experientes consideram resultados limítrofes como estatisticamente significantes.

— A substituição de σ por s também é válida no teste de hipóteses?

— Sim, e já vimos isso anteriormente neste capítulo. Quando em um teste de hipóteses não

conhecemos o desvio-padrão populacional, nós o substituímos pelo desvio-padrão amostral, tal como acabamos de fazer acima para o cálculo do intervalo de confiança. Mas, lembre-se de que, quando em um teste de hipóteses ou ao calcularmos um intervalo de confiança, não conhecermos o desvio-padrão populacional, a distribuição apropriada não será a Z , e sim a T , como discutiremos no próximo capítulo. Lá você verá também que, se o n for suficientemente grande, essas duas distribuições se assemelharão muito.

Mais uma vez, lembre-se de que os procedimentos para inferência estatística nos auxiliam apenas a avaliar se o(s) resultado(s) obtido(s) pode(m) ser explicado(s) pela variação de resultados em diferentes amostras. Não nos indicam absolutamente nada sobre vieses importantes ocorridos no estudo, nem os corrigem.

Outra coisa: não fique desesperado(a) e sem conseguir dormir toda vez que seus resultados não forem estatisticamente significantes. Resultados não estatisticamente significantes são tão importantes quanto os estatisticamente significantes, pois nos ajudam a verificar que, p. ex., uma nova droga não é mais eficaz do que uma droga convencional. Sabermos isso é tão importante quanto estabelecermos que uma nova droga é mais eficaz. Você ainda tem alguma dúvida disso?

Precisamos também, urgentemente, convencer os editores e revisores de revistas científicas, de que devem publicar trabalhos cientificamente válidos e que chegaram a resultados “negativos”, para evitarem o “viés de publicação”, erro muito grave que dificulta ou impede o acesso dos pesquisadores a artigos que deixam de ser publicados porque seus resultados não foram estatisticamente significantes.

Este é o capítulo mais importante deste livro, pois explica detalhadamente o raciocínio que fundamenta qualquer procedimento de inferência estatística não-bayesiana, seja este paramétrico ou não-paramétrico. Os princípios que você aprendeu ou revisou neste capítulo lhe serão úteis na leitura sobre qualquer técnica estatística neste ou em outros livros, na avaliação crítica de artigos publicados na literatura e na realização adequada de suas próprias pesquisas.

Como já dissemos, no próximo capítulo abordaremos um outro teste de significância estatística denominado “teste t ”, muito semelhante ao teste z que acabamos de ver. Lá detalharemos também quando utilizar um ou outro.

CAPÍTULO 11

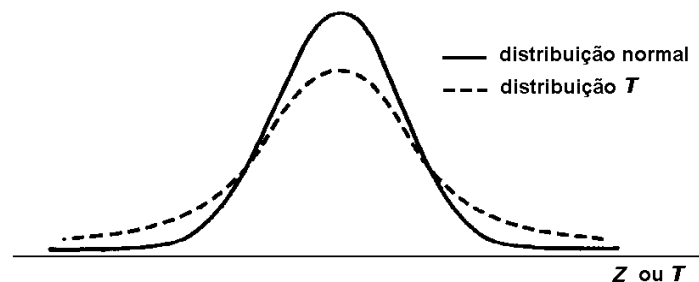
- Quando devemos aplicar o teste z ou o t ?
 - Quais as diferenças entre as distribuições Z e T ?
 - Como realizamos o teste t ?
 - Como calculamos intervalos de confiança com valores de T ?
 - Existem outras aplicações dos testes z e t ?
-



— Quando devemos aplicar o teste z ou o t ?

— No capítulo anterior, abordamos a aplicação do teste z para inferência estatística sobre uma média. No exemplo discutido, assumimos que a distribuição populacional da variável testada era normal, que o n era grande ($n \geq 30$), e que o desvio-padrão populacional, σ , era conhecido, lembra-se? Quando alguns desses pressupostos deixarem de ser atendidos, não poderemos utilizar o teste z para realizarmos inferência estatística.

Para atender a essa situação, um estatístico chamado Gosset (*W S Gosset. The probable error of a mean. Biometrika, 6:1-25, 1908*), apresentou uma alternativa à distribuição Z , que foi chamada de **distribuição T** . Ela tem algumas semelhanças, mas também diferenças em relação à curva Z . A distribuição T é simétrica em torno de zero e seus valores também variam de $-\infty$ a $+\infty$. Contudo, a variância dos valores de T não é 1 e sim maior do que 1, embora se aproxime de 1 à medida que o tamanho da amostra aumenta. Outra diferença é que há uma distribuição T específica para cada tamanho de amostra, sendo este expresso em termos do número de graus de liberdade. Este número é dado por $n - 1$, já que usaremos essa distribuição para inferência sobre uma média, e sabemos que perdemos um grau de liberdade ao calcularmos uma média. A distribuição T aproxima-se da distribuição Z à medida que $n - 1$ se aproxima de infinito, ou seja, quando o n é muito grande. No diagrama abaixo você pode ver que a distribuição T tem um ápice menos pontiagudo e tem caudas mais altas:



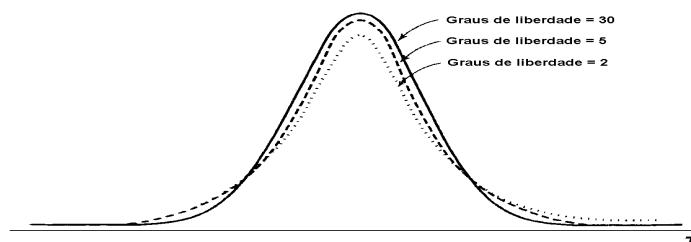
— Por que tenho de trocar a distribuição quando alguns dos pressupostos mencionados acima não são atendidos?

— Porque, por exemplo, sendo o σ desconhecido, ao substituirmos esse parâmetro por s , que é o desvio-padrão obtido na única amostra estudada, evidentemente obteremos uma variabilidade dos valores em torno da média um pouco diferente, já que esta será calculada por s/\sqrt{n} e não por σ/\sqrt{n} . Como consequência disso, o formato da distribuição será um pouco diferente também, implicando em áreas sob essa curva (a serem utilizadas para inferência estatística), diferentes daquelas encontradas na distribuição

normal. Por isso, temos de trabalhar com a distribuição T , que é apropriada para essa situação.

Quanto menor o tamanho da amostra maior será esta diferença e quanto maior esse tamanho menor a diferença, porque quanto maior a amostra mais suas estatísticas se assemelharão aos parâmetros populacionais, e vice-versa.

O diagrama a seguir apresenta distribuições T para alguns tamanhos de amostra: $n = 31$, $n = 6$ e $n = 3$:



Na maioria das vezes o pesquisador desconhece o desvio-padrão populacional (e certamente a média também), porque geralmente é inviável estudar-se toda a população. Quando pudermos assumir que a distribuição populacional da variável testada é normal, σ for desconhecido, e o n pequeno, σ será substituído por s . Neste caso, utilizaremos a distribuição T , e faremos a inferência estatística com base em um valor de T , que será calculado por $t = (\bar{x} - \mu) / (s / \sqrt{n})$.

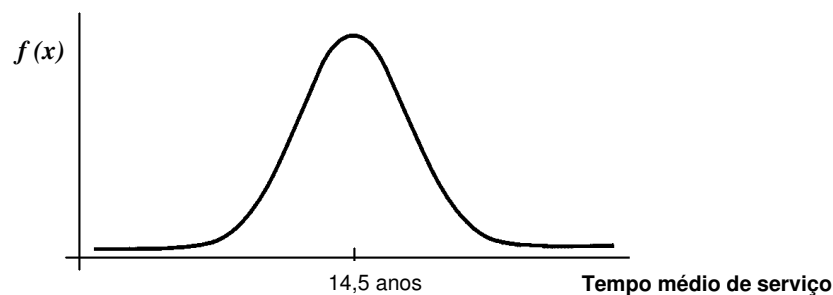
Considere agora que possamos assumir que a distribuição populacional da variável testada seja normal, que o σ seja desconhecido e que o n seja suficientemente grande. Nesse caso, vamos calcular um valor de $t = (\bar{x} - \mu) / (s / \sqrt{n})$, mas esse valor será muito semelhante ao valor de $z = (\bar{x} - \mu) / (\sigma / \sqrt{n})$, porque o n é grande. Olharmos a curva T ou Z vai nos levar à mesma conclusão. Isso porque, na situação que estamos considerando, o desconhecimento de σ nos levaria necessariamente à realização do teste t , mas como ao mesmo tempo o n seria grande, a distribuição T se aproximaria muito da distribuição Z .

Quando a distribuição populacional da variável testada puder ser assumida como normal, σ for desconhecido, mas o n for suficientemente grande, poderemos aplicar o teste z ou o t .

Vamos retornar ao nosso exemplo sobre o tempo de serviço em uma Refinaria de Petróleo. Teremos que fazer algumas modificações, porque agora desconhecemos o valor de σ , que será substituído pelo valor de s . Já vimos que $s = 5,23$ anos. Esse valor, portanto, substituirá o valor 5,53 anos que era o valor conhecido de σ . Considere que o nosso n seja de 51 trabalhadores. Mais adiante explicaremos porque modificamos o n de 50 para 51. Encontramo-nos, então, em uma situação na qual estamos assumindo (baseados em nossa experiência com esse tipo de variável e/ou nos poucos estudos anteriores realizados

sobre essa variável em populações de trabalhadores) que o tempo de serviço é uma variável distribuída normalmente na população de 1.000 trabalhadores da Refinaria; o n é grande, pois $51 > 30$; mas, o desvio-padrão populacional não é conhecido. Nesse caso, já vimos que podemos utilizar como referência para o nosso teste estatístico a distribuição Z ou T . Como ainda temos a opção de usar a distribuição Z , vamos fazer primeiramente um teste z .

Novamente, nosso objetivo é termos uma idéia de qual deve ser o verdadeiro tempo médio de serviço (μ) na Refinaria, levando em conta o resultado obtido em uma única amostra. Por isso, temos que escolher possíveis valores de μ , testando um de cada vez. Vamos supor que desejamos inicialmente verificar se o tempo médio populacional é 14,5 anos. A distribuição a ser considerada, que é traçada tendo como referência a média esperada para a população, é apresentada abaixo:



Do mesmo modo e pelo mesmo motivo apresentado no capítulo anterior (páginas 159 e 160), assumimos que o tempo médio estabelecido na hipótese nula, que é aquele esperado na população, seria a média dos tempos médios de serviço, $\bar{x}_{\bar{x}}$, caso tivéssemos estudado infinitas amostras de tamanho n , retiradas aleatoriamente de uma população infinita de trabalhadores da Refinaria. A distribuição acima nos permitirá avaliar quais valores de tempo médio de serviço, abaixo ou acima dessa média, admitiremos como decorrentes apenas de variação de resultados amostrais e, conseqüentemente, quais valores abaixo ou acima dessa média consideraremos como estatisticamente diferentes de 14,5 anos. Nos permitirá avaliar se o valor obtido na amostra investigada, 13,73 anos, desvia-se suficientemente ou não do valor escolhido para ser testado, 14,5 anos, a ponto de nos permitir concluir, com pequena probabilidade de erro, se tais valores diferem ou não estatisticamente. Tentaremos responder à seguinte pergunta:

Supondo que o tempo médio de serviço na Refinaria seja 14,5 anos, qual a probabilidade (valor- p) do tempo médio de serviço obtido na amostra estudada ser maior do que 13,73 anos ou menor do que o seu simétrico na distribuição?

Na situação atual, as etapas de um teste z bicaudado são:

1ª) Definição do nível de significância: $\alpha = 0,05$;

2ª) Definição das hipóteses: $H_0 : \mu = 14,5$ anos e $H_A : \mu \neq 14,5$ anos;

Em última instância, estamos avaliando se uma média populacional de 14,5 anos é compatível com a observação de uma média amostral de 13,73 anos;

3ª) Cálculo do valor de $t \cong z$:

$$t \cong z = \frac{\bar{x} - \mu_o}{\frac{s}{\sqrt{n}}} = \frac{13,73 - 14,50}{\frac{5,23}{\sqrt{51}}} = \frac{-0,77}{\frac{5,23}{7,14}} = \frac{-0,77}{0,73} \cong -1,05;$$

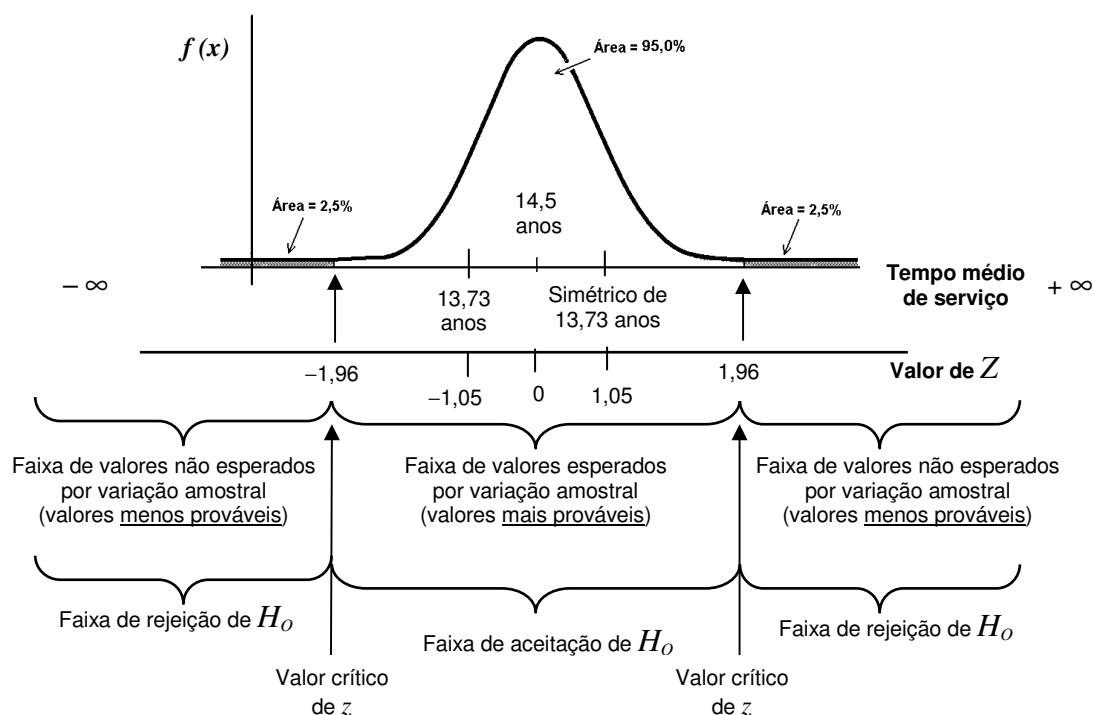
onde, novamente, \bar{x} indica a média obtida na única amostra investigada, μ_o a média estipulada na hipótese nula, e s/\sqrt{n} o erro-padrão das médias amostrais. Como na situação atual o desvio-padrão populacional é desconhecido, no denominador substituímos σ/\sqrt{n} por s/\sqrt{n} , porque estamos utilizando o desvio-padrão amostral, s , que é conhecido, no lugar do desvio-padrão populacional desconhecido, σ . Além disso, como o n é grande, podemos assumir que s é um bom estimador de σ , e o teste z apresentará resultados muito semelhantes àqueles do teste t , que será mostrado mais adiante;

4ª) Obtenção do valor- p : consultamos a parte negativa da tabela da curva normal padrão (página 176), para encontrarmos a probabilidade de obtermos um valor de z entre menos infinito e $-1,05$, ou seja, $P(Z < -1,05)$. Essa probabilidade é de 0,1469 ou 14,69%. Como o teste é bicaudado, temos que multiplicar a probabilidade por dois, porque estamos também testando a possibilidade de 13,73 anos ser estatisticamente maior do que 14,5 anos, sendo essa parte do teste feita considerando a cauda direita da distribuição. Então o valor- p final é igual a $(14,69\%)(2) = 29,38\%$. Veja no diagrama da próxima página a situação encontrada neste exemplo;

5ª) a) Comparação do valor- p ao valor de α e conclusão do teste:

Como $29,38\% > 5,0\%$, concluiremos que, embora 13,73 anos possa ser um tempo médio de serviço obtido em uma amostra retirada de uma população com tempo médio diferente de 14,5 anos, isso seria muito improvável, sendo muito mais provável obter uma média de 13,73 anos em amostras retiradas de uma população de trabalhadores com tempo médio igual a 14,5 anos. Outra interpretação é a de que 13,73 anos é um valor esperado por variação amostral, ou que é um valor localizado na área de aceitação da

hipótese nula, e ainda outra, que a verdadeira média é estatisticamente igual a esse valor. Assim, se a pesquisa realizada não apresentar vieses importantes, poderemos aceitar H_0 e rejeitar H_A , concluindo que a média encontrada na amostra não se afasta muito da média estabelecida pela hipótese nula, sendo muito provável obtê-la, caso estudássemos numerosas amostras. Tais achados estatísticos nos permitem afirmar que obter uma média amostral de 13,73 anos, como a que foi obtida, é compatível com uma média populacional de 14,5 anos. Então, concluiremos, finalmente, que 14,5 anos pode ser a verdadeira média populacional;



ou

b) Comparação do valor observado de z ao valor crítico de z :

Como $-1,05 > -1,96$ e, conseqüentemente, $1,05 < 1,96$, concluiremos que o valor de z correspondente a 13,73 anos está em uma localização muito central da curva, não ultrapassando o valor crítico de z . Na prática, é suficiente fazermos apenas a primeira comparação, porque a distribuição é simétrica, sendo que o que for observado em uma cauda será o mesmo a ser encontrado na outra. Outra interpretação é a de que 13,73 anos é um valor esperado por variação amostral e, ainda outra, a de que esse valor está localizado na área de aceitação da hipótese nula. Assim, se a pesquisa realizada não apresenta vieses importantes, chegaremos à mesma conclusão à qual chegamos na letra **a** da quinta etapa: 14,5 anos

pode ser a verdadeira média populacional.

Outra opção ainda seria calcularmos o intervalo de 95% de confiança, que seria dado por

$$IC(95\%) = \bar{x} \pm z_{(1-\alpha/2)} \sqrt{\frac{s^2}{n}}$$

Intervalo de 95% de confiança

Média obtida no estudo

Valor de z correspondente ao percentil 97,5

Erro-padrão das médias caso tivéssemos feito vários estudos

Colocando na expressão acima os valores do nosso exemplo, teríamos:

$$IC(95\%) = \bar{x} \pm z_{(1-\alpha/2)} \sqrt{\frac{s^2}{n}} = \bar{x} \pm z_{0,975} \frac{s}{\sqrt{n}} = 13,73 \pm 1,96 \left(\frac{5,23}{\sqrt{51}} \right) =$$

$$= 13,73 \pm 1,96 \left(\frac{5,23}{7,14} \right) = 13,73 \pm (1,96)(0,73) = 13,73 \pm 1,43 = (12,30 \text{ a } 15,16).$$

Esses resultados nos indicariam que há uma probabilidade de 95% de que o tempo médio de serviço na população esteja entre 12,3 e 15,16 anos. Qualquer valor dentro desse intervalo seria considerado como aceitável para a verdadeira média. Como o valor 14,5 anos estaria incluído no intervalo, concluiríamos que esse valor poderia ser a média populacional.

Como já foi mencionado, podemos também fazer esse teste utilizando a distribuição T .

Lembra-se de que as células da tabela Z continham valores das áreas sob a curva Z ? Pode-se elaborar também uma tabela de áreas (probabilidades) sob a distribuição T , mas como existem várias distribuições T a depender do tamanho da amostra, há uma tabela de probabilidades sob a curva para cada uma. É mais prático, então, se confeccionar uma única tabela T , contendo valores críticos de T em suas células, e não as probabilidades sob a curva, sendo que cada linha dessa tabela expressa um determinado tamanho de amostra (graus de liberdade), e suas colunas indicam os níveis de significância mais comumente utilizados. Isso não é um problema, porque já vimos que é possível realizar o nosso teste utilizando valores críticos de Z e, do mesmo modo, podemos usar valores críticos de T .

A tabela a seguir apresenta os valores críticos de T para determinadas áreas (probabilidades) sob a

distribuição T e graus de liberdade específicos:

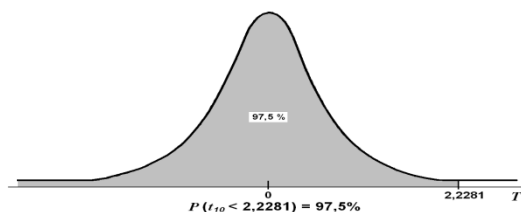


TABELA COM VALORES DE T

Graus de liberdade ($n-1$)	P_{90}	P_{95}	$P_{97,5}$	P_{99}	$P_{99,5}$
1	3,0780	6,3138	12,7060	31,8210	63,6570
2	1,8860	2,9200	4,3027	6,9650	9,9248
3	1,6380	2,3534	3,1825	4,5410	5,8409
4	1,5330	2,1318	2,7764	3,7470	4,6041
5	1,4760	2,0150	2,5706	3,3650	4,0321
6	1,4400	1,9432	2,4469	3,1430	3,7074
7	1,4150	1,8946	2,3646	2,9980	3,4995
8	1,3970	1,8595	2,3060	2,8960	3,3554
9	1,3830	1,8331	2,2622	2,8210	3,2498
10	1,3720	1,8125	2,2281	2,7640	3,1693
11	1,3630	1,7959	2,2010	2,7180	3,1058
12	1,3560	1,7823	2,1788	2,6810	3,0545
13	1,3500	1,7709	2,1604	2,6500	3,0123
14	1,3450	1,7613	2,1448	2,6240	2,9768
15	1,3410	1,7530	2,1315	2,6020	2,9467
16	1,3370	1,7459	2,1190	2,5830	2,9208
17	1,3330	1,7396	2,1098	2,5670	2,8982
18	1,3300	1,7341	2,1009	2,5520	2,8784
19	1,3280	1,7291	2,0930	2,5390	2,8609
20	1,3250	1,7247	2,0860	2,5280	2,8453
21	1,3230	1,7207	2,0796	2,5180	2,8314
22	1,3210	1,7171	2,0739	2,5080	2,8188
23	1,3190	1,7139	2,0687	2,5000	2,8073
24	1,3180	1,7109	2,0639	2,4920	2,7969
25	1,3160	1,7081	2,0595	2,4850	2,7874
26	1,3150	1,7056	2,0555	2,4790	2,7787
27	1,3140	1,7033	2,0518	2,4730	2,7707
28	1,3130	1,7011	2,0484	2,4670	2,7633
29	1,3110	1,6991	2,0452	2,4620	2,7564
30	1,3100	1,6973	2,0423	2,4570	2,7500
35	1,3062	1,6896	2,0301	2,4380	2,7239
40	1,3031	1,6839	2,0211	2,4230	2,7045
45	1,3007	1,6794	2,0141	2,4120	2,6896
50	1,2987	1,6759	2,0086	2,4030	2,6778
60	1,2959	1,6707	2,0003	2,3900	2,6603
70	1,2938	1,6669	1,9945	2,3810	2,6480
80	1,2922	1,6641	1,9901	2,3740	2,6388
90	1,2910	1,6620	1,9867	2,3680	2,6316
100	1,2901	1,6602	1,9840	2,3640	2,6260
120	1,2887	1,6577	1,9799	2,3580	2,6175
140	1,2876	1,6558	1,9771	2,3530	2,6114
160	1,2869	1,6545	1,9749	2,3500	2,6070
180	1,2863	1,6534	1,9733	2,3470	2,6035
200	1,2858	1,6525	1,9719	2,3450	2,6006
∞	1,2820	1,6450	1,9600	2,3260	2,5760

No topo da tabela há um diagrama cuja área sombreada nos orienta sobre como utilizá-la. Essa área nos indica que as células da tabela contêm os valores críticos de T para determinados percentis da

distribuição T e determinados graus de liberdade. Essa porcentagem é indicada no cabeçalho da tabela, onde aparecem vários percentis da distribuição T , e o número de graus de liberdade aparece na coluna indicadora. A notação $P_{97,5}$, por exemplo, representa o valor de t correspondente ao percentil 97,5, ou seja, o valor de t que separa os 2,5% valores mais altos de T dos 97,5% mais baixos. Assim, o número em cada célula da tabela indica o valor crítico de t correspondente ao percentil e graus de liberdade indicados na coluna e na linha, cujo cruzamento originou aquela célula. Por exemplo, se estivéssemos realizando um teste de hipóteses bicaudado com nível de significância, α , de 0,05, e se o nosso n fosse 11, teríamos $n - 1 = 11 - 1 = 10$ graus de liberdade, e iríamos olhar na tabela T a linha correspondente a 10 graus de liberdade e a coluna correspondente a $P_{97,5}$, porque, como já vimos, em um teste bicaudado nosso nível de significância é dividido pelas duas caudas da distribuição, ficando 0,025 ou, equivalentemente, 2,5% para cada uma. Note que 0,975 é o complemento de 0,025, pois $0,975 + 0,025 = 1$. Vemos na tabela que o valor crítico de t nesse caso seria 2,2281. Se o valor de t observado (ainda não vimos como calculá-lo, mas logo adiante você verá que utilizaremos o mesmo modo usado para calcular o valor de z) fosse maior do que 2,2281, concluiríamos que o valor testado não seria aceito como possível valor populacional. Concluiríamos o contrário, se o t observado fosse menor do que 2,2281. Como você pode ver, nosso raciocínio seria também semelhante àquele que desenvolvemos quando aplicamos o teste z .

Podemos também, é claro, concluir o teste utilizando o valor- p . Neste caso, precisaríamos consultar uma tabela T construída contendo em suas células probabilidades sob a curva e não valores críticos. Assim, poderíamos encontrar a probabilidade (valor- p) de obtermos valores maiores do que o t observado, e a compararíamos ao valor de α . Se o valor- p fosse menor do que o de α , concluiríamos que o valor testado não seria aceito como possível valor populacional e, se fosse maior, concluiríamos o contrário.

— Como realizamos o teste t ?

— O teste t é realizado através dos mesmos procedimentos do teste z . Considerando o mesmo exemplo que vem sendo utilizado, temos as seguintes etapas a cumprir:

1ª) Definição do nível de significância: $\alpha = 0,05$;

2ª) Definição das hipóteses:

$$H_O : \mu = 14,5 \text{ anos} \text{ e } H_A : \mu \neq 14,5 \text{ anos};$$

3ª) Cálculo do valor de t :

$$t = \frac{\bar{x} - \mu_o}{\frac{s}{\sqrt{n}}} = \frac{13,73 - 14,50}{\frac{5,23}{\sqrt{51}}} = \frac{-0,77}{\frac{5,23}{7,14}} = \frac{-0,77}{0,73} = -1,05.$$

Observe que realizamos o mesmo cálculo e obtivemos o mesmo resultado do teste z ;

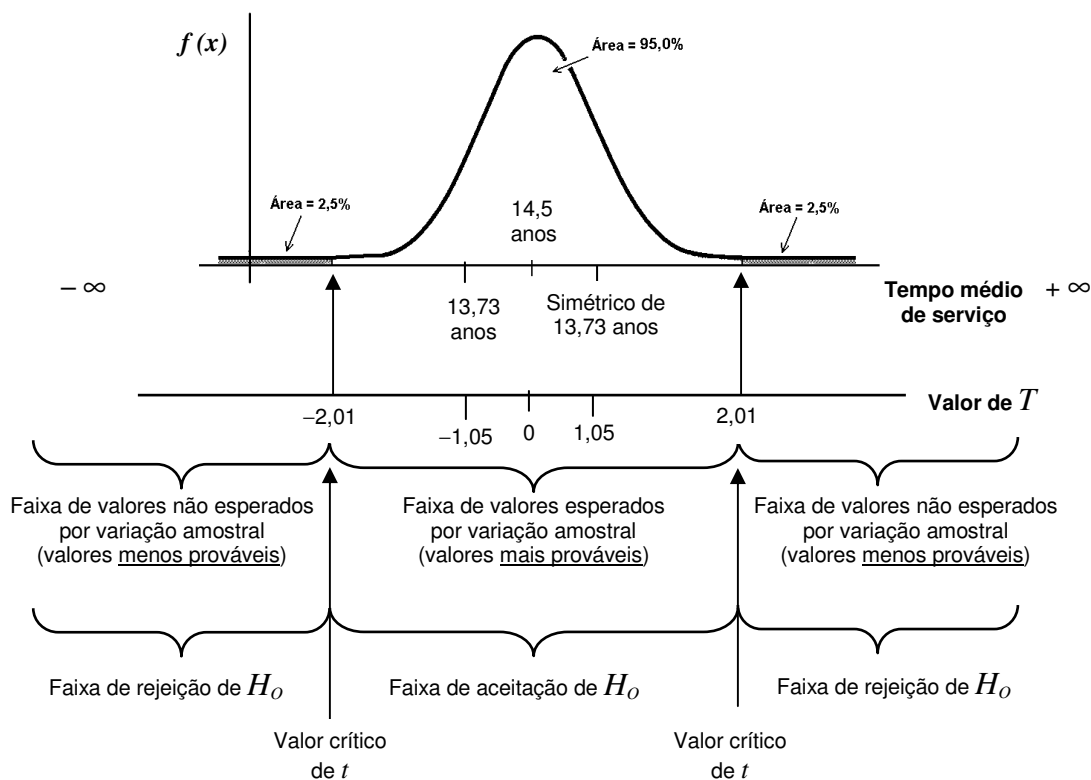
4ª) Obtenção do valor crítico de t :

Consultamos a tabela T para encontrarmos o valor crítico de t correspondente a um teste bicaudado, α de 0,05, e $n - 1 = 51 - 1 = 50$ graus de liberdade. Vemos que o valor é $2,0086 \cong 2,01$. Observe que a tabela T que estamos utilizando, que é mesma utilizada na maioria dos livros de Estatística, é incompleta. Há vários graus de liberdade não contidos na tabela. Se mantivéssemos o n original do nosso exemplo, que era 50, o número de graus de liberdade seria $n - 1 = 50 - 1 = 49$, valor que não aparece na tabela. Foi por isso que modificamos o n para 51.

— Quer dizer que se estudarmos uma amostra de 50, não poderemos fazer o teste t ?

— Poderemos sim. Não se preocupe com isso porque os programas de computador que você utilizará para fazer seus testes estatísticos possuem tabelas completas. Nos livros, para economizar espaço, algumas das tabelas aparecem incompletas, como essa da distribuição T .

Veja no diagrama abaixo a situação encontrada neste exemplo:



Observe que o formato da distribuição acima é semelhante àquele da distribuição Z , porque no nosso exemplo o n é grande. Veja que o valor crítico de t , 2,01, não é muito diferente do valor crítico de z , 1,96;

5ª) a) Comparação do valor observado de t ao valor crítico de t e conclusão do teste:

Como $-1,05 > -2,01$, concluiremos que o valor observado de t , correspondente a 13,73 anos, está em uma localização muito central da curva, não ultrapassando o valor crítico de t . Já que 13,73 anos não ultrapassou o limite de significância, isso nos indica que ele é um valor esperado por variação amostral, localizado, portanto, na área de aceitação da hipótese nula. Assim, chegaremos à mesma conclusão à qual chegamos com o teste z . É claro que nossa conclusão poderia ter sido outra se o valor testado fosse limítrofe, mas, mesmo nesta circunstância, como já vimos que os epidemiologistas experientes consideram valores limítrofes como estatisticamente significantes, terminariamos chegando à mesma conclusão;

b) Podemos também obter o valor- p correspondente e compará-lo ao valor de α . Os programas de computador lhe fornecerão facilmente o valor- p , e você já sabe como interpretar o resultado do teste com base neste valor. Sua conclusão será a mesma mencionada acima. Não faremos isso aqui porque as células da tabela utilizada não expressam áreas sob a curva, e lembre-se de que o valor- p é uma dessas áreas.

Outra opção ainda seria calcularmos o intervalo de 95% de confiança, que seria dado por

$$IC(95\%) = \bar{x} \pm t_{v,(1-\alpha/2)} \sqrt{\frac{s^2}{n}}$$

Intervalo de 95% de confiança

Média obtida no estudo

Valor de t correspondente ao percentil 97,5

Erro-padrão das médias caso tivéssemos feito vários estudos

onde o subscrito v representa o número de graus de liberdade.

Substituindo, teríamos:

$$IC(95\%) = \bar{x} \pm t_{v,(1-\alpha/2)} \sqrt{\frac{s^2}{n}} = \bar{x} \pm t_{50,0,975} \frac{s}{\sqrt{n}} = 13,73 \pm 2,01 \left(\frac{5,23}{\sqrt{51}} \right) = 13,73 \pm 2,01 \left(\frac{5,23}{7,14} \right) = 13,73 \pm (2,01)(0,73) = 13,73 \pm 1,47 = (12,26 \text{ a } 15,20).$$

Tais resultados nos indicariam que há uma probabilidade de 95% de que o tempo médio de serviço na população esteja entre 12,26 e 15,2 anos. Qualquer valor dentro desse intervalo seria considerado como aceitável para a verdadeira média. Como o valor 14,5 anos estaria incluído no intervalo, concluiríamos que ele poderia ser a média populacional.

— E se pudermos assumir que a distribuição da variável na população é normal, o n for pequeno, e o desvio-padrão populacional for conhecido?

— Neste caso você deverá utilizar o teste z que já aprendeu a fazer, sendo o desvio-padrão populacional conhecido, $\sigma = 5,53$ anos, novamente utilizado para o cálculo do valor observado de z .

— E se pudermos assumir que a distribuição da variável na população é normal, o n for pequeno, e o desvio-padrão populacional for desconhecido?

— Nesta situação não podemos mais utilizar o teste z . O teste estatístico que devemos aplicar é o teste t , pois esse utiliza uma distribuição probabilística (a distribuição T) apropriada para amostra pequena ($n < 30$) e σ desconhecido.

Voltando ao nosso exemplo, mas considerando agora que a amostra estudada contém apenas 15 trabalhadores, o que caracteriza um n pequeno, temos as seguintes etapas de realização do teste t :

1ª) Definição do nível de significância: $\alpha = 0,05$;

2ª) Definição das hipóteses:

$$H_O : \mu = 14,5 \text{ anos} \text{ e } H_A : \mu \neq 14,5 \text{ anos},$$

3ª) Cálculo do valor de t :

$$t = \frac{\frac{\bar{x} - \mu_o}{\frac{s}{\sqrt{n}}}}{\frac{5,23}{\sqrt{15}}} = \frac{13,73 - 14,50}{\frac{5,23}{\sqrt{15}}} = \frac{-0,77}{\frac{5,23}{3,87}} = \frac{-0,77}{1,35} = -0,57.$$

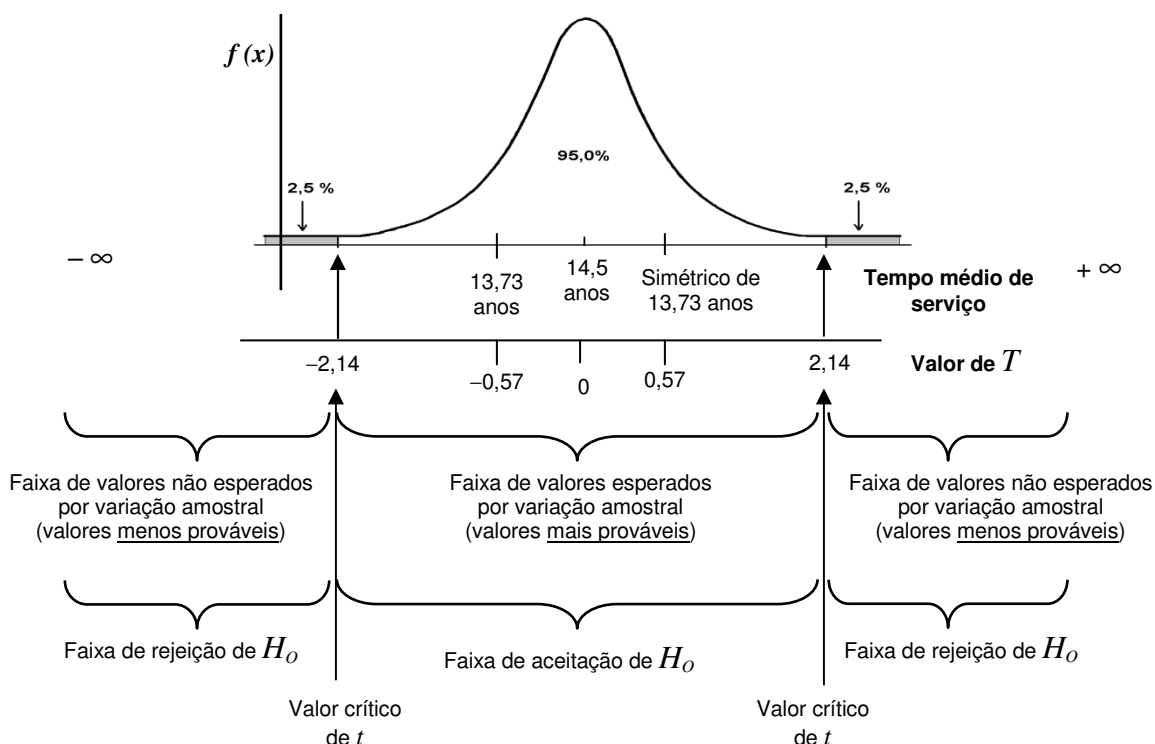
Ao mudarmos o nosso n de 51 para 15, provavelmente a média e o desvio-padrão obtidos seriam outros, mas mantivemos os mesmos valores, para que pudéssemos lhe mostrar a influência dessa mudança do n nos resultados. Veja que com um n pequeno o valor de t muda acentuadamente, de $-1,05$ para $-0,57$;

4ª) Obtenção do valor crítico de t :

Consultamos a tabela T para encontrarmos o valor crítico de t correspondente a um teste bicaudado, α de 0,05, e $n - 1 = 15 - 1 = 14$ graus de liberdade. Vemos que o valor é $2,1448 \cong 2,14$. Observe também que como o n é pequeno o valor crítico de t , que é 2,14, fica mais diferente do valor crítico de z , que é 1,96.

Lembre-se de que existem várias distribuições T , a depender do tamanho da amostra e, conseqüentemente, do número de graus de liberdade. Observe na próxima página que o formato da distribuição não é o mesmo da distribuição Z , embora seja parecido.

Veja no diagrama a seguir a situação no nosso exemplo atual:



5ª) Comparação do valor observado de t ao valor crítico de t e conclusão do teste:

Como $-0,57 > -2,14$ e, conseqüentemente, $0,57 < 2,14$, concluiremos que o valor de t correspondente a 13,73 anos está em uma localização pouco afastada de 14,5 anos, não ultrapassando, portanto, o valor crítico de t . Outra interpretação é a de que 13,73 anos é um valor esperado por variação amostral, localizado na área de aceitação da hipótese nula, e ainda outra que a verdadeira média é estatisticamente igual a 14,5 anos. Assim, se a pesquisa realizada não apresenta vieses importantes, poderemos aceitar H_0 e rejeitar H_A . Caso tivéssemos retirado numerosas amostras de mesmo tamanho $n = 15$, de uma população cujo tempo médio fosse 14,5 anos e desvio-padrão 5,23 anos, seria muito provável obtermos amostras com média de 13,73 anos, sendo 14,5 anos então, um dos valores aceitáveis para a verdadeira média na população, pois o resultado que obtivemos na única amostra estudada, 13,73 anos, seria compatível com isto.

O intervalo de 95% de confiança seria dado por

$$IC(95\%) = \bar{x} \pm t_{v, (1-\alpha/2)} \sqrt{\frac{s^2}{n}}$$

Diagrama de anotações para a fórmula acima:

- $IC(95\%)$: Intervalo de 95% de confiança
- \bar{x} : Média obtida no estudo
- $t_{v, (1-\alpha/2)}$: Valor de t correspondente ao percentil 97,5
- $\sqrt{\frac{s^2}{n}}$: Erro-padrão das médias caso tivéssemos feito vários estudos

Colocando os valores do nosso exemplo, teríamos:

$$IC(95\%) = \bar{x} \pm t_{v, (1-\alpha/2)} \sqrt{\frac{s^2}{n}} = \bar{x} \pm t_{14, 0,975} \frac{s}{\sqrt{n}} = 13,73 \pm 2,14 \left(\frac{5,23}{\sqrt{15}} \right) =$$

$$= 13,73 \pm 2,14 \left(\frac{5,23}{3,87} \right) = 13,73 \pm (2,14)(1,35) = 13,73 \pm 2,89 = (10,84 \text{ a } 16,62).$$

Estes resultados nos indicariam que há uma probabilidade de 95% de que o tempo médio de serviço na população esteja entre 10,84 e 16,62 anos. Qualquer valor dentro desse intervalo seria considerado como aceitável para a verdadeira média. Como o valor 14,5 anos estaria incluído no intervalo, concluiríamos que ele poderia ser o valor da média populacional.

— **E se não pudermos assumir que a distribuição da variável na população é normal?**

— Novamente, o teste a ser aplicado vai depender do tamanho da amostra estudada e do conhecimento, ou não, do desvio-padrão populacional.

Se a distribuição da variável estudada na população de onde a amostra foi retirada não puder ser assumida como normal, mas o n for grande e o desvio-padrão populacional for conhecido, você deverá utilizar o teste z .

— **Por quê?**

— Se a distribuição na população não puder ser assumida como normal, o n for grande e o desvio-padrão populacional for conhecido, podemos utilizar o teste z com base no teorema central do limite, porque se o n for suficientemente grande, a distribuição de médias amostrais vai apresentar uma distribuição normal mesmo que a distribuição na população não seja normal (lembra-se?). E é uma distribuição de médias amostrais que é utilizada na inferência estatística sobre médias. Como o desvio-padrão populacional é conhecido, calculamos o valor de z usando esse valor, e você já sabe fazer esse teste.

Se a distribuição na população não puder ser assumida como normal, o n for grande e o desvio-padrão populacional for desconhecido, vamos utilizar o teste t , mas, também com base no teorema central do limite, como o n é grande, este teste terá resultado muito semelhante ao do teste z . O cálculo do valor de $t \cong z$ será feito usando o desvio-padrão obtido na única amostra estudada. Você também já sabe fazer esse teste.

— **E se não pudermos assumir que a distribuição da variável na população é normal, e o n for pequeno?**

— Nessa situação, como o teorema central do limite não é aplicável porque o n é pequeno, não podemos utilizar nem o teste z nem o t , pois a aplicação de ambos requer que possamos assumir que a distribuição das médias amostrais seja normal.

— **O que devemos fazer então?**

— Quando a distribuição populacional não for normal e o n for pequeno, deveremos aplicar **testes estatísticos não-paramétricos**, independentemente do desvio-padrão populacional ser conhecido ou não. Já mencionamos esses testes anteriormente, mas não os abordaremos neste livro.

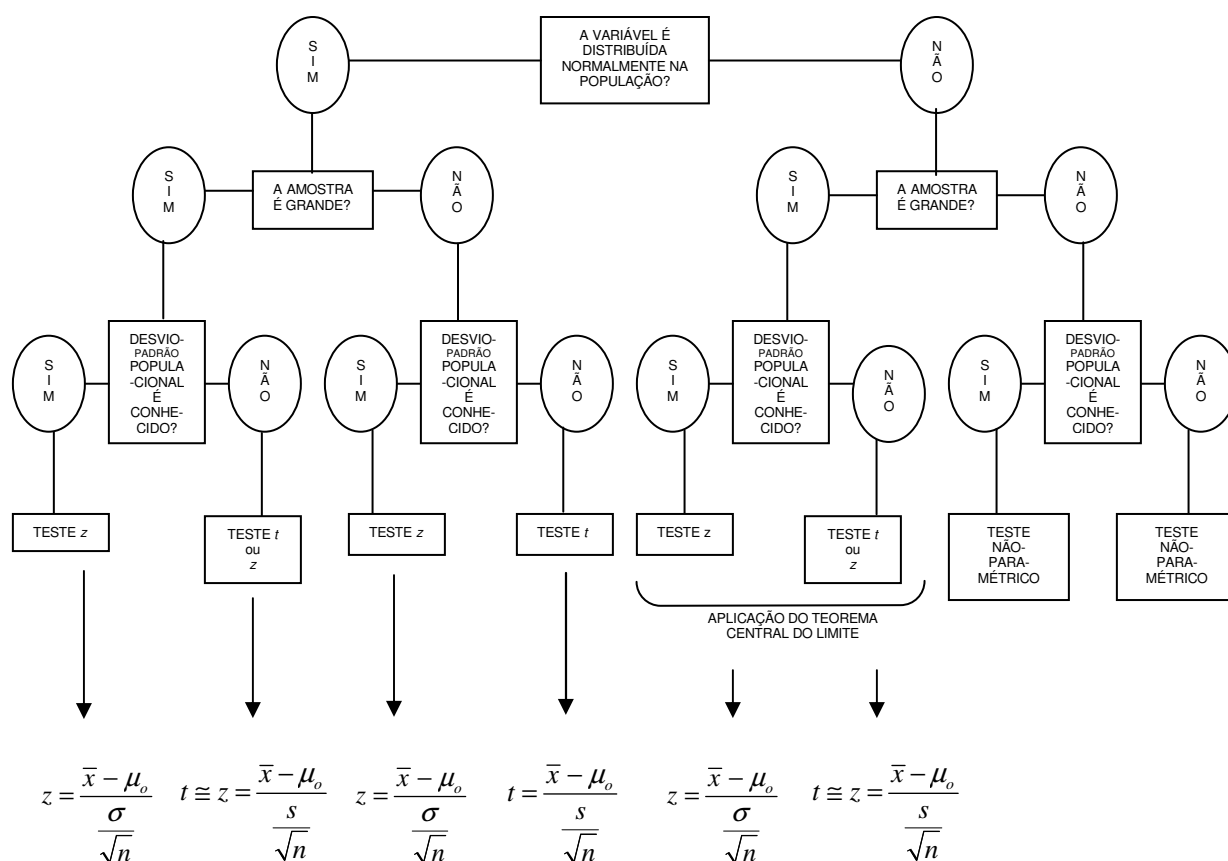
O fluxograma (modificado de *Daniel WW. Biostatistics: a foundation for analysis in the health sciences. 7ª ed. New York(NY): John Wiley;1999*) apresentado no final deste capítulo apresenta um resumo das aplicações dos testes z e t discutidas acima.

— **Existem outras aplicações dos testes z e t ?**

— Sim, há muitas outras aplicações, uma delas para compararmos duas médias obtidas em dois grupos independentes entre si. Lembre-se de que, até o momento, avaliamos se uma determinada média era provável de ser a verdadeira média populacional levando em conta a média obtida em uma única amostra. Como o estudo só obteve uma média, que será comparada a uma outra estabelecida externamente ao estudo, dizemos que foi feita inferência sobre uma média.

No próximo capítulo veremos como fazer inferência estatística sobre duas médias, ou seja, se em um mesmo estudo investigarmos dois grupos obtendo duas médias, uma para cada grupo, estudaremos sobre como avaliar se a diferença entre essas médias diferem estatisticamente.

Antes de prosseguir, diga-nos se você está cuidando adequadamente das suas outras atividades. Trabalho e estudo são apenas algumas das muitas dimensões da nossa vida.



CAPÍTULO 12

-
- Quando a inferência estatística é sobre duas médias e não sobre apenas uma?
 - Quando e como realizamos o teste z ou t para inferência sobre duas médias?
 - Em que outras situações devemos utilizar o teste t ?
-



— **Quando a inferência estatística é sobre duas médias e não sobre apenas uma?**

— No exemplo da Refinaria, podemos querer verificar se o tempo médio de serviço dos trabalhadores do setor de produção é diferente daquele dos trabalhadores do setor administrativo. Os últimos devem apresentar maior estabilidade no emprego e, por isso, maior tempo médio de serviço do que os primeiros. Torna-se então importante avaliarmos se os tempos médios nesses dois grupos de trabalhadores são diferentes, porque se isso ocorrer, sendo então uma das possibilidades a de que, conforme o esperado, a estabilidade no emprego seja menor no setor produtivo, poderíamos pleitear que os trabalhadores desse setor fossem incluídos em um suposto Programa de Aumento da Estabilidade no Emprego, implementado pelo Ministério do Trabalho, em uma suposta sociedade humana, realmente séria e civilizada.

Suponhamos, novamente, que não existam informações disponíveis sobre os tempos de serviço dos 1.000 trabalhadores da Refinaria, e que não tenhamos recursos suficientes para fazer uma coleta dessa informação para todos esses trabalhadores. Podemos estimar a verdadeira diferença entre os tempos médios de serviço no setor produtivo e no administrativo, baseando-nos em resultados obtidos em duas amostras de trabalhadores selecionadas aleatoriamente, uma em cada um desses setores. Para avaliarmos qual deva ser a verdadeira diferença entre as médias de tempo de serviço desses dois grupos na população de trabalhadores da Refinaria, com base em resultados obtidos em uma parte dessa população, teremos de realizar **inferência estatística sobre duas médias**. Nesse caso, vamos comparar duas médias, uma para o setor produtivo e outra para o administrativo, sendo que ambas as médias são obtidas no mesmo estudo. Nosso objetivo agora é estimar a verdadeira diferença populacional entre tais médias, utilizando os resultados obtidos em duas amostras aleatórias, uma de cada setor, investigadas em um único estudo epidemiológico.

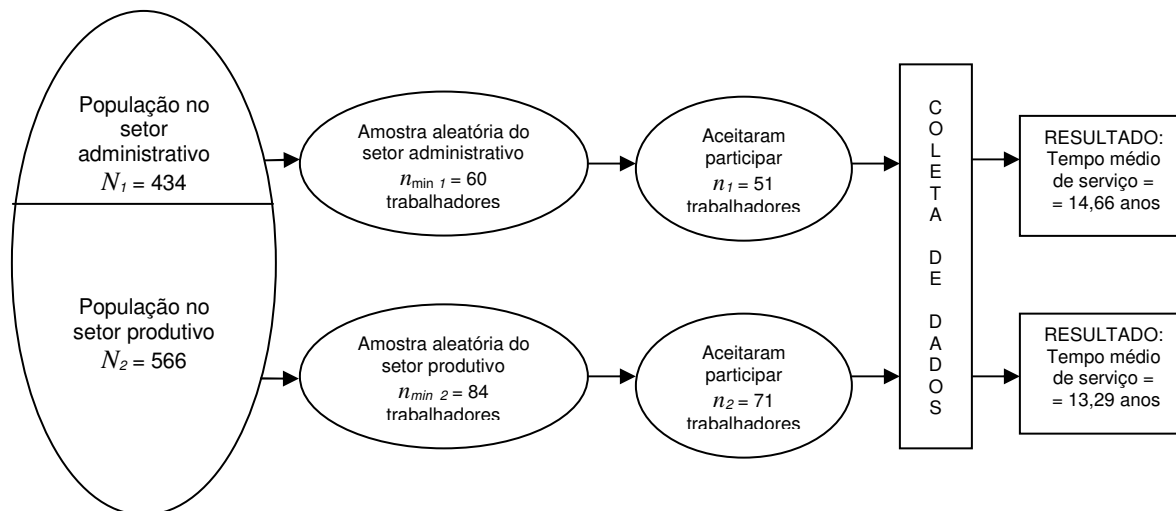
Utilizando uma fórmula adequada, não apresentada neste livro, calculamos o número mínimo (n mínimo) necessário para cada grupo a ser estudado. Suponha que esse número tenha sido de 50 trabalhadores para o setor administrativo, que foi ampliado para 60, porque era viável para nós atingirmos esse número, e para compensarmos possíveis recusas em participar ou perdas; e de 70 para o setor produtivo, que foi ampliado para 84 pelas mesmas razões. O n mínimo da pesquisa foi então de 144. Os 144 trabalhadores sorteados para a pesquisa foram procurados por nossa equipe para aplicação do termo de consentimento informado. Vamos supor que 122 deles tenham aceitado participar da pesquisa, sendo o número estudado no setor administrativo $n_1 = 51$, e no setor produtivo, $n_2 = 71$, tendo sido atingido, portanto, o número mínimo necessário em cada grupo.

De cada trabalhador de cada um dos dois grupos foi obtido o tempo de trabalho na Refinaria. Desse modo, ao final da coleta de dados, foi possível calcularmos o tempo médio de trabalho em cada um dos grupos. Suponha que tenhamos encontrado uma média de $\bar{x}_1 = 14,66$ anos para os trabalhadores do setor

administrativo, e de $\bar{x}_2 = 13,29$ anos para os do setor produtivo.

Veja o desenho do estudo na figura abaixo:

POPULAÇÃO NA REFINARIA
 $N = 1.000$ trabalhadores

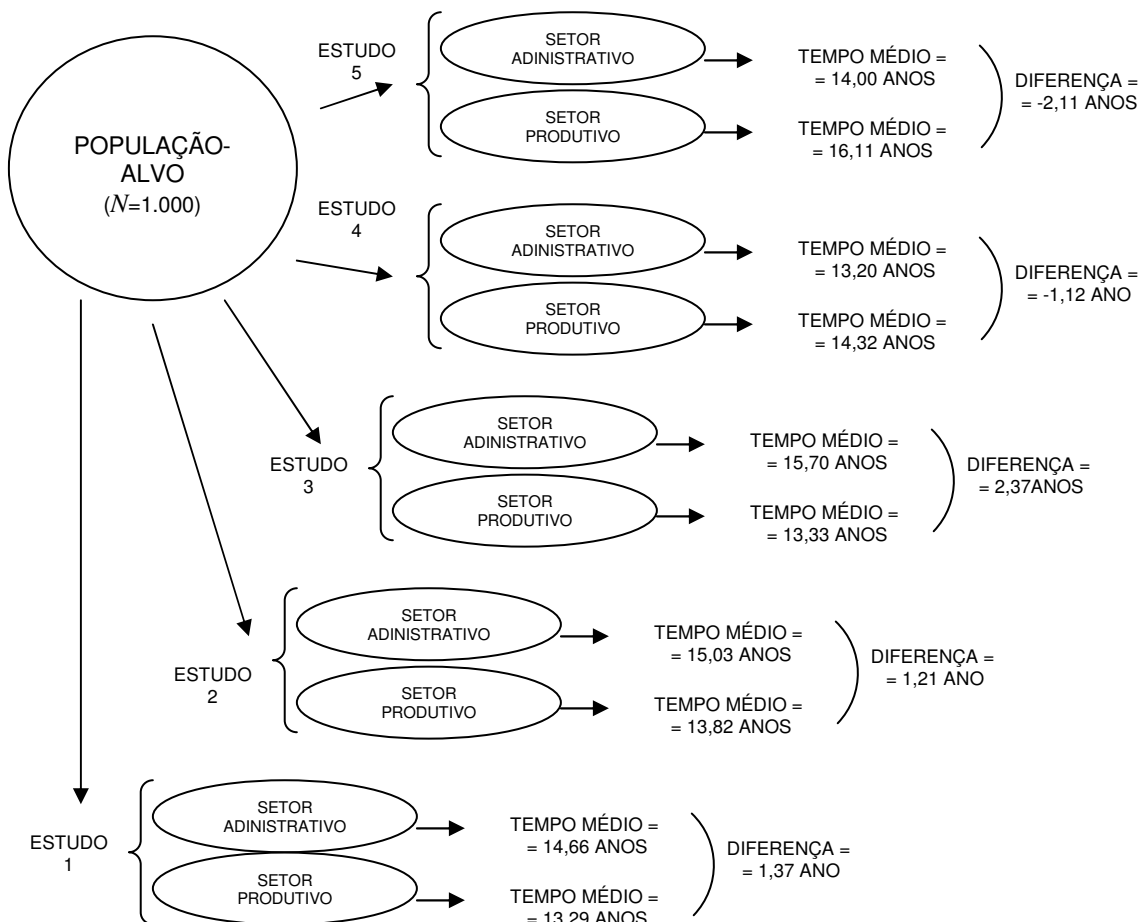


Do ponto de vista epidemiológico, o desenho do nosso estudo é do tipo transversal porque, em um certo momento do tempo, coletamos simultaneamente informações sobre o setor no qual cada trabalhador atuava (administrativo ou produtivo) e sobre seu tempo de serviço na Refinaria. Quanto ao método de amostragem, foi realizada uma amostragem aleatória, estratificada (setor administrativo ou produtivo), proporcional (note que as proporcionalidades entre o número de trabalhadores nos dois setores são aproximadamente as mesmas na amostra e na população). Poderíamos também fazer essa amostragem sem estratificação, selecionando aleatoriamente uma única amostra de 144 trabalhadores, e separando-os posteriormente em dois grupos segundo o setor administrativo ou produtivo. Seria muito provável que o número de trabalhadores em cada setor na amostra apresentasse a mesma proporcionalidade existente na população-alvo, mas isso só estaria garantido com a amostragem estratificada descrita acima.

Nosso objetivo será então verificar se há uma diferença estatisticamente significativa entre esses dois tempos médios de serviço, o que nos ajudará muito a avaliar se devemos solicitar a inclusão do grupo com menor estabilidade no programa governamental.

Geralmente, como já vimos, apenas um estudo é feito de cada vez sendo que, nesse caso, comparamos duas amostras obtidas em um mesmo estudo ou dois subgrupos de uma mesma amostra. Caso tivéssemos realizado numerosos estudos, cada um deles comparando duas amostras, uma constituída por trabalhadores atuando no setor administrativo e a outra no produtivo, sabemos que haveria uma variação dos resultados que seriam obtidos. É isso que nos deixa em dúvida sobre se a diferença entre os dois tempos médios encontradas no único estudo realizado pode ser considerada como a verdadeira diferença que seria obtida caso tivéssemos estudado toda a população, porque outros resultados poderiam ser encontrados se outros estudos semelhantes fossem realizados, cada um com outras duas amostras do mesmo tamanho,

retiradas da mesma população-alvo. Note que a situação é muito parecida com a que já discutimos para inferência sobre uma média, mas agora ao invés de considerarmos a possível variação amostral dos tempos médios caso tivéssemos realizado numerosos estudos, temos de levar em conta a possível variação amostral de diferenças entre dois tempos médios, porque agora estamos investigando duas amostras no mesmo estudo. A figura abaixo pode lhe ajudar a entender essa nova situação:



Observe que resultados diversos uns dos outros poderiam ser obtidos caso tivéssemos feito vários estudos semelhantes. Poderíamos inclusive encontrar tempos médios de serviço menores no setor administrativo, levando a diferenças negativas (amostras 4 e 5).

— Quando e como realizamos os testes z ou t para inferência sobre duas médias?

— Para avaliarmos se a única diferença encontrada, 1,37 ano, no único estudo (estudo 1) realizado por nossa equipe de pesquisa, é estatisticamente significativa, utilizamos também os testes z ou t , com pequenas modificações. Para isso, porém, é necessário que verifiquemos se certos pressupostos estão atendidos:

- a. Os grupos comparados são independentes?
- b. A distribuição da variável testada é normal na população-alvo de trabalhadores do setor administrativo, de onde a amostra desse setor foi aleatoriamente retirada? A distribuição da variável testada é normal na população-alvo de trabalhadores do setor produtivo, de onde a amostra desse setor foi aleatoriamente retirada?
- c. Os tamanhos das amostras investigadas são suficientemente grandes?
- d. Os desvios-padrão populacionais da variável testada para os trabalhadores de cada um dos grupos são conhecidos?
- e. Esses desvios-padrão são iguais?

Até agora não tínhamos chamado sua atenção para o fato de que o tempo médio de serviço observado no setor administrativo e o tempo médio de serviço no setor produtivo são observações independentes uma da outra. O tempo médio obtido em um grupo é completamente independente daquele observado no outro, porque os dois grupos foram constituídos de modo aleatório e independente. Não há nenhuma influência do tempo médio obtido em um grupo sobre o do outro, porque as observações em um e outro são independentes. Isso não ocorreria, por exemplo, se calculássemos o tempo médio para os trabalhadores do setor produtivo e desejássemos compará-lo ao tempo médio nesses mesmos trabalhadores algum tempo depois da implementação do Programa de Estímulo à Estabilidade no Emprego. Como, neste caso, estaríamos comparando os mesmos indivíduos, o tempo médio calculado depois deveria sofrer alguma influência do tempo médio que já existia antes.

Note que os pressupostos **b**, **c** e **d** citados acima são quase os mesmos que já havíamos utilizado para escolhermos entre os testes z , t ou algum teste não-paramétrico, para inferência sobre uma média. Apenas o primeiro (já explicado acima) e o último pressupostos foram acrescentados, este porque agora não um, mas dois desvios-padrão populacionais podem ser conhecidos previamente, já que o estudo investiga dois grupos. Sendo assim, é também importante verificarmos se podemos assumir se esses desvios são iguais ou não, para sabermos qual o teste estatístico adequado.

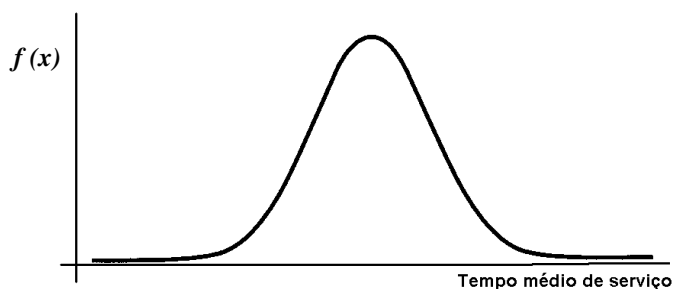
Vamos avaliar o atendimento a esses pressupostos no nosso exemplo: já vimos que os grupos comparados são independentes e que as amostras foram escolhidas aleatoriamente de populações específicas, similares aos grupos a serem comparados; suponha também que possamos assumir, com base em conhecimentos já existentes, que a distribuição do tempo médio de serviço seja normal tanto na população de trabalhadores do setor administrativo quanto do setor produtivo; além disso, os nossos n 's são grandes (grosseiramente, tanto $n_1 = 51$ quanto $n_2 = 71$ são maiores do que 30 e, rigorosamente, supomos que tenhamos calculado o n mínimo necessário para o estudo); passando à verificação do próximo pressuposto, vamos assumir também que os desvios-padrão na população-alvo sejam conhecidos ($\sigma_1 = 6,22$ anos e $\sigma_2 = 4,77$ anos), embora raramente isso seja possível. Nessas condições, torna-se irrelevante avaliarmos se esses desvios-padrão são estatisticamente iguais ou não (quinto pressuposto),

porque independentemente disso o teste adequado será o teste z .

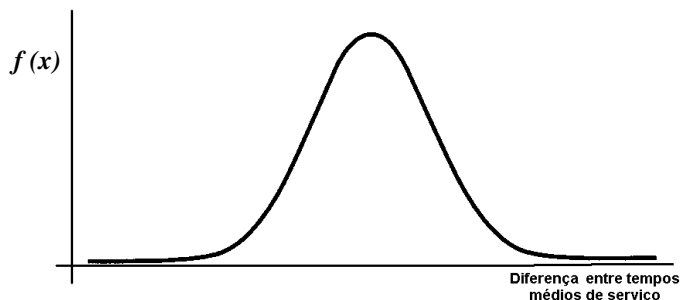
Quando as amostras são independentes; escolhidas aleatoriamente de populações específicas, similares aos grupos a serem comparados; retiradas de populações nas quais a variável estudada possa ser assumida como normalmente distribuída; os n 's são grandes; e os desvios-padrão populacionais são conhecidos, utilizaremos o teste z , independentemente desses desvios serem iguais ou não.

Lembre-se de que agora a distribuição normal a ser utilizada como referência não é mais a distribuição dos tempos médios de serviço dos trabalhadores, mas das diferenças entre os tempos médios de serviço dos trabalhadores nos dois setores da Refinaria em estudo.

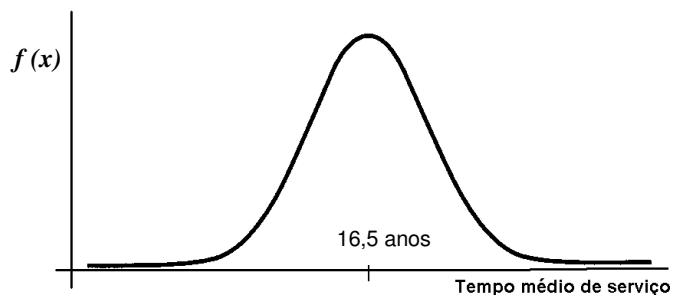
Para inferência sobre uma média (antes):



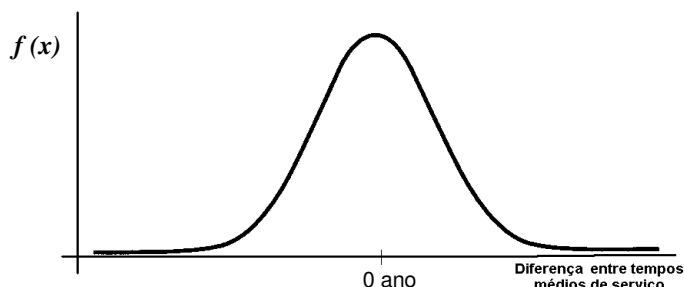
Para inferência sobre duas médias (agora):



Lembre-se também de que antes, como nossa hipótese nula estabelecia que $\mu = 16,5$ anos, considerávamos essa média esperada para a população-alvo como a média da distribuição-modelo utilizada para inferência, conforme mostramos novamente abaixo:



Nosso objetivo agora é avaliar se a diferença obtida no estudo, 1,37 ano, é estatisticamente igual ou diferente de zero, pois uma diferença igual a zero indica que as médias dos dois grupos são iguais, não é? Assim, agora, utilizaremos a seguinte distribuição como modelo para inferência:



Vamos calcular o quanto uma diferença de 1,37 ano se afasta de uma diferença igual a zero. Em seguida, expressaremos esse afastamento, esse desvio, em número de erros-padrão, dividindo-o pelo erro-padrão das diferenças entre médias. Sabendo que se convencionou denotar esse resultado pela letra z , nosso cálculo é então:

$$z = \frac{(\bar{x}_1 - \bar{x}_2) - (\mu_1 - \mu_2)_o}{EP_{(\bar{x}_1 - \bar{x}_2)}}$$

onde $(\bar{x}_1 - \bar{x}_2)$ representa a diferença encontrada no único estudo realizado, $(\mu_1 - \mu_2)_o$ a diferença prevista na hipótese nula (por isso colocamos o subscrito o , ou seja “zero”, nulo), e $EP_{(\bar{x}_1 - \bar{x}_2)}$ o erro-padrão esperado para as diferenças entre as médias amostrais, caso tivéssemos realizado numerosos estudos.

Na inferência sobre uma média, quando σ era conhecido, utilizávamos no denominador o erro-padrão de médias amostrais, que era calculado por $EP_{\bar{x}} = \sqrt{s_{\bar{x}}^2} = \sqrt{\sigma^2/n} = \sigma/\sqrt{n}$, lembra-se? Agora, o denominador é o erro-padrão de diferenças entre duas médias amostrais, ou seja, se tivéssemos realizado numerosos estudos, sabendo que haveria uma variação dos resultados nesses diferentes estudos, qual seria a variabilidade média esperada para as diferenças obtidas entre os tempos médios de serviço dos dois grupos comparados? Se σ_1 for conhecido, a variância para as amostras aleatórias retiradas da população do setor administrativo será dada por $s_{\bar{x}_1}^2 = \sigma_1^2/n_1$. Se σ_2 for conhecido, a variância para as amostras aleatórias retiradas da população do setor produtivo será dada por $s_{\bar{x}_2}^2 = \sigma_2^2/n_2$. A variância para as diferenças que seriam obtidas entre os tempos médios de serviço dessas duas amostras, considerando os numerosos estudos porventura realizados, será então calculada pela soma da variância da primeira população com a

variância da segunda população, como apresentado abaixo:

$$s_{(\bar{x}_1 - \bar{x}_2)}^2 = s_{\bar{x}_1}^2 + s_{\bar{x}_2}^2 = \frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2},$$

sendo então seu erro-padrão dado pela raiz quadrada desta quantidade (o erro-padrão não é a raiz quadrada da variância?):

$$EP_{(\bar{x}_1 - \bar{x}_2)} = \sqrt{s_{(\bar{x}_1 - \bar{x}_2)}^2} = \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}.$$

Podemos agora especificar o cálculo do erro-padrão na fórmula para o cálculo de z que, quando estivermos comparando duas médias, será feito por:

$$z = \frac{(\bar{x}_1 - \bar{x}_2) - (\mu_1 - \mu_2)_o}{EP_{(\bar{x}_1 - \bar{x}_2)}} = \frac{(\bar{x}_1 - \bar{x}_2) - (\mu_1 - \mu_2)_o}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}}.$$

Vamos organizar melhor esse nosso teste de hipóteses, considerando suas diversas etapas:

1ª) Definição do nível de significância: $\alpha = 0,05$;

2ª) Definição das hipóteses:

$$H_O : \mu_1 - \mu_2 = 0 \quad \text{ou} \quad H_O : \mu_1 = \mu_2.$$

Note que se $\mu_1 - \mu_2 = 0$, então $\mu_1 = \mu_2$, já que a diminuição de uma quantidade por outra é zero somente se as quantidades forem iguais, não é?

A hipótese alternativa é:

$$H_A : \mu_1 - \mu_2 \neq 0 \quad \text{ou} \quad H_A : \mu_1 \neq \mu_2.$$

Se $\mu_1 - \mu_2 \neq 0$, então $\mu_1 \neq \mu_2$ porque a diminuição de uma quantidade por outra só será diferente de zero se as quantidades forem diferentes.

Lembre-se de que essas hipóteses nos levam a realizar um teste bicaudado, porque testar que $\mu_1 - \mu_2 \neq 0$ é o mesmo que testar que $\mu_1 - \mu_2 > 0$ ou que $\mu_1 - \mu_2 < 0$.

3ª) Cálculo do valor de z : substituindo os valores do nosso exemplo na equação vista logo acima, temos que

$$z = \frac{(\bar{x}_1 - \bar{x}_2) - (\mu_1 - \mu_2)_0}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}} = \frac{(14,66 - 13,29) - 0}{\sqrt{\frac{(6,22)^2}{51} + \frac{(4,77)^2}{71}}} = \frac{1,37 - 0}{\sqrt{\frac{38,69}{51} + \frac{22,75}{71}}} =$$

$$= \frac{1,37}{\sqrt{0,76 + 0,32}} = \frac{1,37}{\sqrt{1,08}} = \frac{1,37}{1,04} \cong 1,32.$$

Note que, como a hipótese nula estabelece que $\mu_1 - \mu_2 = 0$, e o teste é feito assumindo-se que essa hipótese seja verdadeira, nós substituímos $(\mu_1 - \mu_2)_0$ por 0 nos cálculos acima;

4ª) Obtenção do valor- p :

Olhando na tabela com valores positivos de Z (página 134), encontramos a probabilidade de obtermos um valor de z menor do que 1,32, isto é, $P(Z < 1,32)$. Essa probabilidade é de 0,9066 ou 90,66%. A probabilidade de interesse será então dada por

$$P(Z > 1,32) = (1 - P(Z < 1,32)) = 1 - 0,9066 = 0,0934 \text{ ou } 9,34\%.$$

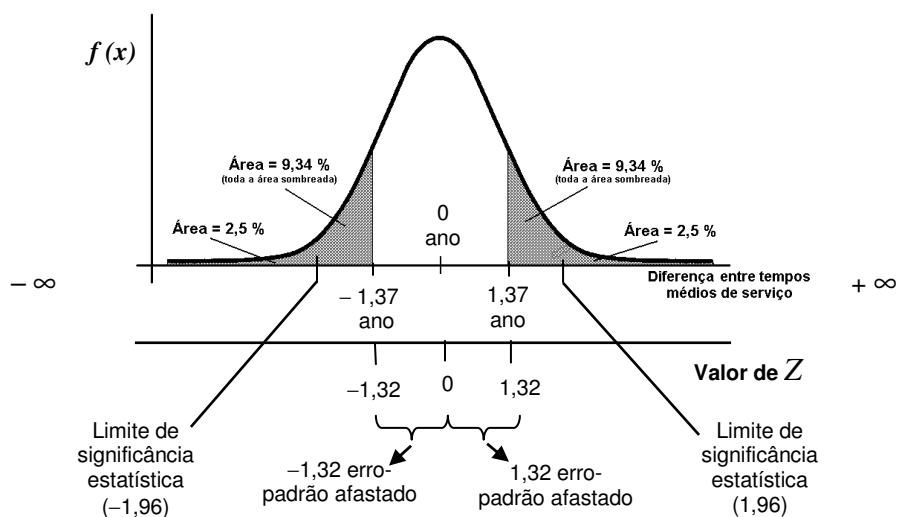
Se tivéssemos realizado numerosos estudos semelhantes, haveria uma probabilidade de 9,34% de obtermos amostras com diferenças entre os tempos médios de serviço maiores do que 1,37 ano, que corresponde ao valor de z igual a 1,32.

Contudo, como o teste é bicaudado, temos ainda que multiplicar esse resultado por dois, para considerarmos a outra cauda da curva, obtendo:

$$P(Z > 1,32)(2) = (0,0934)(2) = 0,1868 \text{ ou } 18,68\%,$$

que é o valor- p final do nosso teste;

A situação atual é mostrada no diagrama abaixo:



5ª) a) Comparação do valor- p ao valor de α e conclusão do teste:

Como $18,68\% > 5,0\%$, concluiremos que seria grande a probabilidade de obtermos uma diferença de 1,37 ano, caso tivéssemos estudado numerosas amostras retiradas de uma população-alvo cuja diferença entre as médias comparadas fosse zero. O pequeno afastamento de 1,37 ano em relação à diferença zero pode ser atribuído à simples variação amostral e não a uma diferença real. A diferença 1,37 ano está localizada na área de aceitação da hipótese nula, $H_0 : \mu_1 - \mu_2 = 0$. Assim, se a pesquisa que realizamos não tivesse apresentado vieses importantes, poderíamos aceitar H_0 e rejeitar H_A , concluindo que a diferença encontrada entre os tempos médios no único estudo realizado não era estatisticamente **significante**, indicando que os tempos médios de serviço nas populações-alvo de trabalhadores dos setores administrativo e produtivo eram, muito provavelmente, iguais, porque a diferença encontrada no único estudo realizado teria sido compatível com isso; ou

b) Comparação do valor observado de z ao valor crítico de z :

Considerando que o nosso teste é bicaudado, devido às hipóteses que foram testadas, e que o nível de significância estabelecido foi 0,05, sabemos que os limites críticos de z são 1,96 e -1,96, certo?

Como $1,32 < 1,96$ e, conseqüentemente, $-1,32 > -1,96$, concluiremos que o limite de significância estatística não foi ultrapassado; que a diferença encontrada está em uma área de aceitação da hipótese nula; em uma área de valores esperados por simples variação amostral, não sendo, portanto, estatisticamente **significante**. Como sempre, essa conclusão pressupõe que os resultados encontrados não tenham sido influenciados por erros (vieses) existentes no estudo.

Observe que, como já esperávamos, as conclusões nos itens **a** e **b** da 5ª etapa do nosso teste foram semelhantes.

Chegaremos também à mesma conclusão se em vez de fazermos um teste de hipóteses, realizarmos o cálculo do intervalo de 95% de confiança, como mostramos a seguir.

Esse intervalo é dado por

$$IC(95\%) = (\bar{x}_1 - \bar{x}_2) \pm z_{(1-\alpha/2)} \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}$$

Diagram illustrating the components of the 95% confidence interval formula:

- $IC(95\%)$: Intervalo de 95% de confiança
- $(\bar{x}_1 - \bar{x}_2)$: Diferença entre as médias obtidas no estudo
- $z_{(1-\alpha/2)}$: Valor de z correspondente ao percentil 97,5
- $\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}$: Erro-padrão das diferenças entre as médias caso tivéssemos feito vários estudos

Como nosso nível de significância é 0,05, temos que

$$z_{(1-\alpha/2)} = z_{(1-0,05/2)} = z_{(1-0,025)} = z_{0,975} = z_{97,5\%}.$$

Então, se queremos calcular um intervalo de 95,0% de confiança, o que equivale a um α de 0,05,

utilizaremos na fórmula o valor na distribuição Z que separa os 97,5% valores mais baixos dos 2,5% valores mais altos. Lembre-se de que esse valor corresponderá ao percentil 97,5 ou 0,975 dessa distribuição.

— **E qual é esse valor?**

— Ora, esse valor é o nosso velho conhecido $z = 1,96$. Não se assuste. Não vimos nada de muito novo aqui. A única novidade é que antes tínhamos calculado o intervalo para uma média e agora para a diferença entre duas médias. Isto resultou em algumas pequenas modificações na fórmula, que são muito fáceis de entender. Quanto à notação utilizada para o valor de z , ela não chega a ser uma novidade, mas apenas um maior detalhamento, $z_{(1-\alpha/2)}$ em vez de simplesmente z .

— **Por que vocês não detalharam isso antes?**

— Nosso maior objetivo antes era que você entendesse conceitualmente a inferência sobre uma média. Agora que você já entendeu o assunto e considerando que, conceitualmente, a inferência sobre duas médias é semelhante, podemos nos dar ao luxo de detalhar alguns aspectos.

Voltando ao cálculo do intervalo de confiança, e substituindo na fórmula acima os valores do nosso exemplo obtemos:

$$IC(95\%) = (\bar{x}_1 - \bar{x}_2) \pm z_{(1-\alpha/2)} \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}} = 1,37 \pm 1,96 \sqrt{\frac{(6,22)^2}{51} + \frac{(4,77)^2}{71}} = 1,37 \pm 1,96 \sqrt{\frac{38,69}{51} + \frac{22,75}{71}} = 1,37 \pm 1,96 \sqrt{0,76 + 0,32} = 1,37 \pm 1,96 \sqrt{1,08} = 1,37 \pm (1,96)(1,04) = 1,37 \pm 2,04 = (-0,67 \text{ a } 3,41).$$

Não se esqueça de que o que acabamos de calcular, foi um intervalo compreendido entre uma diferença igual a 1,37 ano mais ou menos cerca de duas vezes o valor do erro-padrão esperado para as diferenças entre os tempos médios de serviço. Esse resultado nos informa que há uma probabilidade de 95% de que a diferença entre os tempos médios de serviço na população esteja entre $-0,67$ e $3,41$ anos. Assim, qualquer diferença que esteja dentro desse intervalo, será considerada por nós como aceitável para a verdadeira diferença na população.

Vemos que a diferença zero, que indicaria que os tempos médios na população seriam iguais, está dentro do intervalo calculado, nos levando a concluir que é possível que a diferença na população seja zero. É o mesmo que dizermos que a diferença entre 1,37 ano e zero não é estatisticamente significativa.

Considere agora que as amostras a serem comparadas sejam independentes, tenham sido escolhidas aleatoriamente de populações específicas, similares aos grupos a serem comparados; a distribuição do tempo de serviço seja normal nessas populações; e os n 's sejam grandes, mas os desvios-padrão populacionais não sejam conhecidos. Nesse caso, o teste a ser aplicado dependerá da igualdade ou não dos desvios-padrão populacionais. Se pudermos considerar os desvios-padrão populacionais

estatisticamente iguais, aplicaremos o teste t . Se não, usaremos o teste t' (lemos “teste t linha”). Você já sabe que, como na atual situação as amostras são grandes, as conclusões que obteríamos com o teste z seriam muito semelhantes. Não precisaremos mostrar novamente como realizar o teste z .

— **Desvios-padrão populacionais estatisticamente iguais ou diferentes? Quer dizer que temos de utilizar outro teste estatístico apenas para verificar se o pressuposto de igualdade entre os desvios-padrão está atendido?**

— Exatamente, mas esse teste compara duas variâncias e não dois desvios-padrão, o que não se constitui em um problema, porque ao compararmos duas variâncias estamos comparando dois desvios-padrão, já que a variância e o desvio-padrão indicam a mesma coisa (variabilidade média em torno da média) em escalas diferentes, lembra-se? Esse teste utiliza a razão entre duas variâncias para testar se elas são estatisticamente iguais ou diferentes, podendo então ser chamado de teste da razão de variâncias. Uma razão entre variâncias igual a 1 indica que elas são iguais, porque uma quantidade dividida por outra igual tem como resultado a unidade, não é mesmo?

Antes de prosseguirmos com as aplicações dos testes t e t' , vamos apresentá-lo ao teste da razão de variâncias, pois este será fundamental para você saber qual dos testes deve ser aplicado.

— **Espere um pouco! Não entendi uma coisa: se os desvios-padrão populacionais são desconhecidos, como vamos poder utilizar um teste para compará-los?**

— Boa pergunta!

Quando estávamos realizando inferência sobre uma média e o desvio-padrão (variância) populacional era desconhecido, o que fizemos?

— **Utilizamos como estimador do desvio-padrão populacional o desvio-padrão obtido na única amostra estudada.**

— Certo!

Agora faremos o mesmo. Utilizaremos os desvios-padrão obtidos no único estudo realizado como estimadores dos desvios-padrão populacionais desconhecidos. Só que isso torna necessário utilizarmos a distribuição T como modelo para inferência. Entretanto, pode ser demonstrado que, se o n de cada uma das amostras comparadas for suficientemente grande, os desvios obtidos em tais amostras são estimadores válidos dos desvios populacionais, sendo que, nessa circunstância, em última instância, os valores de t serão muito semelhantes aos de z .

Suponha que os desvios-padrão obtidos para as amostras de trabalhadores do setor administrativo e do produtivo, no único estudo realizado, tenham sido 5,28 e 4,28 anos, respectivamente. Tais desvios serão utilizados como estimadores dos verdadeiros desvios nas populações de onde essas amostras foram retiradas aleatoriamente. Vamos então testar se esses desvios são estatisticamente iguais ou diferentes, para

podermos prosseguir com o nosso teste das diferenças entre os dois tempos médios de serviço. Nosso objetivo será responder à seguinte pergunta: a diferença encontrada entre os desvios-padrão no único estudo realizado foi ou não decorrente de variação amostral? Se a resposta for sim, concluiremos que os desvios-padrão populacionais e, portanto, as variâncias são, muito provavelmente, iguais do ponto de vista estatístico. Se não, concluiremos o contrário.

Responderemos a essa questão seguindo as etapas do teste da razão de variâncias, apresentadas abaixo. Contudo, antes disso, devemos verificar se os pressupostos para aplicação desse teste estão atendidos: as amostras são independentes? Cada amostra considerada é uma amostra aleatória retirada de uma população semelhante de indivíduos? Os tempos médios de serviço em cada população são aproximadamente normalmente distribuídos? No nosso exemplo, temos que os trabalhadores de cada amostra foram selecionados de modo independente uns dos outros, de maneira que a observação sobre o tempo de serviço feita em cada trabalhador independe da observação feita nos demais (o valor obtido para o tempo de serviço de um trabalhador não influencia em nada o valor do tempo de serviço de outro); cada amostra estudada (a dos trabalhadores administrativos ou da produção) constitui-se em uma amostra aleatória retirada de uma população semelhante (do setor administrativo ou produtivo, respectivamente) de trabalhadores; e suponhamos que haja evidências na literatura de que os tempos médios de serviço apresentem uma distribuição normal tanto na população de trabalhadores do setor administrativo como do produtivo. Podemos então, considerar que os pressupostos para aplicação do teste da razão de variâncias estão atendidos em nosso exemplo, e passar para as etapas necessárias para sua realização, apresentadas a seguir:

1ª) Definição do nível de significância: $\alpha = 0,05$;

2ª) Definição das hipóteses: denotando as variâncias das populações comparadas por σ_1^2 e σ_2^2 , nossas hipóteses são

$$H_O : \sigma_1^2 = \sigma_2^2 \text{ e } H_A : \sigma_1^2 \neq \sigma_2^2 ;$$

Não se esqueça de que ao compararmos variâncias estamos também comparando desvios-padrão;

3ª) Cálculo do valor de F ;

— **Valor de F ?**

— Até agora já havíamos utilizado valores de Z ou de T , mas agora teremos de utilizar uma distribuição diferente das distribuições Z e T .

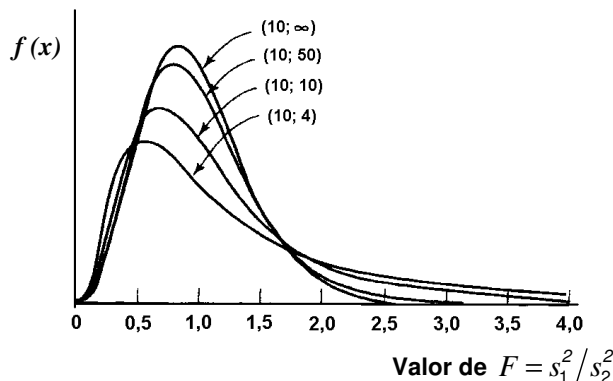
Lembre-se de que estamos discutindo uma situação na qual os desvios-padrão populacionais são desconhecidos e, portanto, também as variâncias populacionais. Temos então de utilizar como estimadores desses parâmetros populacionais, as variâncias amostrais, s_1^2 e s_2^2 , encontradas no único estudo realizado. O teste da razão de variâncias usa, na verdade, uma razão entre duas proporções: a proporção entre a variância encontrada na amostra de trabalhadores administrativos e a verdadeira (e desconhecida) variância populacional nestes trabalhadores; e a proporção entre a variância encontrada na amostra de trabalhadores produtivos e a verdadeira (e desconhecida) variância populacional nestes trabalhadores. Essas proporções são dadas por s_1^2/σ_1^2 , para o primeiro grupo de trabalhadores, e por s_2^2/σ_2^2 , para o segundo. Assim, o teste se baseará na razão: $(s_1^2/\sigma_1^2)/(s_2^2/\sigma_2^2)$. Acontece que, como em todo teste de hipóteses, no teste atual assumimos que a hipótese nula, $H_o: \sigma_1^2 = \sigma_2^2$, é verdadeira e, nessa circunstância, a divisão $(s_1^2/\sigma_1^2)/(s_2^2/\sigma_2^2)$ pode ser simplificada para

$$\frac{s_1^2}{\sigma_1^2} \div \frac{s_2^2}{\sigma_2^2} = \frac{s_1^2}{\sigma_1^2} \div \frac{s_2^2}{\sigma_1^2} = \left(\frac{s_1^2}{\sigma_1^2} \right) \left(\frac{\sigma_1^2}{s_2^2} \right) = \left(\frac{s_1^2}{\sigma_1^2} \right) \left(\frac{\sigma_1^2}{s_2^2} \right) = \frac{s_1^2}{s_2^2}.$$

Podemos fazer essa substituição porque assumimos na H_o que essas variâncias são iguais

Observe que s_1^2/s_2^2 é a razão entre as variâncias obtidas nas duas amostras estudadas. Assim, se a hipótese nula for verdadeira, a quantidade $(s_1^2/\sigma_1^2)/(s_2^2/\sigma_2^2)$ pode ser simplificada para s_1^2/s_2^2 . Esta razão pode ser usada por nós para compararmos as variâncias dos grupos estudados, porque seu resultado reflete a magnitude relativa entre elas. Quando, por exemplo, s_1^2 for igual a s_2^2 , a razão s_1^2/s_2^2 será igual a 1. Mas, se s_1^2 for menor do que s_2^2 , seu resultado será menor do que 1 e, se s_1^2 for maior do que s_2^2 , apresentará um valor maior do que 1.

Se realizarmos numerosos estudos semelhantes em amostras retiradas da mesma população-alvo, e elaborarmos uma distribuição das freqüências das razões entre as variâncias obtidas, obteremos uma curva diferente das distribuições probabilísticas Z e T . Essa nova distribuição é denominada por distribuição F , e apresenta a forma mostrada a seguir (os números entre parêntesis indicam os graus de liberdade do numerador, s_1^2 , e do denominador, s_2^2 , respectivamente):



Observe que estamos tratando aqui também com uma família de distribuições, tal como já havíamos visto para a distribuição T . Assim, a forma específica da distribuição F depende dos graus de liberdade, que equivalem aproximadamente ao número de indivíduos estudados. Só que agora, temos que considerar os graus de liberdade da variância contida no numerador, s_1^2 , e os daquela contida no denominador, s_2^2 , pois ambos influenciarão o formato específico da distribuição F .

O número de graus de liberdade do numerador será calculado por $n_1 - 1$, porque ao calcularmos o tempo médio do grupo administrativo perdemos um grau de liberdade, e por $n_2 - 1$ para o denominador, porque ao calcularmos o tempo médio do grupo da produção perdemos também um grau de liberdade.

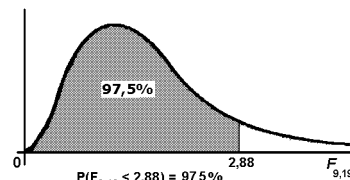
Podemos então prosseguir nosso teste da igualdade entre as variâncias, calculando o valor de F :

$$F = \frac{s_1^2}{s_2^2} = \frac{(5,28)^2}{(4,28)^2} = \frac{27,88}{18,32} \cong 1,52;$$

4ª) Obtenção na tabela F , do valor crítico de F para um nível de significância de 0,05, $n_1 - 1 = 51 - 1 = 50$ graus de liberdade do numerador, e $n_2 - 1 = 71 - 1 = 70$ graus de liberdade do denominador.

Existe uma tabela F para cada um dos diferentes níveis de significância que possamos estabelecer para o nosso teste, de modo que a tabela que vamos consultar é aquela para um teste bicaudado com $\alpha = 0,05$. Sendo assim, nosso α deve ser dividido pelas duas extremidades da distribuição. Como $\alpha/2 = 0,05/2 = 0,025$, utilizamos a tabela que contém os valores críticos de F correspondentes ao percentil 97,5, ou seja, valores que separam as 2,5% razões de variâncias com valores mais altos, das 97,5% razões com valores mais baixos.

Olhando na tabela a seguir, vemos que o valor crítico de F no nosso exemplo é 1,66.

TABELA F 

P _{97,5}																			
Nº de g. l. do denominador	Número de graus de liberdade (g.l.) do numerador																		
	1	2	3	4	5	6	7	8	9	10	15	20	25	30	40	50	100	150	200
1	647,8	799,5	864,2	899,6	921,8	937,1	948,2	956,7	963,3	968,6	984,9	993,1	998	1001	1006	1008	1013	1015	1016
2	38,51	39,00	39,17	39,25	39,30	39,33	39,36	39,37	39,39	39,40	39,43	39,45	39,5	39,46	39,47	39,5	39,5	39,5	39,5
3	17,44	16,04	15,44	15,10	14,88	14,73	14,62	14,54	14,47	14,42	14,25	14,17	14,1	14,08	14,04	14,0	14,0	13,9	13,9
4	12,22	10,65	9,98	9,60	9,36	9,20	9,07	8,98	8,90	8,84	8,66	8,56	8,50	8,46	8,41	8,38	8,32	8,30	8,29
5	10,01	8,43	7,76	7,39	7,15	6,98	6,85	6,76	6,68	6,62	6,43	6,33	6,27	6,23	6,18	6,14	6,08	6,06	6,05
6	8,81	7,26	6,60	6,23	5,99	5,82	5,70	5,60	5,52	5,46	5,27	5,17	5,11	5,07	5,01	4,98	4,92	4,89	4,88
7	8,07	6,54	5,89	5,52	5,29	5,12	4,99	4,90	4,82	4,76	4,57	4,47	4,40	4,36	4,31	4,28	4,21	4,19	4,18
8	7,57	6,06	5,42	5,05	4,82	4,65	4,53	4,43	4,36	4,30	4,10	4,00	3,94	3,89	3,84	3,81	3,74	3,72	3,70
9	7,21	5,71	5,08	4,72	4,48	4,32	4,20	4,10	4,03	3,96	3,77	3,67	3,60	3,56	3,51	3,47	3,40	3,38	3,37
10	6,94	5,46	4,83	4,47	4,24	4,07	3,95	3,85	3,78	3,72	3,52	3,42	3,35	3,31	3,26	3,22	3,15	3,13	3,12
11	6,72	5,26	4,63	4,28	4,04	3,88	3,76	3,66	3,59	3,53	3,33	3,23	3,16	3,12	3,06	3,03	2,96	2,93	2,92
12	6,55	5,10	4,47	4,12	3,89	3,73	3,61	3,51	3,44	3,37	3,18	3,07	3,01	2,96	2,91	2,87	2,80	2,78	2,76
13	6,41	4,97	4,35	4,00	3,77	3,60	3,48	3,39	3,31	3,25	3,05	2,95	2,88	2,84	2,78	2,74	2,67	2,65	2,63
14	6,30	4,86	4,24	3,89	3,66	3,50	3,38	3,29	3,21	3,15	2,95	2,84	2,78	2,73	2,67	2,64	2,56	2,54	2,53
15	6,20	4,77	4,15	3,80	3,58	3,41	3,29	3,20	3,12	3,06	2,86	2,76	2,69	2,64	2,59	2,55	2,47	2,45	2,44
16	6,12	4,69	4,08	3,73	3,50	3,34	3,22	3,12	3,05	2,99	2,79	2,68	2,61	2,57	2,51	2,47	2,40	2,37	2,36
17	6,04	4,62	4,01	3,66	3,44	3,28	3,16	3,06	2,98	2,92	2,72	2,62	2,55	2,50	2,44	2,41	2,33	2,30	2,29
18	5,98	4,56	3,95	3,61	3,38	3,22	3,10	3,01	2,93	2,87	2,67	2,56	2,49	2,44	2,38	2,35	2,27	2,24	2,23
19	5,92	4,51	3,90	3,56	3,33	3,17	3,05	2,96	2,88	2,82	2,62	2,51	2,44	2,39	2,33	2,30	2,22	2,19	2,18
20	5,87	4,46	3,86	3,51	3,29	3,13	3,01	2,91	2,84	2,77	2,57	2,46	2,40	2,35	2,29	2,25	2,17	2,14	2,13
21	5,83	4,42	3,82	3,48	3,25	3,09	2,97	2,87	2,80	2,73	2,53	2,42	2,36	2,31	2,25	2,21	2,13	2,10	2,09
22	5,79	4,38	3,78	3,44	3,22	3,05	2,93	2,84	2,76	2,70	2,50	2,39	2,32	2,27	2,21	2,17	2,09	2,06	2,05
23	5,75	4,35	3,75	3,41	3,18	3,02	2,90	2,81	2,73	2,67	2,47	2,36	2,29	2,24	2,18	2,14	2,06	2,03	2,01
24	5,72	4,32	3,72	3,38	3,15	2,99	2,87	2,78	2,70	2,64	2,44	2,33	2,26	2,21	2,15	2,11	2,02	2,00	1,98
25	5,69	4,29	3,69	3,35	3,13	2,97	2,85	2,75	2,68	2,61	2,41	2,30	2,23	2,18	2,12	2,08	2,00	1,97	1,95
26	5,66	4,27	3,67	3,33	3,10	2,94	2,82	2,73	2,65	2,59	2,39	2,28	2,21	2,16	2,09	2,05	1,97	1,94	1,92
27	5,63	4,24	3,65	3,31	3,08	2,92	2,80	2,71	2,63	2,57	2,36	2,25	2,18	2,13	2,07	2,03	1,94	1,91	1,90
28	5,61	4,22	3,63	3,29	3,06	2,90	2,78	2,69	2,61	2,55	2,34	2,23	2,16	2,11	2,05	2,01	1,92	1,89	1,88
29	5,59	4,20	3,61	3,27	3,04	2,88	2,76	2,67	2,59	2,53	2,32	2,21	2,14	2,09	2,03	1,99	1,90	1,87	1,86
30	5,57	4,18	3,59	3,25	3,03	2,87	2,75	2,65	2,57	2,51	2,31	2,20	2,12	2,07	2,01	1,97	1,88	1,85	1,84
40	5,42	4,05	3,46	3,13	2,90	2,74	2,62	2,53	2,45	2,39	2,18	2,07	1,99	1,94	1,88	1,83	1,74	1,71	1,69
60	5,29	3,93	3,34	3,01	2,79	2,63	2,51	2,41	2,33	2,27	2,06	1,94	1,87	1,82	1,74	1,70	1,60	1,56	1,54
70	5,25	3,89	3,31	2,97	2,75	2,59	2,47	2,38	2,30	2,24	2,03	1,91	1,83	1,78	1,71	1,66	1,56	1,52	1,50
100	5,18	3,83	3,25	2,92	2,70	2,54	2,42	2,32	2,24	2,18	1,97	1,85	1,77	1,71	1,64	1,59	1,48	1,44	1,42
150	5,13	3,78	3,20	2,87	2,65	2,49	2,37	2,28	2,20	2,13	1,92	1,80	1,72	1,67	1,59	1,54	1,42	1,38	1,35
200	5,10	3,76	3,18	2,85	2,63	2,47	2,35	2,26	2,18	2,11	1,90	1,78	1,70	1,64	1,56	1,51	1,39	1,35	1,32
1000	5,04	3,70	3,13	2,80	2,58	2,42	2,30	2,20	2,13	2,06	1,85	1,72	1,64	1,58	1,50	1,45	1,32	1,26	1,23

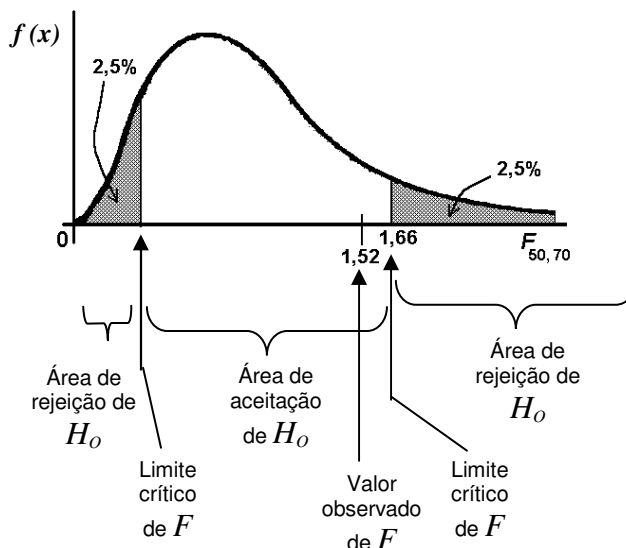
Como existem numerosos formatos para a distribuição F , a depender dos graus de liberdade do numerador e do denominador, o diagrama que aparece logo acima da tabela é apenas ilustrativo, mas do mesmo modo nos ajuda a compreender o conteúdo da tabela.

Achamos o valor crítico 1,66 na célula da tabela F na qual a coluna correspondente aos nossos graus de liberdade do numerador, g.l.=50, se encontra com a linha correspondente aos nossos graus de liberdade do denominador, g.l.=70. Observe que, de propósito, estipulamos os tamanhos das nossas amostras em 51 e 71, justamente porque já sabíamos que ao calcularmos os g.l. obteríamos valores existentes em nossa tabela, 50 e 70. Mas não se preocupe, porque os programas estatísticos que você utilizará no computador terão as tabelas completas de todas as distribuições utilizadas para inferência estatística;

5ª) Comparação do valor observado de F ao valor crítico de F e conclusão do teste:

Como $1,52 < 1,66$, concluiremos que o limite de significância estatística não foi ultrapassado. A razão de variâncias encontrada está em uma área de aceitação da hipótese nula, em uma área, portanto, de valores esperados por variação amostral. O teste nos mostra que a razão entre as variâncias (e, conseqüentemente, entre os desvios-padrão) dos tempos médios de serviço encontrada no único estudo realizado é muito compatível com uma igualdade entre as variâncias na população da qual os trabalhadores

comparados foram retirados. Há uma grande probabilidade dessas variâncias serem iguais nessa população. O diagrama abaixo apresenta a situação nesse teste:



Observe que existe também um limite crítico para a extremidade esquerda da distribuição, já que o teste que estamos realizando é bicaudado. Seu valor não foi apresentado porque a tabela F para o percentil 2,5 não é facilmente encontrada.

Poderíamos também ter calculado o valor- p e tê-lo comparado ao valor de α , sendo essa a estratégia mais utilizada pelos programas estatísticos para computador, mas o usual nos livros de bioestatística são tabelas com limites críticos de F . Certamente o valor- p nesse teste é maior do que 5,0%, o que nos levaria à mesma conclusão acima.

Outra opção ainda seria calcularmos o intervalo de 95% de confiança para razões de variâncias, mas para nosso objetivo nesse momento já é inteiramente suficiente o teste de hipóteses realizado acima. Nossa conclusão também seria a mesma.

Agora podemos assumir que, no nosso exemplo, as amostras são independentes; foram escolhidas aleatoriamente de populações específicas (uma do setor administrativo outra do produtivo), nas quais os tempos de serviço podem ser assumidos como normalmente distribuídos; seus n 's são grandes; e os desvios-padrão dos tempos médios de serviço são desconhecidos, mas estes podem ser considerados estatisticamente iguais, com base no teste que acabamos de realizar. Podemos, então, continuar nosso teste sobre a diferença entre os tempos médios de serviço.

Considerando as circunstâncias acima, o teste apropriado é o teste t , a ser realizado seguindo as seguintes etapas:

- 1ª) Definição do nível de significância: $\alpha = 0,05$;
- 2ª) Definição das hipóteses: $H_0 : \mu_1 - \mu_2 = 0$ e $H_A : \mu_1 - \mu_2 \neq 0$;
- 3ª) Cálculo do valor de t :

Como, nesse caso, os desvios-padrão (variâncias) populacionais são desconhecidos, vamos utilizar em seu lugar os desvios-padrão (variâncias) obtidos nas duas amostras comparadas pelo único estudo realizado. Mas, além disso, como verificamos que, com base nos valores dessas variâncias amostrais,

podemos assumir que as variâncias populacionais são iguais, vamos usar como variância uma variância comum aos dois grupos comparados (já que essas são estatisticamente iguais), e esta variância comum será dada por

$$s_c^2 = \frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{n_1 + n_2 - 2},$$

onde o subscrito c em s_c^2 indica que se trata de uma variância comum aos dois grupos, ou equivalentemente, uma variância combinada, porque combina as variâncias dos dois grupos; $(n_1 - 1)$ indica os graus de liberdade da amostra do setor administrativo; e $(n_2 - 1)$ os graus de liberdade da amostra do setor de produção.

Observe que o estimador acima é uma média ponderada. Lembra-se de quando vimos o assunto “média ponderada” no capítulo 5 (páginas 46 e 47)? Lá nós utilizamos o exemplo do cálculo da nota final de um(a) estudante universitário(a) de uma universidade pública brasileira, que é feito por

$$\text{Nota final} = \frac{6(\text{média nas avaliações parciais}) + 4(\text{nota na prova final})}{6 + 4}.$$

Verifique que é dado peso seis à média nas avaliações parciais; peso quatro à nota na prova final; e no denominador é colocada a soma desses pesos que é $6 + 4 = 10$. Veja que temos os mesmos termos na fórmula para o cálculo da variância combinada: $(n_1 - 1)$ é o peso dado à variância obtida na amostra do setor administrativo; $(n_2 - 1)$ é o peso dado à variância obtida na amostra do setor de produção; e no denominador é colocada a soma desses pesos, já que

$$(n_1 - 1) + (n_2 - 1) = n_1 - 1 + n_2 - 1 = n_1 + n_2 - 2, \text{ certo?}$$

A intenção é calcular uma variância média entre os dois grupos, que leve em conta o número de indivíduos em cada grupo, pois se um grupo tem mais, ou menos, indivíduos, isso irá influenciar sua variância e, portanto, isso deve ser considerado (ponderado) no cálculo.

Faremos então o cálculo do valor de t substituindo σ_1^2 e σ_2^2 por s_c^2 , conforme mostramos a seguir:

$$\text{em vez de } z = \frac{(\bar{x}_1 - \bar{x}_2) - (\mu_1 - \mu_2)_0}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}} \text{ utilizaremos } t = \frac{(\bar{x}_1 - \bar{x}_2) - (\mu_1 - \mu_2)_0}{\sqrt{\frac{s_c^2}{n_1} + \frac{s_c^2}{n_2}}}.$$

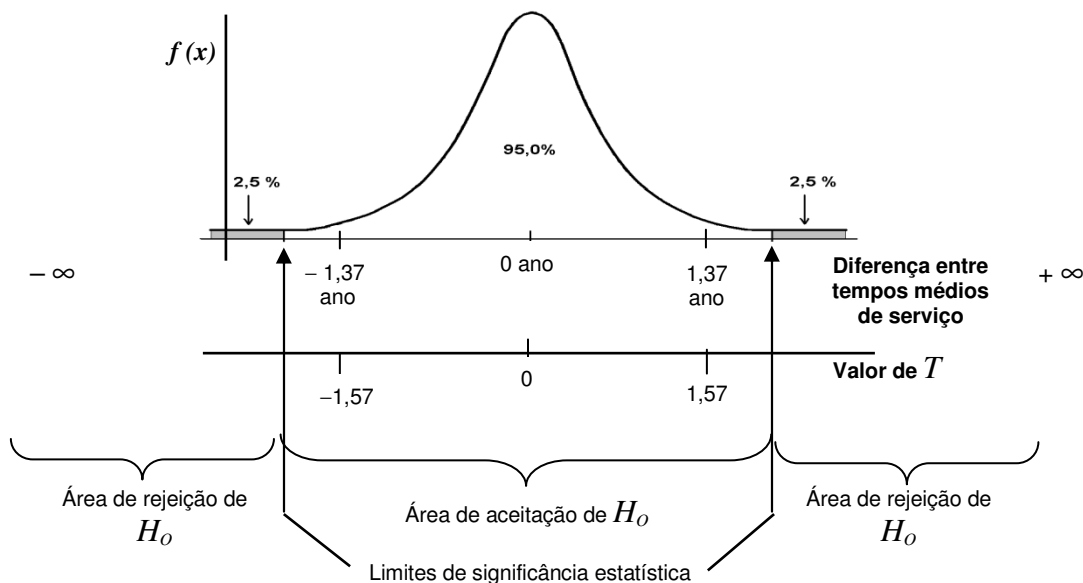
Calculando s_c^2 obtemos:

$$\begin{aligned} s_c^2 &= \frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{n_1 + n_2 - 2} = \frac{(51 - 1)(5,28)^2 + (71 - 1)(4,28)^2}{51 + 71 - 2} = \\ &= \frac{(50)(27,88) + (70)(18,32)}{120} = \frac{1.394,00 + 1.282,40}{120} = \frac{2.676,40}{120} = 22,3 \text{ anos}^2. \end{aligned}$$

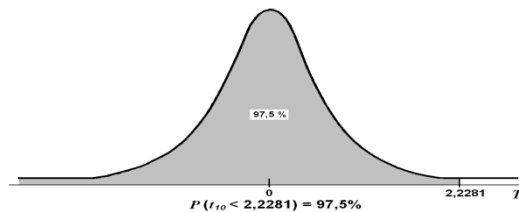
Substituindo-se s_c^2 na expressão utilizada para o cálculo de t encontramos:

$$t = \frac{(\bar{x}_1 - \bar{x}_2) - (\mu_1 - \mu_2)_0}{\sqrt{\frac{s_c^2}{n_1} + \frac{s_c^2}{n_2}}} = \frac{(14,66 - 13,29) - 0}{\sqrt{\frac{22,3}{51} + \frac{22,3}{71}}} = \frac{1,37 - 0}{\sqrt{0,44 + 0,31}} = \frac{1,37}{\sqrt{0,75}} = \frac{1,37}{0,87} \cong 1,57;$$

4ª) Obtenção do valor- p : no capítulo anterior, vimos que a tabela T (reapresentada na próxima página) comumente utilizada, contém valores críticos (porcentis) e não valores- p . Mas, podemos ver indiretamente nessa tabela que o valor- p para $t = 1,57$ está entre 10,0% e 5,0%. Vemos isso, primeiramente verificando a linha da tabela correspondente ao número de graus de liberdade do nosso teste, que é dado por $n_1 + n_2 - 2$, sendo no nosso exemplo igual a $n_1 + n_2 - 2 = 51 + 71 - 2 = 122 - 2 = 120$. Em seguida, olhando nessa linha, da esquerda para a direita, note que o valor de t calculado por nós, 1,57, está entre os valores críticos 1,2887 e 1,6577. Agora, olhando lá em cima da coluna na qual estão estes valores, veja que estão escritos P_{90} e P_{95} . Por isso, podemos afirmar que a probabilidade de obtermos valores maiores do que 1,57 está entre 10,0% e 5,0% ou, como estamos mais acostumados, entre 5,0% e 10,0%. Como o teste é bicaudado, temos que considerar o dobro desse valor, que nesse caso, é o dobro de uma probabilidade cujo valor está entre 5,0% e 10,0%. Podemos afirmar então que o nosso valor- p está entre 10,0 e 20,0%. Acontece que, dessa maneira, não podemos afirmar com precisão qual é o valor- p , porque estamos utilizando uma tabela incompleta. Não fique preocupado com isso, pois você já sabe que os programas estatísticos que utilizará no computador contêm tabelas completas e lhe fornecerão valores- p precisos. No diagrama abaixo, tentamos lhe ajudar a entender nossa situação atual:



Reapresentamos a tabela T a seguir:



Graus de liberdade ($n_1 + n_2 - 2$)	P_{90}	P_{95}	$P_{97,5}$	P_{99}	$P_{99,5}$
1	3,0780	6,3138	12,7060	31,8210	63,6570
2	1,8860	2,9200	4,3027	6,9650	9,9248
3	1,6380	2,3534	3,1825	4,5410	5,8409
4	1,5330	2,1318	2,7764	3,7470	4,6041
5	1,4760	2,0150	2,5706	3,3650	4,0321
6	1,4400	1,9432	2,4469	3,1430	3,7074
7	1,4150	1,8946	2,3646	2,9980	3,4995
8	1,3970	1,8595	2,3060	2,8960	3,3554
9	1,3830	1,8331	2,2622	2,8210	3,2498
10	1,3720	1,8125	2,2281	2,7640	3,1693
11	1,3630	1,7959	2,2010	2,7180	3,1058
12	1,3560	1,7823	2,1788	2,6810	3,0545
13	1,3500	1,7709	2,1604	2,6500	3,0123
14	1,3450	1,7613	2,1448	2,6240	2,9768
15	1,3410	1,7530	2,1315	2,6020	2,9467
16	1,3370	1,7459	2,1190	2,5830	2,9208
17	1,3330	1,7396	2,1098	2,5670	2,8982
18	1,3300	1,7341	2,1009	2,5520	2,8784
19	1,3280	1,7291	2,0930	2,5390	2,8609
20	1,3250	1,7247	2,0860	2,5280	2,8453
21	1,3230	1,7207	2,0796	2,5180	2,8314
22	1,3210	1,7171	2,0739	2,5080	2,8188
23	1,3190	1,7139	2,0687	2,5000	2,8073
24	1,3180	1,7109	2,0639	2,4920	2,7969
25	1,3160	1,7081	2,0595	2,4850	2,7874
26	1,3150	1,7056	2,0555	2,4790	2,7787
27	1,3140	1,7033	2,0518	2,4730	2,7707
28	1,3130	1,7011	2,0484	2,4670	2,7633
29	1,3110	1,6991	2,0452	2,4620	2,7564
30	1,3100	1,6973	2,0423	2,4570	2,7500
35	1,3062	1,6896	2,0301	2,4380	2,7239
40	1,3031	1,6839	2,0211	2,4230	2,7045
45	1,3007	1,6794	2,0141	2,4120	2,6896
50	1,2987	1,6759	2,0086	2,4030	2,6778
60	1,2959	1,6707	2,0003	2,3900	2,6603
70	1,2938	1,6669	1,9945	2,3810	2,6480
80	1,2922	1,6641	1,9901	2,3740	2,6388
90	1,2910	1,6620	1,9867	2,3680	2,6316
100	1,2901	1,6602	1,9840	2,3640	2,6260
120	1,2887	1,6577	1,9799	2,3580	2,6175
140	1,2876	1,6558	1,9771	2,3530	2,6114
160	1,2869	1,6545	1,9749	2,3500	2,6070
180	1,2863	1,6534	1,9733	2,3470	2,6035
200	1,2858	1,6525	1,9719	2,3450	2,6006
∞	1,2820	1,6450	1,9600	2,3260	2,5760

No diagrama acima da tabela, $P(t_{10} < 2,2281) = 97,5$ indica a área sombreada que, por sua vez, representa a probabilidade de obtermos um valor de T menor do que 2,2281, para um teste bicaudado, com 10 graus de liberdade, e $\alpha = 5,0\%$; Como existe uma “família” de tabelas T , esse diagrama específico foi escolhido apenas para nos orientar sobre o conteúdo da tabela;

5ª) a) Comparação do valor p ao valor de α e conclusão do teste:

Já vimos nesse exemplo, que a probabilidade de obtermos, considerando os dois extremos da curva, valores menores ou maiores do que uma diferença entre médias igual a 1,37 ano, encontra-se entre

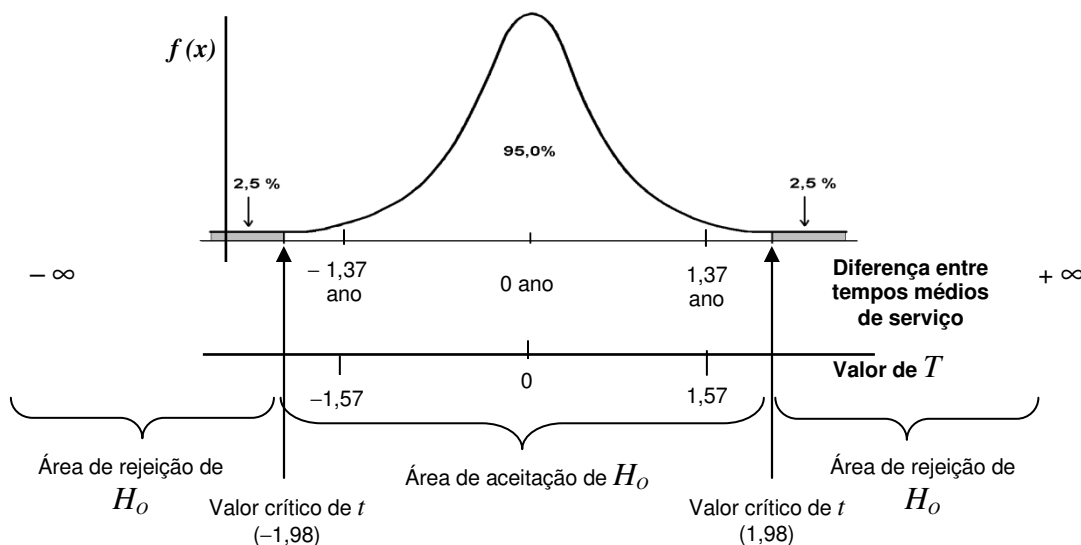
10,0% e 20,0%. Nosso valor- p é então, no mínimo 10,0%. Então nosso valor- p é maior do que 5,0%, que é o nosso α , nos indicando que seria muito alta a probabilidade de obtermos uma diferença igual a 1,37 ano entre os tempos médios de serviço, caso tivéssemos realizado numerosos estudos. Essa diferença será então considerada por nós como podendo ter ocorrido por simples variação amostral, pois encontra-se na área de aceitação da hipótese nula. Assim, se a pesquisa realizada não apresenta vieses importantes, podemos aceitar H_0 e rejeitar H_A , concluindo que, muito provavelmente, a verdadeira diferença entre os tempos médios na população de onde as amostras foram retiradas é zero, pois a diferença obtida no único estudo realizado é muito compatível com isto; ou

b) Comparação do valor observado de t ao valor crítico de t :

Considerando que o nosso teste é bicaudado, devido às hipóteses que foram testadas, que o nível de significância estabelecido foi 5,0%, e o número de graus de liberdade é 120, olhando na tabela T vemos que o valor crítico de t é $1,9799 \cong 1,98$. Note que o valor crítico é aquele para o percentil 97,5, porque se temos de considerar 2,5% em cada cauda da distribuição, temos que 100,0% (que corresponde ao total da área sob a curva) menos 2,5% é igual a 97,5%.

A conclusão que tirarmos para uma cauda da curva será a mesma para a outra cauda, pois o valor crítico na outra extremidade terá o mesmo valor, só mudando o sinal para negativo ($-1,9799 \cong -1,98$).

Como $1,57 < 1,98$ e, conseqüentemente, $-1,57 > -1,98$ (veja diagrama abaixo), concluiremos que o limite de significância estatística não foi ultrapassado, que a diferença encontrada está em uma área de aceitação da hipótese nula, em uma área também de valores esperados por variação amostral, indicando que o afastamento da diferença 1,37 em relação a uma possível diferença populacional igual a zero não é grande o suficiente para ser considerada como pertencente a uma outra população de trabalhadores com diferença entre as médias diferente de zero. Em última instância, o teste nos mostra que não há uma diferença estatisticamente significativa entre a verdadeira diferença de tempos médios e zero, indicando que na população de onde os trabalhadores foram selecionados, há uma grande probabilidade de a diferença ser zero, porque a diferença obtida no único estudo realizado é muito compatível com isso. Essa conclusão pressupõe que os resultados encontrados não sejam devidos a falhas (vieses) existentes no estudo. Observe que, como já esperávamos, a conclusão nos itens **a** e **b** da 5ª etapa do nosso teste foi a mesma.



Chegaremos à mesma conclusão se, em vez de fazermos um teste de hipóteses, realizarmos o cálculo do intervalo de 95% de confiança. Esse intervalo será dado por

$$IC(95\%) = (\bar{x}_1 - \bar{x}_2) \pm t_{(1-\alpha/2)} \sqrt{\frac{s_c^2}{n_1} + \frac{s_c^2}{n_2}}$$

Diagram illustrating the components of the 95% confidence interval formula:

- $IC(95\%)$: Intervalo de 95% de confiança
- $(\bar{x}_1 - \bar{x}_2)$: Diferença entre as médias obtidas no estudo
- $t_{(1-\alpha/2)}$: Valor de t correspondente ao percentil 97,5
- $\sqrt{\frac{s_c^2}{n_1} + \frac{s_c^2}{n_2}}$: Erro-padrão das diferenças entre médias caso tivéssemos feito vários estudos

Como o nosso nível de significância, α , é 0,05, temos que

$$t_{(1-\alpha/2)} = t_{(1-0,05/2)} = t_{1-0,025} = t_{0,975} = t_{97,5\%}.$$

Considerando os valores do nosso exemplo, obtemos o seguinte intervalo:

$$IC(95\%) = (\bar{x}_1 - \bar{x}_2) \pm t_{(1-\alpha/2)} \sqrt{\frac{s_c^2}{n_1} + \frac{s_c^2}{n_2}} = (14,66 - 13,29) \pm t_{0,975} \sqrt{\frac{22,3}{51} + \frac{22,3}{71}} =$$

$$= 1,37 \pm (1,98)(0,87) = 1,37 \pm 1,72 = (-0,35 \text{ a } 3,09).$$

— Como interpretar este resultado?

— Nossa referência para a conclusão da inferência é a diferença entre médias que seria obtida caso elas fossem iguais. É evidente que esse valor é zero, porque se duas quantidades iguais são diminuídas uma da outra, o resultado é zero. Então vamos verificar se o valor zero, que indica ausência de diferença, está ou não contido no intervalo de 95% de confiança que acabamos de calcular. Lembre-se de que há uma probabilidade de 95% da verdadeira diferença populacional dos tempos médios estar entre $-0,35$ ano e $3,09$ anos. Os valores dentro desse intervalo são aceitáveis para essa verdadeira diferença. São valores que não ultrapassam os limites de significância estatística localizados nas duas extremidades da curva. E, como zero é um dos valores dentro desse intervalo, concluímos que a diferença na população pode ser zero.

Resumindo:

Quando compararmos médias de duas amostras que sejam: a) independentes; b) escolhidas de modo aleatório de populações específicas, similares aos grupos a serem comparados; c) retiradas de populações nas quais a variável estudada possa ser assumida como normalmente distribuída; d) grandes; e) selecionadas de populações cujos desvios-padrão da variável investigada não sejam conhecidos; e f) esses desvios possam ser considerados estatisticamente iguais; realizaremos a inferência estatística através do teste t , que se aproximará muito do teste z , porque os n 's são grandes.

— E se os desvios (variâncias) não puderem ser assumidos como estatisticamente iguais?

— Nesse caso teremos de utilizar o teste t' .

Voltando ao nosso exemplo, mostraremos como esse teste é realizado.

Suponha que em vez dos desvios-padrão para as amostras estudadas terem sido 5,28 e 4,28 anos tenham sido 6,43 e 4,28 anos.

O primeiro passo será a verificação da existência de diferença estatística entre esses dois desvios, através do teste da razão de variâncias, seguindo as seguintes etapas:

1ª) Definição do nível de significância: $\alpha = 0,05$;

2ª) Definição das hipóteses:

Denotando as variâncias das populações por σ_1^2 e σ_2^2 , nossas hipóteses são:

$$H_O : \sigma_1^2 = \sigma_2^2 \text{ e } H_A : \sigma_1^2 \neq \sigma_2^2;$$

3ª) Cálculo do valor de F :

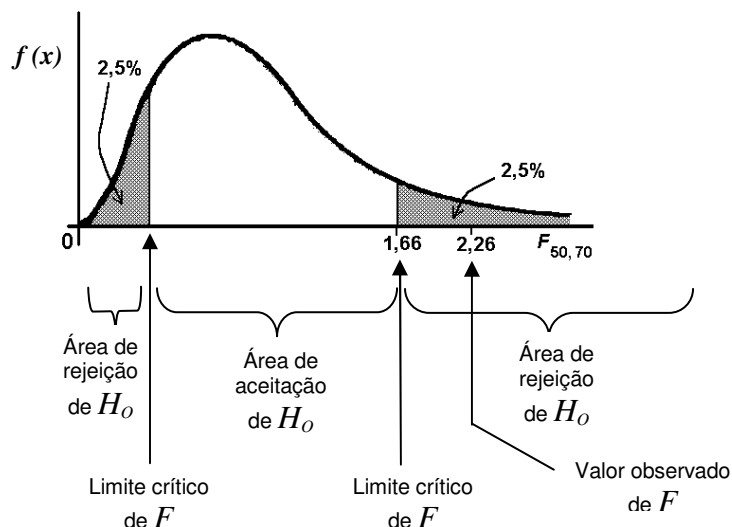
Lembre-se de que estamos discutindo uma situação na qual os desvios-padrão e, portanto, as variâncias populacionais são desconhecidas. Temos então de utilizar como estimadores desses parâmetros populacionais as variâncias, s_1^2 e s_2^2 , encontradas no único estudo realizado.

Podemos então prosseguir nosso teste da igualdade entre as variâncias, calculando o valor de F :

$$F = \frac{s_1^2}{s_2^2} = \frac{(6,43)^2}{(4,28)^2} = \frac{41,34}{18,32} = 2,26;$$

4ª) Obtenção, na tabela F , do valor crítico de F para um nível de significância de 0,05, teste bicaudado, $n_1 - 1 = 51 - 1 = 50$ graus de liberdade do numerador, e $n_2 - 1 = 71 - 1 = 70$ graus de liberdade do denominador. Olhando na tabela (página 216), vemos que o limite crítico de F nesse nosso teste continua sendo 1,66, como esperávamos;

Veja nossa situação atual no diagrama abaixo:



5ª) Comparação do valor observado de F ao valor crítico de F e conclusão do teste:

Como $2,26 > 1,66$, concluiremos que o limite de significância estatística foi ultrapassado. A razão de variâncias encontrada está em uma área de rejeição da hipótese nula, em uma área, portanto, de valores muito improváveis de serem obtidos em amostras retiradas de uma população cujas variâncias sejam iguais. A razão de variâncias obtida é grande o suficiente para concluirmos que há uma grande probabilidade das variâncias na população de onde os trabalhadores estudados foram retirados serem diferentes, porque a razão obtida, 2,26, é muito compatível com isso. Como sempre, essa conclusão pressupõe que os resultados encontrados não possam ser explicados por falhas (vieses) existentes no estudo.

Acabamos de verificar que não podemos assumir que os desvios-padrão sejam iguais e, por isso, temos de aplicar o teste t' , cujas etapas são as seguintes:

1ª) Definição do nível de significância: $\alpha = 0,05$;

2ª) Definição das hipóteses: $H_O : \mu_1 - \mu_2 = 0$ e $H_A : \mu_1 - \mu_2 \neq 0$;

3ª) Cálculo do valor de t' :

Como, nesse caso também, os desvios-padrão (variâncias) populacionais são desconhecidos, vamos utilizar em seu lugar os desvios-padrão (variâncias) obtidos nas duas amostras comparadas pelo único estudo realizado. Mas como, além disso, verificamos que, com base nos valores dessas variâncias amostrais, não podemos assumir que as variâncias populacionais são iguais, não podemos usar uma variância comum aos dois grupos comparados, como fizemos no teste anterior.

Faremos então o cálculo do valor de t' substituindo σ_1^2 e σ_2^2 por s_1^2 e s_2^2 , respectivamente, conforme mostramos a seguir:

$$\text{em vez de } t = \frac{(\bar{x}_1 - \bar{x}_2) - (\mu_1 - \mu_2)_o}{\sqrt{\frac{s_c^2}{n_1} + \frac{s_c^2}{n_2}}} \text{ utilizaremos } t' = \frac{(\bar{x}_1 - \bar{x}_2) - (\mu_1 - \mu_2)_o}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}}.$$

Substituindo com os valores do nosso exemplo temos que

$$\begin{aligned} t' &= \frac{(\bar{x}_1 - \bar{x}_2) - (\mu_1 - \mu_2)_o}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}} = \frac{(14,66 - 13,29) - 0}{\sqrt{\frac{(6,43)^2}{51} + \frac{(4,28)^2}{71}}} = \frac{1,37 - 0}{\sqrt{\frac{41,34}{51} + \frac{18,32}{71}}} = \\ &= \frac{1,37}{\sqrt{0,81 + 0,26}} = \frac{1,37}{\sqrt{1,07}} = \frac{1,37}{1,03} \cong 1,33; \end{aligned}$$

4ª) Obtenção do valor- p :

No capítulo anterior, vimos que a tabela T comumente utilizada contém valores críticos (porcentis) e, não, valores- p . Desse modo, para simplificar, vamos prosseguir o nosso teste sem encontrarmos o valor- p , comparando o valor de t' observado ao valor crítico de t' .

Há, porém, outra dificuldade na situação atual, que é o fato da quantidade

$$t' = \frac{(\bar{x}_1 - \bar{x}_2) - (\mu_1 - \mu_2)_o}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}}$$

não se distribuir conforme a curva T , o que nos impede de utilizar essa curva no nosso teste. Não fique preocupado com essas dificuldades, pois você já sabe que os programas estatísticos que utilizará, contém tabelas completas e lhe fornecerão valores- p precisos e com uma correção adequada, de modo que a curva T possa ser utilizada;

O fato da quantidade $t' = (\bar{x}_1 - \bar{x}_2) - (\mu_1 - \mu_2)_o / \sqrt{(s_1^2/n_1) + (s_2^2/n_2)}$ não seguir uma distribuição T (referido pelos estatísticos como o problema de Behrens-Fisher), tem que ser levado em conta também para encontrarmos o valor crítico adequado. Cochran (*William G. Cochran. Approximate significance levels of the Behrens-Fisher test. Biometrics, 20:191-195, 1964*) sugeriu que esse problema seja resolvido calculando-se

$$t'_{(1-\alpha/2)} = \frac{w_1 t_1 + w_2 t_2}{w_1 + w_2},$$

onde $t'_{(1-\alpha/2)}$ é o valor crítico de t' , $w_1 = s_1^2/n_1$, $w_2 = s_2^2/n_2$, $t_1 = t_{(1-\alpha/2)}$ para $n_1 - 1$ graus de liberdade, e $t_2 = t_{(1-\alpha/2)}$ para $n_2 - 1$ graus de liberdade. Observe que t' é uma média ponderada dos valores de T para os grupos de trabalhadores estudados. Os pesos são w_1 e w_2 , que representam a variância de cada grupo dividida pelo número de trabalhadores em cada um. O denominador é a soma dos pesos, $w_1 + w_2$. Prossiga para você entender melhor.

No nosso exemplo, considerando que

$$w_1 = s_1^2/n_1 = 41,34/51 = 0,81;$$

$$w_2 = s_2^2/n_2 = 18,32/71 = 0,26;$$

$$t_1 = t_{0,975, 50} = 2,0086 \cong 2,01; \text{ (encontramos esse valor na tabela } T)$$

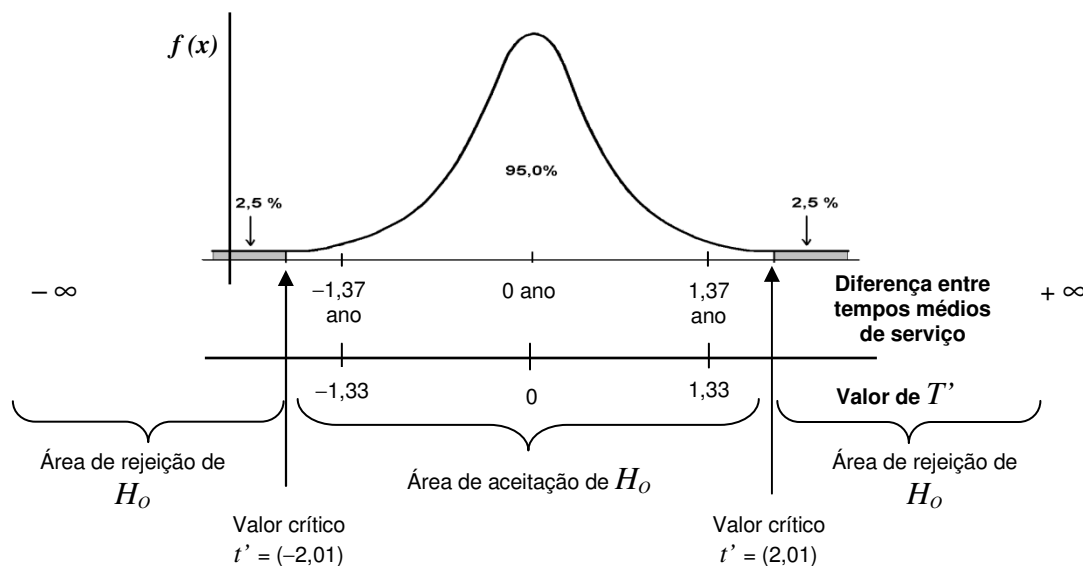
$$\text{e } t_2 = t_{0,975, 70} = 1,9945 \cong 1,99; \text{ (encontramos esse valor na tabela } T); \text{ temos que}$$

$$t'_{(1-\alpha/2)} = \frac{w_1 t_1 + w_2 t_2}{w_1 + w_2} = \frac{(0,81)(2,01) + (0,26)(1,99)}{0,81 + 0,26} = \frac{1,63 + 0,52}{1,07} = \frac{2,15}{1,07} = 2,01,$$

sendo esse o valor crítico de t' para o nosso teste atual.

A conclusão à qual chegarmos para a cauda direita da curva será a mesma para a esquerda, pois o valor crítico na outra extremidade terá o mesmo valor, só mudando o sinal de positivo para negativo ($-2,01$).

No diagrama abaixo tentamos lhe ajudar a entender nossa situação atual:



O problema de Behrens-Fisher tem várias outras soluções, entre estas a de Satterthwaite, que pode ser encontrada nas páginas 293 a 299 do livro *Hosner B. Fundamentals of biostatistics. 5ª ed. Pacific Grove (CA): Duxbury; 2000.*

5ª) Comparação do valor observado de t' ao valor crítico de t' :

Como $1,33 < 2,01$ e, conseqüentemente, $-1,33 > -2,01$, concluiremos que o limite de significância estatística não foi ultrapassado, que a diferença encontrada está em uma área de aceitação da hipótese nula, em uma área também de valores muito prováveis de serem obtidos em uma população cujos tempos médios de serviço dos trabalhadores administrativos e da produção sejam iguais. O afastamento da diferença $1,37$ ano em relação à diferença zero, não é grande o suficiente para que $1,37$ ano possa ser considerado um valor pertencente a uma outra população de trabalhadores, com diferença entre médias diferente de zero. Em última instância, o teste nos mostra que não há uma diferença estatisticamente significativa entre a verdadeira diferença entre os tempos médios nos dois setores da Refinaria e zero, indicando que na população de onde os trabalhadores foram selecionados, há uma grande probabilidade de a diferença ser zero. Essa conclusão pressupõe que os resultados encontrados não sejam devidos a falhas (vieses) existentes no estudo.

Nossa conclusão será a mesma se em vez de fazermos um teste de hipóteses, realizarmos o cálculo do intervalo de 95% de confiança. Esse intervalo será dado por

$$IC(95\%) = (\bar{x}_1 - \bar{x}_2) \pm t'_{(1-\alpha/2)} \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}$$

Intervalo de 95% de confiança

Diferença entre as médias obtidas no estudo

Valor de t' correspondente ao percentil 97,5

Erro-padrão das diferenças entre médias caso tivéssemos feito vários estudos

Note que o valor de t nessa fórmula foi substituído pelo valor crítico de t' .

Utilizando os valores do nosso exemplo, obtemos o seguinte intervalo:

$$\begin{aligned}
 IC(95\%) &= (\bar{x}_1 - \bar{x}_2) \pm t'_{(1-\alpha/2)} \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}} = (14,66 - 13,29) \pm t'_{0,975} \sqrt{\frac{41,34}{51} + \frac{18,32}{71}} = \\
 &= 1,37 \pm 2,01 \sqrt{0,81 + 0,26} = 1,37 \pm 2,01 \sqrt{1,07} = 1,37 \pm (2,01)(1,03) = \\
 &= 1,37 \pm 2,07 = (-0,70 \text{ a } 3,44).
 \end{aligned}$$

Nossa referência para a conclusão da inferência é a diferença entre médias igual a zero, que seria a diferença obtida caso as médias fossem iguais. Verificamos, então, se o valor zero está ou não contido no intervalo de 95% de confiança que acabamos de calcular. Lembre-se de que há uma probabilidade de 95% da verdadeira diferença estar entre $-0,70$ e $3,44$. Os valores dentro desse intervalo são, portanto, aceitáveis para a diferença populacional. São valores que não ultrapassam os limites de significância estatística nas duas extremidades da curva. E, como zero é um dos valores dentro desse intervalo, uma diferença igual a zero poderia ter sido obtida, caso tivéssemos realizado numerosos estudos ou investigado toda a população. Concluiremos que a diferença encontrada, $1,37$ ano, não é estatisticamente diferente de zero. Isto é o mesmo que concluirmos que, do ponto de vista estatístico, muito provavelmente, a diferença na população é zero.

Resumindo:

Quando compararmos médias de duas amostras que sejam: a) independentes; b) escolhidas de modo aleatório de populações específicas, similares aos grupos a serem comparados; c) retiradas de populações nas quais a variável estudada possa ser assumida como normalmente distribuída; d) grandes; e) retiradas de populações cujos desvios-padrão da variável investigada não sejam conhecidos; e f) esses desvios não possam ser considerados estatisticamente iguais; realizaremos a inferência estatística através do teste t' , que também apresentará conclusão semelhante à do teste z , após a aplicação da solução de Cochran, quanto maiores sejam os n 's.

A partir desse ponto revise no fluxograma (modificado de *Daniel WW. Biostatistics: a foundation for analysis in the health sciences*. 7ª ed. New York(NY): John Wiley;1999), no final deste capítulo, as situações existentes para comparação de duas médias já vistas até o momento, e acompanhe aquelas que passaremos a discutir em seguida. Isso lhe ajudará muito a entender qual o teste adequado a cada situação.

Outra situação possível é aquela na qual comparamos médias de duas amostras que: a) são independentes; b) foram selecionadas aleatoriamente de populações específicas, similares aos grupos a serem comparados; c) foram retiradas de populações nas quais a variável estudada pode ser assumida como normalmente distribuída; d) são pequenas; e) e foram retiradas de populações cujos desvios-padrão da variável investigada são conhecidos. Nessas circunstâncias, realizaremos inferência estatística através do teste z , independentemente desses desvios poderem ou não ser assumidos como iguais. Anteriormente, nesse mesmo capítulo, você já aprendeu a realizar esse teste.

Resumindo:

Se as duas amostras forem: a) independentes; b) selecionadas aleatoriamente de populações específicas, similares aos grupos a serem comparados; c) retiradas de populações nas quais a variável estudada possa ser assumida como normalmente distribuída; d) pequenas; e e) retiradas de populações cujos desvios-padrão da variável investigada sejam conhecidos; aplicaremos o teste z , independentemente desses desvios serem ou não iguais.

Outra situação ocorre quando comparamos médias de duas amostras que: a) são independentes; b) foram escolhidas aleatoriamente de populações específicas, similares aos grupos a serem comparados; c) foram retiradas de populações nas quais a variável estudada pode ser assumida como normalmente distribuída; d) são pequenas; e) e foram retiradas de populações cujos desvios-padrão da variável investigada não são conhecidos. Nessa situação, teremos que verificar, através do teste da razão entre variâncias, se podemos ou não assumir que os desvios-padrão sejam iguais. Se pudermos, realizaremos a inferência estatística através do teste t ; se não, aplicaremos o teste t' . Anteriormente, neste mesmo capítulo, você também já aprendeu a realizar tais testes.

Resumindo:

Se as duas amostras forem: a) independentes; b) selecionadas aleatoriamente de populações específicas, similares aos grupos a serem comparados; c) retiradas de populações nas quais a variável estudada possa ser assumida como normalmente distribuída; d) pequenas; e e) retiradas de populações cujos desvios-padrão da variável investigada não sejam conhecidos, aplicaremos o teste t se os desvios-padrão forem iguais, e o teste t' se forem diferentes.

Se compararmos médias de duas amostras que sejam: a) independentes; b) selecionadas aleatoriamente de populações específicas, similares aos grupos a serem comparados; c) retiradas de populações nas quais a variável estudada não possa ser assumida como normalmente distribuída; d) grandes; e e) retiradas de populações cujos desvios-padrão da variável investigada sejam conhecidos; utilizaremos o teste z , independentemente da igualdade ou não desses desvios. Nessa situação, aplica-se o teorema central do limite (página 149), porque o tamanho de cada amostra é grande, o que contorna o problema resultante do fato da variável estudada não se distribuir normalmente na população. Pode ser demonstrado, que o teorema central do limite é válido também para inferência sobre duas médias. Assim, caso o tamanho de cada amostra investigada seja grande ($n \geq 30$), a distribuição das diferenças entre as médias obtidas em numerosos

estudos será aproximadamente normal, com média $\mu_{(\bar{x}_1 - \bar{x}_2)} = \mu_1 - \mu_2$ e variância $\sigma_{(\bar{x}_1 - \bar{x}_2)}^2 = (\sigma_1^2/n_1) + (\sigma_2^2/n_2)$.

Resumindo:

Se as duas amostras forem: a) independentes; b) escolhidas aleatoriamente de populações específicas, similares aos grupos a serem comparados; c) retiradas de populações nas quais a variável estudada não possa ser assumida como normalmente distribuída; d) grandes; e e) retiradas de populações cujos desvios-padrão da variável investigada sejam conhecidos; utilizaremos o teste Z , independentemente da igualdade ou não desses desvios.

Para compararmos médias de duas amostras que: a) sejam independentes; b) tenham sido escolhidas aleatoriamente de populações específicas, similares aos grupos a serem comparados; c) retiradas de populações nas quais a variável estudada não possa ser assumida como normalmente distribuída; d) sejam grandes; e e) tenham sido retiradas de populações cujos desvios-padrão da variável investigada não sejam conhecidos; utilizaremos o teste t , se pudermos assumir que haja igualdade entre esses desvios, e o teste t' , se não pudermos. Nessa situação, aplica-se também o teorema central do limite, porque o tamanho de cada amostra é grande, o que contorna o problema resultante do fato da variável estudada não se distribuir normalmente na população.

Resumindo:

Se as duas amostras forem: a) independentes; b) escolhidas aleatoriamente de populações específicas, similares aos grupos a serem comparados; c) retiradas de populações nas quais a variável estudada não possa ser assumida como normalmente distribuída; d) grandes; e e) retiradas de populações cujos desvios-padrão da variável investigada não sejam conhecidos; utilizaremos o teste t se houver igualdade entre esses desvios, e o teste t' se não houver.

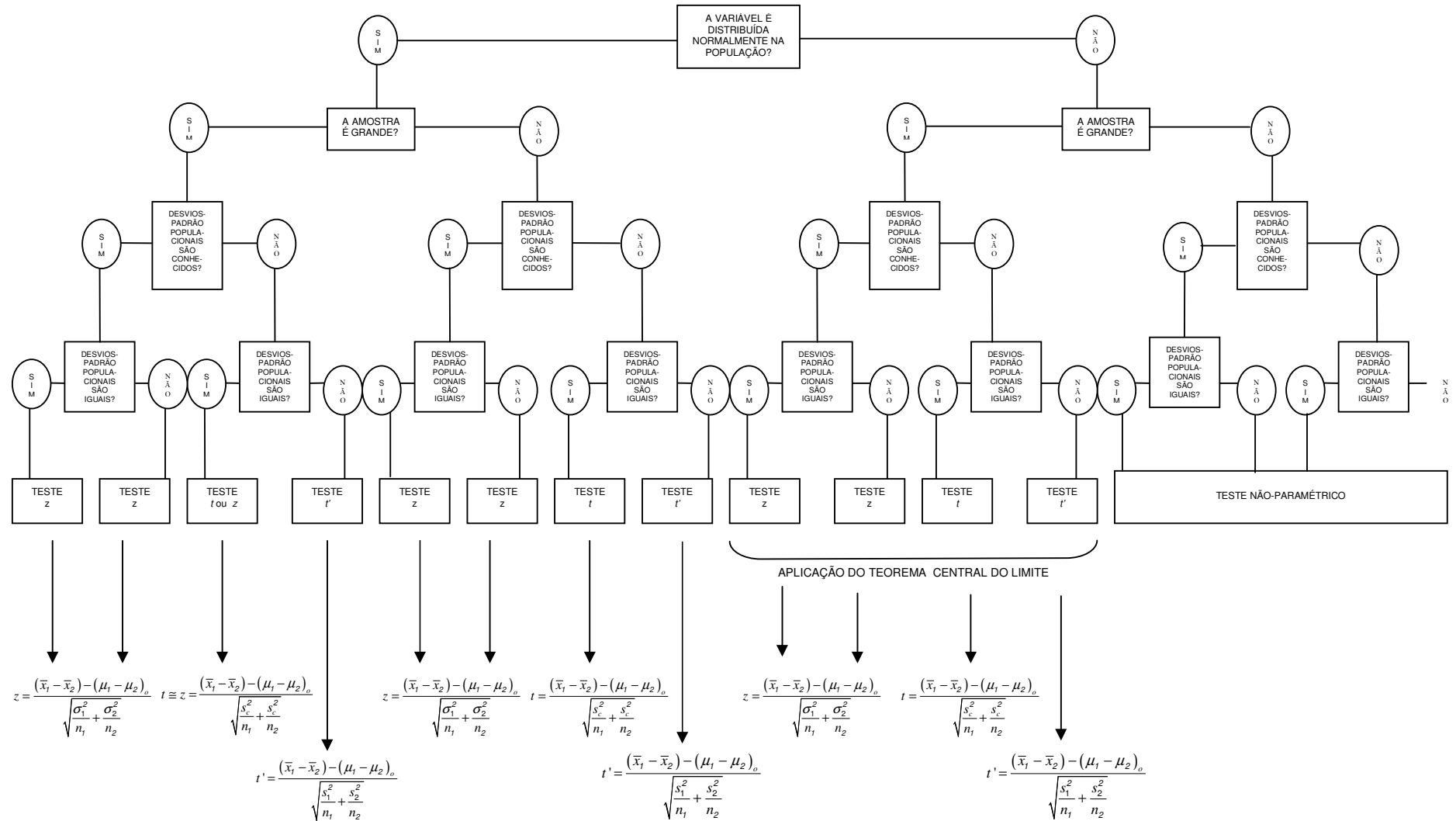
Ao compararmos médias de duas amostras que sejam: a) independentes; b) escolhidas aleatoriamente de populações específicas, similares aos grupos a serem comparados; c) retiradas de populações nas quais a variável estudada não possa ser assumida como normalmente distribuída; d) e pequenas; utilizaremos técnica estatística não-paramétrica, porque não poderemos aplicar o teorema central do limite, já que o tamanho de uma ou das duas amostras seria pequeno, não sendo possível usarmos a distribuição Z nem a distribuição T para fazermos inferência estatística.

Este livro aborda apenas duas técnicas não-paramétricas: o teste qui-quadrado e o teste exato de Fisher. Sugerimos que estude outras técnicas não-paramétricas no livro *Siegel S e Castellan, Jr. NJ. Nonparametric statistics for the behavioral sciences. 2ª ed. New York(NY): McGraw-Hill; 1988.*

— Em que outras situações devemos utilizar o teste t ?

— No próximo capítulo discutiremos como realizar inferência estatística utilizando o teste t quando as amostras comparadas não são independentes.

Não se esqueça de que temos direito à contemplação, ao lazer, ao ócio, à preguiça, e de que devemos reservar sempre um tempo razoável para essas e também para nossas atividades políticas, religiosas e sociais (outras além do trabalho e do pagamento de contas, juros bancários e impostos).



CAPÍTULO 13

- Qual o teste a ser aplicado quando as amostras não forem independentes?
 - Por que com amostras não independentes temos de modificar a forma de fazermos o teste t ?
 - Como realizamos o teste t nessa situação?
 - Como calculamos o intervalo de confiança nessa situação?
-



— **Qual o teste a ser aplicado quando as amostras não forem independentes?**

— Ainda podemos utilizar o teste t com algumas modificações, mas antes de mostrarmos como fazer isso, é necessário discutirmos o que são amostras não-independentes, ou seja, dependentes, também chamadas de emparelhadas.

Quando calculamos a média da tensão arterial sistólica em pacientes recém diagnosticados com hipertensão arterial e, portanto, ainda sem tratamento, e a comparamos com a média tensional nos mesmos pacientes, obtida algum tempo após o início do tratamento, não podemos considerar as médias antes e depois do tratamento como independentes. Isso ocorre porque os níveis tensionais foram observados nas mesmas pessoas e, por isso, estão muito relacionados. É claro que o nível tensional que cada indivíduo pesquisado apresentava antes do tratamento influencia o nível a ser obtido depois do tratamento, porque, p. ex., se alguém antes já tinha um nível muito elevado, após o efeito da droga, se esta for realmente eficaz, aquela pessoa tenderá a exibir nível mais baixo que o seu nível anterior, mas mesmo após essa queda, tenderá ainda a apresentar nível mais alto do que alguém que tinha anteriormente nível mais baixo.

Outra situação de dependência entre os grupos a serem comparados ocorre quando selecionamos um grupo para estudo, e depois outro, com características propositadamente muito semelhantes ao primeiro. Esse procedimento é chamado de **emparelhamento**. Como os dois grupos são muito semelhantes, espera-se que os valores observados nos mesmos sejam mais parecidos do que seriam se não tivesse havido emparelhamento. Por isso, temos de considerar que os dois grupos estabelecidos desse modo são dependentes um do outro.

— **Entendi, mas qual a implicação disso na inferência estatística?**

— A implicação é que, a depender dos grupos comparados serem ou não independentes, os fundamentos matemáticos utilizados na inferência estatística são diferentes. Lembre-se de que, em disciplinas do primeiro e segundo graus, você viu que os procedimentos matemáticos são diferentes quando as observações são independentes ou não. Lembra-se de que, por exemplo, as probabilidades podem ser condicionais ou não-condicionais? As primeiras são utilizadas como fundamento quando os eventos são dependentes, e as últimas quando estes são independentes, e isso influenciará a forma de fazermos os testes estatísticos.

Os testes z e t , tal como você aprendeu a realizá-los até o momento, são apropriados quando os grupos forem independentes, mas é possível utilizarmos o teste t também quando as observações forem dependentes, sendo que, nesse caso, chamamos esse teste de teste t para amostras emparelhadas. Contudo, agora, a forma de realização será um pouco diferente da que você já está acostumado, como mostraremos a seguir. Lembre-se, porém, de que para aplicação do teste, ainda continua sendo necessário que as amostras investigadas tenham sido selecionadas aleatoriamente.

— Por que, simplesmente, não evitamos esse tipo de estudo emparelhado, e utilizamos sempre amostras independentes e os testes já discutidos?

— Porque o emparelhamento é um dos métodos que nos permitem ajustar nossos resultados.

— Ajustar resultados?

— Sim. Considere uma investigação sobre a relação entre hábito de fumar e câncer de pulmão, que tenha encontrado uma forte associação entre essas duas variáveis. Se os fumantes tiverem uma média de idade bem mais elevada do que a dos não-fumantes, pode ser que a associação encontrada entre tabagismo e câncer de pulmão decorra do fato dos fumantes serem mais idosos, e não do hábito de fumar, pois sabemos que a idade está relacionada a uma maior incidência de câncer. Para tirarmos essa dúvida, foram desenvolvidas várias técnicas que podem ser revisadas por você em *Pereira MG. Epidemiologia: teoria e prática. Rio de Janeiro (RJ): Guanabara Koogan; 1995, p. 383 a 392*. Uma destas é o emparelhamento, que consiste em ajustar os resultados, fazendo comparações de grupos de indivíduos bem semelhantes. Ao compararmos indivíduos muito semelhantes, estaremos neutralizando o efeito de diversas características dos indivíduos sobre os resultados do estudo. No nosso exemplo, faríamos com que os grupos comparados tivessem idades semelhantes, de modo que, se encontrássemos associação entre hábito de fumar e câncer de pulmão, não poderíamos atribuir este resultado à variável “idade”, porque os grupos apresentavam a mesma média de idade, concorda? Um ajuste similar seria obtido se comparássemos resultados obtidos antes e depois de um determinado tratamento, já que os mesmos indivíduos (com características iguais, portanto) estariam sendo comparados, neutralizando a influência dessas características nos resultados. Então, muitas vezes, é necessário fazermos emparelhamento nos estudos epidemiológicos, daí a importância de aprendermos a realizar testes adequados a tal situação.

Vamos considerar neste capítulo o exemplo fictício de uma pesquisa realizada por uma equipe de nutricionistas, cujo objetivo tenha sido avaliar a eficácia de uma nova dieta de baixa caloria, em uma amostra aleatória de indivíduos obesos de ambos os sexos.

Suponha que nessa pesquisa tenham sido obtidos os dados apresentados na planilha abaixo:

Número do indivíduo na pesquisa	Peso, em kg, antes da dieta	Peso, em kg, depois da dieta
1	118,3	83,2
2	113,5	84,9
3	97,8	74,8
4	103,4	83,9
5	106,9	82,6
6	81,7	77,4
7	100,4	63,6
8	90,1	69,1
9	77,6	62,9
10	90,1	69,1
11	77,6	62,9
12	118,3	83,2
13	113,5	84,9
14	97,8	74,8
15	103,4	83,9
16	106,9	82,6
17	81,7	77,4
18	100,4	63,6
19	90,1	69,1
20	77,6	62,9
21	118,3	83,2
22	113,5	84,9
23	97,8	74,8
24	103,4	83,9
25	106,9	82,6
26	81,7	77,4
27	100,4	63,6
28	90,1	69,1
29	113,5	84,9
30	97,8	74,8
31	103,4	83,9
32	106,9	82,6
33	81,7	77,4
34	100,4	63,6
35	90,1	69,1
36	98,6	96,7

Se esses pesquisadores realizassem o teste t do mesmo modo como fizemos em capítulos anteriores, a estratégia fundamental deles seria comparar a média de peso antes da dieta com a média depois. Na situação atual, não é essa a comparação a ser feita. Eles vão calcular, para cada indivíduo estudado, a diferença entre seu peso antes e seu peso depois, como mostrado na última coluna da planilha a seguir:

Número do indivíduo na pesquisa	Peso, em kg, antes da dieta	Peso, em kg, depois da dieta	Diferença entre os pesos (antes-depois)
1	118,3	104,2	14,1
2	113,5	105,9	7,6
3	97,8	95,8	2,0
4	103,4	104,9	-1,5
5	106,9	103,6	3,3
6	81,7	98,4	-16,7
7	100,4	84,6	15,8
8	90,1	90,1	0,0
9	77,6	83,9	-6,3
10	90,1	90,1	0,0
11	77,6	83,9	-6,3
12	118,3	104,2	14,1
13	113,5	105,9	7,6
14	97,8	95,8	2,0
15	103,4	104,9	-1,5
16	106,9	103,6	3,3
17	81,7	98,4	-16,7
18	100,4	84,6	15,8
19	90,1	90,1	0,0
20	77,6	83,9	-6,3
21	118,3	104,2	14,1
22	113,5	105,9	7,6
23	97,8	95,8	2,0
24	103,4	104,9	-1,5
25	106,9	103,6	3,3
26	81,7	98,4	-16,7
27	100,4	84,6	15,8
28	90,1	90,1	0,0
29	113,5	105,9	7,6
30	97,8	95,8	2,0
31	103,4	104,9	-1,5
32	106,9	103,6	3,3
33	81,7	98,4	-16,7
34	100,4	84,6	15,8
35	90,1	90,1	0,0
36	98,6	96,7	1,9

— Eles poderiam calcular as diferenças “depois-antes”?

— Poderiam, e chegariam às mesmas conclusões que serão apresentadas ao final desse teste.

Observe que das trinta e seis diferenças “antes-depois”, apenas onze são negativas. Diferenças negativas indicam para os pesquisadores, que o peso depois foi maior do que o de antes, isto é, apesar da dieta instituída, o peso desses indivíduos aumentou. Conseqüentemente, como a maior parte das diferenças foi positiva, isso indica que a maior parte dos pesos depois da dieta foi menor do que a de antes, sugerindo, desde já, que a dieta pode ter sido eficaz. Mas essa evidência, por si só, não seria suficiente para eles concluírem o estudo, pois estariam cientes de que os resultados poderiam variar de estudo para estudo, e só realizaram um único estudo. Seria necessário verificar, através de um teste estatístico, se a magnitude das diferenças encontradas era maior ou menor do que aquela esperada por simples variação de resultados amostrais.

Em seguida, os investigadores calculariam a média aritmética dessas diferenças “antes-depois”, e verificariam se os resultados obtidos naquele único estudo eram compatíveis com uma diferença populacional

igual a zero. Se fossem, isso indicaria que, muito provavelmente, não haveria diferença entre os pesos antes e depois da dieta, sugerindo que esta não era eficaz. Se não fossem, concluiriam o contrário.

O raciocínio acima seria apropriado se o teste fosse bicaudado. No nosso exemplo, a média aritmética das diferenças “antes-depois” foi 1,87 kg, com desvio-padrão de 9,30 kg. O teste poderia ser bicaudado, mas assumiremos que os nutricionistas desejam testar se essa média é, estatisticamente, menor ou igual a zero, ou maior do que zero. Portanto, o teste será monocaudado.

Note que as diferenças “antes-depois”, uma para cada indivíduo estudado, passariam a constituir uma nova variável que seria, a partir daquele momento, a variável a ser testada, e não mais as variáveis originais “peso antes” e “peso depois” da dieta.

Ao procederem assim, os pesquisadores não utilizariam o teste t para comparar duas médias de amostras não-independentes, o que seria incorreto. Contornariam o problema realizando a inferência estatística com base nos pares de informações “peso antes-peso depois”, um par para cada indivíduo selecionado para o estudo, e não mais nas informações originais de peso. Cada um desses pares representaria uma diferença entre pesos e, como já mencionamos, seria a média dessas diferenças (ou diferença média) que seria testada, verificando-se, estatisticamente, se esta média poderia ser zero ou não. É isso que lhes permitiria utilizar o teste t nessa situação, porque evitariam fazer uma comparação direta entre amostras não-independentes, uma dessas representada pelos valores de peso antes da dieta e outra pelos valores de peso depois.

Como os pesos teriam sido obtidos em uma amostra aleatória, eles poderiam assumir que as diferenças entre esses pesos constituiriam também uma amostra aleatória, retirada do conjunto de todas as diferenças entre pesos “antes-depois” que seriam obtidas caso tivéssemos investigado toda a população.

Se pudessem também assumir, com base em conhecimentos prévios (de estudos de outros autores em populações semelhantes, ou na própria experiência da equipe de pesquisa), que essas diferenças seriam distribuídas normalmente na população de onde a amostra foi retirada, poderiam utilizar o valor de t calculado com a expressão abaixo, para testar as hipóteses sobre a média das diferenças de pesos na população:

$$t = \frac{\bar{d} - (\mu_d)_o}{EP_{\bar{d}}}.$$

Na expressão acima, \bar{d} denota a média aritmética das diferenças “antes-depois” obtidas na única amostra estudada; $(\mu_d)_o$, o valor da suposta verdadeira média dessas diferenças na população de onde a amostra foi retirada, estabelecido pela hipótese nula (o subscrito o indica nulidade); e $EP_{\bar{d}}$ o erro-padrão das diferenças médias, se numerosas amostras fossem estudadas.

Por sua vez, $EP_{\bar{d}}$ seria calculado por

$$EP_{\bar{d}} = \frac{s_d}{\sqrt{n}},$$

onde s_d denota o desvio-padrão das diferenças, e n o número de diferenças obtidas na única amostra estudada, que seria igual ao número de indivíduos investigados, porque cada indivíduo daria origem a um par

de pesos (peso antes e peso depois), que daria origem a uma diferença entre um peso “antes” e um peso “depois”. Os subscritos $_d$ e $_{\bar{d}}$ denotam as diferenças e as médias das diferenças (ou diferenças médias), respectivamente.

A fórmula $t = \left[\bar{d} - (\mu_d)_o \right] / EP_{\bar{d}}$ apresentada acima não apresenta muitas novidades para você, pois ao utilizá-la, os pesquisadores estariam calculando o quanto a média das diferenças obtidas no estudo se desvia da média estabelecida na hipótese nula, dividindo esse desvio pelo valor do erro-padrão das diferenças médias, obtendo esse desvio em número de erros-padrão, para que pudessem utilizar a tabela com valores críticos de T , necessários para a realização da inferência estatística. Nessa tabela, procurariam por valores de T correspondentes a $n - 1$ graus de liberdade, porque perderiam um grau de liberdade ao calcularem a média aritmética das diferenças, \bar{d} .

Lembraremos novamente a você a distinção entre desvio-padrão e erro-padrão: o primeiro indica o quanto, em média, as diferenças se desviaram da média das diferenças, na única amostra estudada, e o segundo, o quanto, em média, as médias das diferenças se desviariam da média das diferenças médias, caso tivéssemos realizado numerosos estudos, e não apenas um. Lembre-se também de que, mesmo sem termos estudado várias amostras, é possível calcularmos o erro-padrão utilizando a expressão $EP_{\bar{d}} = s_d / \sqrt{n}$, sendo possível demonstrarmos isso empírica e algebricamente. Se estiver esquecido, revise a demonstração empírica que fizemos nas páginas 144 a 150 deste livro.

Para avaliar se a dieta hipocalórica teria sido eficaz em reduzir o peso dos indivíduos obesos estudados, os nutricionistas realizariam um teste de hipóteses que seguiria as seguintes etapas:

1ª) Definição do nível de significância: $\alpha = 0,05$;

2ª) Definição das hipóteses:

Se já existissem evidências de que a dieta testada seria eficaz, esperariam encontrar uma média das diferenças “antes-depois” estatisticamente maior do que zero, porque para que isso ocorresse, seria necessário que as diferenças fossem maiores do que zero e, para isso, o peso “depois” teria que ser menor do que o peso “antes”, indicando queda no peso, ou seja, eficácia da dieta com redução do peso.

Assim, as hipóteses deles seriam: $H_O : \mu_d \leq 0 \text{ kg}$ e $H_A : \mu_d > 0 \text{ kg}$;

3ª) Cálculo do valor de t :

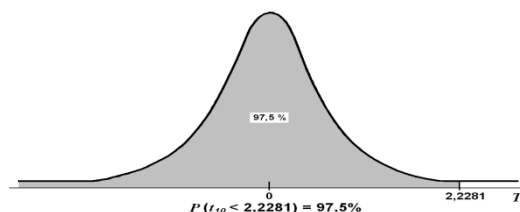
$$t = \frac{\bar{d} - (\mu_d)_o}{\frac{s_d}{\sqrt{n}}} = \frac{1,87 - 0}{\frac{9,30}{\sqrt{36}}} = \frac{1,87}{\frac{9,30}{6,00}} = \frac{1,87}{1,55} = 1,21;$$

4ª) Obtenção do valor crítico de t :

Consultariam a tabela T para encontrar o valor crítico de t correspondente a um teste monocaudado, α de 0,05, e $n - 1 = 36 - 1 = 35$ graus de liberdade.

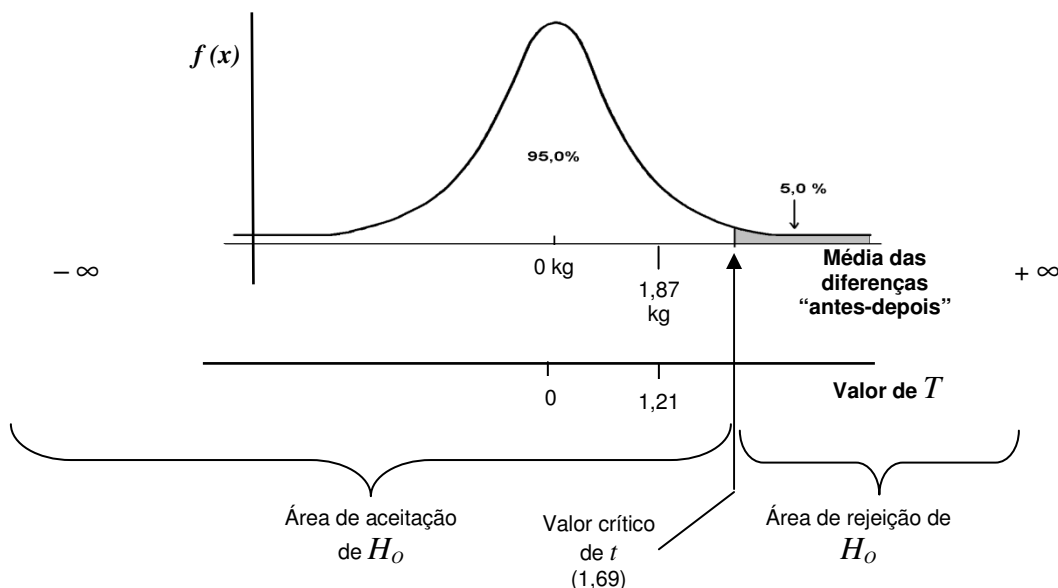
Olhando na tabela T , que reapresentamos a seguir, veriam que o valor crítico seria $1,6896 \cong 1,69$,

situado na célula onde a coluna rotulada P_{95} se encontra com a linha correspondente a 35 graus de liberdade. A coluna utilizada seria essa porque, como o teste seria monocaudado, eles considerariam todo o valor de α em uma das caudas da curva, demarcando duas áreas sob a curva T , delimitadas pelo percentil 95, que é o percentil que separa os 95% valores mais baixos dos 5% valores mais altos.



Graus de liberdade ($n - 1$)	P_{90}	P_{95}	$P_{97,5}$	P_{99}	$P_{99,5}$
1	3,0780	6,3138	12,7060	31,8210	63,6570
2	1,8860	2,9200	4,3027	6,9650	9,9248
3	1,6380	2,3534	3,1825	4,5410	5,8409
4	1,5330	2,1318	2,7764	3,7470	4,6041
5	1,4760	2,0150	2,5706	3,3650	4,0321
6	1,4400	1,9432	2,4469	3,1430	3,7074
7	1,4150	1,8946	2,3646	2,9980	3,4995
8	1,3970	1,8595	2,3060	2,8960	3,3554
9	1,3830	1,8331	2,2622	2,8210	3,2498
10	1,3720	1,8125	2,2281	2,7640	3,1693
11	1,3630	1,7959	2,2010	2,7180	3,1058
12	1,3560	1,7823	2,1788	2,6810	3,0545
13	1,3500	1,7709	2,1604	2,6500	3,0123
14	1,3450	1,7613	2,1448	2,6240	2,9768
15	1,3410	1,7530	2,1315	2,6020	2,9467
16	1,3370	1,7459	2,1190	2,5830	2,9208
17	1,3330	1,7396	2,1098	2,5670	2,8982
18	1,3300	1,7341	2,1009	2,5520	2,8784
19	1,3280	1,7291	2,0930	2,5390	2,8609
20	1,3250	1,7247	2,0860	2,5280	2,8453
21	1,3230	1,7207	2,0796	2,5180	2,8314
22	1,3210	1,7171	2,0739	2,5080	2,8188
23	1,3190	1,7139	2,0687	2,5000	2,8073
24	1,3180	1,7109	2,0639	2,4920	2,7969
25	1,3160	1,7081	2,0595	2,4850	2,7874
26	1,3150	1,7056	2,0555	2,4790	2,7787
27	1,3140	1,7033	2,0518	2,4730	2,7707
28	1,3130	1,7011	2,0484	2,4670	2,7633
29	1,3110	1,6991	2,0452	2,4620	2,7564
30	1,3100	1,6973	2,0423	2,4570	2,7500
35	1,3062	1,6896	2,0301	2,4380	2,7239
40	1,3031	1,6839	2,0211	2,4230	2,7045
45	1,3007	1,6794	2,0141	2,4120	2,6896
50	1,2987	1,6759	2,0086	2,4030	2,6778
60	1,2959	1,6707	2,0003	2,3900	2,6603
70	1,2938	1,6669	1,9945	2,3810	2,6480
80	1,2922	1,6641	1,9901	2,3740	2,6388
90	1,2910	1,6620	1,9867	2,3680	2,6316
100	1,2901	1,6602	1,9840	2,3640	2,6260
120	1,2887	1,6577	1,9799	2,3580	2,6175
140	1,2876	1,6558	1,9771	2,3530	2,6114
160	1,2869	1,6545	1,9749	2,3500	2,6070
180	1,2863	1,6534	1,9733	2,3470	2,6035
200	1,2858	1,6525	1,9719	2,3450	2,6006
∞	1,2820	1,6450	1,9600	2,3260	2,5760

Veja no diagrama abaixo, a situação encontrada nesse exemplo:



O n desse exemplo, apesar de ser maior do que 30, não é grande o suficiente para que a curva utilizada se assemelhe à curva Z . Por isso, no diagrama acima foi desenhada uma distribuição T , mais achatada no centro e com base mais larga;

5ª) a) Comparação do valor observado de t ao valor crítico de t e conclusão do teste:

Como $1,21 < 1,69$, os nutricionistas concluiriam que o valor de t correspondente a uma média das diferenças de $1,87 \text{ kg}$ estaria em uma localização muito central da curva, não ultrapassando o valor crítico de t . Outra interpretação seria a de que $1,87 \text{ kg}$ seria um valor muito provável de ser obtido em amostras retiradas de uma população cuja média das diferenças antes-depois fosse zero, indicando ser muito provável que a verdadeira diferença média na população não seja maior do que zero, porque a diferença média obtida na única amostra investigada é compatível com isto. Observariam ainda, que o valor $1,87 \text{ kg}$ estaria localizado na área de aceitação da hipótese nula. Isso seria o mesmo que dizer que a dieta hipocalórica testada não teria se mostrado eficaz no estudo realizado e isto indicaria que, na população de onde a amostra estudada teria sido retirada, seria muito provável que essa diferença fosse menor ou igual a zero;

b) Poderiam também obter o valor- p correspondente e compará-lo ao valor de α . Programas de computador lhes forneceriam facilmente o valor- p , e você já sabe como eles interpretariam o resultado do teste com base nesse valor. A conclusão seria a mesma vista acima. Neste exemplo, para satisfazer sua curiosidade, fizemos o teste utilizando um programa estatístico para computador e obtivemos um valor- p igual a $0,234$. Como esse valor é maior do que o valor de α , que é $0,05$, nossa conclusão seria a mesma à qual os nutricionistas chegariam ao usarem o valor crítico de t .

Outra opção para os pesquisadores seria calcular o intervalo de 95% de confiança e, se quisessem manter o caráter monocaudado do teste, deveriam calcular o limite superior desse intervalo da seguinte maneira:

$$\text{Limite superior do IC (95\%)} = \bar{d} + t_{v,(1-\alpha)} \frac{s_d}{\sqrt{n}}$$

Diagrama explicativo da fórmula:

- Limite superior do intervalo de 95% de confiança
- Média das diferenças "antes-depois" obtidas no estudo
- Valor de t correspondente ao percentil 95
- Erro-padrão das médias das diferenças "antes-depois", caso tivéssemos feito vários estudos

Na fórmula acima, o subscrito v representa o número de graus de liberdade da distribuição T a ser utilizada, que depende, por sua vez, do tamanho da amostra estudada.

Note que o cálculo do intervalo de confiança seria feito utilizando o valor crítico de t para um α integralmente considerado na cauda direita da curva.

Colocando os valores do nosso exemplo, os pesquisadores obteriam:

$$\begin{aligned} \text{Limite superior do IC (95\%)} &= \bar{d} + t_{v,(1-\alpha)} \frac{s_d}{\sqrt{n}} = \\ &= 1,87 + t_{35, 0,95} \frac{9,30}{\sqrt{36}} = 1,87 + 1,69 \left(\frac{9,30}{6} \right) = \\ &= 1,87 + (1,69)(1,55) = 1,87 + 2,62 = 4,49. \end{aligned}$$

Teoricamente, o limite inferior seria menos infinito. Biologicamente, esse limite seria o valor negativo mais extremo possível para diferenças entre pesos de seres humanos. O diagrama na página 239, embora represente a situação do teste de hipóteses, poderá lhe ajudar a entender o cálculo do intervalo de confiança monocaudado. O intervalo iria, então, do valor negativo mais extremo possível até 4,49 kg. Isso lhes indicaria que haveria uma probabilidade de 95% da verdadeira média das diferenças dos pesos "antes-depois" estar entre esses dois valores. Qualquer valor dentro desse intervalo seria considerado por eles como aceitável para a média populacional. Como o valor zero estaria contido nesse intervalo, não poderiam descartá-lo como possível para a verdadeira média das diferenças na população, e concluiriam que, estatisticamente, os resultados do estudo realizado sugeririam que a dieta testada não seria eficaz para reduzir o peso de pessoas obesas.

Se os(as) nutricionistas desejassem calcular o intervalo de 95% de confiança bicaudado, esse seria calculado por

$$\text{IC (95\%)} = \bar{d} \pm t_{v,(1-\alpha/2)} \frac{s_d}{\sqrt{n}}$$

Diagrama explicativo da fórmula:

- Intervalo de 95% de confiança
- Média das diferenças "antes-depois" obtidas no estudo
- Valor de t correspondente ao percentil 97,5
- Erro-padrão das médias das diferenças "antes-depois", caso tivéssemos feito vários estudos

Na fórmula acima, o subscrito v representa o número de graus de liberdade da distribuição T a ser utilizada, que depende, por sua vez, do tamanho da amostra estudada.

Note que eles fariam o cálculo do intervalo de confiança utilizando o valor crítico de t para um α repartido nas duas caudas da curva. Essa é a maneira que geralmente se usa para calcular intervalos de confiança, e é equivalente a um teste de hipóteses bicaudado.

Utilizando os valores do exemplo, eles obteriam:

$$\begin{aligned} IC(95\%) &= \bar{d} \pm t_{v, (1-\alpha/2)} \frac{s_d}{\sqrt{n}} = 1,87 \pm t_{35, 0,975} \frac{9,30}{\sqrt{36}} = 1,87 \pm 2,03 \left(\frac{9,30}{6} \right) = \\ &= 1,87 \pm (2,03)(1,55) = 1,87 \pm 3,15 = (-1,28 \text{ a } 5,02). \end{aligned}$$

Experimente interpretar sozinho(a) esse resultado.

Já vimos que é muito raro conhecermos os valores das variáveis para toda uma população, mas, se a variância das diferenças de pesos “antes-depois” em uma população finita for conhecida, o teste apropriado seria o teste z , sendo o valor de z calculado por

$$z = \frac{\bar{d} - (\mu_d)_o}{EP_{\bar{d}}},$$

onde $EP_{\bar{d}}$ seria dado por $\frac{\sigma_d}{\sqrt{n}}$.

Quando o pressuposto de que as diferenças são distribuídas normalmente na população não estiver atendido, o teorema central do limite pode ser aplicado, se o n for suficientemente grande. Neste caso, utilizaremos o teste z , com o valor de z sendo calculado pela mesma fórmula acima. Porém, se a variância não for conhecida, utilizaremos o teste t que, como você já sabe, será calculado por

$$t = \frac{\bar{d} - (\mu_d)_o}{\frac{s_d}{\sqrt{n}}},$$

sendo este valor aproximadamente igual a z , se o n for suficientemente grande.

Há ainda várias aplicações para os testes z e t . Até agora, discutimos bastante a comparação estatística de médias. No próximo capítulo você verá que o teste z também pode ser utilizado para a comparação estatística de proporções.

CAPÍTULO 14

- E se estivermos comparando proporções e não médias?
 - Por que podemos usar o teste z também para inferência sobre proporções?
 - Como fazemos inferência sobre uma proporção utilizando o teste z ?
 - Como fazemos inferência sobre duas proporções utilizando o teste z ?
-



— E se estivermos comparando proporções e não médias?

— Teoricamente, ao fazermos inferência estatística sobre proporções, devemos utilizar uma outra distribuição chamada “binomial”, já mencionada na página 122. Mas, como o método baseado na distribuição binomial não é muito utilizado, o mesmo não será abordado neste livro. Se estiver interessado nesse método, leia as páginas 91 a 98 do capítulo 4, e 251 a 253 do capítulo 7, do excelente livro de Hosner B. *Fundamentals of biostatistics*. 5ª ed. Pacific Grove (CA): Duxbury; 2000.

— Como vamos realizar essa inferência, se o método teoricamente mais adequado baseia-se na distribuição binomial, e não vamos vê-lo neste livro?

— Você verá que podemos utilizar o teste z também para fazer inferência sobre proporções.

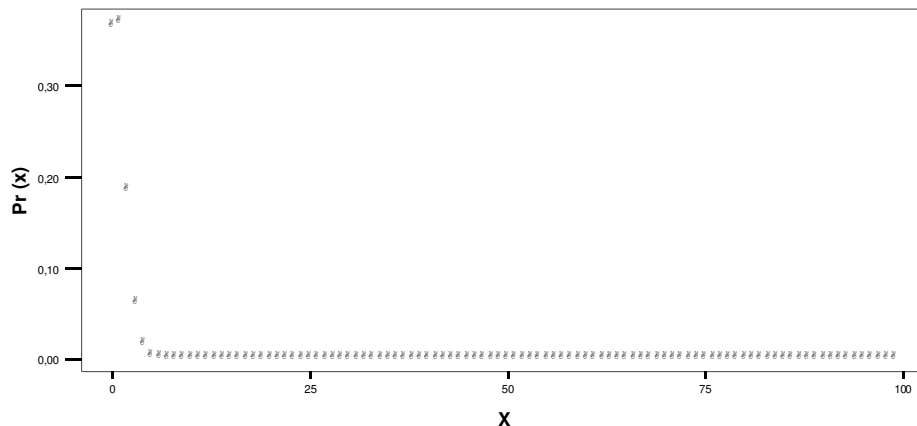
— Por que podemos usar o teste z também para inferência sobre proporções?

— Pode ser demonstrado que, se forem atendidas certas condições relativas ao número de indivíduos estudados e à magnitude da probabilidade de ocorrência do evento de interesse na amostra estudada, a distribuição normal poderá ser utilizada como uma aproximação válida da distribuição binomial. Essa aproximação é mais fácil de ser usada do que a distribuição binomial e, por isso, esta não é muito aplicada.

Veja a seguir uma demonstração empírica de que a aproximação normal da binomial é válida:

Considere uma distribuição binomial para n indivíduos estudados, e uma proporção p de indivíduos acometidos pelo evento de interesse. O número de indivíduos, n , e a proporção de pessoas acometidas, p , são os parâmetros de uma distribuição binomial. Se o n for grande e p próxima de zero (muito pequena), a distribuição binomial será muito desviada para a direita (cauda direita muito longa). Veja isso no diagrama abaixo:

Distribuição binomial para $n=100$ e $p=0,01$.

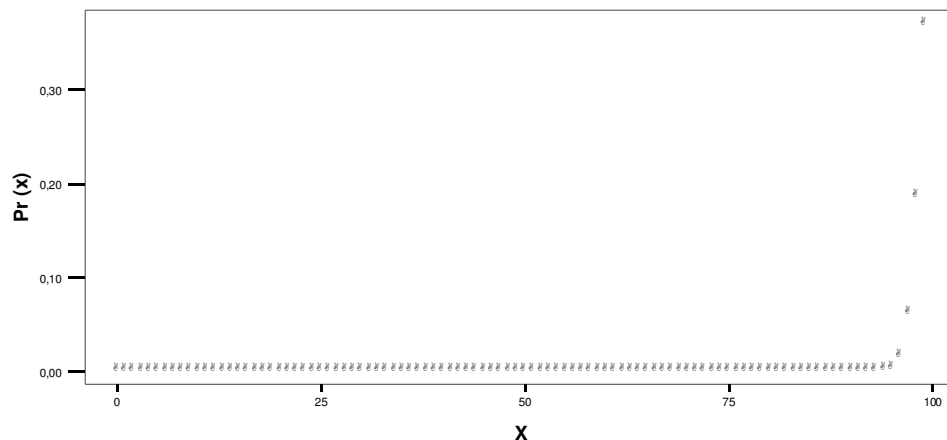


No diagrama acima, a ordenada representa as probabilidades de ocorrerem x eventos, e a abscissa

os valores de X , ou seja, o número de eventos a ocorrerem.

Se o n for grande e p próxima de 1 (muito grande), a distribuição binomial será muito desviada para a esquerda (cauda esquerda muito longa). Veja isso no diagrama abaixo:

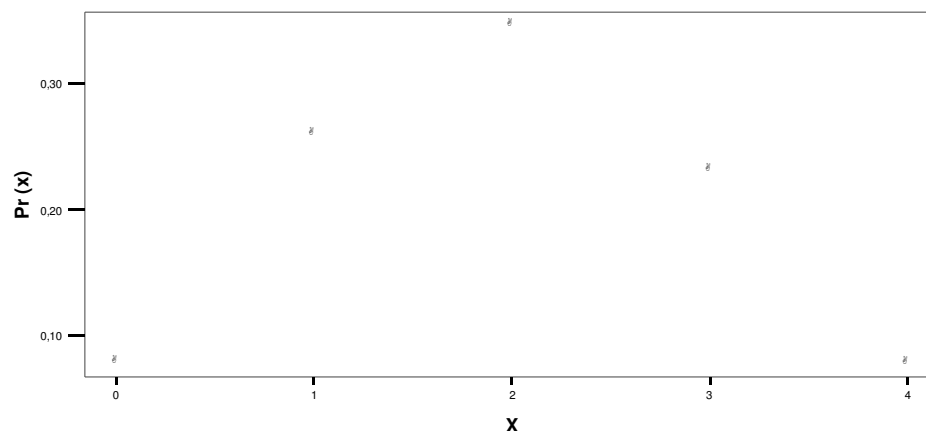
Distribuição binomial para $n=100$ e $p=0,99$.



Então, nas circunstâncias acima (n grande, mas p muito grande ou pequena), as binomiais obtidas não se assemelham à distribuição normal, já que as primeiras não são simétricas, impedindo a utilização da normal como aproximação da binomial.

Se o n for pequeno, qualquer que seja o valor de p , a distribuição binomial continuará diferente da normal. Veja que a distribuição abaixo é muito descontínua, não contém caudas e é menos pontiaguda do que a normal:

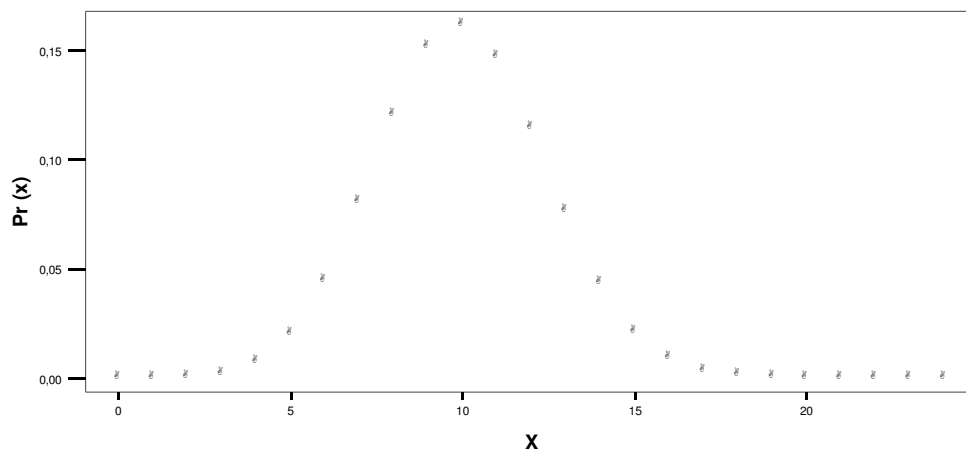
Distribuição binomial para $n=5$ e $p=0,40$.



Agora, veja no primeiro diagrama da próxima página que se o n for grande, mesmo que apenas moderadamente, e p não for muito grande nem muito pequena, a distribuição binomial tenderá a ser

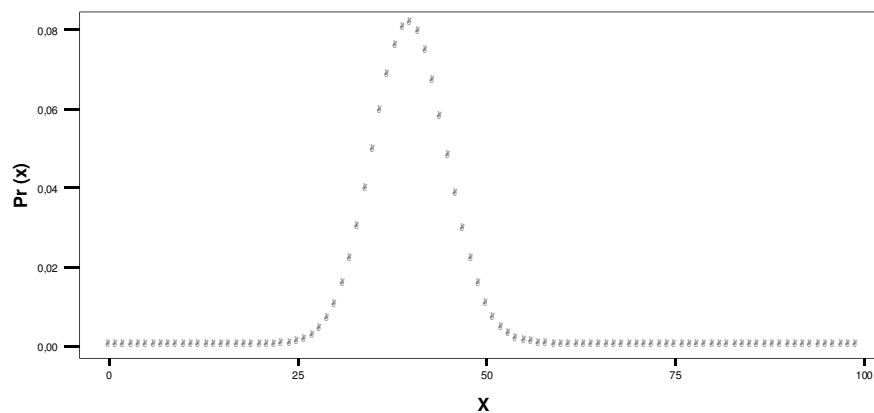
simétrica, embora ainda desviada, mas aproximando-se mais da distribuição normal.

Distribuição binomial para $n=25$ e $p=0,40$.



Se p não for muito grande nem muito pequena, quanto maior o n , maior a adequação da distribuição normal à binomial, como mostramos abaixo:

Distribuição binomial para $n=100$ e $p=0,40$.



Pode também ser demonstrado empiricamente, que essa aproximação será ainda melhor, se fizermos uma **correção para continuidade**. Essa correção é necessária porque há intervalos entre os pequenos círculos nos diagramas mostrados acima, ou seja, a distribuição binomial não é contínua, enquanto a normal é. Considere dois valores quaisquer de X , denotados por x_1 e x_2 . Para que a área sob a curva entre esses dois valores seja melhor aproximada pela distribuição normal, calcularemos a área entre $x_1 - \left(\frac{1}{2}\right)$ e $x_2 + \left(\frac{1}{2}\right)$, e não entre x_1 e x_2 , como fazemos em situações comuns. Com essa correção, trabalharemos com uma área maior sob a curva Z . Contudo, se o n for suficientemente grande, os resultados

obtidos não serão muito diferentes, com ou sem essa correção.

— **Como sabermos se o tamanho da amostra é suficientemente grande para aplicarmos o teste z ?**

— Os estatísticos admitem que essa aproximação seja aceitável se $npq \geq 5$, sendo p a proporção esperada do evento de interesse na população, e q seu complemento, ou seja, $1 - p$.

— **Como fazemos inferência sobre uma proporção utilizando o teste z ?**

— Você verá que os procedimentos serão muito semelhantes àqueles que vimos para inferência sobre uma média.

Considere que tenhamos realizado um estudo epidemiológico sobre intoxicação por chumbo, em crianças de 1 a 9 anos de idade, residentes em uma cidade onde existia uma fundição de chumbo. Foram incluídas aleatoriamente no estudo 250 crianças nessa faixa de idade. Cada criança incluída foi classificada como intoxicada ou não, com base em exame apropriado. Nosso objetivo foi estimar se a prevalência de intoxicação por chumbo na população de onde a amostra foi retirada era 64,0%. Queríamos saber se 64,0% podia ser a verdadeira prevalência populacional, e avaliamos isso com base nos resultados obtidos em uma única amostra de 250 crianças, pois não era viável investigarmos todas as crianças. Note que testamos uma prevalência muito elevada, 64,0%, porque as crianças estudadas moravam em uma área exposta à poluição provocada pela fundição. Considere também que a prevalência encontrada na amostra estudada, denotada por \hat{p} (lê-se p chapéu), tenha sido 70,4%. Desejávamos saber o quanto a prevalência observada, 70,4%, era compatível com uma prevalência populacional de 64,0%.

Pergunta a ser respondida:

Assumindo que a prevalência na população de 1 a 9 anos daquela cidade seja 64,0%, qual a probabilidade (valor- p) de obtermos, na única amostra estudada, uma prevalência de intoxicação por chumbo maior do que 70,4% ou menor do que o simétrico de 70,4%, considerando o teste bicaudado?

Lembre-se de que a prevalência é um indicador calculado dividindo-se o número de indivíduos com uma determinada doença em certo momento e local, pelo número estimado de pessoas expostas ao risco de estarem com aquela doença naquele mesmo momento e local. Sendo assim, matematicamente, a prevalência é uma proporção, pois expressa quantas das pessoas que poderiam estar doentes estavam realmente doentes naquele momento e local, medindo assim uma probabilidade.

As etapas do nosso teste foram então:

1ª) Definição do nível de significância: $\alpha = 0,05$;

2ª) Definição das hipóteses: $H_0 : p = 64,0\%$ e $H_A : p \neq 64,0\%$, onde p denota a prevalência (proporção) de intoxicação por chumbo na população. Estávamos querendo testar, portanto, se 64,0% podia ser a verdadeira prevalência na populacional.

3ª) Cálculo do valor de z :

Como a prevalência esperada, p , era 64,0% ou 0,64, e seu complemento, $1 - p = q$, era 36,0% ou 0,36, verificamos que $npq = (250)(0,64)(0,36) = 57,6$. Como $57,6 > 5$, vimos que nossa pesquisa permitia a utilização da aproximação normal da binomial. Nesse cálculo, usamos p e não \hat{p} , porque um dos pressupostos para a realização desse teste é que a hipótese nula seja verdadeira, e esta estabelecia que $p = 64,0\%$.

Em seguida, calculamos o valor de z , dado por

$$z = \frac{\hat{p} - p_o}{EP_p},$$

onde \hat{p} denota a proporção obtida na única amostra estudada; $p_o = p$ a proporção de intoxicação esperada na população, que foi estabelecida na hipótese nula, estando o subscrito o , como sempre, indicando nulidade; e EP_p o erro-padrão das proporções que seriam obtidas, caso tivéssemos estudado numerosas amostras retiradas de uma população cuja prevalência fosse 64,0%. Relembrando, o erro-padrão nos indica o quanto, em média, essas proporções variariam em torno da média dessas proporções, assumindo que esta última é 64,0%. Se tivéssemos realmente estudado numerosas amostras, esse erro-padrão seria obtido de modo tradicional, calculando-se o desvio de cada proporção amostral em relação à média dessas proporções, elevando-se cada desvio ao quadrado, somando-se todos esses desvios, dividindo-se essa soma por $k - 1$ (sendo k o número de amostras) e, finalmente, extraíndo-se a raiz quadrada do resultado desta divisão. Mas ocorre aqui o mesmo fenômeno que já discutimos com você na inferência sobre médias: na prática não estudamos numerosas amostras, o que nos impede de fazer o cálculo acima. Felizmente, foi demonstrado algébrica e empiricamente que, se multiplicarmos p por q , dividirmos o resultado por n , e extraírmos a raiz quadrada do resultado desta divisão, obteremos o erro-padrão das possíveis proporções amostrais, mesmo sem termos estudado numerosas amostras. Assim,

$$EP_p = \sqrt{\frac{pq}{n}}.$$

Mas, já vimos que p é substituído por p_o , porque este teste é fundamentado no pressuposto de que

H_0 seja verdadeira, já que somente sob essa condição a quantidade $(\hat{p} - p_o)/EP_p$ apresenta uma distribuição normal. Sendo assim, $q = 1 - p$ é também substituído por $q_o = 1 - p_o$. Logo, temos que

$$z = \frac{\hat{p} - p_o}{EP_p} = \frac{\hat{p} - p_o}{\sqrt{\frac{p_o q_o}{n}}}.$$

Substituindo os valores do nosso exemplo na expressão acima, encontramos:

$$z = \frac{\hat{p} - p_o}{\sqrt{\frac{p_o q_o}{n}}} = \frac{0,704 - 0,64}{\sqrt{\frac{(0,64)(0,36)}{250}}} = \frac{0,064}{\sqrt{\frac{0,23}{250}}} = \frac{0,064}{\sqrt{0,00092}} = \frac{0,064}{0,03} = 2,13;$$

4ª) Obtenção do valor- p : consultamos a tabela da curva normal padrão, para encontrarmos a probabilidade de obtermos um valor de Z menor do que 2,13 ($P(Z < 2,13)$). Essa probabilidade era de 0,9834. A área sob a curva na qual estávamos interessados foi encontrada por

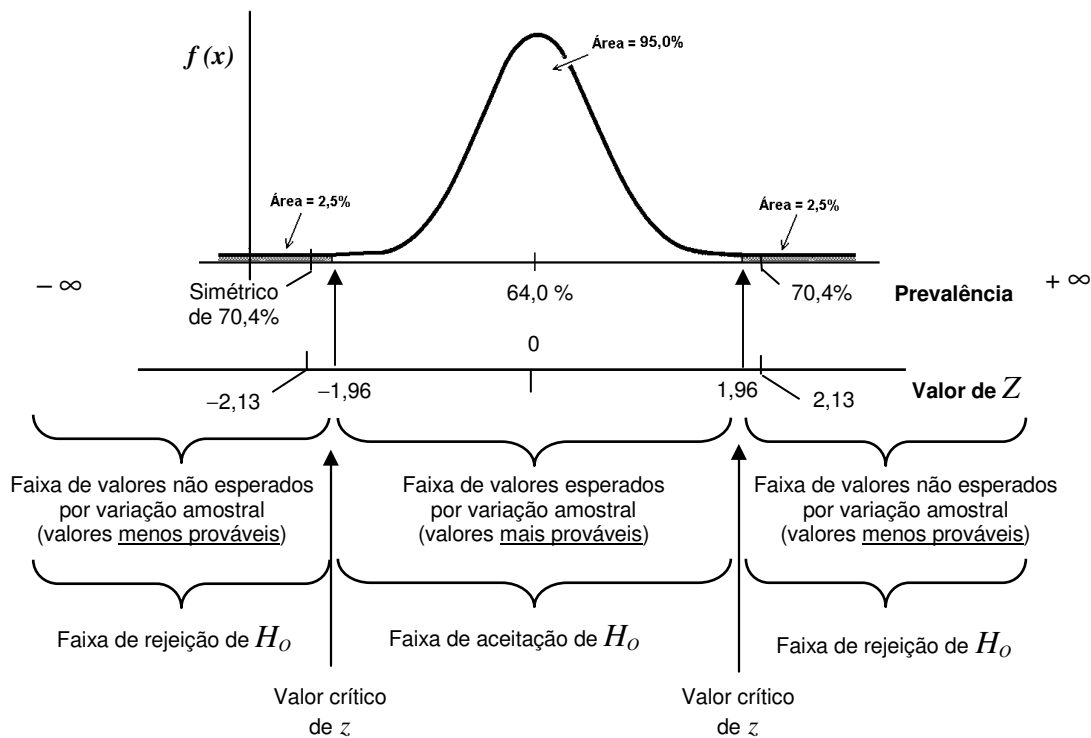
$$P(Z > 2,13) = 1 - P(Z < 2,13) = 1 - 0,9834 = 0,0166 \text{ ou } 1,66\%.$$

Como o teste era bicaudado, tivemos de multiplicar essa probabilidade por dois, obtendo um valor- p igual a $(1,66\%)(2) = 3,32\%$.

Veja no diagrama da próxima página a situação encontrada nesse exemplo;

5ª) a) Comparação do valor- p ao valor de α e conclusão do teste: como $0,0332 < 0,05$ ou, equivalentemente, como $3,32\% < 5,0\%$, concluímos que o resultado era estatisticamente significativo, ou seja, a verdadeira prevalência na população era estatisticamente diferente de 64,0%, sendo muito improvável que 64,0% fosse a verdadeira prevalência de intoxicação por chumbo na população-alvo. Se a prevalência populacional fosse 64,0%, dificilmente obteríamos uma amostra com prevalência de 74,0%, que foi a prevalência que encontramos em nosso estudo. Assim, rejeitamos H_0 e aceitamos H_A .

Não se esqueça de que, para concluirmos nossa investigação, foi fundamental avaliarmos também a ausência de vieses significativos na mesma;



ou

b) Comparação do valor observado de z ao valor crítico de z :

Como $2,13 > 1,96$ e, conseqüentemente, $-2,13 < -1,96$, concluímos que o valor de z correspondente a 70,4% estava em uma localização bastante extrema na curva, ultrapassando o valor crítico de z . Assim, 70,4% era um valor muito improvável de ser obtido em amostras retiradas de uma população cuja prevalência fosse 64,0%. A prevalência populacional, então, não deve ser 64,0%. Outra interpretação foi a de que o valor obtido estava localizado na área de rejeição da hipótese nula. Assim, se a pesquisa que realizamos não tivesse apresentado vieses importantes, seríamos levados à mesma conclusão à qual chegamos na letra **a** da quinta etapa.

Como outra opção ainda, poderíamos ter calculado o intervalo de 95% de confiança, que teria sido dado por

$$IC(95\%) = \hat{p} \pm z_{(1-\alpha/2)} \sqrt{\frac{\hat{p}\hat{q}}{n}}$$

Diagrama de interpretação da fórmula:

- $IC(95\%)$: Intervalo de 95% de confiança
- \hat{p} : Proporção obtida no estudo
- $z_{(1-\alpha/2)}$: Valor de z correspondente ao percentil 97,5
- $\sqrt{\frac{\hat{p}\hat{q}}{n}}$: Erro-padrão das proporções caso tivéssemos estudado numerosas amostras

Observe que agora, como não estaríamos testando hipóteses, e como também quase sempre não disporíamos da proporção do evento na população-alvo, teríamos utilizado a proporção obtida na amostra estudada como referência para nosso cálculo. Lembre-se de que o mesmo aconteceu, quando testamos médias, em capítulos anteriores.

Substituindo os valores do nosso exemplo, teríamos obtido:

$$\begin{aligned} IC(95\%) &= \hat{p} \pm z_{(1-\alpha/2)} \sqrt{\frac{\hat{p}\hat{q}}{n}} = \hat{p} \pm z_{0,975} \sqrt{\frac{\hat{p}\hat{q}}{n}} = 0,704 \pm 1,96 \sqrt{\frac{(0,704)(0,296)}{250}} = \\ &= 0,704 \pm 1,96 \sqrt{\frac{0,21}{250}} = 0,704 \pm 1,96 \sqrt{0,00084} = 0,704 \pm (1,96)(0,03) = \\ &= 0,704 \pm 0,06 = (0,644 \text{ a } 0,764) = (64,4\% \text{ a } 76,4\%). \end{aligned}$$

(Se não quiser mais reler esse tipo de interpretação, pule o próximo parágrafo).

Tais resultados nos teriam indicado que haveria uma probabilidade de 95% da verdadeira prevalência populacional estar entre 64,4% e 76,4%. Qualquer valor dentro deste intervalo teria sido considerado como aceitável para a prevalência populacional, e qualquer valor fora, como inaceitável. Como 64,0% estaria fora do intervalo, teríamos concluído que este valor, muito provavelmente, não seria obtido se numerosas amostras retiradas de uma população com prevalência de intoxicação igual a 64,0% tivessem sido estudadas, indicando que este valor não deveria ser a verdadeira prevalência populacional.

— E a correção para continuidade?

— Se tivéssemos feito o teste de hipóteses com essa correção, teríamos seguido as mesmas etapas já apresentadas, com algumas modificações. Veja as etapas a seguir:

1ª) Definição do nível de significância: $\alpha = 0,05$;

2ª) Definição das hipóteses:

$$H_0 : p = 64,0\% \text{ e } H_A : p \neq 64,0\% ;$$

3ª) Cálculo do valor de z_c :

Considerando x o número de crianças com intoxicação na amostra, teríamos que $\hat{p} = x/n$, donde $x = n\hat{p}$. No nosso exemplo, portanto, $x = n\hat{p} = (250)(0,704) = 176$ crianças intoxicadas. Denotando por x_o , o número esperado de crianças com intoxicação na amostra, caso a prevalência na amostra fosse a mesma da população, teríamos calculado x_o por $x_o = np_o = (250)(0,64) = 160$. Como $x > x_o$, ou seja, como $176 > 160$, teríamos aplicado a correção para continuidade diminuindo $1/2$, ou seja, 0,5, do valor de x ,

e computaríamos o valor de z_c por

$$z_c = \frac{\frac{x - 0,5}{n} - p_o}{EP_p} = \frac{\frac{x - 0,5}{n} - p_o}{\sqrt{\frac{p_o q_o}{n}}} = \frac{\frac{176 - 0,5}{250} - 0,64}{\sqrt{\frac{(0,64)(0,36)}{250}}} = \frac{0,70 - 0,64}{\sqrt{0,0009}} = \frac{0,06}{0,03} \cong 2,00,$$

onde z_c estaria denotando o valor de z corrigido.

Note que o valor de z_c obtido não teria sido muito diferente daquele calculado anteriormente (2,13 comparado a 2,00), mas essa pequena diferença poderia ter nos levado a uma conclusão diferente. Como já vimos, quando o n for suficientemente grande, nossos resultados com e sem correção para continuidade serão muito semelhantes.

— **E se x fosse menor do que x_o ?**

— Se $x < x_o$, faríamos a correção somando $1/2$ ao valor de x ;

4ª) Obtenção do valor- p : teríamos consultado a parte com valores positivos da tabela da curva normal padrão, para encontrarmos a probabilidade de obtermos um valor de Z menor do que 2,00, ou seja, $P(Z < 2,00)$. Essa probabilidade teria sido de 0,9772 ou 97,72%. A probabilidade de interesse teria sido então computada por

$$P(Z > 2,00) = 1 - P(Z < 2,00) = 1 - 0,9772 = 0,0228 \text{ ou } 2,28 \%.$$

Como o teste teria sido bicaudado, multiplicaríamos essa probabilidade por dois, obtendo um valor- p igual a $(2,28\%)(2) = 4,56\%$;

5ª) Comparação do valor- p ao valor de α e conclusão do teste: como $4,56\% < 5,0\%$, teríamos concluído que o resultado obtido era estatisticamente significativo, ou seja, que a verdadeira prevalência e 64,0% eram estatisticamente diferentes, sendo muito improvável que 64,0% fosse a verdadeira prevalência de intoxicação por chumbo na população-alvo.

Assim, teríamos rejeitado H_o e aceitado H_A , sendo que, aqui, nossa conclusão teria sido a mesma do teste sem a correção.

— **Como fazemos inferência sobre duas proporções utilizando o teste z ?**

— Isto ocorrerá quando nossa pesquisa investigar dois grupos e não apenas um. Então, considere agora que tenhamos também classificado as crianças do nosso exemplo como moradoras da área próxima (a menos de 500 metros) ou distante (a 500 metros ou mais) da fundição, e suponha que outro objetivo nosso tenha sido investigar se as crianças residentes na área próxima apresentavam uma prevalência maior de intoxicação do que as da área distante. Nesse caso, então, desejávamos comparar duas proporções, que

eram as prevalências de intoxicação nas áreas próxima e distante. Suponha que as prevalências tenham sido $\hat{p}_1 = 80,4\%$ e $\hat{p}_2 = 62,3\%$, respectivamente. Considere também, no nosso exemplo, que havia $n_1 = 112$ e $n_2 = 138$ crianças, nas áreas próxima e distante, respectivamente.

Como sempre, teria sido possível fazermos a comparação dessas duas prevalências através de um teste de hipóteses ou do cálculo do intervalo de confiança.

Veja abaixo as etapas do teste de hipóteses:

1ª) Definição do nível de significância: $\alpha = 0,05$;

2ª) Definição das hipóteses:

$$H_0 : p_1 - p_2 = 0 \text{ (ou } H_0 : p_1 = p_2) \text{ e } H_A : p_1 - p_2 \neq 0 \text{ (ou } H_A : p_1 \neq p_2).$$

Testar se as proporções populacionais nos grupos comparados eram iguais, teria sido o mesmo que testar se a diferença entre elas era igual a zero, não é?

3ª) Cálculo do valor de z correspondente a uma diferença entre proporções igual à obtida no estudo realizado, que teria sido $80,4\% - 62,3\% = 18,1\%$ ou $0,181$: para esse cálculo precisaríamos computar o erro-padrão das diferenças entre as proporções de intoxicados nos dois grupos a serem comparados. Ou seja, necessitaríamos saber o quanto, em média, essas diferenças entre proporções variariam em torno da média dessas diferenças entre proporções, caso tivéssemos realizado numerosos estudos semelhantes. Se tivéssemos realmente estudado numerosas amostras, teríamos obtido esse erro-padrão pelo modo tradicional, calculando o desvio da diferença entre as proporções obtidas em cada amostra, em relação à média das diferenças entre essas proporções amostrais, elevando cada um desses desvios ao quadrado, somando-os todos, dividindo essa soma por $k - 1$ (sendo k o número de estudos realizados) e, finalmente, extraindo a raiz quadrada do resultado desta divisão. Mas ocorreria aqui o mesmo fenômeno que já discutimos com você na inferência sobre médias: na prática, cada equipe de pesquisa teria estudado apenas um par de amostras a serem comparadas, obtendo apenas uma diferença entre proporções, o que teria impossibilitado o cálculo acima. Felizmente, como já vimos, foi demonstrado algébrica e empiricamente que é possível obtermos a variância das proporções que seriam obtidas em cada grupo comparado, caso tivéssemos estudado numerosas amostras, mesmo sem estudarmos numerosas amostras. Essa variância em um dos grupos seria dada por pq/n_1 e no outro por pq/n_2 . Note que usaríamos p , e não p_1 e p_2 , isso porque se a hipótese nula for verdadeira, $p_1 = p_2 = p$, ou seja, as proporções populacionais nos grupos comparados seriam iguais e, conseqüentemente, iguais à própria proporção na população, que é p . Encontraremos o erro-padrão das diferenças entre as proporções de intoxicados nos dois grupos comparados, extraindo a raiz quadrada da soma dos dois erros-padrão mencionados acima. prossiga para entender melhor isto.

Pelo exposto acima, o erro-padrão das diferenças entre as proporções de intoxicados nos grupos independentes comparados, seria dado por

$$EP_{(\hat{p}_1 - \hat{p}_2)} = \sqrt{\frac{pq}{n_1} + \frac{pq}{n_2}},$$

onde $EP_{(\hat{p}_1 - \hat{p}_2)}$ representa o erro-padrão das diferenças entre as proporções, p a proporção de intoxicados na população, e q , seu complemento, ou seja, a proporção de não-intoxicados na mesma população.

Mas os valores de p e q geralmente são desconhecidos, do mesmo modo que as médias populacionais, porque os estudos populacionais são caros e impraticáveis, lembra-se disso? Necessitamos então de um estimador válido de p . O melhor estimador de p é obtido calculando-se uma média ponderada das proporções de intoxicados encontrados nos dois únicos grupos estudados. Parece-nos claro que, se não dispusermos da proporção na população-alvo, uma boa alternativa para estimá-la é a média entre as duas proporções das quais dispomos. É também fácil entender que esse estimador será melhor se cada uma dessas proporções for ponderada pelo número de crianças em cada grupo, de modo que o grupo com mais crianças influencie mais o resultado a ser obtido ao calcular-se essa média, porque o número de crianças influencia o valor do erro-padrão. Quanto maior esse número, menor o erro-padrão e vice-versa. Assim, o melhor estimador de p , denotada por \hat{p} , teria sido computado em nosso teste por

$$\hat{p} = \sqrt{\frac{n_1 \hat{p}_1 + n_2 \hat{p}_2}{n_1 + n_2}},$$

onde a prevalência na área próxima, \hat{p}_1 , teria sido multiplicada pelo seu peso, n_1 , sendo o resultado somado ao resultado da multiplicação da prevalência na área distante, \hat{p}_2 , pelo seu peso, n_2 e, essa soma, dividida pela soma dos pesos, como ocorre sempre que calculamos uma média ponderada. A raiz quadrada desse resultado nos teria fornecido o estimador de p que precisávamos.

Calculando \hat{p} no nosso exemplo, teríamos obtido:

$$\begin{aligned} \hat{p} &= \sqrt{\frac{n_1 \hat{p}_1 + n_2 \hat{p}_2}{n_1 + n_2}} = \sqrt{\frac{(112)(0,804) + (138)(0,623)}{112 + 138}} = \sqrt{\frac{90,05 + 85,97}{250}} = \\ &= \sqrt{\frac{176,02}{250}} = \sqrt{0,704} = 0,839 \text{ ou } 83,9\%. \end{aligned}$$

Depois desse cálculo, teria sido possível calcularmos o erro-padrão das diferenças entre as proporções, necessário para calcularmos o valor de z , e também para verificarmos se a aproximação normal da binomial se aplicava ao nosso estudo, pois isso também se baseia no valor de p , cuja estimativa seria essa calculada acima.

Primeiramente, então, teríamos verificado a possibilidade de aplicarmos a aproximação normal da binomial. Como

$$n_1 \hat{p} \hat{q} = (112)(0,839)(0,161) \cong 15,13 > 5$$

e

$$n_2 \hat{p} \hat{q} = (138)(0,839)(0,161) \cong 18,64 > 5,$$

podemos utilizar a aproximação normal no nosso exemplo.

Em seguida, teríamos calculado o erro-padrão das diferenças entre as proporções, caso tivéssemos realizado numerosos estudos semelhantes, com amostras independentes de mesmo tamanho, retiradas da mesma população-alvo. Esse erro-padrão teria sido computado por

$$\begin{aligned} EP_{(\hat{p}_1 - \hat{p}_2)} &= \sqrt{\frac{\hat{p} \hat{q}}{n_1} + \frac{\hat{p} \hat{q}}{n_2}} = \sqrt{\frac{(0,839)(0,161)}{112} + \frac{(0,839)(0,161)}{138}} = \sqrt{\frac{0,135}{112} + \frac{0,135}{138}} = \\ &= \sqrt{0,001 + 0,0001} = \sqrt{0,0011} = 0,033. \end{aligned}$$

Teríamos então calculado o valor de z da seguinte forma:

$$z = \frac{(\hat{p}_1 - \hat{p}_2) - (p_1 - p_2)_o}{EP_{(\hat{p}_1 - \hat{p}_2)}} = \frac{(\hat{p}_1 - \hat{p}_2) - (p_1 - p_2)_o}{\sqrt{\frac{\hat{p} \hat{q}}{n_1} + \frac{\hat{p} \hat{q}}{n_2}}},$$

onde $(\hat{p}_1 - \hat{p}_2)$ estaria denotando a diferença entre as proporções encontradas no único estudo realizado; $(p_1 - p_2)_o$ a diferença estabelecida na hipótese nula; e $EP_{(\hat{p}_1 - \hat{p}_2)}$ o erro-padrão das diferenças entre proporções. Como a hipótese nula estabeleceria que $(p_1 - p_2)_o = 0$, e o teste seria realizado assumindo-se que essa hipótese era verdadeira, a expressão acima teria sido simplificada para

$$z = \frac{(\hat{p}_1 - \hat{p}_2) - (p_1 - p_2)_o}{EP_{(\hat{p}_1 - \hat{p}_2)}} = \frac{(\hat{p}_1 - \hat{p}_2) - 0}{\sqrt{\frac{\hat{p} \hat{q}}{n_1} + \frac{\hat{p} \hat{q}}{n_2}}} = \frac{(\hat{p}_1 - \hat{p}_2)}{\sqrt{\frac{\hat{p} \hat{q}}{n_1} + \frac{\hat{p} \hat{q}}{n_2}}}.$$

Substituindo os valores do nosso exemplo nessa expressão, teríamos encontrado:

$$z = \frac{(\hat{p}_1 - \hat{p}_2)}{EP_{(\hat{p}_1 - \hat{p}_2)}} = \frac{(\hat{p}_1 - \hat{p}_2)}{\sqrt{\frac{\hat{p} \hat{q}}{n_1} + \frac{\hat{p} \hat{q}}{n_2}}} = \frac{(0,804 - 0,623)}{0,033} = \frac{0,181}{0,033} = 5,48;$$

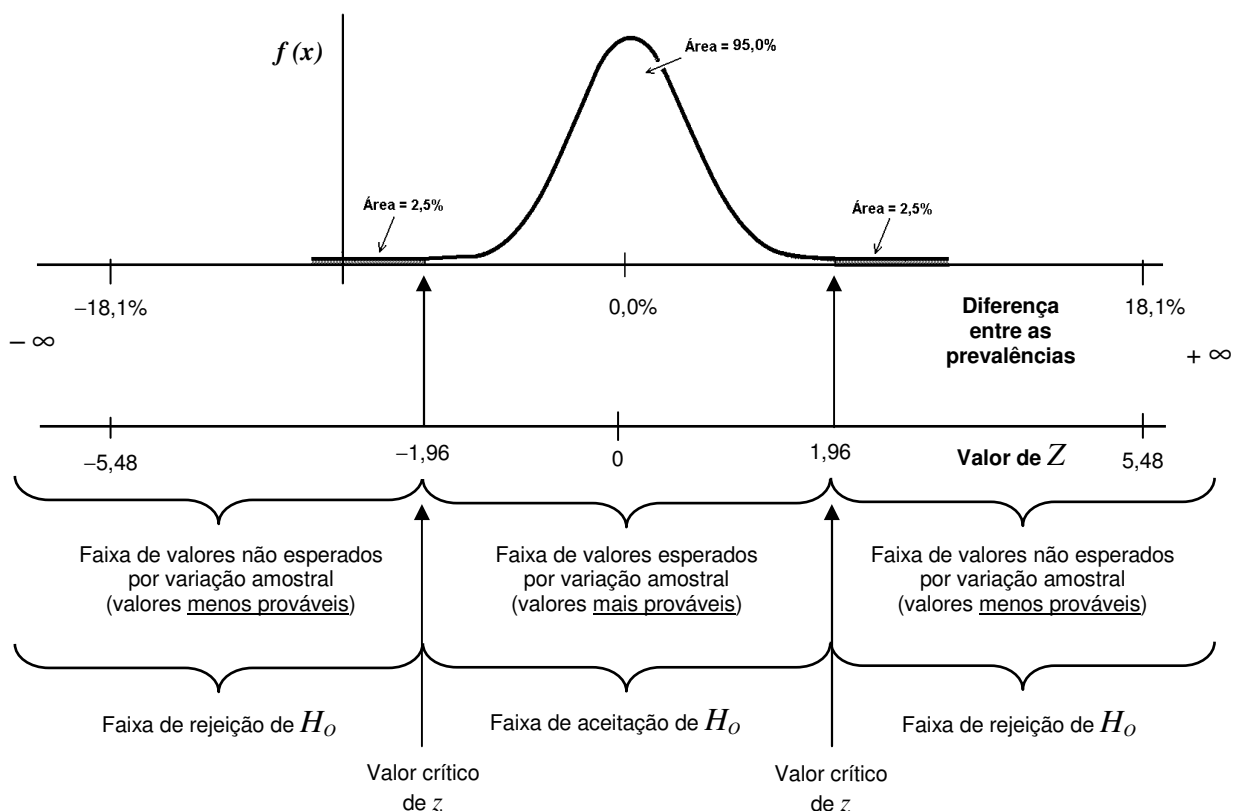
4ª) Obtenção do valor- p : teríamos consultado a parte com valores positivos da tabela da curva normal padrão, para encontrarmos a probabilidade de obter valores de Z menores do que 5,48, ou seja, $P(Z < 5,48)$. Observe que a tabela utilizada não teria sido completa o suficiente para obtermos o valor exato de p , pois o maior valor positivo de z apresentado é 3,89. A probabilidade de obtermos valores menores do que 3,89 é de 0,9999. Então a probabilidade de obtermos valores maiores do que 3,89 é de

$1 - 0,9999 = 0,0001$. Teríamos assim, podido afirmar que a probabilidade de obtermos valores maiores do que 5,48 era ainda menor do que 0,0001, porque 5,48 é um valor mais extremo à direita da curva do que 3,89. Como o teste teria sido bicaudado, teríamos de multiplicar essa probabilidade por dois, obtendo um valor- p menor do que $(0,0001)(2) = 0,0002$.

Veja no próximo diagrama a situação encontrada neste exemplo;

5ª) a) Comparação do valor- p ao valor de α e conclusão do teste:

Como valores menores do que 0,0002 são muito menores do que 0,05 ou, equivalentemente, como valores menores do que 0,02% são muito menores do que 5,0%, teríamos concluído que, estatisticamente, a diferença encontrada no estudo era altamente significante. Assim, teríamos rejeitado H_0 e aceitado H_A . Ou seja, muito provavelmente, a verdadeira diferença entre as prevalências de intoxicação por chumbo nas crianças das áreas próxima e distante era estatisticamente diferente de zero, sendo, portanto, muito improvável que a verdadeira diferença entre essas prevalências na população fosse zero. A diferença igual a zero, foi testada e rejeitada, ao rejeitarmos a hipótese nula. A diferença de 18,1%, encontrada no único estudo realizado, muito provavelmente, não teria sido encontrada em estudos realizados em amostras retiradas de uma população cuja diferença fosse zero. Então, a verdadeira diferença não deve ser zero, sugerindo que as prevalências de intoxicação possam ser diferentes nas áreas próxima e distante, em consequência de diferenças na magnitude da poluição por chumbo nessas áreas. Não se esqueça de que, como sempre, para concluirmos dessa maneira, teria sido fundamental avaliarmos também a ausência de vieses importantes em nossa pesquisa.



ou

b) Comparação do valor observado de z ao valor crítico de z :

Como $5,48 > 1,96$ (e, conseqüentemente, $-5,48 < -1,96$), teríamos concluído que o valor de z , correspondente a uma diferença entre proporções igual a 18,1%, estava em uma localização extrema na curva, ultrapassando o valor crítico de z . Então, uma diferença igual a 18,1% teria sido muito improvável de ser obtida, estando localizada na área de rejeição da hipótese nula. Se a pesquisa realizada não tivesse apresentado vieses importantes, teríamos chegado à mesma conclusão à qual teríamos chegado na letra **a** da quinta etapa.

Outra opção teria sido calcularmos o intervalo de 95% de confiança, dado por

$$IC(95\%) = (\hat{p}_1 - \hat{p}_2) \pm z_{(1-\alpha/2)} \sqrt{\frac{\hat{p}\hat{q}}{n_1} + \frac{\hat{p}\hat{q}}{n_2}}$$

Intervalo de 95% de confiança Diferença entre as proporções obtidas no estudo Valor de z correspondente ao percentil 97,5 Erro-padrão das diferenças entre as proporções, caso tivéssemos estudado numerosas amostras

Utilizando os valores do nosso exemplo, teríamos obtido:

$$IC(95\%) = (\hat{p}_1 - \hat{p}_2) \pm z_{(1-\alpha/2)} \sqrt{\frac{\hat{p}\hat{q}}{n_1} + \frac{\hat{p}\hat{q}}{n_2}} = 0,181 \pm (1,96)(0,033) =$$

$$= 0,181 \pm 0,065 = (0,116 \text{ a } 0,246) = (11,6 \% \text{ a } 24,6 \%).$$

Tais resultados nos teriam indicado que havia uma probabilidade de 95% da verdadeira diferença populacional entre as prevalências estar entre 11,6% e 24,6%. Teríamos considerado qualquer valor dentro desse intervalo como aceitável para a diferença populacional, porque todos os valores dentro desse intervalo teriam sido desviantes, diferentes de 18,1%, apenas por variação amostral de resultados. Qualquer valor fora teria sido considerado como estatisticamente diferente de 18,1%. Como 0,0% (valor equivalente à inexistência de diferença estabelecido na hipótese nula) teria estado fora do intervalo, teríamos concluído que esse valor, muito provavelmente, não teria sido obtido se numerosas amostras tivessem sido estudadas, indicando que esse valor não deveria ser a verdadeira diferença entre as prevalências na população-alvo. A diferença encontrada no único estudo, 18,1%, não teria sido compatível com uma diferença populacional igual a zero. Como esperado, nossa conclusão teria sido igual àquela do teste de hipóteses.

— **Poderíamos ter feito a correção para continuidade também nessa situação?**

— Sim. As etapas teriam sido:

1ª) Definição do nível de significância: $\alpha = 0,05$;

2ª) Definição das hipóteses:

$$H_0 : 80,4\% - 62,3\% = 0 \text{ e } H_A : 80,4\% - 62,3\% \neq 0 ;$$

3ª) Cálculo do valor de z : teríamos aplicado a correção para continuidade subtraindo $\left[(1/2n_1)+(1/2n_2)\right]$ de $|\hat{p}_1 - \hat{p}_2|$, ou seja do módulo de $\hat{p}_1 - \hat{p}_2$. Assim, independentemente do sinal negativo ou positivo da diferença $\hat{p}_1 - \hat{p}_2$ (por isso consideraríamos o módulo dessa diferença), o procedimento teria sido o mesmo mostrado a seguir: teríamos computado o valor de z_c por

$$z_c = \frac{|\hat{p}_1 - \hat{p}_2| - \left(\frac{1}{2n_1} + \frac{1}{2n_2}\right)}{EP_{(\hat{p}_1 - \hat{p}_2)}} = \frac{|\hat{p}_1 - \hat{p}_2| - \left(\frac{1}{2n_1} + \frac{1}{2n_2}\right)}{\sqrt{\frac{\hat{p}\hat{q}}{n_1} + \frac{\hat{p}\hat{q}}{n_2}}} = \frac{|0,804 - 0,623| - \left(\frac{1}{2(112)} + \frac{1}{2(138)}\right)}{0,033} =$$

$$= \frac{0,181 - \left(\frac{1}{224} + \frac{1}{276}\right)}{0,033} = \frac{0,181 - (0,0045 + 0,0036)}{0,033} = \frac{0,181 - 0,0081}{0,033} = \frac{0,173}{0,033} = 5,24,$$

onde z_c , estaria denotando o valor de z corrigido.

Note, novamente, que o valor de z obtido não teria sido muito diferente daquele calculado sem a correção (5,48 comparado a 5,24), mas essa pequena diferença poderia ter nos levado a uma conclusão diferente;

4ª) Obtenção do valor- p : teríamos consultado a parte com valores positivos da tabela da curva normal padrão, para encontrarmos a probabilidade de obter um valor de Z menor do que 5,24, ou seja, a $P(Z < 5,24)$. Não teríamos conseguido obter essa probabilidade na tabela incompleta utilizada, mas a probabilidade de obtermos valores menores do que 3,89 (maior valor positivo de Z na tabela) seria 0,9999, como já vimos. Para calcularmos a área do nosso interesse, teríamos que ter diminuído esse valor de 1, obtendo:

$$P(Z > 3,89) = 1 - P(Z < 3,89) = 1 - 0,9999 = 0,0001.$$

Se esta teria sido a probabilidade de obtermos valores maiores ou iguais a 3,89, a probabilidade de obtermos valores maiores ou iguais a 5,24 teria sido ainda menor. Assim multiplicaríamos 0,0001 por dois, porque o teste teria sido bicaudado, e teríamos podido afirmar que o valor- p do nosso teste era muito menor do que 0,0002;

5ª) Comparação do valor- p ao valor de α e conclusão do teste:

Como valores muito menores do que 0,0002 são muito menores do que 0,05 ou, equivalentemente,

como valores muito menores do que 0,02% são muito menores do que 5,0%, teríamos concluído que a verdadeira diferença na população era estatisticamente diferente de zero. Assim, teríamos rejeitado H_0 e aceitado H_A . Nesse caso, portanto, nossa conclusão teria sido a mesma à qual teríamos chegado sem a correção para continuidade.

— Tenho ouvido falar muito no teste qui-quadrado, que também é usado para a comparação de duas proporções. Vamos estudá-lo neste livro?

— Vamos dedicar o capítulo 16 a esse famoso teste, continuando com o mesmo exemplo, para que possamos compará-lo ao teste que acabamos de realizar. Antes porém, aproveitando que os conhecimentos adquiridos por você nos últimos capítulos ainda devem estar bem nítidos em sua memória, vamos apresentar duas das fórmulas que utilizamos para calcular o número necessário de indivíduos a serem incluídos em nossa amostra.

CAPÍTULO 15

-
- Quais os fundamentos estatísticos para o cálculo do tamanho da amostra?
 - Como calculamos o tamanho da amostra para estimar uma média?
 - Como calculamos o tamanho da amostra para estimar uma proporção?
-



— Quais os fundamentos estatísticos para o cálculo do tamanho da amostra?

— Se desejarmos saber qual a média de glicemia na população de uma das capitais brasileiras, você acha que poderíamos pesquisar apenas um indivíduo residente naquela cidade? Evidentemente não! Seria impossível querermos estimar a média glicêmica em uma população de mais de dois milhões de habitantes, considerando apenas um indivíduo. Nem mesmo seria possível calcularmos uma média com apenas um valor.

E se pesquisássemos dois indivíduos? Ou três, quatro, cinco ou seis? São poucos ainda, não são? Então vamos aumentar mais esse número. Que tal 200, 300, 400 ou quem sabe, 1.000? Provavelmente essas últimas quantidades seriam melhores do que as anteriores, mas será que seriam suficientes? É justamente essa a situação de incerteza na qual nos encontramos quando estamos planejando um estudo epidemiológico a ser realizado com uma parte da população-alvo. Há muito tempo os estatísticos estudam esse problema, e alguns procedimentos têm sido desenvolvidos para diminuir nossa incerteza com relação ao número mínimo suficiente para fazermos um estudo desse tipo.

Em vários momentos neste livro, escrevemos que é necessário calcularmos o número mínimo (n mínimo) de pessoas a serem investigadas em nossas pesquisas, e que existem fórmulas específicas para isso. Essas variam a depender do tipo de estudo epidemiológico realizado e do parâmetro a ser estimado.

Como este é um livro de Bioestatística básica, abordaremos apenas duas dessas fórmulas, uma usada em estudos que pretendem estimar uma média e a outra em estudos para estimar uma proporção. Mas os procedimentos discutidos aqui são os mesmos utilizados para a obtenção de muitas outras fórmulas, e lhe serão úteis em várias situações nas quais esteja calculando o n mínimo requerido para o seu estudo.

Voltando ao exemplo da glicemia, nossos primeiros passos para o cálculo serão: definirmos qual o parâmetro que desejamos estimar, escolhermos a fórmula adequada a esse parâmetro e ao tipo de estudo epidemiológico a ser realizado, e obtermos as informações necessárias exigidas pela fórmula.

— Como calculamos o tamanho da amostra para estimar uma média?

— Se desejarmos estimar uma média, o n mínimo será calculado por

$$n = \frac{z^2 \sigma^2}{d^2},$$

onde z é o valor da distribuição normal padrão correspondente ao nível de confiança desejado para nossa estimação; σ^2 é a variância das glicemias na população-alvo; e d a margem de erro (também chamado de grau de precisão) que nos permitimos ter ao fazer a estimação. Prossiga, pois os elementos dessa fórmula ficarão mais claros mais adiante.

A fórmula acima foi obtida através de algumas etapas de álgebra muito simples e fáceis de entender.

O ponto de partida para estabelecermos o n do nosso estudo, é definirmos o quanto nos permitimos de erro ao estimarmos um parâmetro populacional com base apenas em uma parte dessa população. Ou

seja, temos de decidir sobre qual a margem de erro que admitimos.

— Mas, não queremos erro algum em nossa pesquisa!

— Quando estudamos uma amostra da população, o erro ao qual estamos nos referindo é inevitável. Não existe, pelo menos por enquanto, uma mágica que nos permita obter, com total precisão, um valor populacional com base em apenas uma parte da população. Haverá sempre uma imprecisão, uma margem de erro. Este é um dos erros chamados de aleatórios, porque os indivíduos que compõem a amostra estudada são, geralmente, selecionados por sorteio (aleatoriamente), o que resulta em uma variação ao acaso dos resultados, se estudássemos numerosas amostras, como já vimos em outros capítulos. É esse tipo de erro que tentamos levar em conta quando fazemos inferência estatística. No estágio atual da ciência, a única maneira de evitarmos completamente esse erro seria estudando todos os indivíduos da população-alvo.

No nosso exemplo, vamos supor que estudos prévios similares ao que vamos realizar (populacionais ou amostrais), em indivíduos com características semelhantes aos que desejamos estudar, encontraram uma média glicêmica de 90 mg/dL. Com essa informação, vamos reunir nossa equipe e avaliar qual a margem de erro que toleraremos. Talvez seja muito alta a imprecisão da nossa estimativa, se estudarmos uma amostra de tal tamanho que esse possa se afastar 10 mg/dL, para cima ou para baixo da verdadeira média populacional. Se escolhêssemos tal margem de erro, nossa estimativa poderia variar entre 80 e 100 mg/dL, quando sabemos que a verdadeira média deve ser muito próxima a 90 mg/dL. E um erro de 5 mg/dL? Ao escolhermos tal margem de erro, nossa estimativa poderia variar entre 85 e 95 mg/dL. Se aceitarmos esse grau de imprecisão, substituiremos a notação d na fórmula acima, por 5 mg/dL, e prosseguiremos tentando obter as demais informações necessárias.

Mas, a essa altura da nossa apresentação, já podemos lhe mostrar uma das etapas algébricas para chegarmos a essa fórmula.

Já sabemos que nosso ponto de partida é a definição da margem de erro da nossa estimação, que é denotada por d . Com base no que apresentamos em capítulos anteriores, já sabemos também que na Estatística, sempre que pudermos utilizar a distribuição normal padrão como modelo para inferência, essa margem de erro ou grau de precisão é medida pela quantidade

$$z_{(1-\alpha/2)} EP,$$

onde $z_{(1-\alpha/2)}$ denota o valor da curva normal padrão correspondente a um determinado nível de confiança definido por nós, e EP , como sempre, o erro-padrão do parâmetro a ser estimado. Podemos então expressar nossa margem de erro por

$$d = z_{(1-\alpha/2)} EP.$$

Este erro, como já vimos para o cálculo de intervalos de confiança, deve ser considerado para cima ou para baixo do parâmetro a ser estimado, mas note que aqui não será necessário considerarmos ambas as direções do erro (para baixo e para cima do parâmetro), porque dessa maneira teríamos também de considerar $2d$ e não d apenas, o que nos levaria à mesma fórmula acima, como mostramos a seguir:

$$2d = 2z_{(1-\alpha/2)} EP, \text{ donde resulta que } d = z_{(1-\alpha/2)} EP,$$

após a divisão dos termos da equação por dois. Assim, só é necessário considerarmos uma direção do erro, para determinarmos o n mínimo do nosso estudo, embora alguns livros e programas estatísticos para computador utilizem fórmulas que levam em conta as duas direções. Os resultados obtidos serão os mesmos.

Continuando, podemos substituir apropriadamente, na expressão acima, EP por σ/\sqrt{n} , obtendo:

$$d = z_{(1-\alpha/2)} \left(\frac{\sigma}{\sqrt{n}} \right),$$

já que estamos desenvolvendo a fórmula para determinação do n mínimo de uma pesquisa que visa estimar uma média, e o erro-padrão de uma média, como já cansamos de ver, é dado por σ/\sqrt{n} , quando σ é conhecido.

— **E se for σ desconhecido, como quase sempre ocorre?**

— Nesse caso, teremos que encontrar uma estimativa válida para σ . As formas mais freqüentemente usadas para obtê-la são:

- a) Estimativas de σ podem estar disponíveis em estudos similares já realizados, tais como aqueles que hipoteticamente mencionamos mais acima, e dos quais obtivemos uma estimativa para a média glicêmica esperada para a população;
- b) Realização de um estudo-piloto de tamanho pequeno definido arbitrariamente, e viável de ser investigado por sua equipe em curto espaço de tempo. Usaremos o desvio-padrão obtido nesse estudo como estimador de σ . Se for possível, devemos coletar os dados dos indivíduos incluídos no estudo com o maior rigor possível, de modo a considerá-los como indivíduos do estudo principal, o que evitará desperdício de tempo e de outros recursos;
- c) Se pudermos assumir que a média da variável cuja média queremos estimar tem uma distribuição normal na população de onde a amostra será retirada, e sabendo que a amplitude de variação (valor máximo menos valor mínimo, lembra-se?) de uma distribuição normal é igual a 6σ , temos que

$$Amplitude = 6\sigma, \text{ donde } \sigma = \frac{Amplitude}{6}.$$

Para usarmos esse procedimento, necessitaremos conhecer os valores populacionais máximo e mínimo da variável a ser estudada, o que é mais fácil do que sabermos o desvio-padrão populacional, embora possam também não estar disponíveis.

Retomando o desenvolvimento da fórmula, já vimos como encontrar os valores de d e de σ . Agora

vamos encontrar o valor de $z_{(1-\alpha/2)}$. Tal valor dependerá de qual foi o nível de confiança que escolhemos para a nossa pesquisa. O nível de confiança é o complemento do nível de significância. Do mesmo modo que 0,05 é o nível de significância mais utilizado, 0,95 é o nível de confiança mais usado, pois, $1 - 0,05 = 0,95$ ou 95%. Esse é o mesmo nível de confiança que utilizamos nos cálculos dos intervalos de confiança em capítulos anteriores. Os 5% restantes são divididos em duas partes iguais nas duas extremidades da distribuição. Você deve então se lembrar muito bem de que, se a variável a ser estimada tem uma distribuição normal, o valor da curva normal que separa os 97,5% valores mais baixos dos 2,5% valores mais altos, é igual a $z_{(1-\alpha/2)} = z_{(1-0,05/2)} = z_{(1-0,025)} = z_{0,975} = z_{97,5}$, que é igual a 1,96. Então, ao definirmos que $d = 5$ mg/dL e que $\alpha = 5,0\%$, estamos estabelecendo que desejamos um certo tamanho de amostra de tal modo que tenhamos uma probabilidade de 95% de que a verdadeira média populacional esteja entre 85 e 95 mg/dL. A implicação disso na fórmula é que $z_{(1-\alpha/2)} = 1,96$.

Já temos então, no nosso exemplo, os valores de d e $z_{(1-\alpha/2)}$. Suponha que tenhamos realizado um estudo-piloto com apenas 30 indivíduos (número estabelecido arbitrariamente), no qual encontramos um desvio-padrão $s = 19,7$ mg/dL. Usaremos esse valor como estimativa de σ . Observe a fórmula novamente e veja que só nos falta o valor de n , mas este é justamente o nosso objetivo depois de todo esse esforço: encontrar o valor de n , que nos indicará o número mínimo de indivíduos que precisaremos investigar, de tal modo que haja 95% de probabilidade da verdadeira média populacional estar entre 85 e 95 mg/dL, variação considerada por nós como aceitável.

Nosso próximo passo então, será resolver a equação para n , obtendo:

$$d = z_{(1-\alpha/2)} EP$$

$$d = z_{(1-\alpha/2)} \left(\frac{\sigma}{\sqrt{n}} \right)$$

$$d^2 = z_{(1-\alpha/2)}^2 \left(\frac{\sigma}{\sqrt{n}} \right)^2$$

$$d^2 = z_{(1-\alpha/2)}^2 \left(\frac{\sigma^2}{n} \right)$$

$$d^2 = \frac{z_{(1-\alpha/2)}^2 \sigma^2}{n}$$

$$nd^2 = z_{(1-\alpha/2)}^2 \sigma^2$$

$$n = \frac{z_{(1-\alpha/2)}^2 \sigma^2}{d^2}.$$

Em uma das etapas algébricas, todos os termos da equação foram elevados ao quadrado, com o

objetivo de eliminarmos a raiz quadrada de um dos termos (\sqrt{n}), simplificando a equação.

Veja que, por mais inesperado que pareça, o tamanho da população-alvo, que poderíamos pensar ser muito importante para o cálculo do tamanho da amostra, não foi considerado no cálculo.

Essa é a fórmula para calcularmos o n mínimo de pesquisas que visem estimar uma média, mas só é válida se a amostragem for com reposição e de uma população infinita, ou seja, grande o suficiente para podermos ignorar a correção para população finita.

— Por quê?

— Porque se a amostragem for sem reposição (como ocorre nas pesquisas epidemiológicas) e de população finita (como em muitas dessas pesquisas), foi demonstrado que o erro-padrão não será σ/\sqrt{n} , impedindo-nos de utilizar esse valor na fórmula. Nessas circunstâncias, foi mostrado que o erro-padrão é dado por

$$\frac{\sigma}{\sqrt{n}} \sqrt{\frac{N-n}{N-1}}.$$

Sugerimos que revise o assunto na página 30.

A quantidade $\sqrt{(N-n)/(N-1)}$ é chamada de correção para população finita. Quando essa correção for requerida, o nosso ponto de partida se tornará:

$$d = z_{(1-\alpha/2)} \left(\frac{\sigma}{\sqrt{n}} \sqrt{\frac{N-n}{N-1}} \right) \text{ que, quando resolvida para } n, \text{ resultará em}$$

$$n = \frac{N z_{(1-\alpha/2)}^2 \sigma^2}{d^2 (N-1) + z_{(1-\alpha/2)}^2 \sigma^2}.$$

A álgebra envolvida na resolução dessa equação não foi apresentada, porque é muito semelhante à que será apresentada no APÊNDICE 3 para obtermos o n mínimo, quando desejamos estimar uma proporção.

Resumindo:

Se a amostragem for com reposição e de população infinita, o n mínimo do estudo será calculado por

$$n = \frac{z_{(1-\alpha/2)}^2 \sigma^2}{d^2}.$$

Se a amostragem for sem reposição e de população finita, o n mínimo do estudo será calculado por

$$n = \frac{N z_{(1-\alpha/2)}^2 \sigma^2}{d^2 (N-1) + z_{(1-\alpha/2)}^2 \sigma^2}.$$

Substituindo com os valores do nosso exemplo, temos:

$$n = \frac{z_{(1-\alpha/2)}^2 \sigma^2}{d^2} = \frac{(1,96)^2 (19,7)^2}{(5)^2} = \frac{(3,84)(388,09)}{25} = \frac{1.490,27}{25} \cong 59,6 \cong 60 \text{ indivíduos}.$$

O número foi aproximado para 60 porque representa indivíduos, não existindo frações de indivíduos, pelo menos no sentido matemático. Este será o n mínimo se nossa amostragem for com reposição e em população infinita, uma população residente em uma grande capital brasileira como mencionamos antes. Suponha que essa população seja composta por 2.600.000 habitantes. Lembre-se de que para considerarmos uma população como infinita n/N deve ser menor ou igual a 5% , ou seja, a população-alvo tem que ser muito maior do que o n mínimo calculado. No exemplo, temos que $n/N = 60/2.600.000 = 0,0000231 = 0,00231\%$, que é um número muito menor do que 5%.

Suponha agora que a população-alvo do nosso estudo seja o total de professores de primeiro e segundo grau de escolas particulares, de uma cidade de porte médio do interior de um dos estados brasileiros. Com dados da Secretaria Estadual de Educação e dos sindicatos patronal e de professores, obtivemos um total de 500 professores nessas escolas naquela cidade, no período de planejamento do estudo. Suponha também que nossa amostragem será sem reposição, e que definimos os mesmos valores para d , $z_{(1-\alpha/2)}$ e σ , da situação anterior. Sem a correção para população finita, o n mínimo seria

$$n = \frac{z_{(1-\alpha/2)}^2 \sigma^2}{d^2} = \frac{(1,96)^2 (19,7)^2}{(5)^2} = \frac{(3,84)(388,09)}{25} = \frac{1.490,27}{25} \cong 59,6 \cong 60 \text{ indivíduos},$$

ou seja, o mesmo obtido anteriormente.

Acontece que agora, como $n/N = 60/500 = 0,12 = 12,0\%$, nossa população é finita. Além disto, a amostragem será feita sem reposição e, portanto, não poderemos ignorar a correção para população finita. Desse modo, calcularemos o n mínimo por

$$n = \frac{N z_{(1-\alpha/2)}^2 \sigma^2}{d^2 (N-1) + z_{(1-\alpha/2)}^2 \sigma^2} = \frac{(500)(1,96)^2 (19,7)^2}{(5)^2 (500-1) + (1,96)^2 (19,7)^2} =$$

$$= \frac{(500)(3,84)(388,09)}{(25)(499) + (3,84)(388,09)} = \frac{745.132,80}{12.475 + 1.490,27} = \frac{745.132,80}{13.965,27} = 53,36 \cong 54 \text{ indivíduos}.$$

Este será o número de indivíduos a serem estudados por nós.

Na prática, podemos utilizar apenas o critério de se $n/N > 5,0\%$ ou $n/N \leq 5,0\%$ para decidirmos se faremos ou não a correção para população finita, porque é disso que dependerá o quanto o valor de σ/\sqrt{n} se aproximará do valor de $(\sigma/\sqrt{n})(\sqrt{N-n/N-1})$. Assim, na prática, mesmo que a amostragem seja feita sem reposição, se $n/N \leq 5,0\%$ não será necessário fazermos a correção.

Recomendamos que sempre sejam acrescentados mais alguns indivíduos na amostra, pois é muito comum ocorrerem recusas e/ou perdas no início (não-respostas) ou durante a pesquisa. Observe que arredondamos o número para mais e não para menos, porque sabemos que, quanto maior o n maior a precisão dos nossos resultados, mesmo que o impacto desses arredondamentos seja pequeno.

Devemos aumentar mais ainda o n mínimo calculado, se nosso processo de amostragem prever alguma etapa de amostragem por conglomerados. Se, no nosso exemplo, formos sortear escolas, antes de sortear os professores, devemos estabelecer um “efeito do desenho da amostragem”, que consiste em um fator que será multiplicado pelo n mínimo, aumentando-o. Esse fator é definido pelo pesquisador, não existindo valores previamente estabelecidos. O pesquisador utilizará seu conhecimento sobre o tema investigado, para definir o melhor valor para o fator.

No nosso exemplo, poderíamos avaliar que seria necessário multiplicarmos o n mínimo pelo fator 1,2, o que, na prática, significaria um aumento de 20% no tamanho de nossa amostra. A magnitude desse fator deve ser buscada na literatura, mas muitas vezes não é encontrada. Neste caso, os pesquisadores devem utilizar sua experiência no tema para definir o valor mais apropriado, sempre tendo em mente que quanto maior o tamanho da amostra melhor o poder do estudo.

— Mas, por que teremos de aumentar o tamanho da amostra se houver amostragem por conglomerados?

— Ao sortearmos escolas (conglomerados) e, depois, nas escolas sorteadas, escolhermos os professores, há uma tendência dos professores de uma mesma escola apresentarem características mais próximas entre si, do que se pertencessem a diferentes escolas. Professores que ensinam na mesma escola devem ter seguido caminhos semelhantes para chegar até aquela função, naquela escola. Isso tende a torná-los mais semelhantes. Se os professores selecionados para a amostra são mais semelhantes, as variáveis que queremos estudar tendem a variar menos entre eles, obrigando-nos a estudar um número maior de professores, desde que fenômenos que variam menos são mais difíceis de serem detectados.

Esse aspecto fica fácil também de entender olhando novamente as fórmulas utilizadas acima:

$$n = \frac{z_{(1-\alpha/2)}^2 \sigma^2}{d^2} \text{ e } n = \frac{N z_{(1-\alpha/2)}^2 \sigma^2}{d^2 (N-1) + z_{(1-\alpha/2)}^2 \sigma^2}. \text{ Em ambas, uma maior semelhança entre os indivíduos}$$

a serem estudados resultaria em uma menor variância (σ^2), o que provocaria uma diminuição nos numeradores, diminuindo, conseqüentemente, o resultado final.

Se levássemos em conta em nosso cálculo o efeito do desenho, teríamos:

$$n_{\text{final}} = (n_{\text{inicial}})(fator) = (60)(1,2) = 72 \text{ indivíduos} \text{ ou}$$

$$n_{\text{final}} = (n_{\text{inicial}})(fator) = (54)(1,2) = 64,8 \cong 65 \text{ indivíduos}.$$

— E como calculamos o tamanho da amostra para estimar uma proporção?

— Se quisermos, por exemplo, estimar a prevalência da diabetes em uma determinada população e não a média glicêmica, em primeiro lugar, teremos de encontrar na literatura médica estudos semelhantes ao que desejamos realizar, e que tenham utilizado um escore de corte, para separar os indivíduos diabéticos dos não-diabéticos. Já sabemos que esse escore, para a dosagem da glicemia em jejum, utilizando o método enzimático, é 110 mg/dL. Assim, os indivíduos com glicemia maior do que 110 mg/dL são classificados como diabéticos e os demais como não-diabéticos.

Vamos supor que tais estudos prévios sugeriram que a prevalência da diabetes em capitais brasileiras seja de 13%. Essa prevalência será denotada por p . Se não existirem estudos prévios que nos forneçam esse valor, utilizaremos os dois primeiros procedimentos já explicados quando discutimos o cálculo do n mínimo para estimarmos uma média, na página 264.

Nosso ponto de partida para o cálculo do n mínimo será o mesmo das situações já vistas neste capítulo: vamos avaliar com nossa equipe o quanto aceitaremos de erro na nossa estimação. Se esperarmos uma prevalência de 13% na população-alvo, talvez possamos tolerar uma variação de 3 pontos percentuais (3 pp), para cima ou para baixo desse valor. Logo, esse erro será dado por

$$d = z_{(1-\alpha/2)} EP.$$

No capítulo anterior (página 248), vimos que o erro-padrão de proporções amostrais é calculado por

$\sqrt{\frac{pq}{n}}$, onde $q = 1 - p$. Substituindo na fórmula acima, temos:

$$d = z_{(1-\alpha/2)} \left(\sqrt{\frac{pq}{n}} \right).$$

Resolvendo essa equação para n , obtemos:

$$d = z_{(1-\alpha/2)} \left(\sqrt{\frac{pq}{n}} \right)$$

$$d^2 = z_{(1-\alpha/2)}^2 \left(\sqrt{\frac{pq}{n}} \right)^2$$

$$d^2 = z_{(1-\alpha/2)}^2 \left(\frac{pq}{n} \right)$$

$$d^2 = \frac{z_{(1-\alpha/2)}^2 pq}{n}$$

$$nd^2 = z_{(1-\alpha/2)}^2 pq$$

$$n = \frac{z_{(1-\alpha/2)}^2 pq}{d^2}.$$

Se a correção para população finita não puder ser descartada, o erro-padrão de proporções amostrais terá de ser multiplicado por $\sqrt{(N-n)/N-1}$, e nosso ponto de partida será:

$$d = z_{(1-\alpha/2)} EP = z_{(1-\alpha/2)} \sqrt{\frac{pq}{n}} \sqrt{\frac{N-n}{N-1}}.$$

Resolvendo essa equação para n , obtemos:

$$n = \frac{N z_{(1-\alpha/2)}^2 pq}{d^2 (N-1) + z_{(1-\alpha/2)}^2 pq}.$$

A álgebra necessária para chegarmos até esta fórmula é mostrada no APÊNDICE 3, no final do livro, para satisfazer sua curiosidade e para você aceitá-la sem desconfianças.

Resumindo:

Se a amostragem for com reposição e de população infinita, o n mínimo do estudo será calculado por

$$n = \frac{z_{(1-\alpha/2)}^2 pq}{d^2}.$$

Se a amostragem for sem reposição e de população finita, o n mínimo do estudo será calculado por

$$n = \frac{N z_{(1-\alpha/2)}^2 pq}{d^2 (N-1) + z_{(1-\alpha/2)}^2 pq}.$$

No nosso exemplo, já definimos que o valor de d é 3 pp, o valor de $z_{(1-\alpha/2)}$ é o nosso velho conhecido 1,96 e p é 13%. Substituindo esses valores na primeira fórmula apresentada acima, obtemos:

$$\begin{aligned} n &= \frac{z_{(1-\alpha/2)}^2 pq}{d^2} = \frac{(1,96)^2 (0,13)(0,87)}{(0,03)^2} = \frac{(3,84)(0,13)(0,87)}{0,0009} = \\ &= \frac{0,43}{0,0009} = 477,78 \cong 478 \text{ indivíduos}. \end{aligned}$$

No caso, a correção para população finita pode ser ignorada porque estamos considerando que a amostragem será feita sem reposição, mas $n/N = 478/2.600.000 = 0,000184 = 0,0184\%$, que é bem menor do que 5%.

Suponha que agora nossa população-alvo seja aquela constituída por professores de primeiro e segundo grau de uma cidade do interior, e que esperamos encontrar também nessa população uma prevalência da diabetes de 13%. Além disso, suponha que nossa amostragem seja feita sem reposição e que decidamos manter a mesma margem de erro. Nesse caso, o n inicial obtido será o mesmo já calculado acima: 478 indivíduos. Mas, como $n/N = 478/500 = 0,956 = 95,6\%$, que é muito maior do que 5%, não poderemos ignorar a correção para população finita, e o n terá de ser calculado por

$$n = \frac{N z_{(1-\alpha/2)}^2 pq}{d^2 (N-1) + z_{(1-\alpha/2)}^2 pq} = \frac{(500)(1,96)^2 (0,13)(0,87)}{(0,03)^2 (500-1) + (1,96)^2 (0,13)(0,87)} =$$

$$= \frac{(500)(3,84)(0,13)(0,87)}{(0,0009)(499) + (3,84)(0,13)(0,87)} = \frac{217,15}{(0,45) + (0,43)} =$$

$$= \frac{217,15}{0,88} = 246,76 \cong 247 \text{ indivíduos}.$$

Note que o n caiu bastante com a correção.

Se estivermos realizando amostragem por conglomerados, semelhantemente ao que fizemos em estudos para estimar uma média, estabeleceremos um fator para levar em conta o efeito do desenho sobre o nosso cálculo.

Supondo que tenhamos definido 1,2 como fator, nosso cálculo final seria:

$$n_{\text{final}} = (n_{\text{inicial}})(\text{fator}) = (478)(1,2) = 573,6 \cong 574 \text{ indivíduos} \text{ ou}$$

$$n_{\text{final}} = (n_{\text{inicial}})(\text{fator}) = (247)(1,2) = 296,4 \cong 297 \text{ indivíduos}.$$

No próximo capítulo retomaremos a inferência sobre duas proporções, desta vez utilizando o famoso teste qui-quadrado.

CAPÍTULO 16

-
- Qual a aplicação mais comum do teste qui-quadrado? Por que esse teste tem essa denominação?
 - Como realizamos o teste qui-quadrado para avaliar a independência entre variáveis?
 - Existem outras aplicações para o teste qui-quadrado?
-



— Qual a aplicação mais comum do teste qui-quadrado? Por que esse teste tem essa denominação?

— A aplicação mais comum do teste qui-quadrado é para verificarmos se duas variáveis são independentes ou não, isto é, se a presença de uma está associada, se influencia, portanto, a existência da outra, quando essas variáveis expressam a frequência de determinados eventos.

— Como realizamos o teste qui-quadrado para avaliação da independência entre variáveis?

— Mantendo o mesmo exemplo sobre intoxicação por chumbo em crianças, as hipóteses do teste qui-quadrado desse tipo, são:

H_0 : “área de localização do domicílio” e “intoxicação por chumbo” são independentes, e

H_A : “área de localização do domicílio” e “intoxicação por chumbo” não são independentes.

Outra característica é que o teste qui-quadrado utiliza frequências absolutas em vez de frequências relativas (proporções, prevalências), que foram usadas nos testes feitos no capítulo 14.

Seu ponto de partida é a elaboração de uma tabela de contingência 2×2 (lê-se “tabela de contingência dois por dois”), como a que utilizaremos no nosso exemplo, mas pode ser também uma tabela de contingência maior, $l \times c$, onde l indica o número de categorias da variável apresentada nas linhas, e c o da apresentada nas colunas, sendo $l > 2$ e/ou $c > 2$. Veja, abaixo, a tabela contendo as frequências de indivíduos que foram observadas no nosso exemplo:

Distribuição das crianças, segundo área de localização do domicílio em relação à fundição de chumbo e presença de intoxicação por chumbo.

Área de localização do domicílio	Intoxicação por chumbo		Total
	Sim	Não	
Próxima	90	22	112
Distante	86	52	138
Total	176	74	250

Lembre-se de que a tabela acima é 2×2 porque cada uma das duas variáveis comparadas tem apenas duas categorias, ou seja, a área de localização é classificada em próxima ou distante, e a intoxicação por chumbo admite também apenas duas alternativas, sim ou não. Mais adiante neste capítulo (páginas 284 a 287), veremos como aplicar o teste qui-quadrado em comparações de variáveis com mais de duas categorias.

O próximo passo será calcularmos quais seriam as frequências esperadas de indivíduos, se admitirmos que a hipótese nula é verdadeira, ou seja, que área de residência e intoxicação são fenômenos independentes. Nosso objetivo será comparar as frequências observadas às esperadas. Se não houver muita discrepância entre essas frequências, isso pode indicar que as variáveis comparadas são realmente

independentes, pois as freqüências observadas não diferiram muito daquelas que seriam esperadas se esses eventos fossem independentes. Grandes discrepâncias nos indicarão o contrário. Simples, não é?

Vamos então mostrar como calculamos as freqüências esperadas e, em seguida, como comparamos essas às freqüências observadas.

Com base nos dados disponíveis na pesquisa que realizamos (tabela anterior), podemos afirmar que a probabilidade de uma criança residir em domicílio da área próxima é $112/250 = 0,448$, isto é, das 250 crianças estudadas, 112 realmente moravam nessa área. Lembre-se de que uma probabilidade indica sempre a relação entre o número de eventos que realmente ocorreram (112) e o número total de eventos que poderiam ter ocorrido (250), porque teoricamente seria possível que todas as crianças morassem na área próxima. Do mesmo modo, a probabilidade de uma criança estar intoxicada é $176/250 = 0,704$.

Lembre-se de que, no nível médio de educação, você aprendeu que se $P(A)$ é a probabilidade do evento A ocorrer e $P(B)$ a probabilidade do evento B ocorrer, se A e B forem independentes, a probabilidade de A e B ocorrerem simultaneamente será dada por

$$P(A)P(B).$$

Logo, a probabilidade de uma criança morar em área próxima e estar intoxicada é calculada por

$$P(\text{morar em área próxima}) P(\text{estar intoxicada}) = (0,448)(0,704) = 0,315392.$$

Você também aprendeu que para obtermos o número esperado de crianças residindo em área próxima e com intoxicação na amostra de 250 crianças estudadas, se esses dois eventos forem independentes, procederemos da seguinte maneira:

$$\begin{aligned} \text{Número esperado de intoxicados na área próxima} &= (0,315392)n = (0,315392)(250) = \\ &= 78,848 \cong 78,85 \text{ crianças}. \end{aligned}$$

Veja em que esse cálculo se baseia: como, $0,315392 = 31,5392\%$, encontramos o número esperado, através da seguinte regra de três:

$$\frac{100}{31,5392} = \frac{250}{x},$$

que se lê: “100 por cento está para 31,5392 por cento, assim como 250 está para x ” (que é o número que desejamos encontrar). Se

$$\frac{100}{31,5392} = \frac{250}{x}, \text{ então}$$

$$x = \frac{(250)(31,5392)}{100} = \frac{(31,5392)(250)}{100} = \frac{31,5392}{100}(250) = (0,315392)(250) = 78,848 \cong 78,85 \text{ crianças},$$

que é o número esperado que já tínhamos calculado acima de um modo mais rápido.

Observe que chegaremos ao mesmo resultado obtido acima se multiplicarmos o total marginal da linha correspondente à área próxima ao total marginal correspondente à presença de intoxicação, e dividirmos o resultado pelo total geral da tabela reapresentada no final desta página, como mostramos a seguir:

$$\text{Número esperado de intoxicados na área próxima} = \frac{(112)(176)}{250} = \frac{19.712}{250} = 78,848 \cong 78,85 \text{ crianças}.$$

Isto ocorre porque, como já vimos

$$P(\text{morar em área próxima}) P(\text{estar intoxicada}) = (0,448)(0,704) = \left(\frac{112}{250}\right)\left(\frac{176}{250}\right)$$

e, para obtermos o número esperado temos, como já vimos, de multiplicar a quantidade acima pelo número total de crianças estudadas, ou seja,

$$\text{Número esperado de intoxicados na área próxima} =$$

$$= \left(\frac{112}{250}\right)\left(\frac{176}{250}\right)n = \left(\frac{112}{250}\right)\left(\frac{176}{\cancel{250}}\right)\cancel{250} =$$

$$= \frac{(112)(176)}{250} = 78,848 \cong 78,85 \text{ crianças}.$$

Note também que obtivemos um número fracionário para indicar número de indivíduos, o que biologicamente seria um absurdo, mas manteremos o resultado com dois decimais, já que o número esperado é um conceito teórico e não empírico.

Usaremos o mesmo procedimento para calcular o número esperado para as demais células da tabela, que reapresentamos aqui, para facilitar o seu acompanhamento dos cálculos a seguir:

Distribuição das crianças, segundo área de localização do domicílio em relação à fundição de chumbo e presença de intoxicação por chumbo.

Área de localização do domicílio	Intoxicação por chumbo		Total
	Sim	Não	
Próxima	90	22	112
Distante	86	52	138
Total	176	74	250

$$\text{Número esperado de não-intoxicados na área próxima} = \frac{(112)(74)}{250} = \frac{8.288}{250} \cong 33,15 \text{ crianças}.$$

$$\text{Número esperado de intoxicados na área distante} = \frac{(138)(176)}{250} = \frac{24.288}{250} \cong 97,15 \text{ crianças}.$$

$$\text{Número esperado de não-intoxicados na área distante} = \frac{(138)(74)}{250} = \frac{10.212}{250} \cong 40,85 \text{ crianças}.$$

Apresentamos abaixo a tabela com as freqüências observadas e esperadas:

Distribuição das crianças (freqüências observadas e esperadas, estas últimas aparecendo entre parêntesis), segundo área de localização do domicílio em relação à fundição de chumbo e presença de intoxicação por chumbo.

Área de localização do domicílio	Intoxicação por chumbo		Total
	Sim	Não	
Próxima	90 (78,85)	22 (33,15)	112
Distante	86 (97,15)	52 (40,85)	138
Total	176	74	250

A simples comparação dos números em cada célula da tabela acima, já sugere uma considerável discrepância entre os valores observados e esperados, mas precisamos de uma medida mais objetiva dessa discrepância. Karl Pearson (*On the criterion that a given system of deviations from the probable in the case of a correlated system of variables is such that it can be reasonably supposed to have arisen from random sampling. The London, Edinburgh and Dublin Philosophical Magazine and Journal of Science. 50:157-175, 1900*) propôs que isso fosse feito calculando-se a seguinte estatística:

$$X^2 = \sum_{i=1}^k \left[\frac{(O_i - E_i)^2}{E_i} \right],$$

onde X^2 é lido “qui-quadrado” (mais adiante explicaremos o por quê); O é o número observado de indivíduos em cada célula da tabela analisada, sendo que o subscrito i varia, como indicado no símbolo somatório, $\sum_{i=1}^k$, entre 1 (que indica a primeira célula) e k (que indica a última célula); E representa o número esperado de indivíduos em cada célula, caso “área de residência” e “intoxicação” sejam eventos independentes.

Veja então o que é feito:

Em cada célula comparamos a freqüência observada à freqüência esperada, através de uma operação de diferença (quanto maior a diferença maior a discrepância entre essas freqüências e, portanto, menor a independência entre os eventos investigados, e vice-versa). Como as diferenças podem ser positivas ou negativas, elevamos cada uma ao quadrado, pois o que desejamos obter é o total de discrepância existente entre observado e esperado. Em seguida, dividimos cada diferença pela freqüência esperada em

cada célula, com a finalidade de convertermos essas diferenças às suas unidades originais, já que ao elevarmos ao quadrado sua escala tinha ficado quadrática. Esta é uma das maneiras de retornarmos à escala original, sendo que cada diferença é dividida pela frequência esperada e não pela observada (que era outra possibilidade matematicamente válida para isso), porque a regra é fazermos a divisão pelo valor de referência e, no nosso teste, nossa referência é a frequência esperada, pois esse é o valor esperado caso a hipótese nula seja verdadeira. Finalmente, somamos todos esses termos para encontrarmos o total de discrepância. Note que, se essas variáveis forem totalmente independentes, obteremos frequências observadas exatamente iguais às esperadas, resultando em um qui-quadrado igual a zero, concorda? Então, à medida que essa estatística vai aumentando, afastando-se de zero, maior o grau de dependência, de associação, de influência, entre as variáveis testadas, e maior também a probabilidade dos resultados encontrados serem estatisticamente significantes, assumindo discrepância além daquela esperada por simples variação amostral.

Resumindo:

O X^2 mede o quanto os pares de frequências observadas e esperadas são concordantes.

Pode ser demonstrado que, se a hipótese nula for verdadeira, ou seja, se as variáveis testadas forem independentes, os valores obtidos de $X^2 = \sum_{i=1}^k \left[(O_i - E_i)^2 / E_i \right]$ em diferentes situações encontradas na natureza distribuem-se conforme uma distribuição probabilística chamada também de “qui-quadrado”. Ou seja, os diversos valores assumidos por $X^2 = \sum_{i=1}^k \left[(O_i - E_i)^2 / E_i \right]$ na natureza não seguem uma distribuição normal, que tem sido até agora a distribuição utilizada em nossos testes estatísticos.

A distribuição qui-quadrado, como todas as demais, pode ser representada por uma expressão matemática, que apresentamos abaixo:

$$f(u) = \frac{1}{\left(\frac{\nu}{2} - 1\right)!} \frac{1}{2^{\nu/2}} u^{(\nu/2)-1} e^{-(u/2)}, \quad u > 0,$$

onde o valor variante u substitui o valor variante x da distribuição normal, assumindo valores maiores do que zero; e é o número irracional 2,71828..., que é a base do logaritmo neperiano; e ν o número de graus de liberdade.

Como sempre, não queira olhar para essa fórmula e entendê-la, porque isso é impossível. Você deve confiar em Karl Pearson, estatístico que a desenvolveu e demonstrou, e saber que à medida que formos variando os valores de u e obtendo os valores de $f(u)$ correspondentes, para um determinado número de graus de liberdade ν , e formos colocando os pontos gerados por esses pares de valores de u e $f(u)$ em um

diagrama, obteremos uma distribuição qui-quadrado. Assim, existem distribuições qui-quadrado diferentes, a depender do número de graus de liberdade existente na tabela analisada que, por sua vez, depende do tamanho desta.

O valor variante u é designado por alguns estatísticos pela letra grega *qui* minúscula (χ), sendo por isso, essa distribuição denominada “qui-quadrado”. Observe então, que esses estatísticos utilizam a letra grega maiúscula ao quadrado X^2 para denotar a estatística qui-quadrado, $\sum_{i=1}^k \left[(O_i - E_i)^2 / E_i \right]$, que é uma medida de concordância entre frequências observadas e esperadas, e a letra grega minúscula ao quadrado χ^2 para denotar a distribuição qui-quadrado, que será utilizada como referência para realizarmos esse teste estatístico, em substituição à distribuição normal.

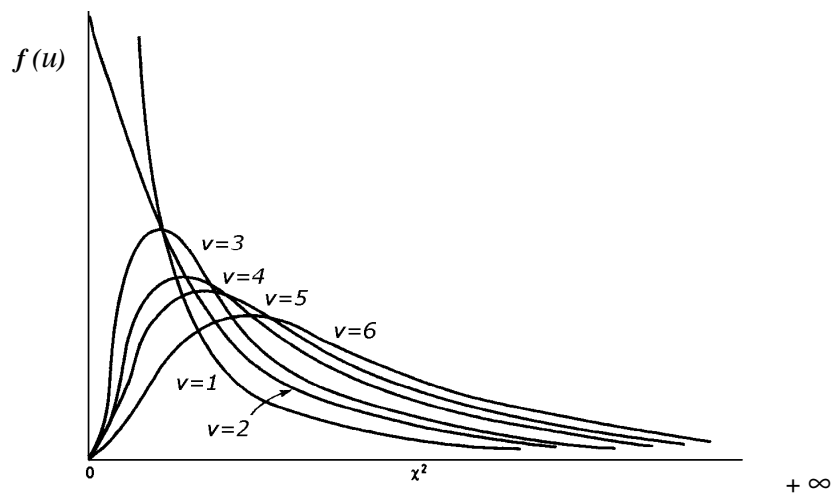
É possível demonstrarmos que quando temos apenas um grau de liberdade, ao elevarmos ao quadrado cada valor da distribuição normal, os valores de Z^2 obtidos ao estudarmos numerosas amostras distribuem-se segundo uma distribuição qui-quadrado. Ou seja,

$$X_{(1)}^2 = z^2,$$

onde o subscrito (1) indica um grau de liberdade. Isto é o que justifica o “quadrado” da denominação “qui-quadrado”. Se quiser ver essa demonstração sugerimos a leitura das páginas 572 e 573 do livro *Daniel WW. Biostatistics: a foundation for analysis in the health sciences. 7ª ed. New York(NY): John Wiley;1999.*

— **Mas, afinal de contas, qual é o formato da distribuição qui-quadrado?**

— Como já mencionamos acima, tal como a distribuição T , e diferentemente da Z , a distribuição qui-quadrado é uma família de distribuições, cada uma com um formato mais ou menos diferente, a depender do número de graus de liberdade da situação estatística sendo testada. Veja abaixo o formato da distribuição para alguns graus de liberdade, escolhidos apenas para ilustração:



Observe que X^2 assume valores entre 0 e $+\infty$, não havendo, portanto, valores negativos de X^2 . Não devemos ficar surpresos com isso, pois, sendo $X_{(1)}^2 = z^2$, os valores negativos de Z , ao serem elevados ao quadrado resultam em valores positivos.

– **Como sabermos quantos graus de liberdade temos?**

– Vimos no capítulo 6 (páginas 62 a 64) e no APÊNDICE 1 que, toda vez que impomos certas condições matemáticas aos possíveis valores de nossas variáveis, diminuimos o número de graus de liberdade, ou seja, a liberdade que esses valores têm de variar. Observe novamente a tabela do nosso exemplo:

Distribuição das crianças (frequências observadas e esperadas, estas últimas aparecendo entre parêntesis), segundo área de localização do domicílio em relação à fundição de chumbo e presença de intoxicação por chumbo.

Área de localização do domicílio	Intoxicação por chumbo		Total
	Sim	Não	
Próxima	90 (78,85)	22 (33,15)	112
Distante	86 (97,15)	52 (40,85)	138
Total	176	74	250

Para calcularmos o X^2 podemos começar calculando a diferença entre 90 e 78,85, elevando-a ao quadrado, dividindo-a por 78,85 e, no final, somando o resultado dessas operações aos resultados das mesmas operações feitas com as frequências observadas e esperadas obtidas para cada uma das demais células da tabela. Acontece que, se escolhermos iniciar o cálculo do X^2 com as frequências da primeira célula da tabela, como sugerimos acima, estaremos fixando as frequências de todas as demais células, porque há restrições matemáticas impostas a essas frequências, que determinam que suas somas marginais sejam 112; 138; 176 e 74, porque esses foram os totais marginais obtidos concretamente no estudo realizado, e não podemos, por nossa livre vontade, alterar esses resultados, sob pena de modificarmos a realidade que estamos investigando.

Veja então que em tabelas de contingência 2×2 , temos apenas um grau de liberdade, isto é, temos a liberdade de escolher apenas uma das frequências constantes da tabela. Feito isso, todas as outras frequências ficam automaticamente fixadas. Se escolhermos começar o cálculo com a frequência de 90 da primeira célula da tabela do nosso exemplo, a frequência da primeira linha, segunda coluna, terá que ser 22 para que o total marginal da primeira linha seja 112. Pela mesma razão, a frequência da segunda linha, primeira coluna, terá que ser 86, para que o total marginal da primeira coluna seja 176, e assim por diante.

Uma maneira prática de encontrar o número de graus de liberdade de tabelas de qualquer tamanho é fazendo o seguinte cálculo:

$$v = (l - 1)(c - 1),$$

onde v , como já vimos, denota o número de graus de liberdade, l o número de categorias da variável expressa nas linhas e c o daquela apresentada nas colunas.

Aplicando essa fórmula ao nosso exemplo, obtemos:

$$v = (l - 1)(c - 1) = (2 - 1)(2 - 1) = (1)(1) = 1,$$

confirmando que só temos um grau de liberdade, em tabelas de contingência 2×2 .

Prosseguindo com o teste, vamos calcular o X^2 :

$$\begin{aligned}
 X^2 &= \sum_{i=1}^k \left[\frac{(O_i - E_i)^2}{E_i} \right] = \frac{(90 - 78,85)^2}{78,85} + \frac{(22 - 33,15)^2}{33,15} + \frac{(86 - 97,15)^2}{97,15} + \frac{(52 - 40,85)^2}{40,85} = \\
 &= \frac{(11,15)^2}{78,85} + \frac{(-11,15)^2}{33,15} + \frac{(-11,15)^2}{97,15} + \frac{(11,15)^2}{40,85} = \\
 &= \frac{124,32}{78,85} + \frac{124,32}{33,15} + \frac{124,32}{97,15} + \frac{124,32}{40,85} = \\
 &= 1,58 + 3,75 + 1,28 + 3,04 = 9,65.
 \end{aligned}$$

Sabemos porém que, quando temos apenas um grau de liberdade, os valores de X^2 não assumem todos os valores possíveis para garantir sua representação por uma distribuição contínua, como apresentamos no diagrama da página 279. Esse desajuste entre os valores reais de X^2 e a distribuição teórica χ^2 para um grau de liberdade, resulta em valores- p subestimados, favorecendo inadequadamente a rejeição da hipótese nula. Para obtermos um melhor ajuste, utilizamos a correção para continuidade proposta por Yates (*Yates F. Contingency tables involving small numbers and the X^2 test. Journal of the Royal Statistical Society, Suplemento 1: 217-235, 1934*), que é feita como mostrado abaixo:

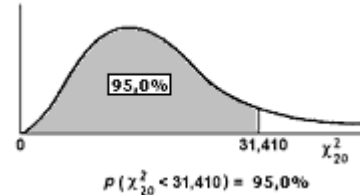
$$\begin{aligned}
 X_c^2 &= \sum_{i=1}^k \left[\frac{(|O_i - E_i| - 0,5)^2}{E_i} \right] = \\
 &= \frac{(|90 - 78,85| - 0,5)^2}{78,85} + \frac{(|22 - 33,15| - 0,5)^2}{33,15} + \frac{(|86 - 97,15| - 0,5)^2}{97,15} + \frac{(|52 - 40,85| - 0,5)^2}{40,85} = \\
 &= \frac{(11,15 - 0,5)^2}{78,85} + \frac{(11,15 - 0,5)^2}{33,15} + \frac{(11,15 - 0,5)^2}{97,15} + \frac{(11,15 - 0,5)^2}{40,85} = \\
 &= \frac{(10,65)^2}{78,85} + \frac{(10,65)^2}{33,15} + \frac{(10,65)^2}{97,15} + \frac{(10,65)^2}{40,85} = \\
 &= \frac{113,42}{78,85} + \frac{113,42}{33,15} + \frac{113,42}{97,15} + \frac{113,42}{40,85} = \\
 &= 1,44 + 3,42 + 1,17 + 2,78 = 8,81.
 \end{aligned}$$

Como podemos ver, o qui-quadrado corrigido, X_c^2 , é menor do que o não-corrigido, X^2 , o que implicará em um valor- p maior, porque o X_c^2 estará situado mais abaixo na curva χ^2 , em uma posição na qual a área sob a curva é maior.

Nosso próximo passo será, então, encontrar na tabela qui-quadrado apresentada a seguir o valor crítico de χ^2 , a partir do qual nosso resultado será considerado estatisticamente significativo. Para isso,

necessitamos definir o nível de significância do nosso teste. Digamos que esse seja 0,05 ou 5,0%. Olhando na tabela vemos que o valor crítico para esse nível de significância é 3,841. Note que este valor foi encontrado na célula da tabela onde ocorre o cruzamento da linha correspondente a um grau de liberdade (denotado por g.l.) com a coluna correspondente a um nível de significância de 0,05 (denotada por χ^2_{95}).

TABELA DE VALORES CRÍTICOS DE χ^2 , SEGUNDO ALGUNS PORCENTIS MAIS UTILIZADOS DA DISTRIBUIÇÃO χ^2 E O NÚMERO DE GRAUS DE LIBERDADE.



g.l.	$\chi^2_{0,5}$	$\chi^2_{2,5}$	χ^2_5	χ^2_{90}	χ^2_{95}	$\chi^2_{97,5}$	χ^2_{99}	$\chi^2_{99,5}$
1	0,0000393	0,000982	0,00393	2,706	3,841	5,024	6,635	7,879
2	0,0100	0,0506	0,103	4,605	5,991	7,378	9,210	10,597
3	0,0717	0,216	0,352	6,251	7,815	9,348	11,345	12,838
4	0,207	0,484	0,711	7,779	9,488	11,143	13,277	14,860
5	0,412	0,831	1,145	9,236	11,070	12,832	15,086	16,750
6	0,676	1,237	1,635	10,645	12,592	14,449	16,812	18,548
7	0,989	1,690	2,167	12,017	14,067	16,013	18,475	20,278
8	1,344	2,180	2,733	13,362	15,507	17,535	20,090	21,955
9	1,735	2,700	3,325	14,684	16,919	19,023	21,666	23,589
10	2,156	3,247	3,940	15,987	18,307	20,483	23,209	25,188
11	2,603	3,816	4,575	17,275	19,675	21,920	24,725	26,757
12	3,074	4,404	5,226	18,549	21,026	23,336	26,217	28,300
13	3,565	5,009	5,892	19,812	22,362	24,736	27,688	29,819
14	4,075	5,629	6,571	21,064	23,685	26,119	29,141	31,319
15	4,601	6,262	7,261	22,307	24,996	27,488	30,578	32,801
16	5,142	6,908	7,962	23,542	26,296	28,845	32,000	34,267
17	5,697	7,564	8,672	24,769	27,587	30,191	33,409	35,718
18	6,265	8,231	9,390	25,989	28,869	31,526	34,805	37,156
19	6,844	8,907	10,117	27,204	30,144	32,852	36,191	38,582
20	7,434	9,591	10,851	28,412	31,410	34,170	37,566	39,997
21	8,034	10,283	11,591	29,615	32,671	35,479	38,932	41,401
22	8,643	10,982	12,338	30,813	33,924	36,781	40,289	42,796
23	9,260	11,688	13,091	32,007	35,172	38,076	41,638	44,181
24	9,886	12,401	13,848	33,196	36,415	39,364	42,980	45,558
25	10,520	13,120	14,611	34,382	37,652	40,646	44,314	46,928
26	11,160	13,844	15,379	35,563	38,885	41,923	45,642	48,290
27	11,808	14,573	16,151	36,741	40,113	43,194	46,963	49,645
28	12,461	15,308	16,928	37,916	41,337	44,461	48,278	50,993
29	13,121	16,047	17,708	39,087	42,557	45,722	49,588	52,336
30	13,787	16,791	18,493	40,256	43,773	46,979	50,892	53,672
35	17,192	20,569	22,465	46,059	49,802	53,203	57,342	60,275
40	20,707	24,433	26,509	51,805	55,758	59,342	63,691	66,766
45	24,311	28,366	30,612	57,505	61,656	65,410	69,957	73,166
50	27,991	32,357	34,764	63,167	67,505	71,420	76,154	79,490
60	35,535	40,482	43,188	74,397	79,082	83,298	88,379	91,952
70	43,275	48,758	51,739	85,527	90,531	95,023	100,425	104,215
80	51,172	57,153	60,391	96,578	101,879	106,629	112,329	116,321
90	59,196	65,647	69,126	107,565	113,145	118,136	124,116	128,299
100	67,328	74,222	77,929	118,498	124,342	129,561	135,807	140,169

— Mas, se o teste é bicaudado não deveríamos olhar na coluna correspondente ao percentil 0,975, já que o nosso alfa se dividiria nas duas caudas da distribuição?

— Boa pergunta! Nossas hipóteses são: as variáveis estudadas são independentes (H_0), e as variáveis estudadas não são independentes (H_A). Note que, indiretamente, estamos testando se a prevalência (proporção) de intoxicação por chumbo em crianças residentes na área próxima à fundição, p_1 , é estatisticamente igual ou diferente daquela em crianças residentes na área distante, p_2 . Sendo assim, nossa

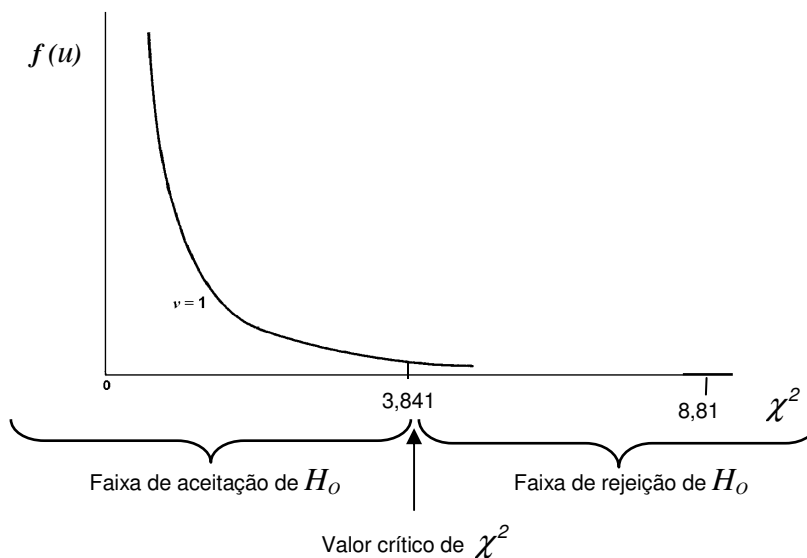
hipótese alternativa que estabelece que as prevalências são diferentes inclui duas possibilidades, como já vimos em outros testes: $p_1 > p_2$ ou $p_1 < p_2$. Acontece que no teste qui-quadrado, se as discrepâncias entre as frequências observadas e esperadas forem grandes, isso resultará em um alto valor da estatística X^2 , não importando se p_1 é maior ou menor do que p_2 . Pelo mesmo raciocínio, valores baixos de X^2 , localizados na parte esquerda da distribuição, representam sempre, nesse teste, evidência em favor da hipótese nula, já que baixos valores de X^2 indicam baixas discrepâncias entre frequências observadas e esperadas.

Por isso, no teste qui-quadrado sempre consideramos o valor integral do nosso alfa na cauda direita da curva.

Queremos chamar sua atenção também para um outro fato: você reparou que o limite crítico encontrado, 3,841, é aproximadamente igual a $1,96^2$? Não tínhamos escrito que $X^2_{(1)} = z^2$?

Continuando o teste, como o valor obtido no nosso exemplo, $X^2_c = 8,81$, é maior do que o limite crítico, $X^2 = 3,841$, concluímos que o resultado é estatisticamente significativo, sendo muito provável que as discrepâncias observadas entre frequências observadas e esperadas não tenham ocorrido por variação amostral. Podemos então rejeitar a hipótese nula e aceitar que, com base nos resultados do nosso estudo, é muito provável que as variáveis “área de localização do domicílio” e “intoxicação por chumbo” não sejam independentes na população de onde as crianças estudadas foram retiradas. Isso sugere que há uma associação estatística entre essas variáveis na população. Para verificarmos qual a direção dessa associação, teremos de nos basear nas prevalências obtidas. Se a prevalência tiver sido maior nas crianças da área próxima, poderemos afirmar que é muito provável que as crianças residentes na área próxima à fundição apresentem uma maior prevalência de intoxicação por chumbo do que as da área distante.

A verificação sobre se essa associação estatística é ou não causal, será feita através de outros procedimentos não abordados neste livro. Veja a situação do teste atual no diagrama a seguir:



Vamos agora realizar esse teste em uma tabela maior, como a que apresentamos abaixo, retirada da mesma pesquisa sobre intoxicação por chumbo:

Distribuição das crianças (frequências observadas e esperadas, estas últimas aparecendo entre parêntesis), segundo raça e presença de intoxicação por chumbo.

Raça	Intoxicação por chumbo		Total
	Sim	Não	
Branco ou mulato claro	35 (34,86)	14 (14,14)	49
Mulato médio	51 (47,66)	16 (19,34)	67
Mulato escuro ou negro	89 (92,48)	41 (37,52)	130
Total	175	71	246

Observe que o total caiu para 246, porque quatro crianças não tiveram a raça classificada pela equipe de pesquisa.

Só para ilustrar, porque você já sabe fazer isso, calcularemos as frequências esperadas nas células da última linha da tabela.

Número esperado de intoxicados entre mulatos escuros ou negros =

$$= \frac{(130)(175)}{246} = \frac{22.750}{246} = 92,479 \cong 92,48 \text{ crianças.}$$

Número esperado de não-intoxicados entre mulatos escuros ou negros =

$$= \frac{(130)(71)}{246} = \frac{9.230}{246} = 37,5203 \cong 37,52 \text{ crianças.}$$

Comparando as frequências observadas às esperadas nas células da tabela, ainda sem o auxílio do teste qui-quadrado, podemos ver que as discrepâncias parecem pequenas, não acha? É provável então que não existam diferenças estatisticamente importantes entre as proporções de intoxicados nas diferentes raças estudadas, e que essas duas variáveis sejam estatisticamente independentes. Mas, para avaliarmos isso de modo mais objetivo, precisaremos realizar o teste qui-quadrado e, para isso, seguiremos as seguintes etapas, já conhecidas por nós:

1ª) Definição do nível de significância: $\alpha = 0,05$;

2ª) Definição das hipóteses:

H_0 : “raça” e “intoxicação por chumbo” são independentes, e

H_A : “raça” e “intoxicação por chumbo” não são independentes;

3ª) Cálculo do valor de X^2 :

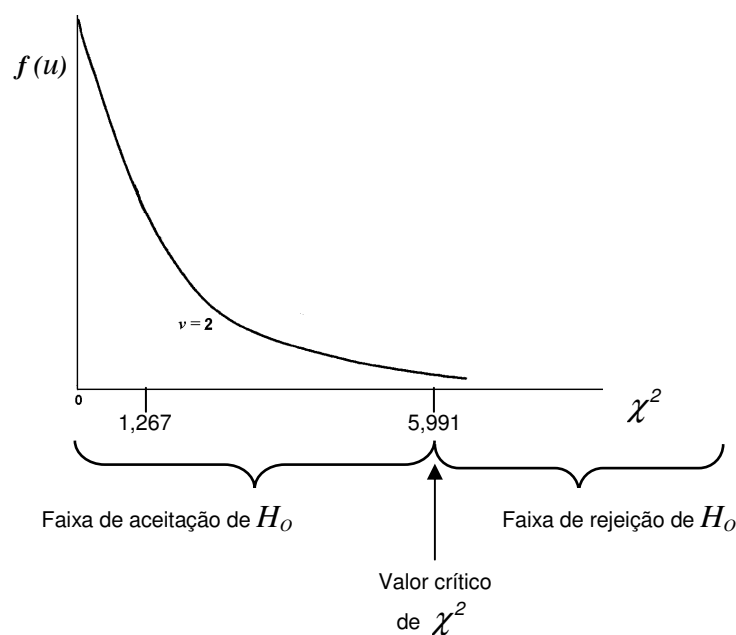
$$X^2 = \sum_{i=1}^k \left[\frac{(O_i - E_i)^2}{E_i} \right] =$$

$$\begin{aligned}
&= \frac{(35-34,86)^2}{34,86} + \frac{(14-14,14)^2}{14,14} + \frac{(51-47,66)^2}{47,66} + \frac{(16-19,34)^2}{19,34} + \frac{(89-92,48)^2}{92,48} + \frac{(41-37,52)^2}{37,52} = \\
&= \frac{(0,14)^2}{34,86} + \frac{(-0,14)^2}{14,14} + \frac{(3,34)^2}{47,66} + \frac{(-3,34)^2}{19,34} + \frac{(-3,48)^2}{92,48} + \frac{(3,48)^2}{37,52} = \\
&= \frac{0,02}{34,86} + \frac{0,02}{14,14} + \frac{11,16}{47,66} + \frac{11,16}{19,34} + \frac{12,11}{92,48} + \frac{12,11}{37,52} = \\
&= 0,0006 + 0,0014 + 0,2342 + 0,5770 + 0,1309 + 0,3228 \cong 1,267;
\end{aligned}$$

4ª) Obtenção do valor crítico do teste na tabela χ^2 , para um alfa de 0,05 e $(l-1)(c-1) = (3-1)(2-1) = (2)(1) = 2$ graus de liberdade. Olhando na tabela vemos que esse valor crítico é 5,991;

5ª) Como $1,267 < 5,991$, verificamos que a discrepância global entre valores observados e esperados nas células da tabela, não é grande o suficiente para ser considerada estatisticamente significativa, sendo muito provável que tenha ocorrido por simples variação amostral. Assim, aceitamos H_0 e rejeitamos H_A , concluindo que raça e intoxicação por chumbo são eventos independentes, ou seja, não estão estatisticamente associados. Indiretamente, podemos concluir que, muito provavelmente, as prevalências de chumbo são iguais nas crianças das diversas raças na população da qual a amostra estudada foi retirada.

A situação desse teste é mostrada graficamente a seguir:



Se o resultado der estatisticamente significativo, perceba que será difícil sabermos quais as proporções

que diferem estatisticamente, entre tantas que podemos calcular em tabelas grandes. No nosso exemplo, as prevalências (proporções) que estão sendo comparadas são:

Distribuição das crianças (frequências observadas e esperadas, estas últimas aparecendo entre parêntesis), segundo raça e presença de intoxicação por chumbo.

Raça	Intoxicação por chumbo		Total
	Sim	Não	
Branco ou mulato claro	35/49 = 71,4%	14/49 = 28,6%	49
Mulato médio	51/67 = 76,1%	16/67 = 23,9%	67
Mulato escuro ou negro	89/130 = 68,5%	41/130 = 31,5%	130

Se o teste qui-quadrado nos mostrasse um resultado estatisticamente significativo, não saberíamos ao certo quais dessas prevalências de intoxicação difeririam, embora pudéssemos afirmar que as maiores diferenças certamente seriam estatisticamente significantes.

O procedimento utilizado nesses casos será fragmentarmos a tabela e fazermos comparações duas a duas (veja isso também em *Glantz SA. Primer of biostatistics. 4ª ed. New York(NY): McGraw-Hill; 1997; páginas 137 a 140*).

Fariamos um teste qui-quadrado para cada uma das tabelas abaixo:

Distribuição das crianças, segundo raça e presença de intoxicação por chumbo.

Raça	Intoxicação por chumbo		Total
	Sim	Não	
Mulato médio	51	16	67
Mulato escuro ou negro	89	41	130
Total	140	57	197

Distribuição das crianças, segundo raça e presença de intoxicação por chumbo.

Raça	Intoxicação por chumbo		Total
	Sim	Não	
Branco ou mulato claro	35	14	49
Mulato médio	51	16	67
Total	86	30	116

Distribuição das crianças, segundo raça e presença de intoxicação por chumbo.

Raça	Intoxicação por chumbo		Total
	Sim	Não	
Branco ou mulato claro	35	14	49
Mulato escuro ou negro	89	41	130
Total	124	55	179

Mas, foi demonstrado que a cada teste estatístico que aplicamos, nosso alfa (que é a probabilidade máxima que nos permitimos de concluir erroneamente que o teste é estatisticamente significativo quando não é) aumenta expressivamente (se desejar, estude isso em *Kleinbaum DG, Kupper LL, Muller KE e Nizam A. Applied regression analysis and other multivariable methods. 3ª ed. Pacific Grove (CA): Duxbury Press; 1998*). Assim, recomenda-se que iniciemos essa série de comparações duas a duas pela tabela na qual a diferença

seja a mais elevada. Em seguida, consideraremos aquela com a segunda maior diferença, e assim por diante. A vantagem desse procedimento é que, assim que obtivermos uma diferença que não seja estatisticamente significativa, podemos interromper a série de testes, evitando que o nosso alfa aumente mais ainda, desnecessariamente.

Outro problema resultante desse procedimento é que, ao fragmentarmos nossa tabela, o número de indivíduos considerados em cada teste será menor do que o tamanho original da nossa amostra, e é conhecido que com números muito pequenos os valores da estatística χ^2 não se ajustam bem à distribuição χ^2 .

— E quando devemos considerar que os números estão pequenos demais para aplicação do teste qui-quadrado?

— Grande parte dos estatísticos segue as seguintes recomendações feitas por Cochran (*Cochran WG. The χ^2 test of goodness of fit. Annals of Mathematical Statistics, 23:315-345, 1952; e Cochran WG. Some methods for strengthening the common χ^2 tests. Biometrics, 10:417-451, 1954*):

Para tabelas com um grau de liberdade, o teste qui-quadrado não deve ser utilizado se:

- $n < 20$;
- $20 < n < 40$ e qualquer das freqüências esperadas for menor do que 5;
- $n \geq 40$ e mais de uma freqüência esperada for igual a 1.

Para tabelas com mais de um grau de liberdade, o teste qui-quadrado não deve ser utilizado se:

- mais de 20% das células tiverem freqüências esperadas menores do que 5.

Para tabelas com mais de um grau de liberdade, podemos utilizar o teste qui-quadrado se não mais do que 20% das células tiverem freqüências esperadas menores do que 5, mesmo que uma dessas tenha freqüência esperada igual a 1.

— E o que faremos quando não pudermos aplicar o teste qui-quadrado?

— As alternativas são diferentes conforme o número de graus de liberdade. Se seus resultados estão dispostos em tabela de contingência 2×2 , substitua o teste qui-quadrado pelo teste exato de Fisher, que será abordado no próximo capítulo. Se estiver em uma situação com mais de um grau de liberdade (tabela maior do que a 2×2), o teste de Fisher não poderá ser utilizado. Nesse caso, podemos tentar aumentar o tamanho da nossa amostra, ou aglutinar categorias das variáveis envolvidas, se isso se apoiar em bases clínicas e epidemiológicas aceitáveis. No nosso exemplo do estudo da associação entre raça e intoxicação por chumbo, poderíamos verificar se seria possível juntarmos os mulatos médios aos mulatos escuros e negros. Assim a comparação passaria a ser entre os brancos e mulatos claros e as demais raças, havendo um aumento dos números nas células dessa nova tabela. Neste caso, é claro, a situação passaria novamente para um grau de liberdade, mas isso não ocorreria com tabelas maiores do que a utilizada no exemplo.

— Existem outras aplicações para o teste qui-quadrado?

— Sim. Observe que, nos exemplos vistos acima, ambas as variáveis estudadas são aleatórias, isto é, o número de crianças na área próxima ou distante, nos diversos grupos raciais, e de intoxicados ou sadios, não foi previamente fixado pelos investigadores. A distribuição das crianças nesses subgrupos resultou do

efeito de vários fatores que caracterizam a própria realidade em que essas crianças viviam, sem interferência dos pesquisadores.

Há pesquisas, entretanto, nas quais os estudiosos fixam previamente o número de crianças a ser investigado em cada grupo. Isso é o que ocorre, por exemplo, em estudos caso-controle, nos quais é muito comum estudarem-se dois grupos de indivíduos de igual tamanho. Nesta situação, a variável que indica a presença ou ausência da doença a ser estudada é denominada de “fixa”, porque suas frequências foram fixadas previamente. O objetivo estatístico dos investigadores será verificar se os dois grupos pré-fixados são homogêneos (semelhantes) ou não, quanto a uma certa exposição de interesse. Esse teste é então denominado de teste de homogeneidade. Como, apesar dessa diferença teórica, esse teste é realizado exatamente do mesmo modo que o teste de independência visto acima, não será necessário abordá-lo aqui.

O teste qui-quadrado é também utilizado para verificarmos se uma determinada distribuição de frequências pode ser assumida como sendo normal, binomial ou de Poisson. Nesse contexto é denominado de teste de qualidade do ajuste, porque visa justamente averiguar se um determinado conjunto de dados se ajusta bem a uma determinada distribuição estatística já conhecida. Tal aplicação do teste qui-quadrado também não será apresentada porque foge aos objetivos deste livro.

Você talvez já tenha ouvido falar em um teste qui-quadrado para tendência. Esse teste deverá ser aplicado quando estivermos analisando dados que possam ser dispostos em uma tabela semelhante à apresentada abaixo:

Distribuição das crianças, segundo raça e presença de intoxicação por chumbo.

Intoxicação por chumbo	Raça					Total
	Branco	Mulato claro	Mulato médio	Mulato escuro	Negro	
Sim	<i>a</i>	<i>b</i>	<i>c</i>	<i>d</i>	<i>e</i>	<i>m₁</i>
Não	<i>f</i>	<i>g</i>	<i>h</i>	<i>i</i>	<i>j</i>	<i>m₂</i>
Total	<i>m₃</i>	<i>m₄</i>	<i>m₅</i>	<i>m₆</i>	<i>m₇</i>	<i>n</i>

Nossa intenção ao analisá-la será verificar se há uma tendência crescente ou decrescente da prevalência de intoxicação, à medida que passamos da raça branca para a negra. Não abordaremos esse teste neste livro. Se estiver interessado, pode aprender a fazê-lo estudando-o no livro *Rosner B. Fundamentals of biostatistics. 5ª ed. Pacific Grove (CA): Duxbury;2000*, páginas 397 a 400.

Existem ainda muitos outros testes estatísticos que não são denominados “qui-quadrado”, mas utilizam a distribuição qui-quadrado, tais como o procedimento de Mantel e Haenszel (para ajustar a verificação de associação entre variáveis nominais por outras variáveis chamadas de confundidoras, pois podem confundir a associação); o teste de McNemar (para verificar associação entre variáveis nominais de grupos não-independentes); o teste da razão de verossimilhanças (para avaliar associações na análise de regressão logística, na análise de regressão de Cox e outras análises). Essas técnicas também fogem aos objetivos deste livro.

Existem ainda outros procedimentos para análise de variáveis nominais (categóricas), que também consistem em organização dos dados em tabelas de contingência. O procedimento a ser utilizado irá depender do tipo de estudo epidemiológico realizado. As estatísticas geradas nesses estudos incluem: razão entre prevalências (*RP*); diferença entre prevalências (abordada no capítulo 14, páginas 252 a 259); risco

relativo (RR) ou razão entre incidências; diferença entre incidências, com vários tipos, como o risco atribuível (RA), o risco atribuível percentual ($RA\%$) ou risco atribuível nos expostos ou fração etiológica; o risco atribuível populacional percentual ($RAP\%$); a razão entre chances (RC); a diferença entre chances (DC); e razão de chances e risco relativo ajustados pelo método de Mantel e Haenszel (RC_{MH} e RR_{MH} , respectivamente). Para cada uma dessas estatísticas podemos calcular o intervalo de confiança, que será interpretado do modo usual aprendido por você neste livro. Você facilmente encontrará boas apresentações desses indicadores em outros livros textos de bioestatística ou de epidemiologia.

Ainda sobre técnicas para análise e inferência para variáveis nominais, existe a estatística Capa ("Kappa" em inglês), também conhecida como índice capa, utilizada na avaliação de concordância entre resultados de um mesmo teste diagnóstico ou classificatório, realizado por observadores diferentes ou em momentos diferentes.

O próximo capítulo é bem curto e abordará, como já antecipamos, o teste exato de Fisher. Até lá!

CAPÍTULO 17

-
- Como realizamos o teste exato de Fisher?
 - Por que este teste é chamado de exato?
-



— Como realizamos o teste exato de Fisher?

— Já vimos que uma das alternativas ao teste qui-quadrado, quando nosso estudo tem pequenos números, é o teste exato de Fisher. Este teste é geralmente aplicado a dados dispostos em tabelas de contingência 2×2 , embora Carr (Carr WE. *Fisher's exact test extended to more than two samples of equal size. Technometrics*, 22:269-270, 1980) apresente uma expansão do mesmo para comparação de mais de dois grupos de igual tamanho.

O teste exato foi proposto em meados dos anos 30 do século passado, quase simultaneamente por Fisher (Fisher RA. *Statistical methods for research workers*. 5ª ed. Edinburg: Oliver and Boyd; 1934; Fisher RA. *The logic of inductive inference. Journal of the Royal Statistical Society Series A*, 98:39-54, 1935), Irwin (Irwin JO. *Tests of significance for differences between percentages based on small numbers. Metron*, 12:83-94, 1935) e Yates (Yates F. *Contingency tables involving small numbers and the χ^2 test. Journal of the royal statistical Society, Suplemento 1*:217-235, 1934). Contudo, tornou-se conhecido como teste exato de Fisher. Mais adiante neste capítulo (página 294), explicaremos porque ele recebe a denominação de exato.

Vamos manter o mesmo exemplo do capítulo anterior, embora retirando várias crianças do estudo, para caracterizar uma situação de números pequenos. Na tabela abaixo apresentamos esses números:

Distribuição das crianças, segundo área de localização do domicílio em relação à fundição de chumbo e presença de intoxicação por chumbo.

Área de localização do domicílio	Intoxicação por chumbo		Total
	Sim	Não	
Próxima	10	2	12
Distante	11	9	20
Total	21	11	32

Nosso estudo agora envolve apenas 32 crianças. Para verificar se as condições para aplicação do teste qui-quadrado estão atendidas, calculamos as frequências esperadas para cada célula, caso os eventos comparados sejam independentes. Veja a tabela a seguir:

Distribuição das crianças (frequências observadas e esperadas, estas últimas aparecendo entre parêntesis), segundo área de localização do domicílio em relação à fundição de chumbo e presença de intoxicação por chumbo.

Área de localização do domicílio	Intoxicação por chumbo		Total
	Sim	Não	
Próxima	10 (7,88)	2 (4,13)	12
Distante	11 (13,13)	9 (6,88)	20
Total	21	11	32

Relembrando as condições para aplicação do teste qui-quadrado, sabemos que para tabelas com um

grau de liberdade o teste qui-quadrado não deve ser utilizado se:

- $n < 20$;
- $20 < n < 40$ e qualquer das freqüências esperadas for menor do que 5;
- $n \geq 40$ e mais de uma das freqüências esperadas forem iguais a 1.

No nosso exemplo atual temos que $20 < n < 40$, porque $20 < 32 < 40$, e uma das freqüências esperadas é menor do que 5. Então, não devemos aplicar o teste qui-quadrado. Como os dados estão dispostos em tabela 2×2 , substituiremos esse teste pelo teste exato de Fisher.

O ponto de partida para a realização deste teste é mantermos fixos os totais marginais observados no estudo (12, 20, 21 e 11). Em seguida, calculamos a probabilidade de obtermos as diversas combinações possíveis para a distribuição das 32 crianças nas células da tabela, considerando fixos os totais marginais observados. Isso é ilustrado na tabela abaixo:

Distribuição das crianças, segundo área de localização do domicílio em relação à fundição de chumbo e presença de intoxicação por chumbo.

Área de localização do domicílio	Intoxicação por chumbo		Total
	Sim	Não	
Próxima	?	?	12
Distante	?	?	20
Total	21	11	32

Manter os totais marginais fixos é uma conveniência matemática, porque se os deixássemos também variar, seria muito difícil calcularmos as probabilidades necessárias para esse teste, mas também uma necessidade estatística, pois o que queremos é, justamente, avaliar as probabilidades das possíveis combinações de distribuição das 32 crianças nas células da tabela, respeitando as proporções observadas na distribuição das crianças segundo as variáveis comparadas, independentemente uma da outra. Observe que estas proporções seriam calculadas com base nos totais marginais. Estamos querendo saber o seguinte: mantendo fixas as proporções observadas na distribuição das crianças sem influência de uma variável sobre a outra, quais as probabilidades das diversas combinações possíveis da distribuição das crianças nas células da tabela, levando em conta agora a possível influência de uma variável sobre a outra?

Desejamos inicialmente, portanto, saber quais e quantas combinações (indicadas por interrogações na tabela acima) são possíveis para a distribuição das 32 crianças nas células da tabela. Obtidas essas combinações, calculamos a probabilidade de encontrarmos cada uma delas, como já mencionamos. No próximo passo, somamos as probabilidades de obtermos combinações tão ou mais extremas do que aquela que concretamente encontramos no nosso estudo.

Diferentemente do teste qui-quadrado, o teste exato de Fisher pode ser feito considerando-se apenas uma ou as duas caudas da distribuição. Se estivermos realizando um teste monocaudado, somaremos apenas as probabilidades de obtermos combinações tão ou mais extremas em uma das caudas da distribuição. Se o teste for bicaudado, somaremos as probabilidades de obtermos combinações tão ou mais extremas em uma das caudas às probabilidades mais extremas na outra cauda. Note que essa soma de probabilidades, tanto no teste mono como no bicaudado, serão os valores- p que utilizaremos para testar se as variáveis dispostas na tabela são independentes ou não. prossiga para entender melhor isso.

— **Por que este teste é chamado de exato?**

— Lembre-se de que temos usado as notações a , b , c e d , para denotar o número de indivíduos (crianças) nas células de uma tabela de contingência 2×2 . Note também que, se mantivermos os totais marginais fixos, ao definirmos o valor de a , automaticamente ficam estabelecidos os valores de b , c e d . Nos capítulos anteriores verificávamos qual a probabilidade de encontrarmos valores maiores ou menores do que certo valor, levando em conta a posição desse valor em uma distribuição teórica já conhecida e com formato pré-definido (a distribuição Z , F , T , ou χ^2), que era utilizada como modelo. Para fazermos isso, lembre-se de que assumíamos o pressuposto de que o valor testado era de uma variável que apresentava aquela distribuição escolhida para o teste, o que não nos garantia que as probabilidades utilizadas fossem totalmente exatas, já que não podíamos garantir uma aderência perfeita da variável àquela distribuição teórica. No teste exato de Fisher a distribuição usada é aquela obtida com o cálculo de probabilidades exatas, calculadas para cada uma das combinações específicas que poderiam ser obtidas no estudo realizado, como veremos mais adiante. Por isso, este teste é chamado de “exato”. Esta distribuição apresenta as probabilidades de obtermos diversos valores de a e, conseqüentemente de b , c e d , em uma tabela de contingência 2×2 . Os estatísticos observaram que os possíveis valores de a , não se distribuem de acordo com as distribuições Z , F , T , ou χ^2 , mas conforme uma distribuição chamada de hipergeométrica.

— **Mas, o que são afinal essas combinações tão ou mais extremas em uma distribuição hipergeométrica?**

— Para você entender isso é necessário, antes de tudo, calcularmos a probabilidade de obtermos a combinação concretamente encontrada em nosso estudo, rerepresentada na tabela abaixo:

Distribuição das crianças, segundo área de localização do domicílio em relação à fundição de chumbo e presença de intoxicação por chumbo.

Área de localização do domicílio	Intoxicação por chumbo		Total
	Sim	Não	
Próxima	10	2	12
Distante	11	9	20
Total	21	11	32

Os elementos dessa tabela podem ser denotados por:

Distribuição dos indivíduos, segundo exposição e doença.

Exposição	Doença		Total
	Sim	Não	
Sim	a	b	$a + b$
Não	c	d	$c + d$
Total	$a + c$	$b + d$	n

A probabilidade de encontrarmos a combinação de valores a , b , c e d na tabela é dada por

$$P(a, b, c, d) = \frac{(a+b)!(c+d)!(a+c)!(b+d)!}{n!a!b!c!d!},$$

onde o sinal de exclamação “!” indica uma operação fatorial, que é um tipo especial de multiplicação. Por

exemplo, $5!$, é lido “cinco fatorial” e significa “cinco vezes quatro vezes três vezes dois vezes um” ou $5 \times 4 \times 3 \times 2 \times 1 = 5 \times 4 \times 3 \times 2$, que pode também ser expresso por $(5)(4)(3)(2)(1) = (5)(4)(3)(2)$.

A expressão para calcularmos $P(a, b, c, d)$ pode ser deduzida da seguinte maneira: considere, por um momento, que estejamos estudando apenas cinco crianças e que escolhamos aleatoriamente três dessas crianças. Suponha também que queiramos saber quantas combinações três a três dessas crianças seriam possíveis, se fizéssemos várias escolhas aleatórias de três crianças, sem reposição das mesmas para o sorteio e sem nos interessarmos em que ordem as três crianças sejam sorteadas. Para simplificar, identificaremos cada criança por um número de 1 a 5, em lugar dos seus nomes. As combinações possíveis dessas cinco crianças três a três, são: crianças 1, 2 e 3; 1, 2 e 4; 1, 2 e 5; 1, 3 e 4; 1, 3 e 5; 1, 4 e 5; 2, 3 e 4; 2, 3 e 5; 2, 4 e 5; e 3, 4 e 5. Veja que existem 10 combinações possíveis. Se quisermos saber o número de combinações em um número muito grande de crianças, seria muito trabalhoso procedermos como fizemos acima. Felizmente, o número de combinações possíveis pode ser calculado por

$$\text{Número de combinações de } n \text{ elementos } x \text{ a } x = \binom{n}{x} = \frac{n!}{x!(n-x)!}.$$

Substituindo na fórmula acima os números que, momentaneamente, estamos utilizando temos

$$\begin{aligned} \text{Número de combinações de 5 crianças 3 a 3} &= \binom{5}{3} = \frac{5!}{3!(5-3)!} = \frac{5!}{3!2!} = \frac{(5)(4)(3)(2)(1)}{(3)(2)(1)(2)(1)} = \\ &= \frac{(5)(4)(\cancel{3})(\cancel{2})}{(\cancel{3})(\cancel{2})(2)} = \frac{(5)(4)}{(2)} = \frac{20}{2} = 10 \text{ combinações.} \end{aligned}$$

Veja que obtivemos o número correto de combinações.

Voltando aos números do nosso exemplo atual, uma das quantidades que desejamos saber é o número de combinações possíveis das 12 crianças da área próxima 10 a 10, porque observamos que 10 destas crianças apresentavam intoxicação por chumbo. Calculamos isso por

$$\text{Número de combinações de 12 crianças de 10 em 10} = \binom{12}{10} = \frac{12!}{10!(12-10)!} = \frac{12!}{10!2!}.$$

Utilizando as notações da tabela obtemos

$$\text{Número de combinações de } (a+b) \text{ crianças de } a \text{ em } a = \binom{a+b}{a} = \frac{(a+b)!}{a!(a+b-a)!} = \frac{(a+b)!}{a!b!}.$$

Outra quantidade que precisamos saber é o número de combinações possíveis das 20 crianças da área distante 11 a 11, porque observamos que 11 destas crianças apresentavam intoxicação por chumbo.

Usando as notações da tabela, calculamos isso por

$$\text{Número de combinações de } (c+d) \text{ crianças de } c \text{ em } c = \binom{c+d}{c} = \frac{(c+d)!}{c!(c+d-c)!} = \frac{(c+d)!}{c!d!}.$$

A probabilidade de obtermos um determinado valor de a (e, conseqüentemente de b , c e d) na tabela, será dada por

$$P(a = \text{determinado valor}) = P(a, b, c, d) = \frac{\binom{a+b}{a} \binom{c+d}{c}}{\binom{n}{a+c}}.$$

Observe atentamente a expressão acima. No seu numerador, temos a multiplicação do número de combinações possíveis de a crianças intoxicadas em $(a+b)$ crianças da área próxima, pelo número de combinações possíveis de c crianças intoxicadas em $(c+d)$ crianças da área distante. Obtemos com isso, o número de combinações da distribuição de crianças intoxicadas de acordo com o seu local de residência, possíveis de serem observadas. No denominador, temos o número de combinações possíveis da distribuição de crianças intoxicadas entre as n crianças estudadas, independentemente do local de residência.

A expressão acima relaciona, portanto, o número de combinações possíveis da distribuição da intoxicação, segundo o local de residência, ao número de combinações possíveis da distribuição da intoxicação, independentemente do local de residência.

Já sabemos que

$$\binom{a+b}{a} = \frac{(a+b)!}{a!b!} \text{ e } \binom{c+d}{c} = \frac{(c+d)!}{c!d!},$$

e que, do mesmo modo

$$\binom{n}{a+c} = \frac{n!}{a+c! [n-(a+c)]!}.$$

Ora! O que é $[n-(a+c)]$? Olhe na tabela com as notações e veja que $[n-(a+c)] = b+d$.

Temos então que

$$\binom{n}{a+c} = \frac{n!}{(a+c)!(b+d)!}.$$

Fazendo substituições na expressão para o cálculo da probabilidade de a ser igual a determinado valor, obtemos

$$P(a = \text{determinado valor}) = P(a, b, c, d) = \frac{\binom{a+b}{a} \binom{c+d}{c}}{\binom{n}{a+c}} = \frac{\left[\frac{(a+b)!}{a!b!} \right] \left[\frac{(c+d)!}{c!d!} \right]}{\frac{n!}{(a+c)!(b+d)!}} =$$

$$= \frac{\frac{(a+b)!(c+d)!}{a!b!c!d!}}{\frac{n!}{(a+c)!(b+d)!}} = \left[\frac{(a+b)!(c+d)!}{a!b!c!d!} \right] \left[\frac{(a+c)!(b+d)!}{n!} \right] = \frac{(a+b)!(c+d)!(a+c)!(b+d)!}{n!a!b!c!d!},$$

que é a fórmula que queríamos deduzir.

Continuando a realização do teste exato de Fisher, utilizamos essa expressão para calcular a probabilidade de encontrarmos a combinação de valores realmente observada no estudo, ou seja, valores tão extremos quanto os obtidos. Apresentamos novamente a tabela do nosso exemplo atual para você acompanhar melhor o cálculo:

Distribuição das crianças, segundo área de localização do domicílio em relação à fundição de chumbo e presença de intoxicação por chumbo.

Exposição	Doença		Total
	Sim	Não	
Sim	$a = 10$	$b = 2$	$a + b = 12$
Não	$c = 11$	$d = 9$	$c + d = 20$
Total	$a + c = 21$	$b + d = 11$	$n = 32$

A probabilidade de obtermos a combinação das frequências 10, 2, 11 e 9 nas células da tabela, mantendo fixos os totais marginais, será calculada por

$$P(10, 2, 11, 9) = \frac{12!20!21!11!}{32!10!2!11!9!} \cong 0,0859.$$

Podemos obter combinações mais extremas do que esta, subtraindo 1 da frequência mais baixa, e recalculando as frequências das outras células da tabela, mantendo os mesmos totais marginais. Fazendo isso, e calculando as probabilidades de obtermos esses resultados, encontramos:

Distribuição possível das crianças, segundo área de localização do domicílio em relação à fundição de chumbo e presença de intoxicação por chumbo.

Área de localização do domicílio	Intoxicação por chumbo		Total
	Sim	Não	
Próxima	11	1	12
Distante	10	10	20
Total	21	11	32

$$P(11, 1, 10, 10) = \frac{12!20!21!11!}{32!11!1!10!10!} \cong 0,0172; \text{ e}$$

Distribuição possível das crianças, segundo área de localização do domicílio em relação à fundição de chumbo e presença de intoxicação por chumbo.

Área de localização do domicílio	Intoxicação por chumbo		Total
	Sim	Não	
Próxima	12	0	12
Distante	9	11	20
Total	21	11	32

$$P(12, 0, 9, 11) = \frac{12!20!21!1!}{32!12!0!9!1!} \cong 0,0013.$$

Não há mais outra combinação extrema nessa cauda da distribuição, porque não podemos baixar a frequência mais baixa, pois esta já chegou ao valor zero, e não existem frequências negativas, não é?

Então, o teste exato de Fisher monocaudado neste exemplo, nos fornece um valor- p igual a $0,0859 + 0,0172 + 0,0013 = 0,1044$. Esta é a probabilidade de obtermos combinações tão ou mais extremas em uma das direções do que a combinação que foi realmente observada em nosso estudo. Nossas hipóteses são: $H_0 : p_1 \leq p_2$ e $H_A : p_1 > p_2$, porque nas combinações acima obtemos valores de p_1 maiores do que os de p_2 , sendo p_1 e p_2 as prevalências de intoxicação nas áreas próxima e distante, respectivamente. Escolhendo um alfa de 0,05, concluiremos que o resultado não é estatisticamente significativo, porque 0,1044 é maior do que 0,05. Muito provavelmente, não podemos rejeitar a hipótese de que na população-alvo a prevalência de intoxicação por chumbo em crianças da área próxima seja menor ou igual àquela em crianças da área distante.

Se quisermos testar as hipóteses $H_0 : p_1 \geq p_2$ e $H_A : p_1 < p_2$, vamos ter de calcular a probabilidade de obtermos combinações extremas na outra cauda da distribuição. Para isso, enumeramos todas as tabelas com as combinações restantes possíveis, calculamos a probabilidade de cada uma, e somamos as probabilidades das combinações cuja probabilidade de ocorrerem forem menores ou iguais àquela da combinação observada no estudo. Encontramos essas combinações restantes, aumentando progressivamente em uma unidade a frequência mais baixa da tabela observada, até que uma das células seja zero. Veja essas combinações e suas probabilidades de ocorrência nas tabelas a seguir:

Área de localização	Intoxicação por chumbo		Total
	Sim	Não	
Próxima	9	3	12
Distante	12	8	20
Total	21	11	32

$$P(9, 3, 12, 8) = \frac{12!20!21!1!}{32!9!3!12!8!} \cong 0,2148$$

Área de localização	Intoxicação por chumbo		Total
	Sim	Não	
Próxima	8	4	12
Distante	13	7	20
Total	21	11	32

$$P(8, 4, 13, 7) = \frac{12!20!21!1!}{32!8!4!13!7!} \cong 0,2974$$

Área de localização	Intoxicação por chumbo		Total
	Sim	Não	
Próxima	7	5	12
Distante	14	6	20
Total	21	11	32

$$P(7, 5, 14, 6) = \frac{12!20!21!1!1!}{32!7!5!14!6!} \cong 0,2379$$

Área de localização	Intoxicação por chumbo		Total
	Sim	Não	
Próxima	6	6	12
Distante	15	5	20
Total	21	11	32

$$P(6, 6, 15, 5) = \frac{12!20!21!1!1!}{32!6!6!15!5!} \cong 0,1110$$

Área de localização	Intoxicação por chumbo		Total
	Sim	Não	
Próxima	5	7	12
Distante	16	4	20
Total	21	11	32

$$P(5, 7, 16, 4) = \frac{12!20!21!1!1!}{32!5!7!16!4!} \cong 0,0297$$

Área de localização	Intoxicação por chumbo		Total
	Sim	Não	
Próxima	4	8	12
Distante	17	3	20
Total	21	11	32

$$P(4, 8, 17, 3) = \frac{12!20!21!1!1!}{32!4!8!17!3!} \cong 0,0044$$

Área de localização	Intoxicação por chumbo		Total
	Sim	Não	
Próxima	3	9	12
Distante	18	2	20
Total	21	11	32

$$P(3, 9, 18, 2) = \frac{12!20!21!1!1!}{32!3!9!18!2!} \cong 0,0003$$

Área de localização	Intoxicação por chumbo		Total
	Sim	Não	
Próxima	2	10	12
Distante	19	1	20
Total	21	11	32

$$P(2, 10, 19, 1) = \frac{12!20!21!1!1!}{32!2!10!19!1!} \cong 0,00001$$

Área de localização	Intoxicação por chumbo		Total
	Sim	Não	
Próxima	1	11	12
Distante	20	0	20
Total	21	11	32

$$P(1, 11, 20, 0) = \frac{12!20!21!1!1!}{32!1!11!20!0!} \cong 0,00000009.$$

Como a probabilidade da tabela observada foi 0,0859, vemos que as últimas cinco tabelas acima apresentam combinações com probabilidades menores do que esta, sendo, portanto, mais extremas. Essas cinco tabelas se constituem na outra cauda da distribuição. Para encontrarmos a probabilidade de obtermos valores tão ou mais extremos do que os valores observados (valor- p para o teste das hipóteses atuais), somamos essas cinco probabilidades à probabilidade de obtermos os valores observados no estudo, obtendo: $p = 0,0297 + 0,0044 + 0,0003 + 0,00001 + 0,00000009 + 0,0859 \cong 0,1203$. Como 0,1203 é maior do que o nosso alfa, concluímos que, muito provavelmente, a prevalência de intoxicação por chumbo na área próxima é maior ou igual àquela na área distante, na população da qual a amostra foi retirada, pois o resultado obtido nesta amostra é muito compatível com isto.

Observe que os valores- p , 0,1044 e 0,1203, foram diferentes nas duas caudas. Como a distribuição hipergeométrica quase sempre não é simétrica, não podemos simplesmente multiplicar um desses valores por dois, para realizarmos um teste bicaudado, como vínhamos fazendo até o momento. Só poderemos fazer isso, quando os totais marginais das linhas ou das colunas da tabela observada forem iguais.

— O que faremos então, quando o teste for bicaudado?

— Quando o teste for bicaudado, somaremos as probabilidades de obtermos combinações tão ou mais extremas em ambas as caudas da distribuição, mas diminuiremos desta soma a probabilidade de obtermos a combinação observada, pois em cada teste monocaudado esta probabilidade foi considerada, sendo que agora só deveremos considerá-la uma única vez. Ao fazermos isso obteremos: $p = 0,1044 + 0,1203 - 0,0859 = 0,1388$. Nesse caso, como nossas hipóteses são $H_0 : p_1 = p_2$ e $H_A : p_1 \neq p_2$, concluiremos que, muito provavelmente, as prevalências comparadas são iguais na população-alvo, porque 0,1388 é maior do que 0,05, indicando que podemos aceitar a hipótese nula.

Note que o teste exato de Fisher tira vantagem justamente do que traz limitações para o teste qui-quadrado, ou seja, de os números da tabela serem pequenos, porque é isso que nos permite examinar todas as possíveis combinações, já que o número destas não é tão grande.

Terminamos nossa jornada aqui. Esperamos que tenham gostado do livro. Até outra oportunidade!



APÊNDICE 1

— O que são graus de liberdade?

— Apresentamos aqui explicações mais detalhadas do conceito de **graus de liberdade** (geralmente denotado por *g.l.*), para aqueles que quiserem se aprofundar mais nesse assunto.

Este apêndice baseia-se no excelente artigo: *Walker HM. Degrees of freedom. The Journal of Educational Psychology 1940; 31: 253- 269.*

O conceito de graus de liberdade já era familiar a Carl F. Gauss (1777-1855) e seus colegas astrônomos, por volta de 1826. Nos trabalhos estatísticos modernos, esse conceito passou a ser encontrado a partir do artigo de William S. Gosset (1876-1937) (que utilizou o pseudônimo de “student”), em 1908 (*Gosset WS. The probable error of a mean. Biometrika 1908; 6:1-25*). O conceito, contudo, só foi primeiramente explicitado por Ronald A. Fisher (1890-1962), em seu artigo de 1915 sobre a distribuição do coeficiente de correlação, só recebendo reconhecimento geral cerca de uma década depois dessa explicitação (*Fisher RA. Frequency distribution of the values of the correlation coefficient in samples from an indefinitely large population. Biometrika 1915; x: 507-521*).

Para você entender melhor o conceito de graus de liberdade, é necessário utilizarmos tanto uma abordagem geométrica como uma abordagem algébrica.

Inicialmente, tente compreender que um trem pode mover-se apenas indo ou voltando de um ponto a outro, aos quais a linha férrea o conduz. Consideramos então que um trem possui um grau de liberdade, porque pode mover-se apenas em uma dimensão do espaço, um caminho unidimensional. A localização de um trem pode ser facilmente determinada sabendo-se a que distância o mesmo se encontra de algum ponto de partida ou de chegada.

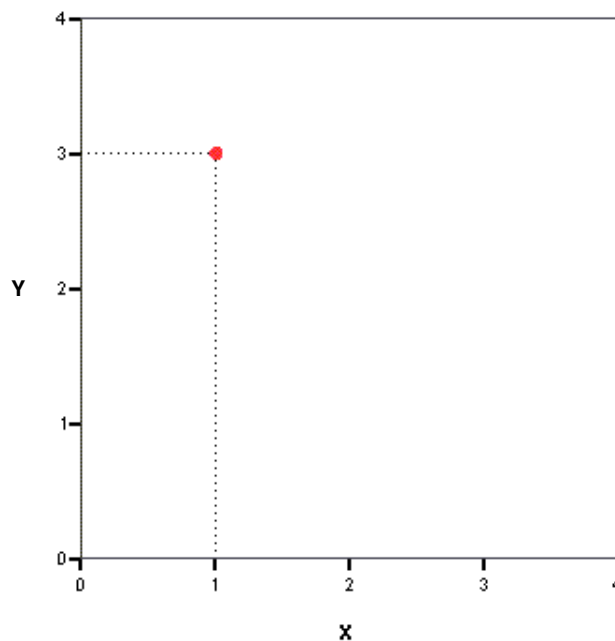
Um automóvel pode mover-se em duas dimensões. Pode ir para adiante ou para trás, e também para a direita ou esquerda. Esse veículo tem, portanto, dois graus de liberdade porque pode movimentar-se em um espaço bidimensional. Para determinarmos a localização de um automóvel precisamos de duas informações: uma que indique sua posição na dimensão para adiante ou para trás e outra que nos informe sobre sua localização na dimensão para a esquerda ou direita.

Um helicóptero (ou um mosquito) tem três graus de liberdade, já que pode movimentar-se em três dimensões diferentes: para adiante ou para trás, para a direita ou esquerda, e para cima ou para baixo. Dizemos então que um helicóptero tem três graus de liberdade, pois pode mover-se em um espaço tridimensional. Sua localização exigirá o conhecimento de três informações diferentes.

Considere que precisamos selecionar aleatoriamente um par de números (x , y). Ao fazer essa seleção teremos dois graus de liberdade, porque estaremos totalmente livres para escolher cada um dos números que compõem o par. O par selecionado poderá ser (0,1), (0,2), (0,3), etc.; ou (1,2), (1,3), (1,4), etc.; ou qualquer combinação possível de dois números. Concorde? Devido também a essa liberdade, consideramos que x e y são independentes, pois, os valores de x podem variar independentemente dos de y , e vice-versa.

Suponha que tenhamos obtido aleatoriamente o par $(1, 3)$. Os números que formam esse par podem ser considerados como coordenadas de um ponto localizado em um espaço bidimensional, ou seja, em um plano que denominaremos por plano xy . Como pode ser visto no gráfico abaixo, esse ponto $(1, 3)$ é livre para mover-se em duas dimensões, como já vimos quando abordamos a situação de um automóvel. Uma dessas dimensões é paralela à ordenada (eixo vertical do gráfico) e a outra paralela à abscissa (eixo horizontal). Por isso, consideramos que o ponto tem dois graus de liberdade.

Representação gráfica do par de números $(1,3)$.

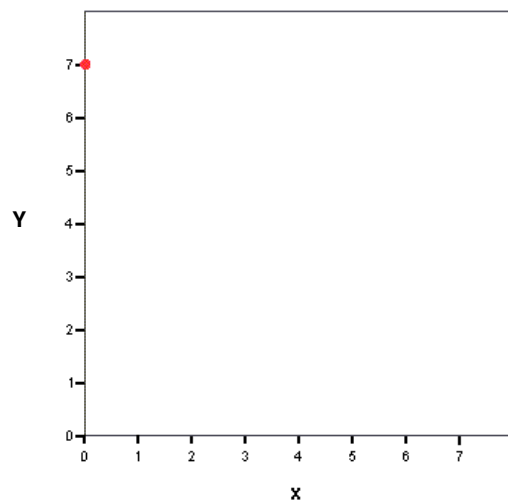


Suponha agora que tenhamos de escolher um par de números cuja soma seja igual a uma constante qualquer c , isto é, $x + y = c$. Considerando, por exemplo, $c = 7$, a condição seria $x + y = 7$, e ao selecionarmos o primeiro número, digamos, p. ex., que tenha sido o número 4, automaticamente o segundo número teria sido fixado como sendo o número 3 para que a soma fosse igual a sete. Haveria dois números sendo escolhidos, mas apenas um teria a liberdade de variar. Ou x ou y (aquele número que fosse o último a ser selecionado) não seria independente (não teria a liberdade de variar) porque iria depender do valor do número escolhido primeiro. Os pontos gerados por esses pares de números cuja soma seria sete, não poderiam mais se mover para qualquer lugar no plano xy . Concorda? Se você ainda não entendeu isso, observe que se $x + y$ tem que ser igual a sete, ao considerarmos diversos valores de x , obteremos os valores de y necessários para que essa condição seja atendida, como exposto em seguida.

Lembre-se de que se $x + y = 7$, então $y = 7 - x$.

Então, para $x = 0$, $y = 7 - x = 7 - 0 = 7$. Esse par $(x = 0, y = 7)$ está representado graficamente no diagrama a seguir:

Representação gráfica do par de números $(0,7)$.



Repetindo o mesmo procedimento teríamos:

para $x = 1, y = 6$; para $x = 2, y = 5$;

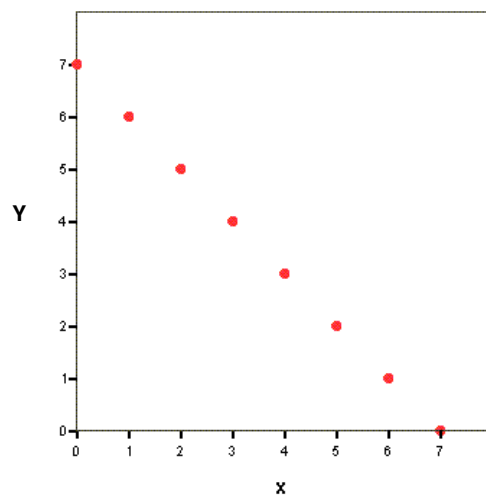
para $x = 3, y = 4$; para $x = 4, y = 3$;

para $x = 5, y = 2$; para $x = 6, y = 1$;

para $x = 7, y = 0$;

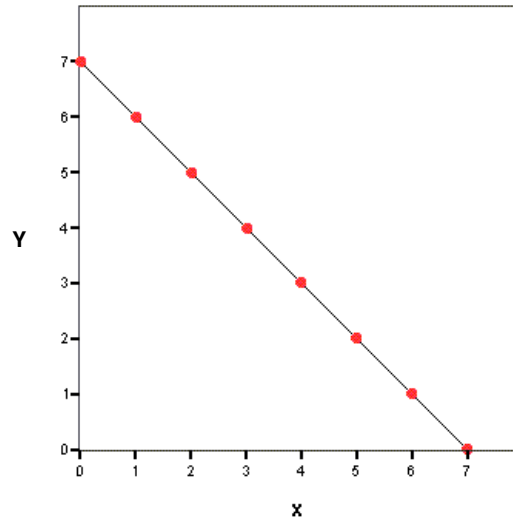
Ao allocarmos também esses pares $(1, 6)$; $(2, 5)$; $(3, 4)$; $(4, 3)$; $(5, 2)$; $(6, 1)$ e $(7, 0)$ no diagrama teríamos:

Representação gráfica dos pares de números $(0, 7)$, $(1, 6)$, $(2, 5)$, $(3, 4)$, $(4, 3)$, $(5, 2)$, $(6, 1)$ e $(7, 0)$.



Note que se considerássemos todos os valores possíveis de x (inclusive os valores fracionários e negativos) obteríamos uma linha reta. Podemos também obter essa linha unindo os pontos gerados pelos pares já considerados acima. Tal linha reta foi traçada no diagrama abaixo:

Linha obtida alocando-se todos os pares de números (x, y) que atendam a condição $x + y = 7$.



O que acabamos de demonstrar graficamente foi que, ao estabelecermos uma condição para os valores de x e de y , os pontos possíveis gerados por esses pares de números só poderiam estar localizados na linha desenhada acima e em nenhum outro lugar do espaço bidimensional. Ou seja, os pontos que antes tinham liberdade para mover-se em duas dimensões, agora só poderiam ir para adiante ou para trás naquela mesma linha, em uma única dimensão. Isso que dizer que esses pontos, que antes possuíam dois graus de liberdade, com a condição que lhes seria imposta ($x + y = 7$) perderiam um grau de liberdade, ficando portanto com apenas um grau de liberdade, representado por sua movimentação sobre a linha reta mostrada acima.

Foi mostrado também que a equação $y = 7 - x$ expressa uma linha reta específica desenhada no diagrama acima. A expressão genérica de uma linha reta é $y = c + bx$, onde c é uma constante qualquer. A equação $y = 7 - x$ contém os mesmos elementos de $y = c + bx$, só que na primeira equação $c = 7$ e $b = -1$. Note que substituindo os valores $c = 7$ e $b = -1$ em $y = c + bx$, obtemos: $y = 7 + (-1)x = 7 - 1x = 7 - x$, ou seja, $y = 7 - x$, como queríamos mostrar.

Se denotarmos por n o número de observações a serem feitas, o que no exemplo atual corresponde à quantidade de números a serem selecionados, teremos que nosso n é igual a dois. E se denotarmos por r o número de condições a serem atendidas por esses números, que no nosso exemplo atual é apenas uma, ($x + y = 7$), teremos que o número de graus de liberdade será dado por $n - r$, ou seja, o número de

observações menos o número de condições impostas a essas observações. No nosso exemplo atual temos:

$$n - r = 2 - 1 = 1 \text{ grau de liberdade,}$$

como já tínhamos demonstrado graficamente.

Suponha agora que os pares (x, y) tenham que atender à condição de que a soma dos seus quadrados seja 25. Para cada valor de x teríamos um valor de y que satisfaria a condição $x^2 + y^2 = 25$.

$$\text{Se } x^2 + y^2 = 25, \quad y^2 = 25 - x^2 \text{ e } y = \pm\sqrt{25 - x^2}.$$

Assim, para $x = 0$,

$$y = \pm\sqrt{25 - x^2} = \pm\sqrt{25 - 0^2} = \pm\sqrt{25 - 0} = \pm\sqrt{25} = \pm 5.$$

Utilizando o mesmo procedimento acima, teríamos:

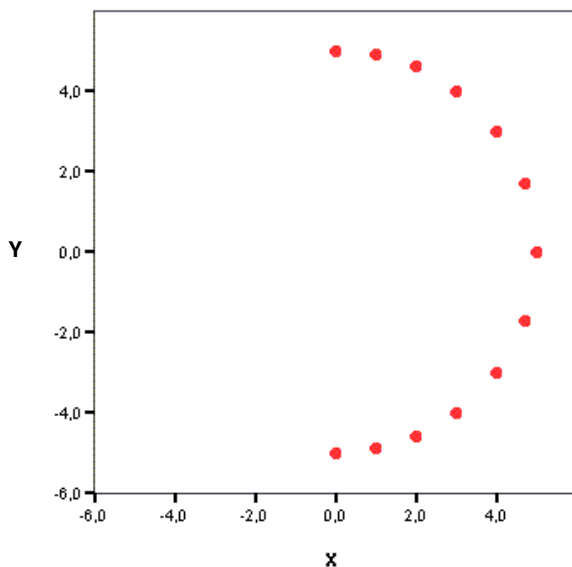
para $x = 1, y \cong \pm 4,9$; para $x = 2, y \cong \pm 4,6$;

para $x = 3, y = \pm 4$; para $x = 4, y = \pm 3$;

para $x = 4,7, y \cong \pm 1,7$; para $x = 5, y = 0$;

Alocando esses pontos em um diagrama obteríamos:

Representação gráfica dos pares de números $(0; 5), (0; -5), (1; 4,9), (1; -4,9), (2; 4,6), (2; -4,6), (3; 4), (3; -4), (4; 3), (4; -3)$ e $(5; 0)$.



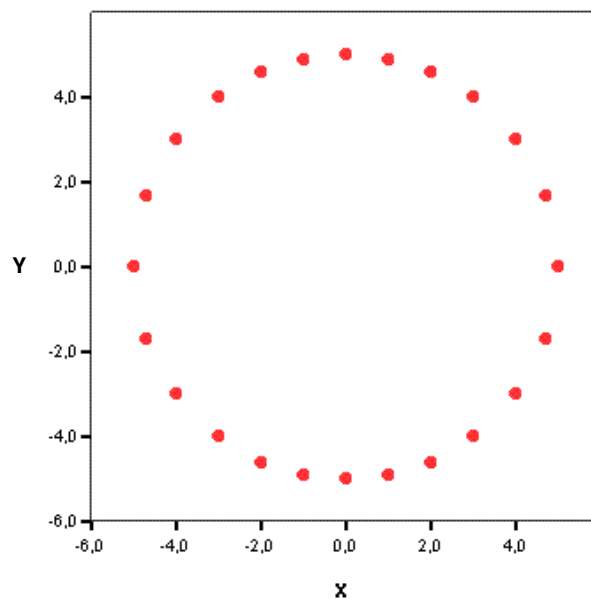
Veja que estaria sendo formada uma semicircunferência no plano xy . Se assumíssemos valores negativos de x , obteríamos uma circunferência completa com raio igual a 5.

Assim,

para $x = -1, y \cong \pm 4,9$; para $x = -2, y \cong \pm 4,6$;
 para $x = -3, y = \pm 4$; para $x = -4, y = \pm 3$;
 para $x = -4,7, y \cong \pm 1,7$; para $x = -5, y = 0$;

Alocando esses novos pontos juntamente com os já desenhados teríamos:

Representação gráfica de todos os pares de números calculados até este momento.



Só assumimos até agora os valores fracionários $x = 4,7$ e $x = -4,7$. Se considerássemos todos os valores fracionários de x , positivos ou negativos, obteríamos uma circunferência completa de raio igual a $\sqrt{25} = 5$. Essa circunferência pode ser visualizada unindo-se os pontos obtidos no diagrama acima.

— **Por que não consideramos os valores de x acima de 5 ou abaixo de -5?**

— Porque esses valores nos dariam resultados inválidos. Por exemplo, para $x = 6$,

$$y = \pm\sqrt{25 - x^2} = \pm\sqrt{25 - 6^2} = \pm\sqrt{25 - 36} = \pm\sqrt{-11} = \text{valor inexistente}.$$

Acabamos de demonstrar graficamente que $x^2 + y^2 = 25$ é a equação de uma circunferência. Já vimos anteriormente que $y = 7 - x$ é a equação de uma linha reta. Lembra-se? Acabamos de ver também que a condição $x^2 + y^2 = 25$ obriga os pontos gerados pelos pares de números que a atendam, a caírem em uma circunferência, que é um espaço unidimensional assentado sobre o plano bidimensional original. Os pontos podem mover-se apenas para adiante ou para trás ao longo dessa circunferência, tendo, portanto, apenas um grau de liberdade.

Há dois números a serem escolhidos, $n = 2$, que estão submetidos a apenas uma condição (relação) limitante, $r = 1$, sendo o número de graus de liberdade dado por

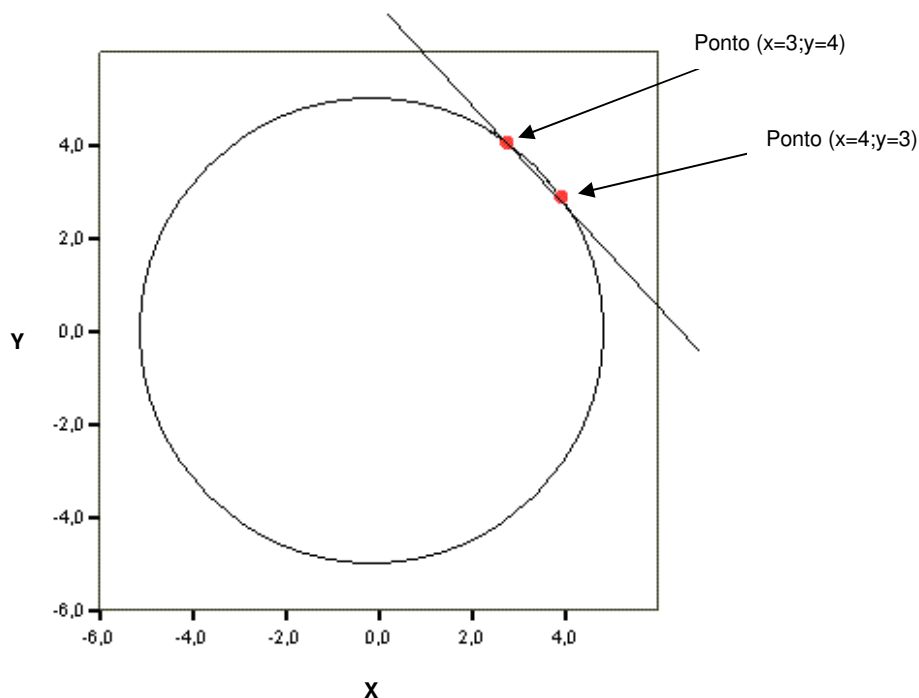
$$n - r = 2 - 1 = 1,$$

como já havíamos demonstrado graficamente.

— E se os números tivessem que atender a duas condições simultaneamente?

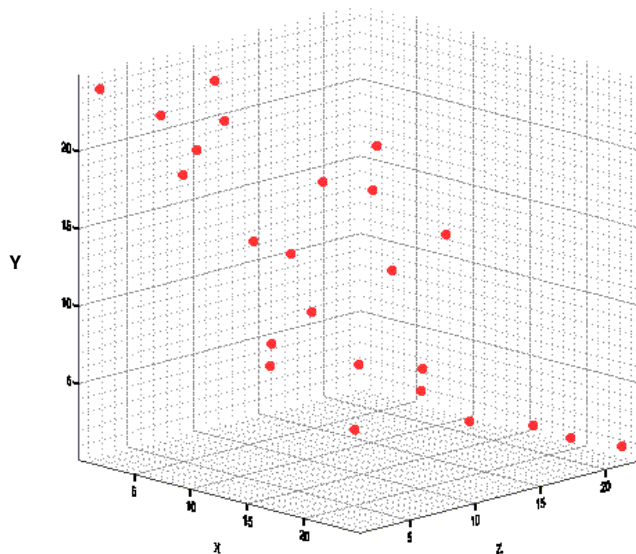
— Se os números, por exemplo, tivessem que atender simultaneamente às duas condições referidas acima, $x + y = 7$ e $x^2 + y^2 = 25$, obteríamos algebricamente apenas duas soluções possíveis: $x = 3$ e $y = 4$ ou $x = 4$ e $y = 3$ (no APÊNDICE 2 apresentamos a resolução algébrica dessas duas equações). Nesse caso, os pontos ficariam limitados pela condição $x + y = 7$ a moverem-se ao longo de uma linha reta, pela condição $x^2 + y^2 = 25$ a moverem-se ao longo de uma circunferência e, pelas duas condições juntas, estariam confinados à interseção dessa linha com essa circunferência. Não haveria, portanto, liberdade de movimento para os pontos. Nesta situação, temos que $n = 2$ e $r = 2$, sendo o número de graus de liberdade $n - r = 2 - 2 = 0$. Veja isso no diagrama abaixo:

Representação gráfica da interseção entre a linha e o círculo do exemplo.



Se tivéssemos de escolher três números, x , y e z , em vez de dois, e nenhuma condição fosse imposta a esses números, cada ponto agora gerado em um espaço tridimensional, tendo esses números como coordenadas, poderia mover-se em três dimensões, tendo portanto três graus de liberdade. Como $n = 3$ e $r = 0$, o número de graus de liberdade seria $n - r = 3 - 0 = 3$. O diagrama a seguir mostra a distribuição de alguns pontos, gerados, cada um, por três números quaisquer escolhidos aleatoriamente, que foram utilizados como coordenadas:

Representação gráfica de combinações livres de três números.



Se impuséssemos a esses números a condição $x + y + z = c$, onde c fosse uma constante qualquer, apenas dois, dos três números, poderiam ser livremente escolhidos, ou seja, apenas dois números seriam observações (variáveis) independentes. Nesse caso, como $n = 3$ e $r = 1$, o número de graus de liberdade seria igual a $n - r = 3 - 1 = 2$. Considerando, por exemplo, $c = 10$, verifique que se os dois primeiros números escolhidos fossem 2 e 3, obrigatoriamente o terceiro teria que ser 5 para que a condição $x + y + z = 10$ fosse atendida ($2 + 3 + 5 = 10$). Concorda? Os pontos, que inicialmente poderiam mover-se em três dimensões, portanto, em um espaço tridimensional, com essa condição que lhes seria imposta só poderiam movimentar-se em duas dimensões, isto é, em um espaço bidimensional. A equação $x + y + z = c$ é a equação de um plano, que é um espaço bidimensional. Se você não identificou em $x + y + z = c$ a equação de um plano, lembre-se de que usando-se álgebra se

$$x + y + z = c, \text{ então}$$

$$y = c - x - z.$$

Lembre-se também de que a equação de uma linha reta é $y = c + bx$, e que $x + y = 7$, que

algebricamente é equivalente a $y = 7 - x$, expressa a equação de uma linha reta para a qual $c = 7$ e $b = -1$. Para expressarmos mais um número, z , temos de colocar mais um termo na equação de uma linha reta:

$$y = c + b_1x + b_2z.$$

↑
termo acrescentado para representar a dimensão z

Assim, vemos que $y = c + b_1x + b_2z$ é a equação de um plano e que, ao estabelecermos a condição $x + y + z = 10$, que é algebricamente igual a $y = 10 - x - z$, limitamos a movimentação dos pontos a um plano, pois, $y = 10 - x - z$ é a expressão de um plano para o qual $c = 10$, $b_1 = -1$ e $b_2 = -1$. Observe que substituindo esses valores em $y = c + b_1x + b_2z$, obtemos: $y = 10 + (-1)x + (-1)z$, donde $y = 10 - x - z$.

Agora suponha que as coordenadas dos pontos (x, y, z) precisassem atender à condição $x^2 + y^2 + z^2 = 36$. Nesse caso, os pontos que atendessem a essa condição seriam obrigados a cair na superfície de uma esfera cujo centro está na origem, ou seja, em zero, e cujo raio seria igual a $\sqrt{36} = 6$. A superfície de uma esfera é um espaço bidimensional. Assim, os pontos só poderiam mover-se em duas dimensões. O número de graus de liberdade seria calculado por $n - r = 3 - 1 = 2$.

Se essas duas últimas condições, $x + y + z = 10$ e $x^2 + y^2 + z^2 = 36$, fossem impostas simultaneamente aos números a serem escolhidos por nós, os pontos só poderiam cair na interseção do plano com a superfície da esfera mencionados acima. A figura resultante dessa interseção seria uma circunferência. Assim os pontos só poderiam mover-se ao longo dessa circunferência, que seria um espaço unidimensional. O número de graus de liberdade seria calculado por $n - r = 3 - 2 = 1$.

Discutimos até aqui os graus de liberdade existentes quando precisamos escolher dois ou três números quaisquer. O raciocínio desenvolvido para essas duas situações pode ser generalizado para quando a quantidade de números a serem escolhidos for maior do que três. Contudo, quando $n > 3$, é impossível visualizarmos a situação graficamente e a generalização é necessariamente matemática. Qualquer conjunto de n números determina um único ponto em um espaço n -dimensional, cada número sendo uma das n coordenadas desse ponto. Se nenhuma condição for imposta aos números, cada um estaria livre para variar independentemente dos demais, e o número de graus de liberdade seria n . Cada condição que seja imposta aos números reduz em uma unidade o número de graus de liberdade. Se r dessas condições forem impostas simultaneamente, os pontos ficarão confinados à interseção das várias locações possíveis, que será um espaço com $n - r$ dimensões dentro de um espaço n -dimensional original.

— Será que vocês poderiam agora aplicar o exposto acima na prática científica em saúde?

— Certo! Vamos parar de falar em trens, automóveis, helicópteros, mosquitos e de números

abstratos, porque esses não são objetos de estudo na pesquisa em saúde, embora alguns agravos à saúde ou óbitos de seres humanos (estes são os nossos objetos) possam resultar do uso ou contato impróprio com algum desses veículos.

Se n números quaisquer podem ser representados por um único ponto em um espaço de n dimensões, os valores de uma variável qualquer, obtidos em uma amostra aleatória de n indivíduos podem também ser assim representados.

Suponha que você tenha estudado a variável altura em uma amostra aleatória de $n = 5$ indivíduos e que os resultados obtidos tenham sido os seguintes:

Número da criança na pesquisa	Valores de altura (em metros, dois decimais)
1	1,14
2	0,86
3	1,24
4	1,17
5	0,94

Esses valores de altura poderiam ser considerados como coordenadas de um ponto que representaria graficamente o conjunto desses valores de altura em um espaço de cinco dimensões. Se nenhuma condição for imposta às suas coordenadas (que são os cinco valores de altura, um para cada um dos cinco indivíduos estudados), esse ponto teria $n = 5$ graus de liberdade. Qualquer valor de altura seria válido para os indivíduos a serem sorteados, tendo a altura total liberdade para variar seus valores, dentro do espectro de valores possíveis de altura para seres humanos.

Considere agora os valores dos desvios entre cada valor de altura e a média das cinco alturas. É evidente que para obtermos esses desvios temos de calcular antes a média. Acontece que ao calcularmos a média nós impomos a seguinte condição aos valores de altura: a soma dos valores de altura dividida pelo número de indivíduos estudados teria que ser igual a um determinado valor. No nosso exemplo temos que

$$\bar{x}_{altura} = \frac{\sum_{i=1}^n x_i}{n} = \frac{1,14 + 0,86 + 1,24 + 1,17 + 0,94}{5} = \frac{5,35}{5} = 1,07 \text{ metros},$$

ou seja, a soma desses cinco valores de altura dividida por cinco tem que ser 1,07 m.

— **Mas, como tem que ser? Por que não poderia ser outro valor?**

— Porque não podemos alterar o fato de que a média dessas cinco alturas é 1,07 m. Se arbitrariamente decidíssemos mudar essa média para 2,13 m ou outro valor qualquer, estaríamos alterando o verdadeiro resultado obtido na amostra estudada, o que nos levaria a conclusões completamente incorretas e cientificamente inválidas para a população ou amostra que estivéssemos estudando. Ao calcularmos a média das alturas, estabelecemos uma condição limitante para os valores de altura: a soma desses valores dividida por cinco tem que ser 1,07 m e nenhum outro valor, porque se fizéssemos isso estaríamos alterando um dos resultados do nosso estudo.

— **E por que a imposição dessa condição nos faz perder um grau de liberdade?**

— Por tudo que apresentamos neste apêndice, podemos dizer que, ao estabelecermos uma condição limitante para os valores de altura, fazemos com que, durante o cálculo da média, após a escolha dos quatro primeiros valores de altura, o quinto e último seja obrigado a ter um determinado valor, para que sua soma com os demais dividida por cinco seja 1,07 m.

No exemplo, se escolhermos os valores de altura na ordem na qual aparecem na planilha acima, efetuaremos a seguinte soma para calcularmos a média:

$$1,14 + 0,86 + 1,24 + 1,17 + 0,94 .$$

Acontece que se já escolhemos as alturas 1,14; 0,86; 1,24 e 1,17, a última altura não tem liberdade para variar, pois, tem que ser 0,94 para que essa soma, quando dividida por cinco, seja 1,07 m, valor que, como já vimos, não podemos alterar.

Podemos escolher uma outra ordem para entrada dos valores de altura no cálculo da média, porque temos liberdade para variar os quatro primeiros valores. Por exemplo, podemos efetuar a soma na seguinte ordem:

$$0,94 + 1,17 + 1,24 + 0,86 + 1,14 .$$

Neste caso, pelo mesmo motivo, o último número considerado no cálculo da média não teve a liberdade de variar, pois teve que ser 1,14 para que o resultado final da média seja 1,07 m.

Outras ordens para efetuação da soma das alturas poderiam também ser consideradas. Por exemplo, poderíamos somar $1,14 + 0,86 + 1,17 + 0,94 + 1,24$. Novamente, os quatro primeiros números tiveram a liberdade de variar seu valor, mas o último teve seu valor automaticamente fixado pela escolha dos quatro valores anteriores e não teve, portanto, a liberdade de variar.

Uma regra universal é que o número de graus de liberdade é sempre igual ao número de observações menos o número de relações necessárias entre estas observações.

Em termos geométricos, o número de observações equivale ao número de dimensões do espaço original, e cada relação existente entre as observações representa uma seção através desse espaço multidimensional, que restringe os pontos gerados utilizando as observações como coordenadas, a um espaço com uma dimensão a menos.

Podemos afirmar também que o número de graus de liberdade é igual ao número de observações independentes (que é o número original de observações) menos o número de parâmetros (uma média, p. ex.) estimados a partir dessas observações. Assim, estimar um parâmetro a partir das observações, equivale a estabelecer uma relação entre valores que antes eram totalmente independentes. É o que acontece quando estimamos (calculamos) a média das alturas no nosso exemplo. A necessidade de calcularmos a média das alturas para obtermos o desvio-padrão, impõe uma relação, uma interdependência entre os valores de altura, que tira de um dos valores de altura a liberdade de variar. Por isso, a fórmula para o cálculo do desvio-padrão

tem no seu denominador $n - 1$ e não n . Se, no desvio-padrão, queremos dividir o total dos desvios dos valores de altura em relação à média das alturas, pelo número de indivíduos, com o objetivo de obtermos a quantidade de desvio por indivíduo (ou a média do desvio), temos que considerar no denominador apenas aqueles indivíduos cuja altura teve a liberdade de variar, concorda? Como uma das alturas (a última a ser considerada no cálculo da média) não teve a liberdade de variar, temos que retirá-la do cálculo, e por isso, o denominador é $n - 1$.

Os estatísticos recomendam que, para descrever dados, não importa muito se o denominador é n ou $n - 1$, mas para inferência estatística importa, porque o uso do primeiro denominador resulta em um estimador viciado do erro-padrão, enquanto o uso do segundo não.

Existem várias outras situações em estatística nas quais perdemos graus de liberdade. Os fundamentos genéricos para você entender essas perdas foram colocados neste apêndice. Ao longo deste livro explicamos especificamente as perdas de graus de liberdade para cada técnica estatística na qual essas perdas ocorram.



APÊNDICE 2

– Quais os valores de x e de y que satisfazem as condições $x + y = 7$ e $x^2 + y^2 = 25$ simultaneamente?

– Veja a seguir a resolução dessas equações para x e y :

Resumo:

$$x^2 + x^2 - 14x + 49 = 25$$

$$x^2 - 7x + 12 = 0$$

$$S = 7 \quad P = 12 \quad x = 3 \text{ ou } x = 4.$$

Resolução:

Se $x + y = 7$, então $y = 7 - x$.

Substituindo y na outra equação $x^2 + y^2 = 25$ temos

$$x^2 + (7 - x)^2 = 25.$$

O próximo passo é resolvermos o termo $(7 - x)^2$ da equação acima. Há duas opções:

a) 1ª opção:

$$(7 - x)^2 = (x - 7)^2, \text{ pois}$$

$$(7 - x)^2 = (7 - x)^2 (1) =$$

$$= (7 - x)^2 (-1)^2 =$$

$$= [(7 - x)(-1)]^2 = (x - 7)^2.$$

Sabe-se que $(a + b)^2 = a^2 + 2ab + b^2$ e que $(a - b)^2 = a^2 - 2ab + b^2$, então

$$(x - 7)^2 = x^2 - 14x + 49.$$

Dedução:

$$(a - b)^2 = (a - b)(a - b) = a^2 - ab - ba + (-b)(-b) = a^2 - 2ab + b^2.$$

b) 2ª opção:

Como $(a - b)^2 = a^2 - 2ab + b^2$, sendo isso válido para quaisquer valores de a e de b , podemos substituir a por 7 e b por x , obtendo:

$$(7 - x)^2 = 7^2 - 2(7)x + x^2 = 49 - 14x + x^2 = x^2 - 14x + 49.$$

Substituindo agora $(7 - x)^2$ por $x^2 - 14x + 49$ na equação $x^2 + (7 - x)^2 = 25$ temos

$$x^2 + x^2 - 14x + 49 = 25$$

$$2x^2 - 14x + 49 - 25 = 0$$

$$2x^2 - 14x + 24 = 0.$$

Dividindo todos os termos por 2 chegamos a

$$x^2 - 7x + 12 = 0.$$

Os valores de x que satisfazem a equação acima são as raízes do polinômio do segundo grau $ax^2 - bx + c$, no qual os coeficientes são $a = 1$, $b = -7$ e $c = 12$. Note que se substituirmos estes valores de a , b e c no polinômio $ax^2 - bx + c$, obteremos $x^2 - 7x + 12 = 0$, que é a equação à qual chegamos logo acima. Como o polinômio é do segundo grau, sabemos que serão duas raízes que satisfarão esta equação. Só falta agora obtermos os valores dessas raízes. Existem duas opções:

a) Opção prática:

As raízes são tais que sua soma S é igual a menos b sobre a , ou seja, $S = -\frac{b}{a}$, e seu produto P é igual a c sobre a , isto é, $P = \frac{c}{a}$. Logo, as duas raízes serão dois números tais que $S = -\frac{(-7)}{1} = 7$ e $P = \frac{12}{1} = 12$. Então, as raízes são 3 e 4, que são os valores de x e de y que satisfazem aquelas duas equações iniciais $x + y = 7$ e $x^2 + y^2 = 25$.

Observação: quando as raízes não são números inteiros, é difícil usar este método.

b) Opção tradicional:

$$x = \frac{-b \pm \sqrt{\Delta}}{2a}, \text{ onde } \Delta = b^2 - 4ac. \text{ Logo } \Delta = (-7)^2 - (4)(1)(12) = 49 - 48 = 1.$$

Substituindo Δ na equação temos

$$x = \frac{-(-7) \pm \sqrt{1}}{(2)(1)} = \frac{7 \pm 1}{2}, \text{ donde}$$

$$x = \frac{7-1}{2} = 3 \text{ ou } x = \frac{7+1}{2} = 4.$$



APÊNDICE 3

— Como resolver a equação $d = z_{(1-\alpha/2)} EP = z_{(1-\alpha/2)} \sqrt{\frac{pq}{n}} \sqrt{\frac{N-n}{N-1}}$ para n ?

— Há dois modos de fazermos isso: um modo curto e um longo.

Modo mais curto:

$$d = z_{(1-\alpha/2)} \sqrt{\frac{pq}{n}} \sqrt{\frac{N-n}{N-1}}$$

$$d^2 = z_{(1-\alpha/2)}^2 \left(\sqrt{\frac{pq}{n}} \right)^2 \left(\sqrt{\frac{N-n}{N-1}} \right)^2$$

$$d^2 = z_{(1-\alpha/2)}^2 \left(\frac{pq}{n} \right) \left(\frac{N-n}{N-1} \right)$$

$$d^2 = \frac{z_{(1-\alpha/2)}^2 pq (N-n)}{n(N-1)}$$

$$d^2 (N-1)n = z_{(1-\alpha/2)}^2 pq (N-n)$$

$$d^2 (N-1)n = z_{(1-\alpha/2)}^2 pq N - z_{(1-\alpha/2)}^2 pq n$$

$$d^2 (N-1)n + z_{(1-\alpha/2)}^2 pq n = z_{(1-\alpha/2)}^2 pq N$$

Colocando o fator comum em evidência, obtemos:

$$\left[d^2 (N-1) + z_{(1-\alpha/2)}^2 pq \right] n = N z_{(1-\alpha/2)}^2 pq$$

$$n = \frac{N z_{(1-\alpha/2)}^2 pq}{d^2 (N-1) + z_{(1-\alpha/2)}^2 pq}$$

Modo mais longo:

$$d = z_{(1-\alpha/2)} \sqrt{\frac{p(1-p)}{n}} \sqrt{\frac{N-n}{N-1}}$$

$$d^2 = z_{(1-\alpha/2)}^2 \left(\sqrt{\frac{p(1-p)}{n}} \right)^2 \left(\sqrt{\frac{N-n}{N-1}} \right)^2$$

$$d^2 = z_{(1-\alpha/2)}^2 \left(\frac{p(1-p)}{n} \right) \left(\frac{N-n}{N-1} \right)$$

$$d^2 = \frac{z_{(1-\alpha/2)}^2 p(1-p)(N-n)}{n(N-1)}$$

$$d^2 n(N-1) = (z_{(1-\alpha/2)}^2 p - z_{(1-\alpha/2)}^2 p^2)(N-n)$$

$$d^2 n(N-1) = z_{(1-\alpha/2)}^2 pN - z_{(1-\alpha/2)}^2 pn - z_{(1-\alpha/2)}^2 p^2 N + z_{(1-\alpha/2)}^2 p^2 n$$

$$d^2 n(N-1) + z_{(1-\alpha/2)}^2 pn - z_{(1-\alpha/2)}^2 p^2 n = z_{(1-\alpha/2)}^2 pN - z_{(1-\alpha/2)}^2 p^2 N.$$

Colocando o fator comum em evidência, obtemos:

$$\left[d^2 (N-1) + z_{(1-\alpha/2)}^2 p - z_{(1-\alpha/2)}^2 p^2 \right] n = z_{(1-\alpha/2)}^2 pN - z_{(1-\alpha/2)}^2 p^2 N$$

$$n = \frac{z_{(1-\alpha/2)}^2 pN - z_{(1-\alpha/2)}^2 p^2 N}{d^2 (N-1) + z_{(1-\alpha/2)}^2 p - z_{(1-\alpha/2)}^2 p^2}$$

$$n = \frac{z_{(1-\alpha/2)}^2 pN(1-p)}{d^2 (N-1) + z_{(1-\alpha/2)}^2 p(1-p)}.$$

Como $1-p = q$ e, considerando a propriedade comutativa da multiplicação, obtemos finalmente:

$$n = \frac{N z_{(1-\alpha/2)}^2 pq}{d^2 (N-1) + z_{(1-\alpha/2)}^2 pq}.$$

ÍNDICE REMISSIVO

A

Amostragem, 24-26

- aleatória, 26
- não-aleatória, 2, 142
- representatividade, 25
- tamanho mínimo
 - correção p/ população finita, uma média, 266-268
 - correção p/ população finita, uma proporção, 270-272
 - efeito do desenho, uma média, 268, 269
 - efeito do desenho, uma proporção, 272
 - formas p/ estimar desvio-padrão, 264
 - fundamento estatístico, 262
 - para estimar uma média, 262-269
 - para estimar uma proporção, 269-272

Amostragem, 24-26

- aleatória, 26
 - estratificada, 27-29, 203
 - por conglomerados, 27
 - proporcional, 27-29
 - simples, 26, 28
 - sistemática, 26, 27
- com reposição, 26, 266, 267, 271
- correção p/ população finita, 30, 266-268, 270-272
- não-aleatória, 26, 29, 143
 - de voluntários, 29
 - por conveniência, 29
 - por auto-seleção. *Consulte Amostragem não-aleatória de voluntários*
- sem reposição, 26, 29, 30, 266-268, 271, 295
- unidade amostral, 26, 27

Análise

- bivariável, 5
- de agrupamento, 6
- de componentes principais, 6
- de contingência, 6
- de correlação
 - canônica, 6
 - de Cronbach, 5, 6
 - de Kendall, 6
 - de Pearson, 4-6
 - de Spearman, 4-6
 - intraclasse, 5, 6
 - parcial múltipla, 6
- de correspondência, 6
- de escala multidimensional, 6
- de fator, 6
- de homogeneidade, 6
- de regressão
 - binomial negativa, 6
 - de Cox, 6
 - de Poisson, 6
 - de Weibull, 6
 - hierárquica, 6
 - linear, 5, 6
 - logística, 6
 - log-linear, 6

- de variância, 5, 6
 - multinomial, 6
- discriminante, 6
- estratificada, 4, 5, 17
- exploratória, 4, 6
- multivariável, 4, 5
- por redes neurais artificiais, 6

B

Bioestatística. 9. 32. 46. 262. 289

C

Confundimento. *Consulte* Variável independente secundária
Contagem, 3, 8, 24, 32-34, 40
Covariável. *Consulte* Variável independente secundária

D

Distribuição
de frequências, 7, 40, 102, 118-115, 121, 128,
129, 144, 146, 150, 288
binomial, 122, 244-247
obtenção de áreas sob a, 123, 133
de Gauss. Consulte Distribuição normal
de Poisson, 122
de médias amostrais, 149, 150, 178, 198
F, 214-216
hipergeométrica, 294, 300
normal, 8, 115, 125, 127, 129, 130-132, 136,
137, 143, 144, 146, 149, 150-152, 156,
157, 159, 160, 161, 163, 171, 174,
182, 198, 206, 213, 244-246, 249, 264,
265, 278, 279
aproximação da binomial, 244, 245,
248, 254
aproximação da binomial, correção p/
continuidade, 246, 251, 252, 257-
259
propriedades da, 127-132
normal padrão, 130-131, 158, 161, 262, 263
desvio-padrão e erro-padrão da, 161
média da, 161
obtenção de áreas sob a, 133-137
qui-quadrado, 137, 278, 279, 288
real, 124, 132
T, 137, 150, 186, 187, 191-194, 196, 212, 225,
229, 239, 240, 241, 279
propriedades da, 186
probabilística, 8, 122, 125, 136, 137, 144, 152,
159, 196, 278

E

Emparelhamento, 232, 233
Estatística, 2, 3, 6, 8-11, 32, 35, 86, 263
 analítica, 2-5, 86, 137
 bayesiana, 10, 11
 descritiva, 2-4, 40, 68, 86
 etapas iniciais, 34

organização dos dados, 35
 inferencial, 2, 3, 5, 40, 86, 128, 137
 moderna, 8
 não-paramétrica, 7, 8, 184, 229
 paramétrica, 7, 8, 132, 184
 teste de significância, 2, 3, 25, 167, 169, 184
 bicaudado, 156, 165, 172, 173, 177, 181, 189, 193, 194, 196, 208-210, 215, 217, 219-221, 223, 236, 241, 249, 252, 256, 258, 282, 293, 300
 da mediana, 5, 6
 da razão de variâncias, 5, 212, 213, 214, 223
 de Friedman, 5, 6
 de Kruskal-Wallis, 5, 6
 de Mann-Whitney, 5, 6
 de McNemar, 5, 6
 de uma variância, 5, 6
 de Wilcoxon, 5, 6
 do sinal, 5, 6
 etapas do, 150-166
 exato de Fisher, 5, 229, 287, 289, 292-300
 livre de distribuição. *Consulte* teste de significância não-paramétrico
 monocaudado, 171-173, 175-177, 236-239, 293, 298, 300
 não-paramétrico, 8, 184, 198, 199, 205, 229, 230
 paramétrico, 7, 8, 184
 poder do, 169
 qui-quadrado, 3, 5, 6, 229, 259, 272, 274-289
 correção p/ continuidade, 281
 graus de liberdade, 278, 280, 281, 285, 287
 outras aplicações, 287-289
 quando não utilizar, 287
 teste de homogeneidade, 288
 teste de independência, 274-288
 teste de qualidade do ajuste, 288
 teste de tendência, 288
 qui-quadrado de Mantel e Haenszel, 5, 6
 t^2 , 212, 223-227
 t p/ amostras emparelhadas, 5, 229, 232-239
 t p/ duas médias, 5, 47, 199, 202-205, 212, 217-221, 228, 229
 t p/ uma média, 5, 186, 187, 191-198
 z p/ duas médias, 5, 47, 199, 202-210, 228, 229
 z p/ duas proporções, 5, 252-259
 z p/ uma média, 5, 150-166, 171-177, 187-191, 196, 198
 z p/ uma proporção, 5, 247-252

Estudo
 caso-controle, 9, 16, 19, 29, 288
 de agregados, 9
 de coorte, 9
 de incidência, 9
 de prevalência, 9
 experimental, 9, 29
 transversal, 2, 9, 203

F

Frequência

relativa, 4, 36-40
 relativa acumulada, 4, 36, 38-40
 simples, 4, 36-38, 40
 simples acumulada, 4, 36, 38, 40

G

Gráfico, 4, 78, 86, 93-115, 118

cartograma, 4, 93
 diagrama, 4, 5, 7, 9, 40, 78, 79, 93-115, 126
 de barras, 4, 93, 95-98
 de barras de erro, 5, 93, 98-100
 de caixa, 4, 93, 109-113
 de dispersão, 4, 93, 107-109
 de linha, 93, 105, 106
 de linhas de afastamento, 93, 113, 114
 de pontos, 93, 105
 de setores, 4, 93-95
 de talo e folha, 4, 93, 104, 105
 histograma, 4, 93, 100-102
 polígono de frequências, 93, 102-104

Grau de liberdade, 64, 186, 191-196, 214-216, 218-221, 223, 225, 237, 238, 240, 241, 278-282, 285, 287, 293

I

Inferência estatística, 5, 8, 10, 65, 73, 83, 98, 99, 102, 119, 132, 137, 140-144, 147, 148, 150, 153, 159, 167, 178, 182, 184, 186, 187, 198, 199, 202-204, 216, 227-229, 232, 236, 237, 244, 263

definição, 142
 erros na, 166-170
 intervalo de confiança, 3, 99, 150, 154, 177-183
 interpretação, 181-182
 monocaudado, 239, 240
 p/ amostras não independentes, 239-241
 p/ duas médias, 210, 211, 222, 226, 227
 p/ duas proporções, 257
 p/ uma média, 180-182, 191, 195, 197, 198
 p/ uma proporção, 250, 251, 257
 maneiras de fazer, 150
 intervalo de confiança, 178-183
 teste de hipóteses, 150-166
 sobre duas médias, 5, 202-212, 217-222, 224-230, 232-241
 sobre duas proporções, 5, 252-259
 sobre duas proporções ou mais, 5, 259, 272, 274-288
 sobre uma média, 5, 156, 186-198
 sobre uma proporção, 5, 214, 247-252
 valor de F, 213, 215, 223
 valor de z , 133, 135, 136, 158-162
 valor-p, 159, 163, 164
 variação amostral e, 141, 142, 144, 151, 152, 164-166, 170

Inferência não-estatística, 25, 142

Interação, 17, 18, 55
 antagonismo, 17
 sinergismo, 17

M

Medição, 3, 8, 24, 32-34

Medida

- de associação, 4, 6, 182, 288, 289
 - coeficiente de correlação, 4-6, 107, 108
 - coeficiente de regressão, 4-6, 107, 153, 288
 - diferença de chances (risco atribuível), 4, 289
 - diferença de incidências (risco atribuível), 4, 289
 - diferença de prevalências, 4, 288
 - razão de chances, 4, 6, 182, 289
 - razão de incidências (risco relativo), 4, 6, 182, 289
 - razão de prevalências, 3, 4, 182, 288
- de concordância, 279
 - índice capá, 5, 6, 289
- de dispersão, 4, 40, 55, 58-68, 80, 92
 - amplitude, 4, 40, 58-60, 66, 68, 80, 82, 83, 264
 - amplitude interquartil, 58-60, 80-82
 - amplitude interquartil porcentual, 82, 83
 - coeficiente de variação, 4, 40, 58, 59, 66, 67
 - desvio médio, 4, 40, 58, 59, 60-62, 65, 66, 68
 - desvio-padrão, 4, 33, 40, 58-60, 65, 66-68
 - erro-padrão, correção p/ população finita, 30, 266-268, 270-271
 - erro-padrão da diferença entre duas médias, 207, 208
 - erro-padrão da diferença entre duas médias de amostras não independentes, 236, 237
 - erro-padrão da diferença entre duas proporções, 253-255
 - erro-padrão de uma média, 147-150
 - erro-padrão de uma proporção, 248, 249, 269
 - finalidade, 158-160
 - variância, 4, 5, 40, 58, 62-66, 68, 147-150, 186, 207, 208, 212-218, 223-225, 241, 253, 262, 269
- de frequência,
 - prevalência, 3, 125, 247-249-256, 269, 271, 274, 282, 283, 285, 286, 288, 298, 300
- de posição, 4, 40, 68, 70-83
 - aplicação dos percentis, 78-83
 - percentis, 4, 40, 71-83
 - quartis, 72-78
- de tendência central, 4, 40, 42-55
 - média aritmética, 4, 44-46
 - média geométrica, 44, 47
 - média harmônica, 44, 47
 - média ponderada, 4, 44, 46, 47
 - mediana, 4, 40, 42, 44, 47-52
 - moda, 4, 40, 42-44
 - vantagens e limitações, 52-55
- Modificação de efeito. *Consulte* interação

N

Normalidade, 33

- definição clínica, 127
- definição estatística, 125-127
- definição terapêutica, 127

P

Pesquisa

- qualitativa ou quantitativa, 11

População,

- alvo, 25-29, 45, 140-144
- amostrada, 25, 26
- finita, 30, 266-268, 270, 271
- infinita, 8, 64, 121, 124, 129, 178, 179, 188, 266, 267-268, 267

Q

Quadro, 5, 86, 87

T

Tabela, 4, 9, 86-92, 115

- c/ valores de F, 216
- c/ valores de qui-quadrado, 282
- c/ valores de T, 192, 220
- c/ valores negativos de Z, 176
- c/ valores positivos de Z, 134
- de contingência, 4, 274-277, 280-281, 287, 288, 292, 294

Teorema central do limite, 149, 150, 178, 198, 199, 228, 229, 241

Transformação em valor de z, 131, 134-137

V

Variável, 4, 5, 10, 14-22

- aleatória, 19, 20
- categoria de uma, 14, 18-22, 27, 32-34, 40, 55, 79, 94-96, 99, 101, 106, 113, 274, 280, 287
- classificação, 14-22
- codificação, 32
- confundidora. *Consulte* Variável independente secundária
- contínua, 18, 19, 55, 107, 121
- de razão, 20-22, 55, 94, 99, 102, 113
- dependente, 15-18
- dicotômica, 19
- discreta, 18, 19, 55, 122
- escore de corte de uma, 33, 79, 115, 126, 269
- fixa, 19, 20
- independente, 15
 - interveniente, 16-18
 - principal, 15-18, 20, 24
 - secundária, 15, 17-19
- intervalar, 20-22
- nominal, 20-22, 32, 53, 54, 94, 100, 109, 113, 288, 289
- ordinal, 20-22, 94, 100, 113
- policotômica, 19
- qualitativa, 14, 15, 19
- quantitativa, 14, 15, 18
- valores perdidos de uma, 71

Viés, 24, 26, 27, 29, 184