Capítulo

10

Introdução à Inferência Estatística

Gilberto Pereira Sassi
gilberto.sassi@ufba.br

Objetivo do Capítulo

Este capítulo tem o objetivo de apresentar métodos de inferência estatística. Em Informática na Educação, usamos inferência estatística para analisar uma base de dados com os resultados de uma pesquisa e construir afirmações válidas para toda população. Ao final da leitura deste capítulo, você deve ser capaz de:

- Entender o conceito de probabilidade.
- Entender a distribuição normal
- Entender o Teorema Central do Limite.
- Construir intervalos de confiança para a média quando conhecemos a variância da população
- Construir intervalos de confiança para a média quando não conhecemos a variância da população, mas a variável tem distribuição normal
- Construir intervalos de confiança para a média quando não conhecemos a variância e a variável não tem distribuição normal usando *bootstrap*
- Checar a independência entre duas variáveis qualitativas
- Checar a independência entre duas variáveis quantitativas
- Checar a normalidade de uma variável quantitativa usando os testes de hipóteses de Shapiro-Wilks, Anderson-Darling e Kolmogorov-Smirnov
- Construir teste de hipóteses para média quando não conhecemos a variância da população e a variável tem distribuição normal
- Construir teste de hipóteses para média quando conhecemos a variância da população
- Uso e interpretação do do p-valor

Era uma vez... *Heloísa é uma aluna de um programa de pós-graduação em Informática na Educação que decidiu analisar se uma versão gamificada de um ambiente virtual de aprendizagem engajava mais os alunos nas atividades da comunidade. Para tanto, ela elaborou um estudo experimental em que comparou a versão gamificada com a versão não-gamificada desse ambiente. Para verificar o engajamento dos alunos nas atividades, ela coletou várias informações sobre os alunos através do log do ambiente, como tempo para terminar uma atividade proposta, número de interações com outros alunos, notas nos testes semanais, idade do aluno, e outros. Heloísa agora precisa responder a seguinte pergunta: "As mudanças propostas melhoram o desempenho dos alunos? ". Será que podemos ajudar Heloísa?*

1 Introdução

O processo de generalização de alguns casos para todos os casos possíveis é um processo essencial e recorrente na pesquisa científica e a área de Informática na Educação não é a exceção. Apresentaremos a você conceitos e métodos para fazer afirmações sobre todos os casos (ou indivíduos) alvo do seu estudo a partir de uma base de dados. Chamamos esse processo de inferir sobre o todo a partir de alguns casos de inferência estatística ou estatística inferencial.

Para facilitar a sua leitura, organizamos este capítulo em três partes: dedicamos a primeira seção sobre probabilidade, pois para construirmos a generalização da parte para o todo vamos usar probabilidade; em seguida, na seção 2, explicamos e mostramos como construir intervalo de confiança; e, finalmente, na seção 3 construímos testes de hipóteses usando o procedimento de Neyman-Pearson e aprendemos a calcular o p-valor. Ilustramos os conceitos com um exemplo na seção 4 "Cenário Ilustrativo" onde mostramos como construir intervalos de confiança e testes de hipóteses usando o *software* R.

2 Probabilidade e a distribuição normal

Nesta seção, vamos procurar um modelo matemático adequado para uma característica numérica dos indivíduos e com esse modelo matemático queremos fazer afirmações sobre toda população. Chamamos este modelo matemático de probabilidade. Em estatística descritiva, se temos uma variável quantitativa contínua X, podemos calcular as frequências relativas $f_1, f_2, \dots, f_k$ para faixas de valores $(a_0, a_1], (a_1, a_2], \dots, (a_{k-1}, a_k]$. Aqui, desejamos substituir as frequências relativas (válidas para a uma amostra) por probabilidades (válidas para toda população). De forma análoga à frequência relativa, probabilidades sempre estão entre zero e um. Quando estamos lidando com probabilidades e uma variável, substituímos a palavra "quantitativa" pela palavra "aleatória" para deixar claro ao leitor que estamos lidando com probabilidades ao invés de frequências relativas. Se os valores possíveis de uma variável aleatória X são número inteiros, dizemos que X é uma variável aleatória discreta. Por

exemplo, a idade do aluno (em anos) é uma variável aleatória discreta. Se os valores de uma variável aleatória X podem ser qualquer número real, dizemos que X é uma variável aleatória contínua. Por exemplo, a nota (qualquer valor real entre 0 e 10) de um aluno é uma variável aleatória contínua.

Existem vários modelos de probabilidades para variáveis aleatórias discretas, mas o foco desta capítulo será variável aleatória contínua. Caso você tenha interesse ou necessidade de usar variáveis aleatórias discretas, deixaremos algumas referências para consulta ao final deste capítulo. Vamos focar em apenas um modelo de probabilidade, conhecido como distribuição normal. Esse é o principal modelo de probabilidade e sem dúvida o modelo de probabilidade mais relevante para a área de Informática na educação. Este modelo é essencial para entendermos intervalo de confiança e teste de hipótese.

Para uma variável quantitativa contínua X, podemos construir o histograma. O histograma é válido para uma amostra específica. Para uma outra amostra da variável quantitativa contínua X na mesma população, podemos obter um histograma ligeiramente diferente. A ideia é obter uma curva ou uma função que aproxime bem o formato de todos histogramas, como ilustrado na figura 1. Chamamos esta curva ou função que representa o formato dos histogramas de função densidade de probabilidade ou simplesmente função densidade. De forma análoga ao histograma, a probabilidade da variável aleatória contínua X estar entre a e b é a área da função densidade no intervalo $(a, b]$, como ilustrado na Figura 2. Consequentemente, ao identificarmos a função densidade de probabilidade para uma variável aleatória contínua X, identificamos as probabilidades para intervalos e temos um modelo probabilístico. Geralmente chamamos esse modelo probabilístico de distribuição ou distribuição de probabilidade.

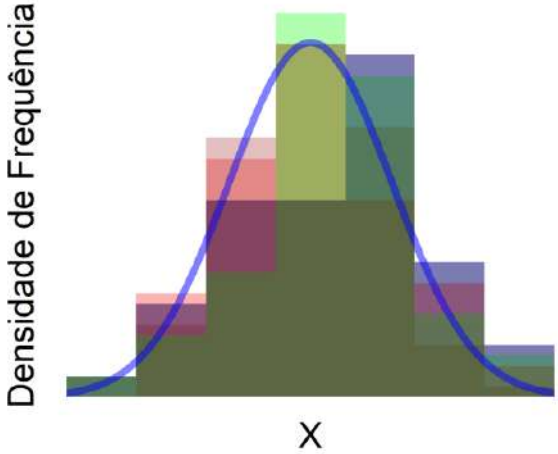


Figura 1: Histogramas sobrepostos para diversas amostras da variável X. A linha azul é o modelo de probabilidade para X. Chamamos esta curva azul de função densidade de probabilidade.

Como pesquisador da área de Informática na Educação, você certamente vai se deparar com a distribuição normal. Podemos afirmar, sem medo, que o modelo probabilístico mais usado é a distribuição normal e sua função densidade é descrita matematicamente por

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

em que μ é a média da população e σ é o desvio padrão da população. A probabilidade de X estar entre a e b é a área da função densidade de probabilidade delimitada por a e b, conforme ilustrado na Figura 3. Quando $\mu = 0$ e $\sigma = 1$, dizemos que a X tem distribuição normal padrão. Representamos matematicamente a probabilidade de X estar no intervalo $(a, b]$ por

$$P(a < X \leq b).$$

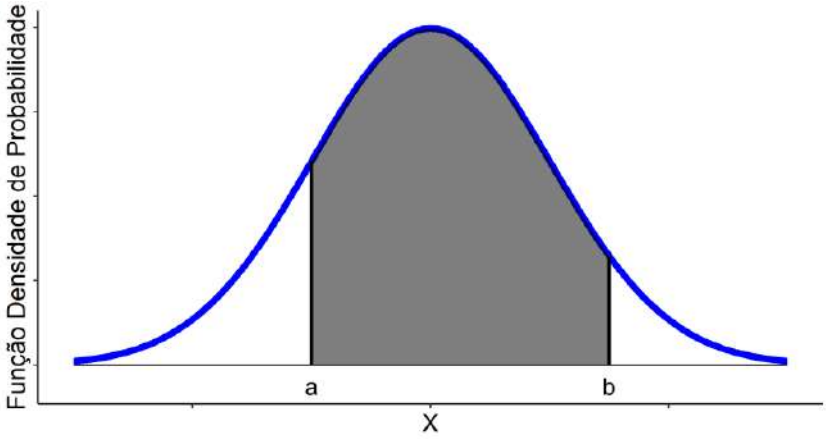


Figura 2: A probabilidade de X estar entre a e b é a área da função densidade de probabilidade delimitada por a e b. Usamos a notação $P(a < X \leq b)$ para a área demarcada no gráfico.

Um dos motivos da importância e popularidade da distribuição normal é o Teorema Central do Limite. Vamos ilustrar a ideia do Teorema Central do Limite com um exemplo. Imagine que Heloísa, a aluna de pós-graduação na área de Informática da Educação da seção **Era uma vez...**, está ensinando Estatística para uma turma de 30 alunos que realizaram a primeira prova com as notas na Tabela 1. Heloísa decidiu selecionar aleatoriamente 5 alunos. Existem 142.506 formas de selecionar estes cincos alunos. Na tabela 2, ilustramos 10 amostras com alunos que Heloísa poderia ter escolhido. Podemos calcular a média dos cinco alunos para cada amostra como ilustrado na tabela 2. Você pode notar que as médias das amostras mudam se Heloísa seleciona alunos diferentes, e você pode considerar que essas médias são valores observados de uma variável aleatória contínua que vamos representar por $\bar{X}$. O Teorema Central do Limite afirma que o modelo de probabilidade de $\bar{X}$ pode ser aproximado pela distribuição normal com média μ e desvio padrão $\frac{\sigma}{\sqrt{n}}$, em que μ é a média populacional, σ é o desvio padrão populacional e n é o tamanho da amostra. Ilustramos a ideia do Teorema Central do Limite na Figura 3.

Tabela 1: Nota dos 30 alunos na primeira prova.

5.80	7.27	7.04	6.83	7.64	7.11	7.51	7.30	6.89	6.30
5.60	8.80	7.05	6.13	5.42	6.24	8.54	5.74	5.43	5.95
6.82	5.73	7.84	5.90	7.80	5.48	8.59	5.76	3.90	7.10

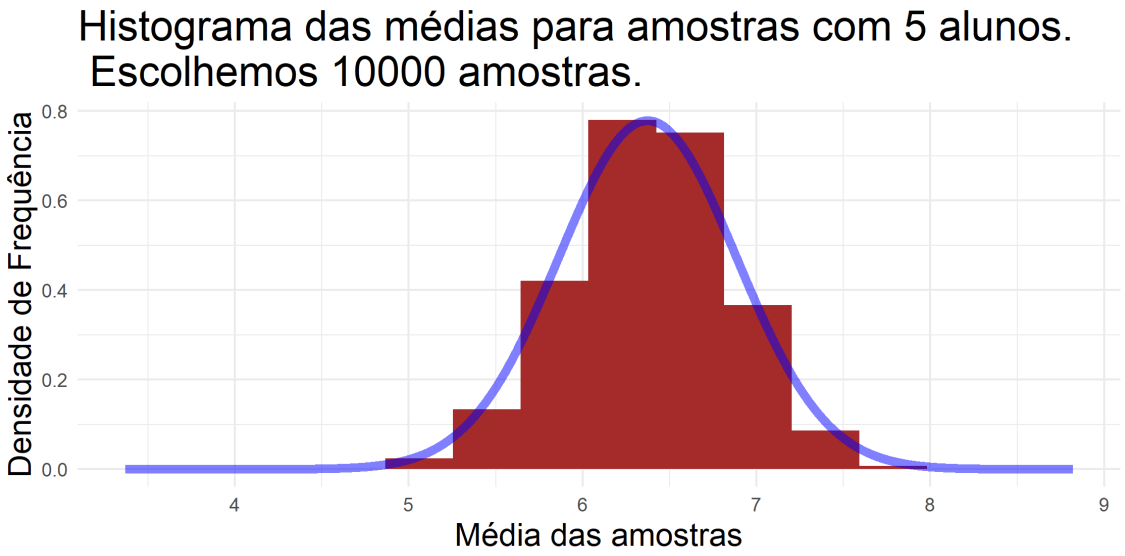


Figura 3: Histograma das médias de amostras com cinco alunos dos alunos da tabela 1, e o modelo de probabilidade segundo o Teorema Central do Limite (curva azul).

Tabela 2: Dez amostras com cinco alunos provenientes dos 30 alunos da Tabela 1.

Amostras	Notas dos cinco indivíduos da amostra de Heloísa					Média
Amostra 1	6,83	7,30	6,89	7,51	3,90	6,49
Amostra 2	5,76	6,13	8,59	6,89	6,82	6,84
Amostra 3	5,80	5,48	6,13	6,82	6,24	6,09
Amostra 4	8,59	5,73	5,80	7,10	6,83	6,81
Amostra 5	6,82	6,30	6,13	7,10	8,80	7,03
Amostra 6	7,51	8,59	5,73	3,90	6,82	6,51
Amostra 7	5,95	6,89	5,73	3,90	7,30	5,95
Amostra 8	8,59	5,74	6,30	7,10	6,24	6,79
Amostra 9	6,83	7,80	7,51	6,89	5,95	7,00
Amostra 10	5,43	7,80	6,89	5,80	7,30	6,64

3 Intervalo de confiança

Nesta seção, queremos usar probabilidade, a distribuição normal padrão e o Teorema Central do Limite e as informações da base de dados para construir um limite inferior e um limite superior para o parâmetro média μ. Mais precisamente, queremos encontrar dois números reais a e b tal que a média populacional μ está entre a e b. Geralmente, denominamos o intervalo (a, b) de intervalo de confiança e acreditamos que este intervalo de confiança está correto (o parâmetro μ está neste intervalo) com alguma probabilidade que vamos chamar de coeficiente de confiança. Vamos dividir nossa discussão em duas situações: quando conhecemos a variância da população (variância conhecida) e quando desconhecemos a variância da população (variância desconhecida).

3.1 Intervalo de confiança para média: variância conhecida

Heloísa, nossa aluna de pós-graduação, selecionou n alunos e salvou o tempo que estes alunos ficaram logados no ambiente virtual de aprendizagem: vamos denotar essa variável (tempo) por X. Imagine um cenário em que Heloísa conhece o desvio padrão (ou a variância) da variável X para a população de todos os alunos do ambiente virtual de aprendizagem. Se os valores observados por Heloísa para a variável X foram $x_1, \dots, x_n$ com média $\bar{x} = \frac{x_1 + \cdots + x_n}{n}$, então os estatísticos demonstraram que o intervalo

$$IC(\mu; \gamma) = \left(-z_\gamma \frac{\sigma}{\sqrt{n}} + \bar{x}; z_\gamma \frac{\sigma}{\sqrt{n}} + \bar{x} \right),$$

contém a média da população μ em $100 \cdot \gamma\%$ das amostras, em que z_γ é um número que satisfaz $P(Z \leq z_\gamma) = \frac{1+\gamma}{2}$, σ é o desvio padrão da população e n é o tamanho da amostra. Chamamos γ de coeficiente de confiança, e denominamos z_γ de quantil. Na figura 4, ilustramos a ideia do quantil z_γ.

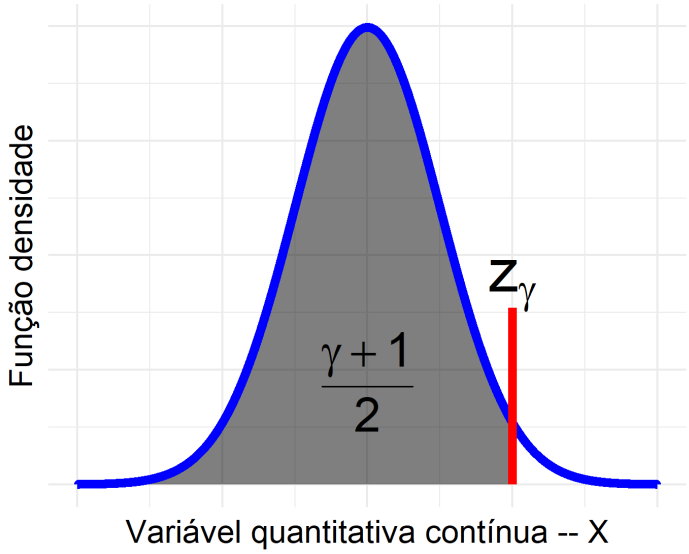


Figura 4: Quantil z_γ para a distribuição normal padrão.

Gostaríamos de reforçar que Heloísa realizou uma inferência indutiva, ou seja, Heloísa fez uma afirmação sobre a média populacional usando uma amostra. Tenha em mente que para diferentes amostras Heloísa obtém intervalos diferentes, e existe a possibilidade de a média populacional não estar no intervalo, como ilustrado na Figura 5.

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O
1	População					Média Populacional		Amostras						Intervalo de Confiança	Média populacional no intervalo?
2	7,29	7,15	5,93	6,44	8,16	6,373		Amostra 1	5,77	6,57	6,44	8,16	7,93	(6,46; 7,49)	Não
3	4,84	6,11	3,37	6,7	6,31			Amostra 2	6,14	7,36	3,37	6,7	7,1	(5,62; 6,65)	Sim
4	5,89	7,77	5,24	5,77	6			Amostra 3	8,82	6,31	7,36	5,77	5,72	(6,28; 7,31)	Sim
5	7,19	5,54	5,53	6,27	5,72			Amostra 4	7,1	7,93	4,84	6,44	5,93	(5,94; 6,96)	Sim
6	4,63	7,1	7,36	5,7	7,64			Amostra 5	3,37	4,84	7,15	5,93	6,11	(4,97; 5,99)	Não
7	8,82	7,93	6,08	6,57	6,14			Amostra 6	4,63	7,93	7,19	6,27	5,77	(5,85; 6,87)	Sim
8															

Figura 5: População com notas de 30 alunos e intervalos de confianças com coeficiente de confiança $\gamma = 95\%$ para 6 amostras de cinco alunos.

Na Figura 5, Heloísa percebeu que o intervalo de confiança com coeficiente de confiança $\gamma = 0{,}95$ pode ou não conter a média populacional. Mas Heloísa não precisa se preocupar, pois 95% dos intervalos de confiança vão conter a média populacional já que $\gamma = 0{,}95$. Ou seja, de 100 intervalos de confiança de amostras distintas, 95 destes intervalos de confiança contêm a média populacional. Na Figura 6, ilustramos a ideia do coeficiente de confiança.

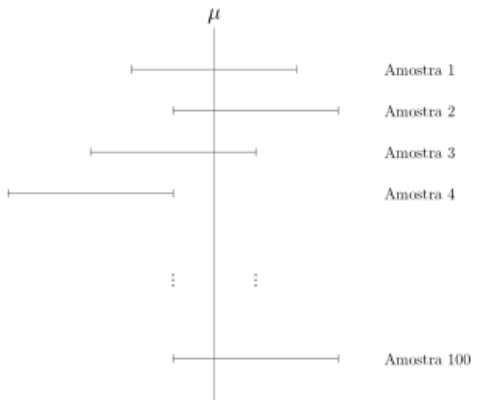


Figura 6: Interpretação do coeficiente de confiança. De 100 intervalos de confiança com coeficiente de confiança γ de diferentes amostras, $100 \cdot \gamma\%$ destes intervalos de confiança conterão a média populacional μ.

3.2 Intervalo de confiança para média: variância desconhecida

É bastante comum não conhecermos a variância (ou desvio padrão) da população. Mas não se preocupe! Se o modelo de probabilidade para a variável aleatória contínua X é a distribuição normal, então o intervalo de confiança para a média da população com coeficiente de confiança γ é

$$IC(\mu, \gamma) = \left(-t_\gamma \frac{s}{\sqrt{n}} + \bar{x}; t_\gamma \frac{s}{\sqrt{n}} + \bar{x}\right),$$

em que $s = \sqrt{\frac{(x_1-\bar{x})^2+(x_2-\bar{x})^2+\cdots+(x_n-\bar{x})^2}{n-1}}$ é o desvio padrão da amostra, n é o tamanho da amostra e $P\left(t_{n-1} \leq t_\gamma\right) = \frac{1+\gamma}{2}$. t_{n-1} é uma variável aleatória contínua com modelo de probabilidade distribuição t-Student com $n-1$ graus de liberdade, cuja função densidade de probabilidade tem formato em sino, como ilustrado na Figura 7.

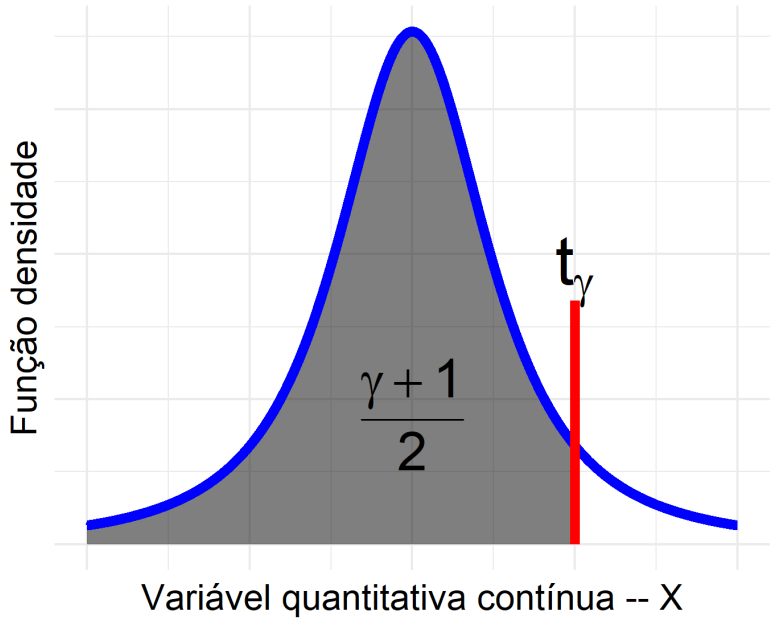


Figura 7: Quantil t_γ para a distribuição t-Student com $n-1$ graus de liberdade. A linha azul é a função densidade da distribuição t-Student com $n-1$ graus de liberdade.

E se não conhecemos a variância (ou desvio padrão) da população e o modelo de probabilidade para a variável aleatória contínua X não é a distribuição normal? Não se preocupe, podemos usar uma técnica denominada *bootstrap*.

A ideia do intervalo de confiança *bootstrap* é simples e resumimos nos passos a seguir:

Passo 1) Retire uma amostra com n indivíduos da população;

Passo 2) Repita o **Passo 1)** para um número grande B, geralmente $B \geq 200$. Para cada amostra, calculamos a média: $\bar{x}_1^*, \bar{x}_2^*, \ldots, \bar{x}_B^*$. Usamos a notação $\bar{x}^*$ para deixar explícito que estamos falando das médias destas B amostra;

Passo 3) Calculamos as diferenças entre as médias das amostras e a média população: $d_1 = \mu - \bar{x}_1^*, d_2 = \mu - \bar{x}_2^*, \ldots, d_B = \mu - \bar{x}_B^*$.

Passo 4) Para um coeficiente de confiança γ $(0 \leq \gamma \leq 1)$, calculamos os quantis para as diferenças: $q(\frac{1-\gamma}{2})$ (quantil de ordem $\frac{1-\gamma}{2}$) e $q(\frac{1+\gamma}{2})$ (quantil de ordem $\frac{1+\gamma}{2}$) das diferenças $d_1, d_2, \ldots, d_B$. Ou seja, $100 \cdot \gamma\%$ das B amostras satisfazem a seguinte desigualdade

$$q\left(\frac{1-\gamma}{2}\right) \leq \mu - \bar{x}^* \leq q\left(\frac{1+\gamma}{2}\right),$$

isto é, $100 \cdot \gamma\%$ das amostras satisfazem a seguinte desigualdade

$$q\left(\frac{1-\gamma}{2}\right) + \bar{x}^* \leq \mu \leq q\left(\frac{1+\gamma}{2}\right) + \bar{x}^*.$$

Se substituirmos $\bar{x}^*$ pela média amostral $\bar{x}$, temos o intervalo de confiança com coeficiente de confiança γ usando *bootstrap* dado por

$$IC(\mu, \gamma) = \left(q\left(\frac{1-\gamma}{2}\right) + \bar{x};\ q\left(\frac{1+\gamma}{2}\right) + \bar{x}\right),$$

em que $q\left(\frac{1-\gamma}{2}\right)$ e $q\left(\frac{1+\gamma}{2}\right)$ são os quantis de ordem $\frac{1-\gamma}{2}$ e $\frac{1+\gamma}{2}$, respectivamente, das diferenças entre a média da população e as médias das B amostrais: $d_1 = \mu - \bar{x}_1^*, d_2 = \mu - \bar{x}_2^*, \ldots, d_B = \mu - \bar{x}_B^*$.

Um pouco confuso? Vamos resolver suas dúvidas construindo um intervalo de confiança para a média populacional da variável Notas da população de 30 alunos de Heloísa. Para exemplificar, vamos pegar um número pequeno de amostras: vamos coletar dez amostras com cinco alunos e calcular a média para cada uma das amostras, como mostramos na Tabela 3.

Tabela 3: Dez amostras com cinco alunos e suas respectivas médias.

Amostras	Valores da amostra					Média $\bar{x}^*$	Diferença entre média populacional e amostral $d = \mu - \bar{x}^*$
Amostras *bootstrap*	7,19	8,82	5,54	6,70	6,11	6,872	-0,499
	7,36	4,84	6,31	6,27	7,29	6,414	-0,041
	5,77	6,57	6,44	8,16	7,93	6,974	-0,601
	6,44	7,29	5,77	3,37	6,11	5,796	0,577
	6,14	7,36	3,37	6,70	7,10	6,134	0,239
	8,82	6,31	7,36	5,77	5,72	6,796	-0,423
	7,10	7,93	4,84	6,44	5,93	6,448	-0,075
	7,10	5,72	7,36	5,77	4,84	6,158	0,215
	6,44	6,14	7,64	6,08	5,70	6,400	-0,027
	6,00	7,77	5,53	5,24	7,15	6,338	0,035
Amostra original	7,77	5,54	6,08	6,27	6,31	6,394	--

Na tabela 3, temos dez amostras com cinco alunos e a média para cada uma das amostras. Vamos construir um intervalo de confiança com o coeficiente de confiança $\gamma = 95\%$ usando *bootstrap*. O primeiro passo é encontrar os quantis de ordem $\frac{1-\gamma}{2}$ e $\frac{1+\gamma}{2}$ das diferenças entre a média populacional e as médias amostras $d = \mu - \bar{x}^*$: $q\left(\frac{1-\gamma}{2}\right) = -0{,}58$ e $q\left(\frac{1+\gamma}{2}\right) = 0{,}5$, então o intervalo de confiança com o coeficiente de confiança é dado por

$$IC(\mu; 95\%) = (-0{,}58 + 6{,}394; 0{,}5 + 6{,}394) = (5{,}814; 6{,}894).$$

Lembre que a média da população é 6,373, então o intervalo de confiança $IC(\mu; 95\%) = (5{,}814; 6{,}894)$ contém a média populacional.

Precisamos reforçar dois pontos do intervalo de confiança construído usando *bootstrap*. O primeiro ponto é: geralmente não conhecemos a média populacional μ e substituímos a média da população pela média da amostra pra calcular as diferenças $d_{1,}$ $d_{2,}, \ldots, d_{B,}$. O segundo ponto é: nem sempre é simples coletar várias amostras de uma população. Mas não se preocupe, pois é possível simular esse processo de coletar várias amostras de uma população usando uma única amostra. Em nome da simplicidade e concisão, não vamos descrever como você pode simular a coleta de várias amostras de uma população usando uma única amostra, mas você pode consultar EFRON e TBSHIRANI (1994) para detalhes.

4 Teste de Hipóteses

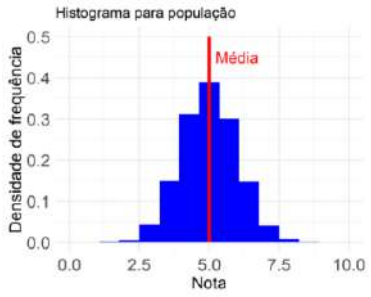
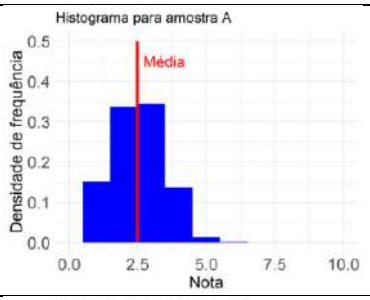
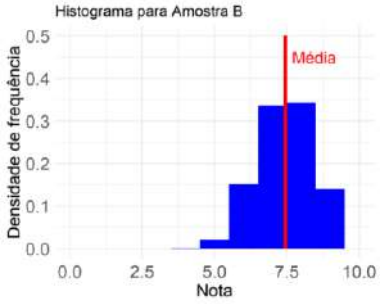
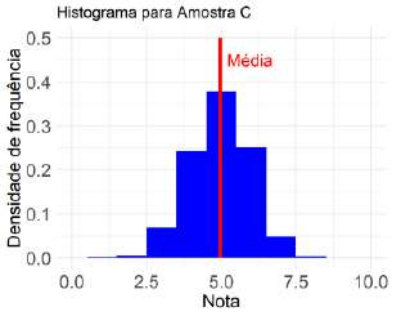
Nessa seção, vamos explicar a você o conceito de teste de hipóteses e como usá-lo. Em um teste de hipótese, estabelecemos duas hipóteses científicas complementares H_0 e H_1 e desejamos tomar a melhor decisão entre as hipóteses H_0 e H_1 usando informações da amostra. Em estatística, chamamos H_0 de hipótese nula, chamamos H_1 de hipótese alternativa e H_1 é a negação de H_0.

O objetivo é fazer afirmações sobre a população, usando a amostra. Vamos ilustrar a ideia do teste de hipótese com um exemplo. Imagine que o ambiente virtual de aprendizagem de Heloísa, a aluna de pós-graduação na área de Informática em Educação da seção ***Era uma vez...***, tem 10.000 alunos. Heloísa implementou algumas melhorias e deseja verificar se o desempenho dos alunos melhorou, e ela estabeleceu duas hipóteses

H_0: *A nota média dos alunos se manteve a mesma,*
H_1: *A nota média dos alunos foi alterada,*

ou seja, $H_0: \mu = 5$ e $H_1: \mu \neq 5$. Heloísa precisa decidir entre H_0 e H_1, e coletou três amostras com 1.000 alunos: amostra A, amostra B e amostra C. A amostra A tem média $2{,}5$, a amostra B tem média 7,5 e amostra C tem média 5. Com a amostra A, Heloísa afirmaria que a média da população é menor que 5 (pois a média da amostra A $(2{,}5)$ menor que 5) e Heloísa decidiria por H_1. Com a amostra B, Heloísa afirmaria que a média da população é maior que 5 (pois a média da amostra B $(7{,}5)$ maior que 5) e Heloísa decidiria por H_1. Com a amostra C, Heloísa afirmaria que a média da população se manteve a mesma (pois a média da amostra C (5) é aproximadamente igual a 5) e Heloísa decidiria por H_0. No quadro 1, ilustramos as decisões que Heloísa tomou para cada amostra e mostramos o histograma da população e das três amostras com uma linha vermelha indicando a média.

Quadro 1: Histograma para a população e para três amostras e a decisão de Heloísa para cada amostra.

População	Amostras	Média	Decisão de Heloísa
		2,5	Média da população é menor que 5, ou seja, rejeitamos H_0.
		7,5	Média da população é menor que 7,5, ou seja, rejeitamos H_0.
		5	Média da população é igual a 5, ou seja, não rejeitamos H_0.

Observe que Heloísa coletou três amostras diferentes (amostra A, amostra B e amostra C), e tomou decisões diferentes para cada amostra. Então, em um teste de hipóteses, existe a possibilidade de Heloísa estar errada. Ao decidir, você pode errar de duas maneiras:

- Erro tipo I: você decidiu por H_1, mas a hipótese H_0 é a verdadeira;
- Erro tipo II: você decidiu por H_0, mas a hipótese H_1 é a verdadeira.

No quadro 2, sintetizamos os tipos de erros que você pode cometer ao decidir entre duas hipóteses complementares H_0 e H_1.

Quadro 2: Tipos de erros: Erro tipo I e Erro tipo II.

	A verdade.		
A decisão que você tomou.		H_0	H_1
	H_0	Você acertou!	Erro Tipo II
	H_1	Erro tipo I	Você acertou!

O ideal é tomar uma decisão que minimiza ao mesmo tempo a probabilidade de cometer erro tipo I e erro tipo II. Usamos a seguinte notação matemática para a probabilidade do erro tipo I e para a probabilidade do erro tipo II:

- $\alpha = P(Erro\ Tipo\ I)$. Geralmente consideramos o erro tipo I mais grave e chamamos α de nível de significância;
- $\beta = P(Erro\ tipo\ II)$.

Infelizmente, é impossível tomar uma decisão que minimiza simultaneamente as probabilidades do erro tipo I (α) e a probabilidade do erro tipo II (β), conforme ilustrado na figura 8. A estratégia é fixar a probabilidade do erro tipo I e tomar a decisão que minimiza a probabilidade do erro tipo II.

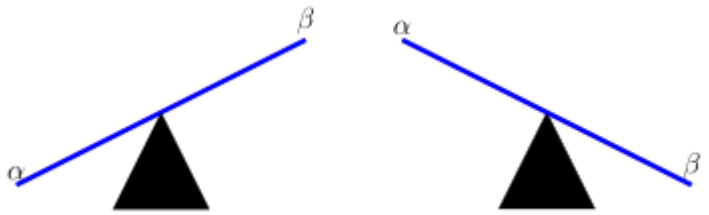


Figura 8: Interpretação das probabilidades do erro tipo I e tipo II. Não é possível minimizar estas duas probabilidades simultaneamente.

Para encontrar a decisão que minimiza $P(Erro\ tipo\ II)$ quando $P(Erro\ tipo\ I) = \alpha$, usamos o procedimento de Neyman-Pearson que pode ser resumido em cinco passos:

Passo 1) Estabeleça as duas hipóteses nula (H_0) e alternativa (H_1).

Passo 2) Estabeleça o nível de significância α (geralmente menor que 5%).

Passo 3) Encontre a "ideia da decisão".

Passo 4) Encontre ou calcule a "ideia da decisão" do passo 3. Este passo envolve encontrar o valor crítico. Não se preocupe em saber o que é valor crítico, iremos explicar esse conceito pra você na próxima seção.

Passo 5) Decida entre H_0 e H_1.

Por razões históricas em Estatística, ao decidirmos por H_0 falamos "não rejeitamos H_0", e ao decidirmos por H_1 dizemos que "rejeitamos H_0".

Agora vamos apresentar alguns testes de hipóteses que você provavelmente usará como pesquisador na área de Informática na Educação.

4.1 Teste bilateral para média populacional

Suponha que você coletou uma amostra com *n* indivíduos da variável X, e imagine que você deseja checar se a média populacional de X é μ_0. Em um primeiro momento, vamos supor que você conhece o desvio padrão (ou variância) populacional σ da variável X. Vamos construir o teste de hipóteses usando o procedimento de Neyman-Pearson:

Passo 1) O primeiro passo é estabelecer as hipóteses nula e alternativa. Desejamos decidir entre as duas hipóteses $H_0: \mu = \mu_0$ e $H_1: \mu \neq \mu_0$.

Passo 2) Vamos fixar um nível de significância α. Fixamos um valor pequeno para o nível de significância. Na Informática na Educação, geralmente usamos 5%.

Passo 3) Neste momento vamos estabelecer a "ideia de decisão". Suponha que a média $\bar{X}$ é uma aproximação de μ_0. Decidimos por H_0 se $\bar{X}$ for aproximadamente μ_0 e decidimos por H_1 se $\bar{X}$ for diferente μ_0. Resumindo:

- Se $\bar{X}$ está perto de μ_0, decidimos por H_0. Em outras palavras, se $|\bar{X} - \mu_0| < x_c$, não rejeitamos $H_0: \mu = \mu_0$;
- Se $\bar{X}$ está longe de μ_0, decidimos por H_1. Em outras palavras, se $|\bar{X} - \mu_0| \geq x_c$, rejeitamos $H_0: \mu = \mu_0$;

Mais precisamente, decidimos por H_1 se $\bar{X}$ está no conjunto $\{\bar{x}\ tal\ que\ |\bar{x} - \mu_0| \geq x_c\}$. Em estatística, chamamos o valor x_c de valor crítico e o conjunto $RC = \{\bar{x}\ tal\ que\ |\bar{x} - \mu_0| \geq x_c\}$ é chamado de Região Crítica.

Passo 4) O valor crítico x_c é a solução da seguinte equação

$$\alpha = P\left(|Z| \geq \frac{x_c\sqrt{n}}{\sigma}\right),$$

em que Z é uma variável aleatória contínua com distribuição normal padrão. Para detalhes, consulte CASELLA e BERBER (2002).

Passo 5) Estamos prontos para tomar uma decisão usando a seguinte regra de decisão:

- Se $|\bar{X} - \mu_0| < x_c$, não rejeitamos H_0 (decidimos por H_0);
- Se $|\bar{X} - \mu_0| \geq x_c$, rejeitamos por H_0 (decidimos por H_1).

Observe que se você mudar o nível de significância α, o valor crítico x_c será diferente e sua decisão pode ser diferente.

Um procedimento ligeiramente diferente é calcular uma medida de evidência. Note que o procedimento de Neyman-Pearson sempre produz uma decisão dicotômica (ou H_0 ou H_1), e essas decisões dependem do nível de significância α (níveis de significância diferentes podem induzir decisões diferentes). De uma forma diferente, medidas de evidência resultam em valores entre zero e um, valores perto de zero indicam que a hipótese H_1 é verdadeiro e valores perto de 1 indicam que a hipótese H_0 é verdadeira, ou seja, com medidas de evidência temos valores entre zero e um em vez de uma decisão dicotômica. Ilustramos essa ideia na Figura 9.

A medida de evidência que consideramos nesse capítulo é o p-valor. Note que o p-valor é sem dúvida a medida de evidência mais usada em Estatística, mas não é a única (o s-valor proposto por PATRIOTA (2013) e o e-valor proposto por DE BRAGANÇA PEREIRA e STERN (1999) são exemplo de outras medidas de evidências).

Em nome da brevidade e da clareza, não vamos dar detalhes de como calcular o p-valor neste capítulo. Caso você tenha interesse, você pode consultar LEHMANN e ROMANO (2006) para detalhes em como calcular o p-valor.

Usamos e interpretamos o p-valor da seguinte forma:

- Se p for grande, a amostra **não** tem indícios suficientes para rejeitar H_0 e você deveria decidir por H_0;
- Se p for pequeno, a amostra tem indícios suficientes para rejeitar H_0 e você deveria decidir por H_1

Você deve estar se perguntando o que é um valor pequeno e o que é um valor grande para o p-valor. A ideia é comparar o nível de significância α com o p-valor p. Então, se $p < \alpha$ rejeitamos H_0, se $p \geq \alpha$ não rejeitamos H_0. Ilustramos o uso do p-valor na Figura 10.

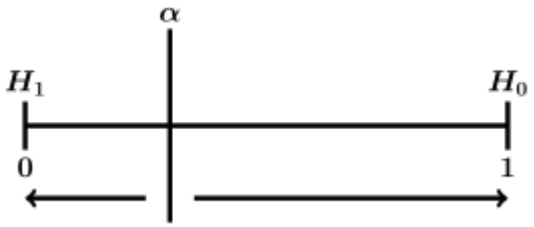


Figura 9: Uso prático p-valor. Se o p-valor estiver abaixo do nível de significância α, rejeito H_0 (decido por H_1). Se o p-valor estiver acima do nível de significância α, não rejeito por H_0 (decido por H_0).

É bastante comum não conhecermos o desvio padrão (ou a variância) da população. Se a variável aleatória X tem distribuição normal, ainda podemos construir um teste de hipóteses. Os passos **Passo 1)**, **Passo 2)** e **Passo 3)** permanecem inalterados.

Passo 1) Imagine que você não conhece o desvio padrão (ou variância) da população e sabe que a variável aleatória X tem distribuição normal. Nesse contexto, o valor crítico é a solução da seguinte equação

$$\alpha = P\left(|T| \geq \frac{x_c\sqrt{n}}{S}\right),$$

em que T é uma variável aleatória com distribuição t-Student com $n-1$ graus de liberdade. Para detalhes de como resolver a equação acima, consulte CASELLA e BERGER (2002).

Passo 5) Estamos prontos para tomar uma decisão usando a seguinte regra de decisão:

- Se $|\bar{X} - \mu_0| < x_c$, não rejeitamos H_0 (decidimos por H_0);
- Se $|\bar{X} - \mu_0| \geq x_c$, rejeitamos por H_0 (decidimos por H_1).

Também podemos calcular o p-valor p, e tomamos a decisão segundo a seguinte regra:

- Se $p < \alpha$, então rejeitamos H_0;
- Se $p \geq \alpha$, então não rejeitamos H_0.

4.1.1 Teste unilateral para média populacional

Imagine que você tem uma variável X e que você deseja checar se a média de X é no máximo (ou no mínimo) μ_0. Em um primeiro momento, suponha que você conhece o desvio padrão (ou a variância) da variável X na sua população.

Nesta seção, vamos desenvolver o teste de hipótese para checar se a média populacional de X é no máximo μ_0, pois o procedimento para testar se a média de X é no mínimo μ_0 é bastante semelhante. Primeiro vamos construir o teste de hipóteses usando o procedimento de Neyman-Pearson e, em seguida, explicamos como usar o p-valor.

Passo 1) Desejamos verificar se a média da população é no máximo μ_0, então temos duas hipóteses:
- H_0 : a média da população é no máximo μ_0, isto é, $H_0: \mu \leq \mu_0$;
- H_0 : a média da população é maior do que μ_0, isto é, $H_0: \mu > \mu_0$.

Passo 2) Nesse passo, você fixa o nível de significância (geralmente menor que 5%);

Passo 3) Agora está na hora de encontrar a "ideia da decisão", e para decidir vamos usar a estimativa da média da população: $\bar{x}$.
- Se $\bar{x}$ é muito menor que μ_0, não rejeitamos por H_0 (decidimos por H_0), ou seja, se $\bar{x} - \mu_0 \leq x_c$;
- Se $\bar{x}$ é muito maior que μ_0, rejeitamos por H_0 (decidimos por H_1), ou seja, se $\bar{x} - \mu_0 > x_c$.

Então a região crítica é $RC = \{\bar{x} | \bar{x} - \mu_0 > x_c\}$.

Passo 4) O valor crítico x_c é a solução da seguinte equação

$$P\left(Z > \frac{x_c\sqrt{n}}{\sigma}\right) = \alpha,$$

em que Z é uma variável aleatória com distribuição normal padrão. Para detalhes, você pode consultar CASELLA e BERGER (2002).

Passo 5) Finalmente estamos prontos para tomar a decisão entre H_0 e H_1, usando a região crítica. Primeiro calcular a média $\bar{x}_{obs}$ dos valores observados $x_1, x_2, \ldots, x_n$, da variável X, e em seguida checamos se $\bar{x}_{obs}$ está na região crítica
- Se $\bar{x}_{obs}$ está na região crítica $RC = \{\bar{x} | \bar{x} - \mu_0 > x_c\}$, rejeitamos H_0 (decidimos por H_1);
- Se $\bar{x}_{obs}$ está na região crítica $RC = \{\bar{x} | \bar{x} - \mu_0 \leq x_c\}$, não rejeitamos H_0 (decidimos por H_0).

Você também pode usar o p-valor p para decidir entre H_0 e H_1:

- Se p for grande, não temos evidência para rejeitar $H_0: \mu \leq \mu_0$, ou seja, não rejeitamos H_0 (decidimos por H_0). Na prática, se $p \geq \alpha$ não rejeitamos H_0;
- Se p for pequeno, temos evidência para rejeitar $H_0: \mu \leq \mu_0$, ou seja, rejeitamos H_0 (decidimos por H_1). Na prática, se $p < \alpha$ rejeitamos H_0.

É bastante comum não conhecermos o desvio padrão (ou a variância) da população. Se a variável aleatória X tem distribuição normal, ainda podemos construir um teste de hipóteses. Os passos **Passo 1)**, **Passo 2)** e **Passo 3)** permanecem inalterados.

Passo 4) I magine que você não conhece o desvio padrão (ou variância) da população e sabe que a variável aleatória X tem distribuição normal. Nesse contexto, o valor crítico é a solução da seguinte equação

$$\alpha = P\left(T \geq \frac{x_c\sqrt{n}}{S}\right),$$

m que T é uma variável aleatória com distribuição t-Student com $n-1$ graus de liberdade. Para detalhes de como resolver a equação acima, consulte CASELLA e BERGER (2002).

Passo 5) Estamos prontos para tomar uma decisão usando a seguinte regra de decisão:

- S e $|\bar{X} - \mu_0| < x_c$, não rejeitamos H_0 (decidimos por H_0);
- Se $|\bar{X} - \mu_0| \geq x_c$, rejeitamos por H_0 (decidimos por H_1).

Você também pode usar o p-valor p para decidir entre H_0 e H_1:

- Se p for grande, não temos evidência para rejeitar $H_0: \mu \leq \mu_0$, ou seja, não rejeitamos H_0 (decidimos por H_0). Na prática, se $p \geq \alpha$ não rejeitamos H_0;
- Se p for pequeno, temos evidência para rejeitar $H_0: \mu \leq \mu_0$, ou seja, rejeitamos H_0 (decidimos por H_1). Na prática, se $p < \alpha$ rejeitamos H_0.

4.2 Teste de independência entre duas variáveis qualitativas

Nessa seção, temos interesse em verificar se duas variáveis qualitativas estão associadas. Por exemplo, durante o século XX os profissionais de saúde estavam preocupados em estudar a associação entre o hábito de fumar (tabagismo) e câncer.

Antes de estabelecer as hipóteses científicas, vamos estabelecer o que entendemos por associação entre duas variáveis qualitativas. Para facilitar, imagine que Heloísa, a aluna de pós-graduação na área de informática na educação da seção **Era uma vez..**, leciona em uma grande escola com 2000 alunos. Heloísa conseguiu na secretaria o desempenho dos alunos em Música e em Matemática conforme tabela 4. Se Heloísa

selecionar um aluno chamado João, este aluno terá baixo desempenho em Música com probabilidade $\frac{244}{2000} \cdot 100\% = 12{,}2\%$. Um tempo depois Heloísa consultou o histórico escolar de João e descobriu que ele teve um alto desempenho em Matemática, então a probabilidade do João ter baixo desempenho em Música diminui para $\frac{1}{218} \cdot 100\% = 0{,}45\%$. Ou seja, o conhecimento do desempenho em Matemática do João alterou a probabilidade de João ter baixo desempenho em Música. De uma forma geral, dizemos que duas variáveis qualitativas X e Y estão associadas, se o conhecimento do valor da variável Y para um indivíduo altera as plausibilidades dos valores de X para este indivíduo.

Tabela 4: Desempenho em Matemática e desempenho em Música para 2000 alunos da escola de Heloísa.

	Música			
Matemática	Baixo	Médio	Alto	Total
Baixo	88	167	3	258
Médio	155	1215	154	1524
Alto	1	146	71	218
Total	244	1528	228	2000

Nesta seção, queremos decidir entre hipóteses $H_0: X\ e\ Y\ \boldsymbol{n}\text{ã}\boldsymbol{o}\ são\ associadas$ e $H_1: X\ e\ Y\ são\ associadas$. De novo, vamos construir um teste de hipóteses usando os cincos passos do procedimento de Neyman-Pearson, e em seguida explicaremos como usar o p-valor.

Passo 1) A primeira coisa que você precisa fazer é estabelecer sua hipótese científica. No estudo da associação entre duas variáveis qualitativas X e Y, temos as seguintes hipóteses:

$$H_0: X\ e\ Y\ \textbf{não}\ estão\ associadas$$
$$H_1: X\ e\ Y\ estão\ associadas$$

Passo 2) Nesse passo, você fixa o seu nível de significância (geralmente menor ou igual a 5%).

Passo 3) Nesse passo, vamos entender qual é a ideia da decisão. Para facilitar o entendimento, imagine que você tem duas variáveis qualitativas:

$$X: com\ valores\ possíveis\ A_1, A_2, A_3,$$
$$Y: com\ valores\ possíveis\ B_1, B_2, B_3.$$

Considere uma tabela conjunta de distribuição de frequência para X e Y como a tabela 5.

Tabela 5: Tabela conjunta de distribuição de frequência conjunta de X e Y.

X	Y			Total
	A_1	A_1	A_1	
A_1	n_{11}	n_{12}	n_{13}	$n_{1.} = n_{11} + n_{12} + n_{13}$
A_2	n_{21}	n_{22}	n_{23}	$n_{2.} = n_{21} + n_{22} + n_{23}$
A_3	n_{31}	n_{32}	n_{33}	$n_{3.} = n_{31} + n_{32} + n_{33}$
Total	$n_{.1}$=$n_{11} + n_{21} + n_{31}$	$n_{.1}$=$n_{11} + n_{21} + n_{31}$	$n_{.1}$=$n_{11} + n_{21} + n_{31}$	$n_{..} = n_{1.} + n_{2.} + n_{3.}$ $n_{..} = n_{.1} + n_{.2} + n_{.3}$

Se H_0 é verdade, ou seja, se X e Y **não** estão associadas, então temos a seguinte propriedade mostrada na tabela 6: o número de indivíduos com $X = A_1$ e $Y = B_1$ é igual ao número de indivíduos com $X = A_1$ multiplicado pelo número de indivíduos com $Y = B_1$ dividido pelo tamanho amostral. Para uma elucidação desta propriedade, você consultar BARBETTA (2008).

Tabela 6: Propriedade importante sob a hipótese de não associação entre X e Y. É comum chamarmos os valores desta tabela de distribuição de frequência esperada.

X	Y			Total
	B_1	B_2	B_3	
A_1	$n_{11}^* = \frac{n_{1.} \times n_{.1}}{n_{..}}$	$n_{12}^* = \frac{n_{1.} \times n_{.3}}{n_{..}}$	$n_{13}^* = \frac{n_{1.} \times n_{.3}}{n_{..}}$	$n_{1.}$
A_2	$n_{21}^* = \frac{n_{2.} \times n_{.1}}{n_{..}}$	$n_{22}^* = \frac{n_{2.} \times n_{.2}}{n_{..}}$	$n_{23}^* = \frac{n_{2.} \times n_{.3}}{n_{..}}$	$n_{2.}$
A_3	$n_{31}^* = \frac{n_{3.} \times n_{.1}}{n_{..}}$	$n_{32}^* = \frac{n_{3.} \times n_{.2}}{n_{..}}$	$n_{33}^* = \frac{n_{3.} \times n_{.3}}{n_{..}}$	$n_{3.}$
Total	$n_{.1}$	$n_{.2}$	$n_{.3}$	$n_{..}$

Então, temos a seguinte ideia de decisão: Se a tabela 5 (tabela conjunta de distribuição de frequência) e a tabela 6 (tabela conjunta de distribuição de frequência esperada) tiverem valores bem próximos, então decidimos por H_0. Ou seja,

- Se a tabela 5 (tabela de distribuição de frequência) e tabela 6 (tabela de distribuição de frequência esperada) tiverem valores bem próximo, não rejeitamos H_0 (decidimos por H_0);
- Se a tabela 5 (tabela de distribuição de frequência conjunta) e tabela 6 (tabela de distribuição de frequência esperada) tiverem valores diferentes, rejeitamos H_0(decidimos por H_1);

Em vez de você olhar célula-a-célula na tabela 5 e na tabela 6, você pode usar uma medida de distância entre tabelas denominada qui-quadrado calculada pela seguinte equação:

$$\chi^2 = \frac{(n_{11} - n_{11}^*)^2}{n_{11}^*} + \frac{(n_{12} - n_{12}^*)^2}{n_{12}^*} + \frac{(n_{13} - n_{13}^*)^2}{n_{13}^*} + \frac{(n_{21} - n_{21}^*)^2}{n_{21}^*} + \frac{(n_{22} - n_{22}^*)^2}{n_{22}^*} + \frac{(n_{23} - n_{23}^*)^2}{n_{23}^*} +$$

$$\frac{(n_{31}-n_{31}^*)^2}{n_{31}^*}+\frac{(n_{32}-n_{32}^*)^2}{n_{32}^*}+\frac{(n_{33}-n_{33}^*)^2}{n_{33}^*}.$$

Então, temos que a seguinte regra de decisão

- Se χ^2 for pequeno, decidimos por H_0. Ou seja, se $\chi^2 < x_c$, não rejeitamos H_0;
- Se χ^2 for grande, decidimos por H_1. Ou seja, se $\chi^2 \geq x_c$, rejeitamos H_0.

Passo 4) O valor crítico x_c é a solução da seguinte equação

$$\alpha = 1 - P(\chi^2 < x_c),$$

em que χ^2 é uma variável aleatória contínua com distribuição qui-quadrado com $l = (3-1)\cdot(3-1)$ graus de liberdade. Para detalhes sobre a derivação da equação acima e sobre a distribuição qui-quadrado consulte CASELLA e BERGER (2002).

Passo 5) Finalmente estamos prontos para tomar a decisão entre H_0 e H_1, usando a região crítica. Primeiro calculamos χ^2 e em seguida checamos se

- Se χ^2 está na região crítica decidimos por H_1. Ou seja, se $\chi^2 \geq x_c$, rejeitamos H_0;
- Se χ^2 não está na região crítica decidimos por H_0. Ou seja, se $\chi^2 < x_c$, não rejeitamos H_0.

De forma semelhante aos testes de hipóteses que construímos para média, podemos calcular o p-valor p e decidimos segundo a seguinte regra

- Se $p \geq \alpha$, não rejeitamos H_0;
- Se $p < \alpha$, rejeitamos H_0.

4.3 Teste de independência entre duas variáveis quantitativas

Antes de construir o teste de hipóteses, vamos explicar o que entendemos por associação entre duas variáveis quantitativas com um exemplo. Imagine que Heloísa, a aluna de pós-graduação na área de Informática na Educação na seção **Era uma vez...**, lecionou para uma turma de 15 alunos as disciplinas Estatística Básica I e Estatística Básica II e as notas finais nesses dois cursos está na tabela 7. Podemos construir o gráfico de dispersão entre as variáveis: X – nota final em Estatística Básica I e Y – nota final em Estatística Básica II, conforme ilustrado na Figura 10. Neste gráfico de dispersão, notamos que alunos com notas maiores em Estatística Básica I tendem a ter com notas maiores em Estatística Básica II, ou seja, quando o valor de X aumenta o valor de Y tende a aumentar. De uma forma geral, dizemos que duas variáveis quantitativas X e Y estão positivamente associadas se indivíduos com maiores valores de X tendem a ter valores maiores em Y, e X e Y estão negativamente associadas se maiores de X tendem a ter valores menores de Y.

Tabela 7: Nota para a turma de dez alunos com notas finais em Estatística Básica I (X) e em Estatística Básica II (Y).

X	2,7	5,1	5,1	3,7	7,4	6,5	4,9	5,2	5,0	6,3	3,9	5,6	5,2	5,3	4,4
Y	2,9	7,0	5,1	5,6	6,1	6,1	4,7	5,6	4,5	5,8	3,5	5,4	5,1	5,8	5,3

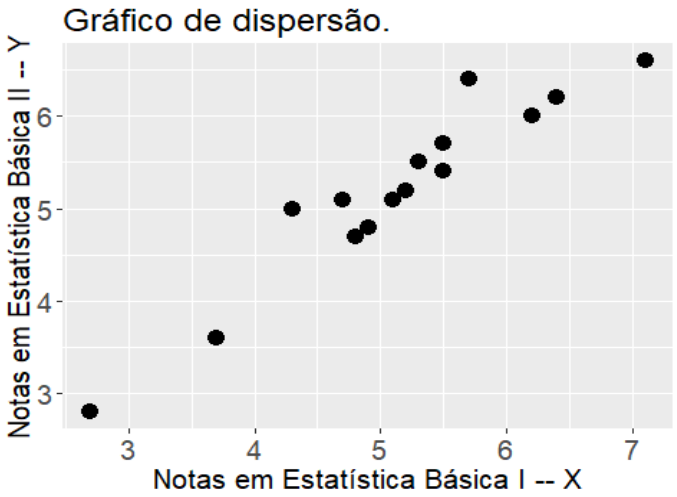


Figura 10: Gráfico de dispersão entre as notas em Estatística Básica I e as notas em Estatística Básica II.

Agora imagine que você tem duas variáveis quantitativas, X e Y, e você deseja testar se X e Y estão associadas. Você pode estabelecer os cinco passos do procedimento de Neyman-Pearson:

Passo 1) Vamos estabelecer as hipóteses. O objetivo é testar se existe associação entre as duas variáveis quantitativas e então temos as seguintes hipóteses:

$$H_0: X \text{ e } Y \textbf{ não } \text{estão associadas},$$
$$H_1: X \text{ e } Y \text{ estão associadas}.$$

Considere a correlação linear de Pearson entre duas variáveis quantitativas X e Y. Então, as hipóteses H_0 e H_1 podem ser reescritas nas seguintes hipóteses

$$H_0: \rho = 0,$$
$$H_1: \rho \neq 0,$$

em que ρ é o coeficiente de correlação linear de Pearson entre X e Y na população. Para uma descrição e detalhes sobre o coeficiente de correlação linear de Pearson consulte BARBETTA (2008).

Passo 2) Estabeleça o nível significância α (geralmente menor que 5%).

Passo 3) Vamos agora estabelecer a ideia da decisão entre H_0 e H_1.

Primeiro calculamos o coeficiente de correlação linear de Pearson r entre X e Y.

- Se r está perto de zero decidimos por H_0, ou seja, se $|r| \leq x_c$ não rejeitamos H_0
- Se r está longe de zero decidimos por H_1, ou seja, se $|r| > x_c$ rejeitamos por H_0.

Geralmente, usamos uma padronização do coeficiente de correlação linear de Pearson, obtendo a estatística de teste $t^* = \frac{r\sqrt{n-2}}{\sqrt{1-r^2}}$. Então, usamos a seguinte regra de decisão:

- se $|t^*| \leq s_c$, não rejeitamos H_0 (decidimos por H_0);
- se $|t^*| > s_c$, rejeitamos por H_0 (decidimos por H_1).

E a região crítica é $RC = \{t^* tal\ que\ |t^*| > s_c\}$.

Passo 4) Agora podemos encontrar o valor crítico resolvendo a equação a seguir:

$$P(\ t_{n-2} \leq s_c) = 1 - \frac{\alpha}{2},$$

em que t_{n-2} é uma variável aleatória com distribuição t-Student com $n-2$ graus de liberdade. Para detalhes, consulte BARBETTA (2008).

Passo 5) Agora estamos prontos para decidir. Primeiro você calcula o valor de t^*, e então

- se t^* está dentro da região crítica *RC,* ou seja, se $|t^*| > s_c$, então rejeitamos H_0 (decidimos por H_1);
- se t^* está fora da região crítica *RC,* ou seja, se $|t^*| \leq s_c$, então não rejeitamos H_0 (decidimos por H_0).

Alternativamente, podemos calcular o p-valor usando a equação

$$p = 2\big(1 - P(T \leq t^*)\big),$$

em que *T* é uma variável aleatória com distribuição t-Student com *n-2* graus de liberdade. Para detalhes em como calcular o p-valor consulte CASELLA e BERGER (2002). O importante é a interpretação e o uso da medida de evidência p-valor p:

- se $p \geq \alpha$, não rejeitamos H_0 (decidimos por H_0);
- se $p < \alpha$, rejeitamos H_1 (decidimos por H_1).

4.4 Testes de normalidade

Agora apresentamos a você três abordagens usando teste de hipóteses para checar se a distribuição normal é um modelo de probabilidade adequado para uma variável aleatória X. Em todos os testes de hipóteses desta seção, temos as mesmas hipóteses nula e alternativa:

H_0: *A distribuição normal* é *adequada para a variável X,*
H_1: *A distribuição normal não* é *adequada para a variável X.*

Vamos apresentar o teste de Shapiro-Wilks, Anderson-Darlling e Kolmogorov-Smirnov. Estes testes de hipóteses têm modelos de probabilidade que estão fora do escopo deste capítulo e vamos focar no uso e interpretação do p-valor. Caso você tenha interesse, vamos deixar referência desses três testes de Normalidade ao final deste capítulo.

O primeiro teste é denominado de Shapiro-Wilks, pois Shapiro e Wilks propuseram este teste de hipótese em um artigo de 1965. Se for para escolher um teste de hipótese para checar a normalidade de uma variável *X*, os estatísticos aconselham a usar o teste de hipóteses proposto por Shapiro e Wilks. A ideia de decisão (**Passo 3)**) neste teste é simples: calcule as estatísticas de ordem da variável X e as estatísticas de ordem se a hipótese nula é verdadeira. Geralmente, chamamos as estatísticas de ordem de uma variável quantitativa X de *estatísticas de ordem amostrais*, e as estatísticas de ordem calculados assumindo que a hipótese nula é verdadeira de *estatísticas de ordem teóricas*. Então se as *estatísticas de ordem amostrais* e *estatísticas de ordem teóricas* estiverem próximas não rejeitamos H_0 e se estiverem afastadas rejeitamos H_0.Adicionalmente, você pode checar visualmente a normalidade dos dados usando o gráfico de dispersão entre *os quantis amostrais* e *quantis teóricos*. Chamamos este gráfico de gráfico Quantil-Quantil.

Quadro 3: Testes de hipóteses para checar se uma variável quantitativa X tem distribuição normal.

Hipóteses	Nome do teste	Pacote no R	Função
$H_0: X\ tem\ distribuição\ normal$ $H_1: X\ não\ tem\ distribuição\ normal$	Shapiro-Wilks	Não precisa de pacote	`shapiro.test`
	Anderson-Darling	`nortest`	`ad.test`
	Kolmogorov-Smirnov	`nortest`	`lillie.test`

Vamos agora apresentar o teste de Anderson-Darling e o teste de Kolmogorov-Smirnov. Ambos testes de hipóteses usam probabilidades no estilo $F(z) = P(Z \leq z)$, em que Z é uma variável aleatória com modelo normal padrão. Chamamos $F(z)$ de função de distribuição acumulada. A ideia de decisão (**Passo 3)**) é construir uma estimativa $\hat{F}(z)$ para a probabilidade $F(z) = P(X \leq z)$ usando os valores observados $x_1, x_2, \ldots, x_n$ da variável quantitativa contínua X. Se H_0 é verdadeira, então a variável aleatória contínua $Z = \frac{X-\mu}{\sigma}$ tem função de distribuição acumulada $F(z)$. Se $\hat{F}(z)$ e F(z) estão próximos não rejeitamos H_0 e se $\hat{F}(z)$ e F(z) estiverem afastados rejeitamos H_0.

No quadro 3, apresentamos as hipóteses nula e alternativas nos testes de Shapiro-Wilks, Anderson-Darling e Kolmogorov-Smirnov, e mostramos qual pacote você deve carregar no R e qual função usar para obter o p-valor.

Antes de passarmos a próxima seção, vamos ilustrar os testes de hipóteses desta seção. Imagine que a Heloísa, a aluna de pós-graduação na área de Informática na Educação, leciona em uma escola com dois mil alunos. Esta escola guarda a altura de todos os alunos matriculados por causa das aulas de educação física. Vamos verificar se a variável altura tem distribuição normal. Na figura 11, mostramos o gráfico quantil-quantil. Neste gráfico quantil-quantil, percebemos que os pontos estão sobre a reta

vermelha indicando a normalidade dos dados. Podemos usar os testes de Shapiro-Wilks, Anderson-Darling e Kolmogorov-Smirnov para testar as hipóteses

$$H_0: A\ distribuição\ normal\ é\ adequada\ para\ a\ variável\ altura,$$
$$H_1: A\ distribuição\ normal\ não\ é\ adequada\ para\ a\ variável\ altura.$$

Na tabela 8, os p-valores para os testes de hipóteses usando os testes de Shapiro-Wilks, Anderson-Darling e Kolmogorov são maiores que 5% e não rejeitamos H_0, ou seja, nos três testes de hipóteses decidimos que a distribuição normal é um modelo de probabilidade adequado para a variável altura.

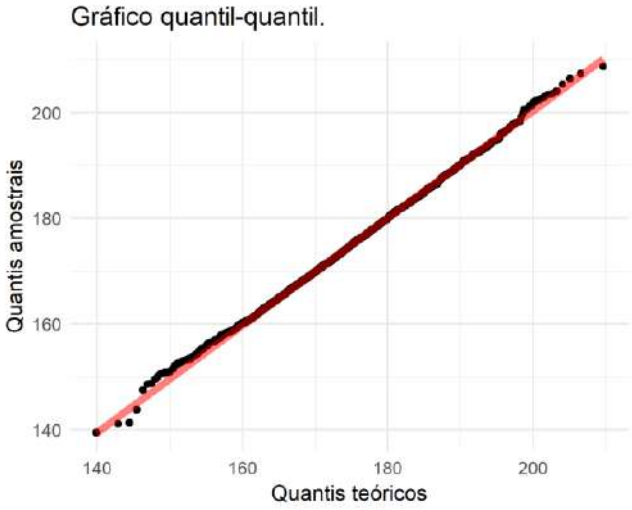


Figura 11: Gráfico quanti-quantil para variável altura.

Tabela 8: p-valor para testar a normalidade: Shapiro-Wilks, Anderson-Darling e Kolmogorov-Smirnov.

Teste de hipóteses	p-valor
Shapiro-Wilks	0,51
Anderson-Darling	0,75
Kolmogorov-Smirnov	0,71

5 Cenário ilustrativo

Nesta seção, vamos usar o que aprendemos para ajudar Heloísa, a aluna de pós-graduação em Informática na Educação da seção **Era uma vez...** Heloísa decidiu acompanhar cem alunos de um ambiente de aprendizagem virtual e coletou as seguintes variáveis para cada um dos alunos:

- ID: rótulo usado para identificar os alunos na ambiente virtual de aprendizagem;
- Nota: nota em matemática;

- Tempo: tempo (em minutos) que o aluno ficou logado no ambiente virtual de aprendizagem na semana;
- Gênero: gênero declarado pelo aluno;
- Localização: variável qualitativa com dois valores possíveis – Capital e Interior. Capital indica que o aluno mora em uma capital ou região metropolitana, e Interior indica que o aluno não mora em uma capital ou região metropolitana.

Estas variáveis foram armazenadas em um arquivo excel denominado `"data_plataforma.xlsx"`. Heloísa tem as seguintes hipóteses:

(1) Nas regiões metropolitanas tem predominância maior de alunos do sexo feminino, ou seja, as variáveis Localização e Gênero estão associadas;
(2) Quanto mais tempo o aluno fica logado na ambiente virtual de aprendizagem, melhor a sua nota em matemática, isto é, as variáveis Tempo e Nota estão positivamente associadas;
(3) Os estudantes estão ficando mais tempo logados na ambiente virtual de aprendizagem, após as melhorias propostas por Heloísa.

Para responder estas hipóteses vamos o R e vamos precisar dos seguintes pacotes: `TeachingDemos`, `nortest`, `tidyverse` e `readxl`. No quadro 4, listamos os pacotes usados que Heloísa vai usar e uma breve descrição.

Quadro 4: Pacotes que Maria irá usar para responder as três hipóteses científicas.

Pacote	Descrição do pacote
`tidyverse`	Oferece algumas ferramentas que podem simplificar e economizar tempo no R.
`readxl`	Lê arquivos excel no R
`TeachingDemos`	Constrói testes de hipóteses e intervalo de confiança quando conhecemos o desvio padrão populacional
`nortest`	Calcula o teste de Anderson-Darling e o teste de Kolmogorov-Smirnov.

Vamos começar carregando os pacotes e arquivo `"data_plataforma.xlsx"`.

```
# Carregando os pacotes
library(tidyverse)
library(TeachingDemos)
library(nortest)
library(readxl)

# lendo o arquivo
dados_heloisa <- read_xlsx("data_plataforma.xlsx", sheet = "dados",
                           col_names = TRUE)

# as cinco primeiras observações
head(dados_heloisa, n = 5)
```

```
## # A tibble: 5 x 5
##      ID  Nota tempo genero    localizacao
##   <dbl> <dbl> <dbl> <chr>     <chr>
## 1 71573   8.2 118.  Feminino  Interior
## 2 88855   6.8  91.2 Feminino  Capital
## 3 52826   6.7  75.6 Feminino  Capital
## 4 14692   7.7 102.  Feminino  Capital
## 5 28539   6.3  81.4 Masculino Interior
```

Vamos primeiro fazer uma análise descritiva das variáveis: Nota, Tempo, Gênero e Localização. Vamos começar com a variável Gênero. A sua primeira tarefa é construir a tabela de distribuição de frequência, conforme código abaixo.

```
# Tabela de distribuição de frequência
dados_heloisa %>% group_by(genero) %>%
  summarise(frequencia = n()) %>%
  mutate(frequencia_relativa = frequencia / sum(frequencia),
         porcentagem = 100 * frequencia_relativa)
```

```
## # A tibble: 3 x 4
##   genero    frequencia frequencia_relativa porcentagem
##   <chr>          <int>               <dbl>       <dbl>
## 1 Feminino          49                0.49          49
## 2 Masculino         47                0.47          47
## 3 Outro              4                0.04           4
```

Podemos representar a tabela de distribuição de frequência usando um gráfico de barras conforme figura 12. No gráfico da figura 12, notamos que existe uma minoria com sexo “outro” e um equilíbrio entre o número de alunos do sexo masculino e do sexo feminino.

```
# Gráfico de barras: Gênero
ggplot(dados_heloisa) +
  geom_bar(aes(x = genero, y = ..prop.., group = 1),
           fill = "blue") +
  xlab("Gênero") + ylab("Frequência relativa") +
  ylim(c(0,1))
```

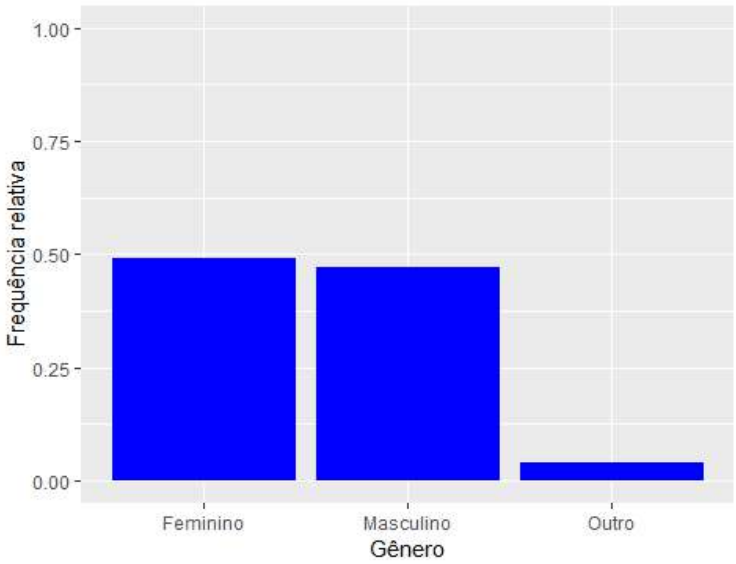


Figura 12: Gráfico de barras para variável Gênero.

Você pode repetir o mesmo processo para a variável localização. Começamos com a tabela de distribuição de frequência.

```
# Tabela de distribuição de frequência: localização
dados_heloisa %>% group_by(localizacao) %>%
  summarise(frequencia = n()) %>%
  mutate(frequencia_relativa = frequencia / sum(frequencia),
         porcentagem = 100 * frequencia_relativa)
```

```
## # A tibble: 2 x 4
##   localizacao frequencia frequencia_relativa porcentagem
##   <chr>            <int>               <dbl>       <dbl>
## 1 Capital             74                0.74          74
## 2 Interior            26                0.26          26
```

Apesar da tabela de distribuição de frequência ser útil, aconselhamos sempre a usar uma representação gráfica. Para variáveis qualitativas, recomendamos o uso do gráfico de barras. Você pode usar o código abaixo, com as devidas alterações, para construir um gráfico de barras no R. Mostramos o gráfico da variável localização na Figura 13 e notamos que a maioria dos estudantes moram em regiões metropolitanas.

```
# Gráfico de barras: localização
ggplot(dados_heloisa)+
  geom_bar(aes(x = localizacao, y = ..prop.., group = 1),
           fill = "blue") +
  xlab("Localização") + ylab("Frequência relativa") +
  ylim(c(0,1))
```

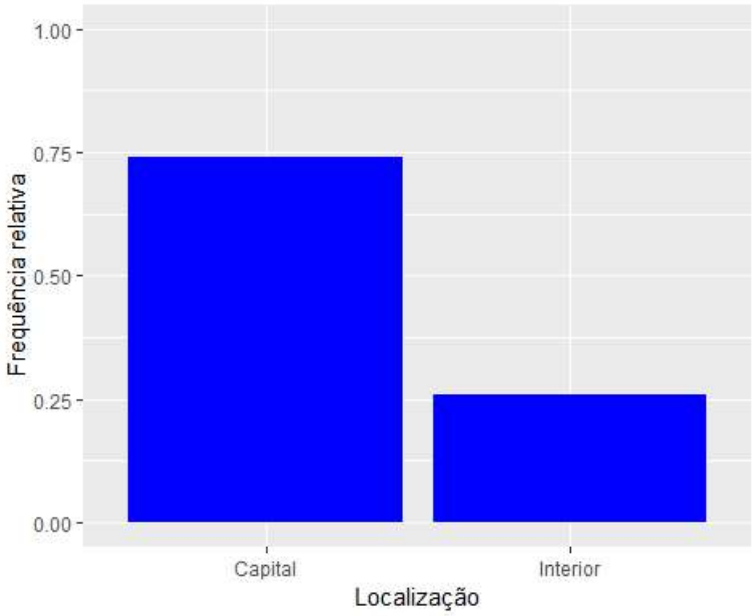


Figura 13: Gráfico de barras para a variável Localização.

A primeira hipótese científica de Heloísa é: Localização e Gênero estão associadas. O primeiro passo para estudar a associação entre duas variáveis qualitativas é construir um gráfico de barras usando o código abaixo.

```
# Associação: gráfico de barras
ggplot(dados_heloisa) +
  geom_bar(aes(x=genero, fill=localizacao),
           position = "fill") +
  xlab("Gênero") + ylab("Frequência relativa")
```

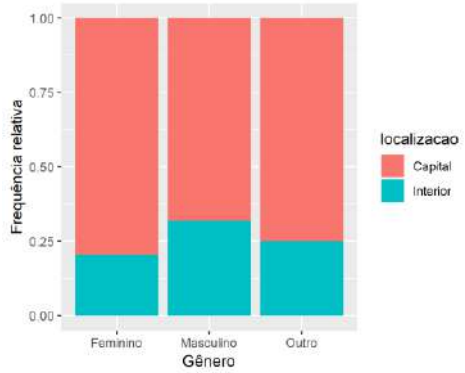


Figura 14: Associação entre Localização e Gênero.

No gráfico da figura 14, as barras para Feminino, Masculino e Outro são semelhantes. Isso já te indica a não associação entre as variáveis Localização e Gênero.

Você pode checar a associação entre Localização e Gênero usando o teste de independência, ou seja, você pode testar as hipóteses

$$H_0: \textit{Localização e Gênero } \textbf{não } \textit{estão associadas,}$$
$$H_1: \textit{Localização e Gênero estão associadas.}$$

Podemos obter o p-valor do teste de independência usamos a função `chisq.test` conforme código abaixo. No output, temos a valor do qui-quadrado $\chi^2 = 1{,}65$, os graus de liberdade $l = 2$, e o p-valor $p = 0{,}44$. Fixe um nível de significância α menor 5%. Ao nível de significância α, não rejeitamos H_0, ou seja, ao nível de significância α Localização e Gênero não estão associadas. E com isso respondemos a primeira hipótese de Heloísa.

```
# Testando a associação entre Gênero e Localização
chisq.test( dados_heloisa$genero,
            dados_heloisa$localizacao)
```

```
##
##  Pearson's Chi-squared test
##
## data:  dados_heloisa$genero and dados_heloisa$localizacao
## X-squared = 1.6531, df = 2, p-value = 0.4376
```

Antes de responder a segunda hipótese científica de Heloísa, vamos fazer uma análise descritiva das variáveis quantitativas contínuas Tempo e Nota. No R, você pode usar a função `summarise` do pacote `tidyverse`.

```
# Medidas resumo para Nota
dados_heloisa %>%
  summarise(media = mean(Nota), mediana = median(Nota),
            desvio_medio = (Nota-media) %>% abs() %>% mean(),
            desvio_padrao = (Nota-media)^2  %>% mean() %>% sqrt())
```

```
## # A tibble: 1 x 4
##   media mediana desvio_medio desvio_padrao
##   <dbl>   <dbl>        <dbl>         <dbl>
## 1  8.12    8.72         1.31          1.87
```

```
# Medidas resumo para tempo
dados_heloisa %>%
  summarise(media = mean(tempo), mediana = median(tempo),
            desvio_medio = (tempo-media) %>% abs() %>% mean(),
            desvio_padrao = (tempo-media)^2  %>% mean() %>% sqrt())
```

```
## # A tibble: 1 x 4
##   media mediana desvio_medio desvio_padrao
##   <dbl>   <dbl>        <dbl>         <dbl>
## 1 100.0    98.6         18.1          22.1
```

Essas medidas de resumo são úteis para entender o que está acontecendo com variáveis Tempo e Nota. A nota média dos estudantes foi 8,12 e tempo médio logado foi 100 minutos.

Depois de estudar cada variável quantitativa individualmente, está na hora responder a segunda hipótese da Heloísa: as variáveis tempo (logado na plataforma digital) e Nota (em matemática) estão associadas? Primeiro construímos o gráfico de dispersão, mostrado na figura 15. Observamos uma tendência: os alunos que gastam mais tempo no ambiente virtual de aprendizagem têm notas maiores. Ou seja, as variáveis estão positivamente associadas.

```
# gráfico de dispersão
ggplot(dados_heloisa)+
  geom_point(aes(x=Nota, y = tempo), color = "blue")
```

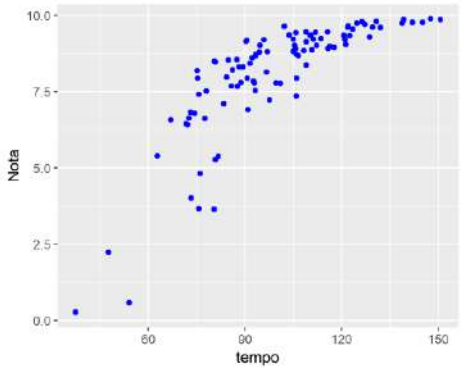


Figura 15: Gráfico de dispersão entre as variáveis quantitativas Notas e tempo.

Podemos testar as hipóteses H_0: *Tempo e nota* ***não*** *são associadas* e H_1: *Tempo e nota são associadas*. Para calcular o p-valor, podemos usar a função `cor.test`. Vamos fixar um nível de significância α menor que 5%. O p-valor virtualmente é zero e rejeitamos H_0, ou seja, as variáveis Nota e tempo estão associadas.

A função `cor.test` também calcula o intervalo de confiança para coeficiente de correlação linear de Pearson ρ da população. Usamos o coeficiente de confiança 95%: $IC(\rho; 95\%) = (0{,}71; 0{,}86)$, ou seja, o coeficiente de correlação linear de Pearson para a população ρ está entre 0,71 e 0,86 com coeficiente de confiança 95%. Além disso, o valor do coeficiente de correlação linear de Pearson também é 0,79. Então, podemos concluir que as variáveis Nota (em matemática) e Tempo (logado na plataforma) estão positivamente associadas, ou seja, alunos que ficam mais tempo logado na plataforma digital tendem a ter notas maiores em Matemática.

```
# coeficiente de correlação linear de Pearson
cor.test(dados_heloisa$Nota,
         dados_heloisa$tempo, conf.level = 0.95)

##
##  Pearson's product-moment correlation
##
## data:  dados_maria$Nota and dados_maria$tempo
## t = 12.97, df = 98, p-value < 2.2e-16
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  0.7092202 0.8574468
```

```
## sample estimates:
##       cor
## 0.7949051
```

Finalmente, vamos responder a terceira hipótese de Heloísa: os estudantes estão ficando mais tempo logado no ambiente virtual de aprendizagem? O administrador do sistema informou à Heloísa que os estudantes ficavam em média 70 minutos por semana logados. Então desejamos checar se o tempo médio logado está maior que 70 minutos. Vamos construir o teste de hipótese usando os cinco passos do procedimento de Neyman-Pearson e também vamos calcular o p-valor.

Passo 1) Vamos estabelecer as hipóteses nula e alternativa. Geralmente colocamos nossa hipótese científica ou nossa pergunta na hipótese alternativa, porque controlamos da probabilidade de erroneamente decidir pela hipótese alternativa: $\alpha = P(Erro\ tipo\ I)$. No caso de Heloísa, a hipótese científica é: O tempo médio logado na plataforma digital é maior que 70 minutos. Então, temos as seguintes hipóteses:

H_0: *o tempo logado dos estudantes não aumentou,*
H_1: *o tempo médio logado dos estudantes aumentou,*

As hipóteses podem ser reescritas em termos da média da população μ

$$H_0: \mu \leq 70,$$
$$H_0: \mu > 70.$$

Passo 2) Heloísa decidiu usar o nível de significância $\alpha = 0{,}05$.

Passo 3) Aqui vamos ajudar Heloísa a encontrar a ideia de decisão:

- Se $\bar{x} - 70$ é menor ou igual a 0, não rejeitamos por H_0. Ou seja, se $\bar{x} - 70 \leq x_c$, não rejeitamos H_0;
- Se $\bar{x} - 70$ é maior que 0, rejeitamos por H_0. Ou seja, se $\bar{x} - 70 > x_c$, rejeitamos H_0.

Então a região crítica é $RC = \{\bar{x}\ tal\ que\ \bar{x} - 70 > x_c\}$.

Passo 4) Note que não conhecemos a variância da população, então Heloísa precisa checar se a variável quantitativa Tempo tem distribuição normal. Primeiro você pode construir o gráfico Quantil-Quantil conforme o código abaixo. Note que os pontos do gráfico da figura 16 estão alinhados em uma reta indicando a Heloísa que a variável Tempo tem distribuição normal.

```
# gráfico quantil-quantil
media <- dados_heloisa$tempo %>% mean()
dp <- dados_heloisa$tempo %>% sd()
dados_heloisa %>%
  ggplot(aes(sample = tempo))+
  geom_qq(distribution = qnorm,
          dparams = list(mean = media,sd = dp)) +
  geom_qq_line(distribution = qnorm,
               dparams = list(mean = media,sd = dp),
               size = 1.5, color = "red", alpha = 0.6) +
```

```
  labs(x = "Quantis teóricos",
       y = "Quantis amostrais")
```

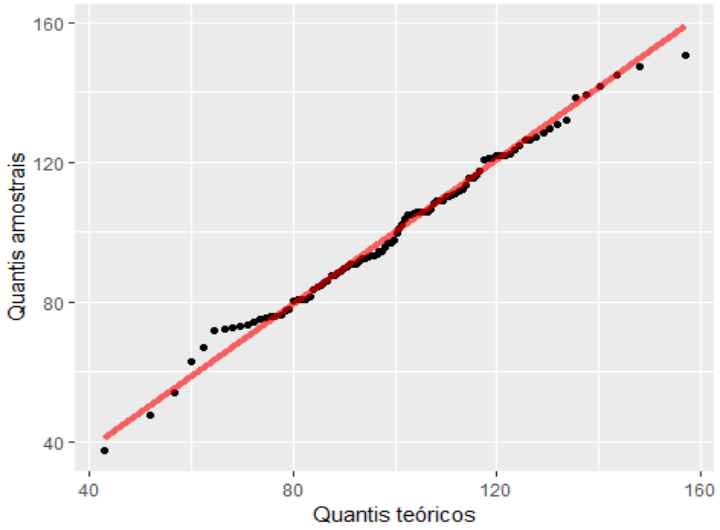


Figura 16: Gráfico quantil-quantil da variável quantitativa tempo.

Para testar se a variável quantitativa Tempo tem distribuição normal usamos os testes de normalidade de Shapiro-Wilks, Anderson-Daling e Kolmolgorov-Smirnov. Usamos as funções `shapiro.test`, `ad.test`, e `lillie.test`.

```
# shapiro-wilks test
dados_heloisa$tempo %>% shapiro.test()
```

```
##
##  Shapiro-Wilk normality test
##
## data:  .
## W = 0.99101, p-value = 0.7458
```

```
# Anderson-Darling test
dados_heloisa$tempo %>% ad.test()
```

```
##
##  Anderson-Darling normality test
##
## data:  .
## A = 0.28302, p-value = 0.6276
```

```
# kolmogorov-smirnov test
dados_heloisa$tempo %>% lillie.test()
```

```
##
##        Lilliefors (Kolmogorov-Smirnov) normality test
##
```

```
##                     data:                                              .
## D = 0.056104, p-value = 0.6135
```

Os p-valores para o teste de Shapiro-Wilks, Anderson-Darling e Kolmogorov-Smirnov são maiores que 0,05, e não rejeitamos H_0 ao nível de significância 0,05, ou seja, a variável quantitativa Tempo tem distribuição normal e agora vamos encontrar o valor crítico x_c usando a seguinte equação

$$P\left(T \leq \frac{x_c\sqrt{n}}{S}\right) = 1 - \alpha = 0{,}95,$$

em que T tem distribuição t-Student com $n - 1 = 99$ graus de liberdade. No R, podemos encontrar o valor crítico usando o código abaixo.

```
# Valor crítico
alpha <- 0.05 #nível de significância
n <- nrow(dados_heloisa)  # tamanho amostral
x_c <- qt(1-alpha, df=n-1)*sd(dados_heloisa$tempo) / sqrt(n)
x_c
```

```
## [1] 3.681274
```

Passo 5) Está na hora da Maria decidir. Primeiro calculamos o tempo médio dos estudantes na plataforma digital: $\bar{x} = 99{,}99$. Note que $99{,}99 - 70 = 29{,}99$ é maior que 3,68 e rejeitamos H_0. Então, ao nível de significância $\alpha = 5\%$, Maria pode afirmar que os estudantes estão ficando mais tempo no ambiente virtual de aprendizagem.

Podemos calcular o p-valor para testar as hipóteses $H_0: \mu \leq \mu_0$ e $H_1: \mu > \mu_0$ com a função `t.test`, conforme código abaixo. O p-valor é virtualmente zero e rejeitamos H_0 ao nível de significância $\alpha = 0{,}05$. Então, ao nível de significância $\alpha = 0{,}05$, Heloísa pode afirmar que os estudantes estão ficando mais tempo logado no ambiente virtual de aprendizagem.

```
# t.test
dados_heloisa$tempo %>% t.test(alternative = "greater", mu = 70)
```

```
##
##  One Sample t-test
##
## data:  .
## t = 13.528, df = 99, p-value < 2.2e-16
## alternative hypothesis: true mean is greater than 70
## 95 percent confidence interval:
##  96.31243      Inf
## sample estimates:
## mean of x
##   99.9937
```

Em resumo, depois de analisar a base de dados, Heloísa tem as respostas para as hipóteses científicas:

(1) As variáveis qualitativas Localização e Gênero não estão associadas. Ou seja, a proporção de homens e mulheres que usam a plataforma digital é igual nas regiões metropolitanas e no interior;
(2) As variáveis quantitativas Nota (em matemática) e Tempo (logado na plataforma digital) estão positivamente associadas, ou seja, quanto mais tempo logado na plataforma digital maior a Nota em matemática do estudante;
(3) As melhorias propostas por Heloísa aumentaram o tempo que os estudantes ficam logados no ambiente virtual de aprendizagem.

6 Resumo

Neste capítulo, aprendemos conceitos de probabilidade, intervalos de confiança e testes de hipóteses. Começamos descrevendo um modelo de probabilidade chamado distribuição normal e o intrigante teorema central do limite que estabelece que para amostras suficientemente grandes a distribuição normal é um modelo de probabilidade adequado para a média. Em seguida, mostramos como construir intervalos de confiança. Existe duas divisões principais: se conhecemos o desvio padrão da população e se não conhecemos o desvio padrão. Se não conhecemos o desvio padrão da população e a variável aleatória tem distribuição normal, então podemos usar a distribuição t-Student para construir o intervalo de confiança. Se não conhecemos o desvio padrão da população e não podemos assumir que a variável aleatória tem distribuição normal, podemos usar uma técnica denominada *bootstrap*. Intervalos de confiança construídos usando *bootstrap* usam diversas amostras da população para construir o intervalo de confiança e não assume modelo de probabilidade para a variável aleatória. Na última seção deste capítulo, estávamos preocupados em decidir entre duas hipóteses complementares: a hipótese nula H_0 e a hipótese alternativa H_1 (que é a negação de H_0). Associado ao conceito de teste de hipóteses, temos o conceito de nível de significância, valor crítico e p-valor. O p-valor é um medida entre zero e um que auxilia a decidir entre H_0 e H_1: valores pequenos do p-valor indicam que você deveria decidir por H_1 e valores grandes de do p-valor indicam que você deveria por H_0. Encerramos o capítulo ilustrando os conceitos deste capítulo com um exemplo real em que usamos o R para ajudar Heloísa, uma aluna de pós-graduação em Informática na Educação, a responder três

hipóteses para compor o seu trabalho de dissertação.

Figura 17: Mapa mental para inferência estatística.

7 Leituras Recomendadas

- **Estatística Básica** (BUSSAB; MORETTIN, 2014). Este livro é uma referência bastante usada para cursos introdutórios de Estatística Descritiva e Inferencial. O livro cobre os tópicos deste capítulo com mais detalhes, além de apresentar outros tópicos de estatística incluindo teste de hipóteses para comparar médias de amostras diferentes, regressão linear e probabilidade. O autor disponibiliza os conjuntos de dados usados no livro no endereço eletrônico: https://www.ime.usp.br/~pam/EstBas.html.

- **Statistical Inference** (CASELLA, 2008). Este livro também é bastante usado em cursos de inferência e probabilidade, mas o foco é inferência estatística. O autor detalha e motiva cada um dos modelos e métodos de inferência estatística e ao final de cada capítulo tem vários exercícios para fixar os conceitos e métodos.

- **Estatística Aplicada às Ciências Sociais** (BARBETTA, 2008). Esta é outra referência bastante usada em cursos de Estatística Básica. Este livro foi construído para alunos de graduação em Ciências Humanas e existe uma preocupação especial em deixar a linguagem simples.

- **R for Data Science** (WICKHAM; GROLEMUND, 2016). Este livro apresenta desde estatística descritiva até modelagem usando o R. O autor diz que explicitamente que está preocupado em difundir o uso de R para análise de dados. O livro tem uma leitura fluída e existe a opção de você ler o livro gratuitamente no endereço eletrônico: http://r4ds.had.co.nz/ .

- **An Introduction to the Bootstrap** (JOHNSON, 2001). Este artigo explica de simples e concisa o conceito de *bootstrap* para construir intervalos de confiança e testes de hipóteses. Esse método é especialmente útil quando a variável aleatória contínua não tem distribuição normal e não conhecemos o desvio padrão.

8 Artigos Exemplos

- **Reduced GUI for an interactive geometry software: Does it affect students' performance?** (BORGES et al., 2015). Este artigo foi publicado na revista *Computers in Human Behaviour*. O artigo usa medidas resumo, diagrama de caixa e teste de hipótese.
- **What do students do on-line? Modeling students' interactions to improve their learning experience** (PAIVA et al., 2016). Este artigo analisa a interação de estudantes com um ambiente de aprendizado on-line (MeuTutor). Os autores construíram histogramas e gráficos de barras, e testaram a normalidade usando o teste de Shapiro-Wilks e Anderson-Darling.

9 Checklist

i. Identificar o tipo de variável de cada coluna de sua base de dados: qualitativa ordinal, qualitativa nominal, quantitativa discreta e quantitativa contínua.

ii. Fazer uma análise descritiva de cada uma das variáveis de interesses: calcule medidas resumo, desenhe gráficos de barra e/ou histogramas.

iii. Precisamos estabelecer as hipóteses científicas (ou perguntas) que precisamos responder com as informações em sua base de dados.

iv. Construímos intervalos de confiança e testes de hipóteses para checar ou responder as perguntas de **iii.**

v. Interpretar os resultados.

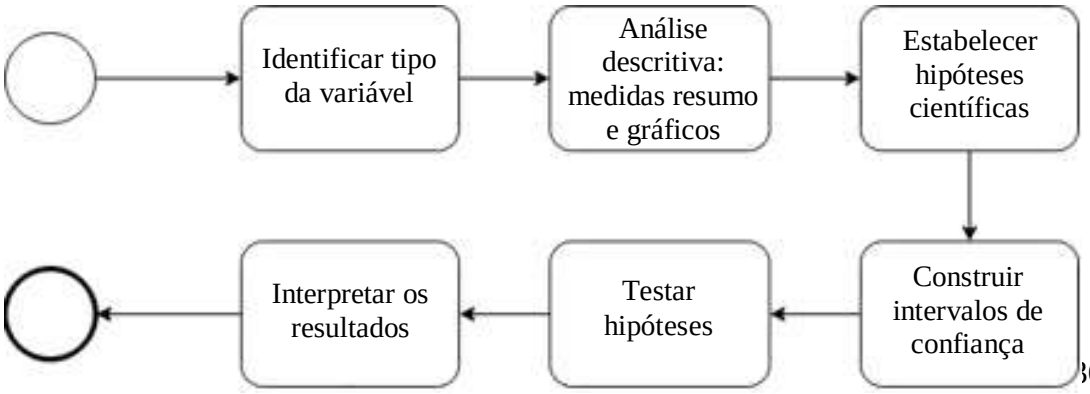


Figura 18: Fluxograma para análise descritiva e inferencial.

10 Exercícios

A) Um professor de estatística coletou as notas finais e a idade de turma com 15 alunos ao final do semestre corrente. Ele salvou as informações no arquivo `"notas_finais.xlsx"`.

a. Construa um intervalo de confiança para a média das variáveis idade e notas finais usando coeficiente de confiança $\gamma = 95\%$.

b. Você acha que a variável nota e idade estão associadas? Justifique sua resposta.

c. A turma do atual semestre passou teve uma nota média final 7. Ao nível de significância 5%, você acha que a nota média final melhorou?

B) Considere o conjunto de dados `"nlschool"` do pacote MASS do R. Esse conjunto de dados tem as variáveis quantitativas QI e a nota em linguagem para 2287 crianças holandesas.

a. Ao nível de significância 5%, as variáveis IQ e lang tem distribuição normal? Use o teste de Shapiro-Wilks, Anderson-Darling e Kolmogorov-Smirnov e desenhe o gráfico quantil-quantil.

b. Construa intervalo de confiança para as variáveis IQ e lang usando *bootstrap* com coeficiente de confiança $\gamma = 95\%$.

C) Considere o conjunto de dados `"Caschool"` do pacote Ecdat do R. Esse conjunto contém informações de 420 escolas na Califórnia incluindo a nota média de testes em leitura e matemática em 1999.

a. Verifique se as variáveis computer, mathscr e readscr tem distribuição normal. Construa os gráficos quantil-quantil e use o teste de Shapiro-Wilks.

b. Você acha que as computer e mathscr estão associadas? E as variáveis mathscr e readscr? Justifique a sua resposta.

11 Referências

BARBETTA, Pedro Alberto. **Estatística aplicada às ciências sociais**. Ed. UFSC, 2008.

BORGES, Simone S. et al. Reduced GUI for an interactive geometry software: Does it affect students' performance? **Computers in Human Behavior**, v. 54, p. 124-133, 2016.

BUSSAB, W. O.; MORETTIN, P. A. Estatística Básica, Saraiva. **São Paulo-SP. 526p**, 2003.

CASELLA, George; BERGER, Roger L. **Statistical inference**. Pacific Grove, CA: Duxbury, 2002.

D'AGOSTINO, Ralph B.; BELANGER, Albert; D'AGOSTINO JR, Ralph B. A suggestion for using powerful and informative tests of normality. **The American Statistician**, v. 44, n. 4, p. 316-321, 1990.

DE BRAGANÇA PEREIRA, Carlos Alberto; STERN, Julio Michael. Evidence and credibility: full Bayesian significance test for precise hypotheses. **Entropy**, v. 1, n. 4, p. 99-110, 1999.

EFRON, Bradley; TIBSHIRANI, Robert J. **An introduction to the bootstrap**. CRC press, 1994.

GHASEMI, Asghar; ZAHEDIASL, Saleh. Normality tests for statistical analysis: a guide for non-statisticians. **International journal of endocrinology and metabolism,** v. 10, n. 2, p. 486, 2012.

JOHNSON, Roger W. An introduction to the bootstrap. **Teaching Statistics**, v. 23, n. 2, p. 49-54, 2001.

LEHMANN, Erich L.; CASELLA, George. **Theory of point estimation**. Springer Science & Business Media, 2006.

LEHMANN, Erich L.; ROMANO, Joseph P. **Testing statistical hypotheses**. Springer Science & Business Media, 2006.

LEHMANN, Erich L. The Fisher, Neyman-Pearson theories of testing hypotheses: one theory or two?. **Journal of the American statistical Association**, v. 88, n. 424, p. 1242-1249, 1993.

LEHMANN, Eric L. Some principles of the theory of testing hypotheses. In: **Selected Works of EL Lehmann**. Springer, Boston, MA, 2012. p. 139-164.

MOOD, Alexander McFarlane; GRAYBILL, Franklin A.; BOES, Duane C. **Introduction to the Theory of Statistics**. McGraw-Hill Kogakusha, 1974.

PAIVA, Ranilson et al. What do students do on-line? Modeling students' interactions to improve their learning experience. **Computers in Human Behavior**, v. 64, p. 769-781, 2016.

PATRIOTA, Alexandre G. A classical measure of evidence for general null hypotheses. **Fuzzy Sets and Systems**, v. 233, p. 74-88, 2013.

THADEWALD, Thorsten; BÜNING, Herbert. Jarque–Bera test and its competitors for

testing normality–a power comparison. **Journal of Applied Statistics**, v. 34, n. 1, p. 87-105, 2007.

TUKEY, John Wilder; MOSTELLER, Frederick; HOAGLIN, David Caster (Ed.). **Understanding robust and exploratory data Analysis**. Wiley, 1983.

WICKHAM, Hadley; GROLEMUND, Garrett. **R for data science: import, tidy, transform, visualize, and model data**. " O'Reilly Media, Inc.", 2016.

YAP, Bee Wah; SIM, Chiaw Hock. Comparisons of various types of normality tests. **Journal of Statistical Computation and Simulation**, v. 81, n. 12, p. 2141-2155, 2011.

Sobre o Autor

Gilberto Pereira Sassi

Doutor em Estatística e mestre em Ciência da Computação e Matemática Computacional e graduado em Matemática pela Universidade de São Paulo, Brasil. Atualmente, Gilberto é Professor Adjunto na Universidade Federal da Bahia, e sua pesquisa está focada em Análise de Dados Funcionais.