FGV EPGE
ESCOLA BRASILEIRA DE ECONOMIA E FINANÇAS

Notas de Aula de Estatística para o Doutorado em Economia

EPGE/FGV — Professor Eduardo Campos

Vitor Wilher[1]

Versão preliminar — compilada ao longo do curso. Baseada nos slides da disciplina (Prof. Eduardo Campos), no programa de 2026.1 e nas referências do curso.

[1]Bacharel e Mestre em Economia pela UFF, Candidato ao PhD em Economia pela EPGE/FGV. É Especialista em Ciências de Dados e Inteligência Artificial Generativa pela PUC-Rio. Atualmente, exerce a função de Data Tech Lead na Análise Macro (http://analisemacro.com.br). Saiba mais em https://github.com/vitorwilher.

Índice

1 Introdução

Estas notas acompanham a disciplina de **Estatística** do Doutorado em Economia da EPGE/FGV, ministrada pelo Professor Eduardo Campos na turma de 2026.1. Elas seguem a ordem dos doze módulos do programa do curso e tomam os slides da disciplina como fonte primária: a notação, os exemplos numéricos e as convenções são as do professor, ainda que o texto seja de minha autoria e acrescente derivações, comentários e implementações computacionais que os slides deixam implícitos.

1.1 O que a estatística faz

A definição que abre o curso é econômica em espírito: *a estatística é a ciência que permite obter informações sobre um fenômeno a partir do registro de observações desse fenômeno.* O movimento é sempre o mesmo —

$$\text{DADOS} \longrightarrow \text{INFORMAÇÃO}$$

— e a disciplina se divide, classicamente, em duas metades que organizam também estas notas:

- a **estatística descritiva** (ou análise exploratória de dados), que resume um conjunto de dados por meio de tabelas, gráficos e medidas-resumo, com o objetivo de facilitar sua visualização e compreensão;
- a **inferência estatística**, que é o conjunto de técnicas para, a partir de uma *amostra* selecionada de um universo, formular conclusões sobre esse universo.

O exemplo canônico da primeira é o coeficiente de rendimento de um aluno — uma média ponderada das notas que resume seu desempenho acadêmico. O da segunda é a pesquisa eleitoral, que estima as intenções de voto de todo o eleitorado a partir de uma amostra de, digamos, duas mil pessoas. Entre uma coisa e outra há um salto lógico considerável, e é esse salto que a maior parte do curso se dedica a tornar rigoroso.

1.2 O fio condutor: da descrição à inferência

O caminho que percorremos tem uma arquitetura clara, e vale enunciá-la desde já porque cada capítulo ocupa um lugar preciso nela.

Começamos **descrevendo** dados (Seção 2): medidas de posição e de dispersão, box-plot, análise bidimensional. Nada aqui é ainda probabilístico — são resumos de números observados. Mas duas ideias plantadas nesse capítulo atravessam todo o curso: a de que **dispersão é risco** e a de que uma medida-resumo só se interpreta em contexto.

Para dar o salto à inferência precisamos de um modelo do acaso, e é o que a **probabilidade** oferece (Seção 3): espaço amostral, eventos, os axiomas, a probabilidade condicional e o Teorema de Bayes. A **variável aleatória** (Seção 4) é a ponte formal entre os dois mundos — ela traduz resultados de experimentos em números e permite reescrever, agora em termos de esperança e variância, exatamente as medidas de posição e dispersão do primeiro capítulo.

Munidos disso, catalogamos os modelos probabilísticos mais úteis — as distribuições **discretas** (Seção 5) e **contínuas** (Seção 6) — e estudamos o que acontece quando **somamos e combinamos** variáveis aleatórias (Seção 7). É aqui que aparece o **Teorema Central do Limite**, o resultado que torna a inferência possível: ele nos diz que a média amostral é aproximadamente Normal, quase independentemente de como os dados individuais se distribuem. Esse mesmo capítulo colhe os frutos financeiros do maquinário, com risco-retorno, árvores binomiais e carteiras.

A última terça parte é a inferência propriamente dita. **Estimação pontual** (Seção 8) pergunta qual número usar como palpite para um parâmetro desconhecido, e com que propriedades. **Intervalos de confiança** (Seção 9) reconhecem que um palpite pontual quase certamente erra e propõem, em seu lugar, uma faixa de valores plausíveis. **Testes de hipóteses** (Seção 10) invertem a pergunta: dada uma afirmação sobre o parâmetro, os dados a contradizem?

1.3 Como ler estas notas

Cada capítulo abre com a **motivação** do problema, desenvolve a teoria com **derivações passo a passo** e fecha com duas seções fixas:

- um **Laboratório computacional em R**, que ou reproduz numericamente um resultado do capítulo ou gera a figura que o ilustra. O código é executado na própria compilação do PDF, de modo que os números impressos são os que o R de fato produziu;
- uma lista de **Exercícios**, no espírito das listas da disciplina.

Resultados centrais aparecem em caixa, assim. Hipóteses e definições que não podem ser esquecidas vão em caixas de destaque. Quando um conceito matemático é mobilizado — uma integral, uma otimização, uma expansão de Taylor —, uma caixa *Ferramenta matemática* o nomeia e remete ao capítulo correspondente das **Notas de Aula de Matemática** (Profa. Silvia Matos), escritas para o mesmo primeiro ano.

> **Uma convenção do curso que merece atenção**
>
> No capítulo de estatística descritiva, o professor trabalha com a **notação populacional** (μ, σ^2, σ) e com o divisor n na variância — inclusive quando os dados são, a rigor, uma amostra. Só no capítulo de estimação (Seção 8) o divisor $n-1$ aparece, e aí com justificativa precisa: S^2 com $n-1$ é o estimador **não viesado** de σ^2. Mantenho a convenção do curso em cada capítulo, mas sinalizo a distinção onde ela importa — é uma fonte clássica de confusão.

1.4 Software

Os laboratórios usam **R**, com `ggplot2` e `dplyr`, e nada além do que vem na instalação padrão dessas duas bibliotecas. A escolha acompanha o curso, que apresenta os comandos `dbinom`, `pnorm`, `integrate`, `qexp` e afins ao lado de cada distribuição, e por vezes os equivalentes em Excel. Onde o professor cita a função, eu a reproduzo; onde ele resolve na tabela, mostro a tabela **e** o comando, porque a prova é feita com tabela mas a vida profissional, não.

1.5 Referências

A bibliografia principal do curso é o material do professor. Como auxiliares, o programa lista Keller (2011), Larson (1982), Meyer (1983), Bussab e Morettin (2010) e Hoffmann (2006). Onde estas notas avançam além do slide — em particular nas propriedades de estimadores e na construção dos intervalos —, apoio-me também em Casella e Berger (2002) e DeGroot e Schervish (2012).

2 Estatística Descritiva

Trinta filiais de uma rede de varejo faturaram, no mês passado, os seguintes valores em milhões de reais:

11,8	3,6	16,6	13,5	4,8	8,3
8,9	9,1	7,7	2,3	12,1	6,1
10,2	8,0	11,4	6,8	9,6	19,5
15,3	12,3	8,5	15,9	18,7	11,7
6,2	11,2	10,4	7,2	5,5	14,5

Que conclusões se pode tirar? Olhando assim, na **forma bruta**, praticamente nenhuma — e esse é exatamente o ponto de partida do curso. Trinta números não cabem na cabeça de ninguém ao mesmo tempo; precisamos de técnicas que os resumam sem destruir o que eles têm de informativo. É disso que trata a estatística descritiva.

Este capítulo percorre três famílias de resumos: as **medidas de posição**, que dizem em torno de que valor os dados se concentram; as **medidas de dispersão**, que dizem o quanto eles se espalham; e a **análise bidimensional**, que descreve a associação entre duas variáveis. A segunda família é, economicamente, a mais importante das três — dispersão é outro nome para risco —, e é dela que o curso extrai seu fio condutor.

> **!** Convenção de notação deste capítulo
>
> Seguindo os slides da disciplina, todo este capítulo usa **notação populacional** — μ para a média, σ^2 para a variância, σ para o desvio padrão — e o divisor n (não $n-1$) na variância. A distinção entre descrever uma população e estimar seus parâmetros a partir de uma amostra só se torna essencial em Seção 8, onde o divisor $n-1$ aparece com justificativa. Por ora, tratamos os dados como *o* conjunto de interesse, não como amostra de algo maior.

2.1 Medidas de posição

Uma **medida de posição** é um valor em torno do qual os dados se concentram. São sinônimos *medida de localização* e *medida de tendência central*. As três principais são a média, a mediana e a moda.

2.1.1 Média

A média é a soma das observações dividida pelo seu número:

$$\mu = \frac{\sum_{i=1}^{n} x_i}{n} = \frac{x_1 + x_2 + \cdots + x_n}{n}$$

Para os trinta faturamentos, $\sum x_i = 307{,}7$ e portanto $\mu = 307{,}7/30 \approx 10{,}3$ milhões. Note que o valor 10,3 **não ocorre** entre os dados observados. Nenhum problema: a média de um conjunto de dados não precisa ser um dos valores observados — ela é um resumo, não uma observação.

A média tem, porém, uma fragilidade decisiva. Considere os salários iniciais, em milhares de reais, de cinco economistas recém-formados:

$$2{,}8 \qquad 6{,}0 \qquad 2{,}6 \qquad 3{,}1 \qquad 3{,}0$$

A média é $\mu = 3{,}5$, ou R$ 3.500. Esse número é representativo? Não: ele está **acima de quatro dos cinco valores**. O responsável pela distorção é o 6,0 — um valor atípico, discrepante, um **outlier**. Uma única observação extrema arrasta a média para longe do grosso dos dados, porque cada observação entra na soma com seu valor integral.

2.1.2 Mediana

A alternativa é ordenar os dados e tomar o valor central. A **mediana** Md é

$$Md = \begin{cases} x_{(\frac{n+1}{2})}, & n \text{ ímpar (a observação central)}, \\ \dfrac{x_{(n/2)} + x_{(n/2+1)}}{2}, & n \text{ par (média das duas centrais)}. \end{cases}$$

Nos salários, ordenando: 2,6; 2,8; **3,0**; 3,1; 6,0, de modo que $Md = 3{,}0$ — bem mais representativo da tendência central do que os 3,5 da média.

A razão é estrutural: a mediana depende apenas da *ordem* das observações, não de seus valores. Se o maior salário fosse 60,0 em vez de 6,0, a mediana continuaria 3,0 enquanto a média saltaria para 14,3. Por isso dizemos que a mediana é uma medida **robusta** ou **resistente** — ela resiste, mantém seu valor, na presença de outliers.

2.1.3 Moda

A **moda** Mo é o valor que ocorre com maior frequência. É a medida adequada quando a pergunta é sobre o valor *mais comum*, e não sobre o valor *típico*: um gerente de loja de calçados que precisa decidir qual numeração estocar quer a moda dos tamanhos vendidos, não a média (que poderia ser um número quebrado, inútil para a decisão).

Para as notas 9, 7, 8, 6, 3, 8, 7, 8, a moda é 8 (frequência 3). Se o aluno que tirou 6 conseguisse revisar sua nota para 7, teríamos 7 e 8 com frequência 3 cada, e a distribuição passaria a ser **bimodal**. Conforme o número de modas, um conjunto é *bimodal* (duas), *multimodal* (mais de duas) ou *amodal* (nenhum valor se repete).

2.1.4 Média ponderada

Quando as observações têm importâncias distintas, usa-se a **média ponderada**

$$\boxed{\mu_p = \frac{\sum_{j=1}^{k} \omega_j x_j}{\sum_{j=1}^{k} \omega_j} = \frac{\omega_1 x_1 + \omega_2 x_2 + \cdots + \omega_k x_k}{\omega_1 + \omega_2 + \cdots + \omega_k}}$$

em que ω_j é o peso do j-ésimo valor distinto. Se os pesos são as frequências, recupera-se a média simples escrita de outro jeito. Numa empresa em que 5 funcionários ganham R$ 2.500, 2 ganham R$ 3.000, 3 ganham R$ 4.000 e 2 ganham R$ 4.500, o salário médio é

$$\mu_p = \frac{5(2500) + 2(3000) + 3(4000) + 2(4500)}{12} = \frac{39.500}{12} = 3.291{,}67.$$

Uma aplicação em economia. Um investidor tem R$ 8.000 distribuídos em R$ 1.000 em ações, R$ 3.000 em um fundo multimercado e R$ 4.000 na poupança, com rentabilidades projetadas de 15%, 10% e 5% respectivamente. A rentabilidade esperada da carteira é a média ponderada das rentabilidades, com pesos iguais às frações do capital:

$$\mu_p = \frac{1000(15) + 3000(10) + 4000(5)}{8000} = \frac{65.000}{8000} = 8{,}125\%.$$

Esse cálculo reaparecerá, praticamente idêntico, na análise de carteiras de Seção 7 — só que lá as rentabilidades serão variáveis aleatórias e a média ponderada será uma esperança.

2.2 Medidas de dispersão

Uma pessoa com metade do corpo em um forno e a outra metade em um freezer está, *na média*, bem. A piada do professor tem uma moral séria: a posição sozinha pode ser gravemente enganosa. Considere dois fornecedores e seus prazos de entrega (em dias) aos últimos cinco clientes:

						Média
Fornecedor A	18	10	17	3	2	10
Fornecedor B	9	10	10	9	12	10

Os dois têm prazo médio de 10 dias e são, por esse critério, indistinguíveis. Mas qualquer gerente prefere B: com A, a entrega pode levar 2 dias ou 18, e essa **imprevisibilidade** é um custo real. Precisamos medi-la.

2.2.1 Da amplitude à variância

A medida mais simples é a **amplitude total**, $A = x_{\max} - x_{\min}$, que para A vale 16 e para B vale 3. Ela capta a intuição certa, mas usa apenas duas observações e ignora todas as outras.

A ideia melhor é olhar os **desvios** em relação à média, $(x_i - \mu)$. O problema é que sua soma é sempre nula:

$$\sum_{i=1}^{n}(x_i - \mu) = \sum_{i=1}^{n} x_i - n\mu = n\mu - n\mu = 0,$$

qualquer que seja o conjunto de dados — os desvios positivos cancelam exatamente os negativos. A saída é eliminar os sinais, tomando módulos ou quadrados. A estatística escolhe os quadrados (deriváveis, e algebricamente muito mais tratáveis), e define a **variância**:

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

com a forma alternativa de cálculo, em geral mais rápida à mão,

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

Para os fornecedores, $\sigma_A^2 = 45{,}2$ e $\sigma_B^2 = 1{,}2$: o fornecedor B tem **menor risco**, ou menor dispersão inerente ao prazo de entrega. A escolha se justifica.

> **i** Ferramenta matemática — somatório e a soma dos desvios
>
> A demonstração de que $\sum(x_i - \mu) = 0$ usa apenas a linearidade do somatório e a definição de média, $\sum x_i = n\mu$. Essa mesma linearidade — $\mathbb{E}$ é um operador linear — reaparecerá em Seção 4 como propriedade do valor esperado, e é o que torna a variância uma forma quadrática bem comportada. *(Propriedades do somatório e manipulação algébrica: capítulo* Nivelamento *das Notas de Matemática.)*

2.2.2 Desvio padrão e coeficiente de variação

A variância tem um defeito prático: sua unidade é o **quadrado** da unidade original — dias ao quadrado, reais ao quadrado —, em geral algo que sequer faz sentido interpretar. A correção é tomar a raiz, obtendo o **desvio padrão**

$$\sigma = \sqrt{\sigma^2},$$

que preserva a unidade original dos dados e por isso é a medida de dispersão que se reporta. Sob normalidade (Seção 6), σ ganha interpretação precisa: cerca de 68% das observações caem em $[\mu - \sigma, \mu + \sigma]$ e praticamente todas em $[\mu - 3\sigma, \mu + 3\sigma]$ — o que fornece, aliás, um primeiro critério de outlier: são atípicos os valores fora dessa faixa de três desvios.

Resta um problema. Suponha que gerentes ganhem em média R$ 5.000 e auxiliares de escritório, R$ 500, e que **ambos** os grupos tenham desvio padrão de 100. A mesma dispersão absoluta significa coisas opostas nos dois casos: para os auxiliares, R$ 100 é uma variação substancial; para os gerentes, é ruído. Comparações entre grupos de magnitudes (ou unidades) diferentes exigem uma medida **relativa**, o **coeficiente de variação**:

$$\boxed{\mathrm{CV} = \frac{\sigma}{\mu}}$$

Aqui, $\mathrm{CV}_{\mathrm{aux}} = 100/500 = 0{,}20$ (20%) contra $\mathrm{CV}_{\mathrm{ger}} = 100/5000 = 0{,}02$ (2%). A dispersão relativa dos salários gerenciais é dez vezes menor, embora a absoluta seja idêntica.

2.3 Quartis, box-plot e outliers

Além da posição e da dispersão, interessa a **forma** da distribuição — em particular sua **assimetria**. Uma distribuição é simétrica quando as duas metades em torno do centro se espelham; é **assimétrica à direita** (ou positiva) quando a cauda longa aponta para valores altos, e **à esquerda** (negativa) no caso oposto. Distribuições de renda e de preços de ativos são o exemplo clássico de assimetria positiva.

O instrumento que organiza tudo isso é a família dos **percentis** (ou quantis). O p-ésimo percentil é o valor abaixo do qual está p% das observações. Três deles recebem nome próprio:

$$Q_1 = 25^{\mathrm{o}} \text{ percentil}, \qquad Q_2 = 50^{\mathrm{o}} \text{ percentil} = Md, \qquad Q_3 = 75^{\mathrm{o}} \text{ percentil}.$$

A distância entre os quartis externos é a **amplitude interquartílica**

$$\Delta Q = Q_3 - Q_1,$$

que mede a dispersão dos 50% centrais dos dados e é, como a mediana, robusta a outliers.

O **box-plot** resume graficamente essas cinco quantidades. A caixa vai de Q_1 a Q_3, com um traço na mediana; as hastes ("bigodes") se estendem até a observação mais extrema que ainda esteja dentro dos limites

$$\boxed{LI = Q_1 - 1{,}5\,\Delta Q \qquad \text{e} \qquad LS = Q_3 + 1{,}5\,\Delta Q}$$

e as observações fora de $[LI, LS]$ são assinaladas individualmente, em geral com um asterisco: são os **outliers** pelo critério do box-plot.

O box-plot serve a três propósitos, nesta ordem de importância:

1. **comparar distribuições** lado a lado (medianas, dispersões, presença de extremos) — é a razão de ele ser desenhado em painéis, um por grupo;
2. **identificar assimetria**, por um macete útil: se a mediana está mais próxima de Q_1, a assimetria é **positiva**; se está mais próxima de Q_3, é **negativa**; distâncias iguais sugerem simetria;

3. **detectar outliers**, pelo critério de $1{,}5\,\Delta Q$ acima.

Considere as idades de um grupo de mulheres, com $Q_1 = 49$, $Md = 54$, $Q_3 = 63$, mínimo 40 e máximo 71. Temos $\Delta Q = 14$, e portanto $LI = 49 - 21 = 28$ e $LS = 63 + 21 = 84$. Como todas as observações estão em $[40, 71] \subset [28, 84]$, **não há outliers** — as hastes vão de 40 a 71. E como $Md - Q_1 = 5$ é menor que $Q_3 - Md = 9$, a mediana está mais próxima de Q_1: a distribuição tem assimetria positiva.

2.4 Análise bidimensional

Até aqui, uma variável de cada vez. Mas as perguntas economicamente interessantes quase sempre relacionam **duas** variáveis: escolaridade e renda, faltas e notas, risco e retorno. O instrumento gráfico é o **diagrama de dispersão**, que plota os pares (x_i, y_i) e permite responder se existe associação, de que tipo (linear ou não), em que direção (positiva ou negativa) e com que força.

Para quantificar a direção, define-se a **covariância**:

$$\boxed{\sigma_{XY} = \frac{\sum_{i=1}^{n}(x_i - \mu_X)(y_i - \mu_Y)}{n}}$$

A lógica é transparente. Cada parcela é positiva quando as duas variáveis se desviam de suas médias no *mesmo* sentido (ambas acima ou ambas abaixo) e negativa quando em sentidos opostos. Se predominam os pares concordantes, $\sigma_{XY} > 0$: associação positiva. Se predominam os discordantes, $\sigma_{XY} < 0$.

O problema da covariância é que sua magnitude não significa nada isoladamente — ela depende das unidades de X e Y, e trocar reais por centavos a multiplica por cem. A correção é dividir pelos desvios padrão, obtendo o **coeficiente de correlação** (de Pearson):

$$\boxed{\rho_{XY} = \frac{\sigma_{XY}}{\sigma_X\,\sigma_Y}, \qquad -1 \le \rho_{XY} \le 1}$$

que é adimensional e limitado. Os extremos correspondem à associação linear **perfeita** ($\rho = \pm 1$: todos os pontos exatamente sobre uma reta), e $\rho = 0$ indica ausência de associação *linear*. Valores como 0,95 indicam relação positiva forte; $-0{,}1$, relação negativa fraca.

Exemplo. Três alunos, com número de faltas X e nota Y:

	X (faltas)	Y (nota)
Aluno 1	4	3
Aluno 2	4	4
Aluno 3	1	8

Temos $\mu_X = 3$ e $\mu_Y = 5$. A covariância é

$$\sigma_{XY} = \frac{(1)(-2) + (1)(-1) + (-2)(3)}{3} = \frac{-9}{3} = -3,$$

e com $\sigma_X = \sqrt{2} \approx 1{,}41$ e $\sigma_Y = \sqrt{14/3} \approx 2{,}16$,

$$\rho_{XY} = \frac{-3}{1{,}41 \times 2{,}16} \approx -0{,}98.$$

A associação linear entre faltas e nota é negativa — como se esperaria — e bastante forte.

> **!** Correlação não é causalidade
>
> O coeficiente de correlação mede associação linear e **nada diz** sobre causação: não informa se X causa Y, se Y causa X, ou se ambas respondem a uma terceira variável omitida. Correlações **espúrias** — entre séries que apenas compartilham uma tendência temporal, por exemplo — são abundantes. Conclusões causais exigem uma justificativa teórica ou um modelo estrutural subjacente; é matéria de econometria, não de estatística descritiva.

2.5 Laboratório computacional em R

Reproduzimos os cálculos do capítulo com os trinta faturamentos do Exemplo 1.1, definidos no preâmbulo destas notas como o vetor `faturamento`. Note o uso de `var_p()` e `dp_p()`, funções auxiliares com divisor n — as nativas `var()` e `sd()` do R usam $n-1$, a convenção de Seção 8, não a deste capítulo.

```
n <- length(faturamento)
c(n = n, soma = sum(faturamento), media = mean(faturamento),
  mediana = median(faturamento),
  variancia_n = var_p(faturamento), desvio_n = dp_p(faturamento),
  CV = dp_p(faturamento) / mean(faturamento),
  variancia_n1 = var(faturamento))   # a do R, divisor n-1: compare
```

```
          n         soma        media      mediana  variancia_n     desvio_n
 30.0000000  307.7000000   10.2566667    9.9000000   17.8077889    4.2199276
          CV variancia_n1
  0.4114327   18.4218506
```

A média confere com os 10,3 milhões do slide. Veja que a variância com divisor $n-1$ difere da com divisor n — pouco, com $n = 30$, mas difere; com n pequeno a diferença é substancial.

O critério de outlier do box-plot, aplicado aos mesmos dados:

```r
Q <- quantile(faturamento, c(0.25, 0.50, 0.75))
dQ <- unname(Q[3] - Q[1])
LI <- unname(Q[1] - 1.5 * dQ); LS <- unname(Q[3] + 1.5 * dQ)
c(Q1 = unname(Q[1]), Md = unname(Q[2]), Q3 = unname(Q[3]),
  deltaQ = dQ, LI = LI, LS = LS,
  n_outliers = sum(faturamento < LI | faturamento > LS))
```

```
      Q1       Md       Q3   deltaQ       LI       LS n_outliers
  7.3250   9.9000  12.2500   4.9250  -0.0625  19.6375   0.0000
```

Nenhum outlier: os trinta faturamentos estão todos dentro dos limites. Como $Md - Q_1$ e $Q_3 - Md$ são próximos, a distribuição é aproximadamente simétrica.

O contraste entre média e mediana sob contaminação por outlier — a lição central da primeira seção — fica evidente ao substituir uma observação por um valor extremo:

```r
salarios <- c(2.8, 6.0, 2.6, 3.1, 3.0)
contaminado <- c(2.8, 60.0, 2.6, 3.1, 3.0)   # o outlier dez vezes maior
rbind(original    = c(media = mean(salarios),    mediana = median(salarios)),
      contaminado = c(mean(contaminado), median(contaminado)))
```

```
            media mediana
original      3.5       3
contaminado  14.3       3
```

A média salta de 3,5 para 14,3 — quadruplica — enquanto a mediana **não se move**. É a robustez da mediana em uma linha.

```r
d <- data.frame(x = faturamento)
g1 <- ggplot(d, aes(y = x)) +
  geom_boxplot(fill = az, alpha = 0.15, colour = az, width = 0.5) +
  geom_hline(yintercept = c(LI, LS), linetype = "dashed", colour = cz) +
  scale_x_continuous(breaks = NULL) +
  labs(x = "box-plot", y = "faturamento (R$ milhoes)")
g2 <- ggplot(d, aes(x = x)) +
  geom_histogram(bins = 8, fill = az, alpha = 0.55, colour = "white") +
  geom_vline(xintercept = mean(faturamento), colour = vm, linewidth = 0.8) +
  geom_vline(xintercept = median(faturamento), colour = vd,
             linewidth = 0.8, linetype = "dashed") +
  labs(x = "faturamento  |  media (vermelho), mediana (verde, tracejada)",
       y = "frequencia")
library(patchwork)
g1 + g2 + plot_layout(widths = c(1, 2))
```

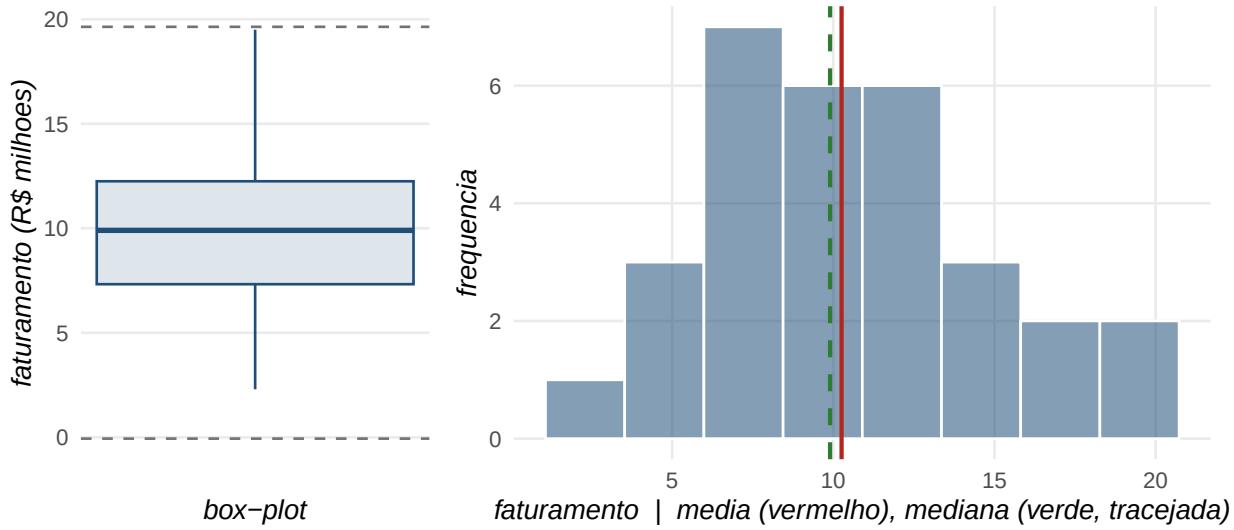


Figura 1: Os trinta faturamentos: box-plot (esquerda) e histograma com média e mediana (direita). A caixa cobre os 50% centrais; as linhas tracejadas marcam LI e LS.

Por fim, a covariância e a correlação do exemplo faltas × notas, com as funções auxiliares de divisor n:

```
faltas <- c(4, 4, 1); notas <- c(3, 4, 8)
cov_p <- function(x, y) mean((x - mean(x)) * (y - mean(y)))
c(cov_n = cov_p(faltas, notas),
  dp_faltas = dp_p(faltas), dp_notas = dp_p(notas),
  rho = cov_p(faltas, notas) / (dp_p(faltas) * dp_p(notas)),
  rho_nativo = cor(faltas, notas))
```

```
     cov_n  dp_faltas   dp_notas        rho rho_nativo
-3.0000000  1.4142136  2.1602469 -0.9819805 -0.9819805
```

Note que ρ calculado "à mão" com divisor n coincide com o `cor()` nativo: a correlação é **invariante** ao divisor, porque ele aparece no numerador e no denominador e se cancela. Já a covariância, não.

2.6 Exercícios

1. **Média e mediana.** Para as notas $9, 7, 8, 6, 3, 8, 7, 8$, calcule média, mediana e moda. (a) Qual das três melhor representa o desempenho da turma? (b) O que acontece com cada uma se a nota 3 for corrigida para 9?

2. **Sensibilidade da média.** Mostre algebricamente que, ao substituir a maior observação $x_{(n)}$ por $x_{(n)} + c$ com $c > 0$, a média aumenta em c/n e a mediana não se altera (supondo $n \geq 3$). Qual o "ponto de ruptura" de cada medida?

3. **Formas da variância.** Demonstre a identidade $\frac{1}{n}\sum(x_i-\mu)^2 = \frac{1}{n}\sum x_i^2 - \mu^2$. Em seguida, use os dados dos fornecedores A e B para verificar que ambas as formas dão 45,2 e 1,2.

4. **Coeficiente de variação.** Duas ações têm retorno médio de 2% e 8% ao mês, com desvios padrão de 3% e 10%. (a) Qual é mais arriscada em termos absolutos? (b) E em termos relativos? (c) Discuta em que situação cada critério é o relevante para um investidor.

5. **Box-plot.** Um conjunto ordenado tem $Q_1 = 20$, $Q_2 = 25$, $Q_3 = 28$, com mínimo 5 e máximo 45. (a) Calcule ΔQ, LI e LS. (b) Quais das observações extremas são outliers? (c) Que tipo de assimetria os quartis sugerem?

6. **Assimetria pelos quartis.** Justifique o macete: se a mediana está mais próxima de Q_1, a assimetria é positiva. Construa um conjunto de dados pequeno em que o macete acerte e outro em que ele falhe — o que a falha revela sobre suas limitações?

7. **Covariância e unidades.** Mostre que, se $X' = aX$ e $Y' = bY$ com $a, b > 0$, então $\sigma_{X'Y'} = ab\,\sigma_{XY}$ mas $\rho_{X'Y'} = \rho_{XY}$. Por que essa invariância torna a correlação a medida reportável?

8. **Correlação espúria.** Duas séries temporais crescentes — o PIB nominal brasileiro e o número de usuários de internet no mundo — têm correlação próxima de 1. (a) Por que isso não indica causação? (b) Que procedimento mínimo você adotaria antes de interpretar correlações entre séries temporais?

3 Probabilidade

O capítulo anterior terminou com trinta faturamentos resumidos em uma média, um desvio padrão e um box-plot. Todos os números que produzimos ali eram **fatos sobre os dados observados**: a média era 10,3 porque aquelas trinta filiais faturaram aquilo, e ponto final. Nenhuma dessas afirmações se estende um milímetro além da amostra.

Mas as perguntas que interessam ao economista quase nunca são dessa forma. Não queremos saber qual foi o faturamento médio das trinta filiais observadas; queremos saber qual é o faturamento médio da **rede**, ou qual será o do próximo mês. Não queremos descrever as intenções de voto de duas mil pessoas entrevistadas; queremos as de todo o eleitorado. Esse salto — da amostra para a população, do observado para o não observado — é o objeto da inferência estatística, e ele é **logicamente inválido** enquanto não tivermos um modelo do que separa uma coisa da outra.

O que separa é o **acaso**. Se as trinta filiais tivessem sido outras trinta, a média teria dado outro número. A inferência só se torna possível quando conseguimos dizer algo preciso sobre *quanto* e *como* esse número teria mudado — quando o acaso deixa de ser desculpa e vira objeto de cálculo. A probabilidade é a linguagem que faz isso: ela atribui números aos resultados possíveis de um experimento cujo desfecho não controlamos, e o faz com regras suficientemente rígidas para que se possa demonstrar teoremas.

Este capítulo constrói essa linguagem do zero. Começamos com o vocabulário — experimento, espaço amostral, evento — e com os axiomas que qualquer atribuição de probabilidade deve respeitar. Em seguida chegamos ao conceito que é, de longe, o mais importante para a inferência: a **probabilidade condicional**, que formaliza a ideia de *atualizar* uma avaliação de incerteza à luz de informação nova. Dela saem a independência, a Lei da Probabilidade Total e o Teorema de Bayes — e é o Teorema de Bayes, no fundo, que descreve o que um pesquisador faz quando olha para os dados.

3.1 Experimento aleatório, espaço amostral e eventos

Um **experimento aleatório** é qualquer procedimento cujo resultado não pode ser previsto com certeza antes de sua realização, mas cujo conjunto de resultados possíveis é conhecido. Lançar um dado é o exemplo de manual; o retorno de uma ação no próximo pregão, a duração de um período de desemprego ou o resultado de uma eleição são os exemplos que nos interessam. O que os une é a combinação de duas coisas: **incerteza sobre qual resultado ocorrerá** e **conhecimento sobre quais resultados podem ocorrer**.

> **!** Espaço amostral e evento
>
> O **espaço amostral** S é o conjunto de *todos* os resultados possíveis de um experimento aleatório. Cada elemento de S é um **ponto amostral**.
> Um **evento** é qualquer subconjunto de S — isto é, uma coleção de resultados possíveis. Dizemos que o evento A *ocorreu* quando o resultado observado do experimento pertence a A.
> Seguimos aqui a notação dos slides da disciplina, que usa S (de *sample space*) para o espaço amostral, e não o Ω mais comum em textos de teoria da medida.

No lançamento de um dado, $S = \{1, 2, 3, 4, 5, 6\}$. O evento $A =$ "sai face par" é o subconjunto $A = \{2, 4, 6\}$, e o evento $B =$ "sai face maior que 3" é $B = \{4, 5, 6\}$. Repare que um evento não é

uma frase, é um **conjunto** — e é essa tradução, da frase em português para o subconjunto de S, que faz todo o trabalho pesado nos exercícios.

Como eventos são conjuntos, as operações da teoria dos conjuntos ganham imediatamente leitura probabilística:

- a **união** $A \cup B$ é o evento "A ocorre **ou** B ocorre" (ou ambos);
- a **interseção** $A \cap B$ é o evento "A **e** B ocorrem simultaneamente";
- o **complementar** A^c é o evento "A **não** ocorre".

Para o dado, $A \cup B = \{2, 4, 5, 6\}$ e $A \cap B = \{4, 6\}$. Note que o "ou" da união é **inclusivo**: o resultado 4 pertence a $A \cup B$ porque é par *e* maior que 3, e isso não o desqualifica. Essa observação parece trivial, mas é exatamente a origem do termo de correção na Lei da Adição, que veremos em instantes.

> **i** Ferramenta matemática — álgebra de conjuntos
>
> Toda a manipulação de eventos deste capítulo é álgebra de conjuntos: comutatividade e associatividade de $\cup$ e $\cap$, distributividade, e sobretudo as **leis de De Morgan**,
>
> $$(A \cup B)^c = A^c \cap B^c, \qquad (A \cap B)^c = A^c \cup B^c,$$
>
> que traduzem "não ocorre nem A nem B" em "não ocorre A e não ocorre B". Elas resolvem, sozinhas, boa parte dos exercícios em que o enunciado vem na negativa. *(Conjuntos e operações: capítulo* Nivelamento *das Notas de Matemática.)*

3.2 Os axiomas e suas consequências imediatas

Uma **probabilidade** é uma função P que associa a cada evento $A \subseteq S$ um número $P(A)$. Não qualquer função: para merecer o nome, ela precisa satisfazer três exigências.

> **!** Axiomas da probabilidade
>
> Para todo evento $A \subseteq S$:
>
> 1. $0 \leq P(A) \leq 1$ — probabilidades são números entre zero e um;
> 2. $P(S) = 1$ — o espaço amostral inteiro é o evento certo;
> 3. se A e B são **disjuntos** ($A \cap B = \emptyset$), então $P(A \cup B) = P(A) + P(B)$ — probabilidades de eventos que não podem ocorrer juntos se somam.

Vale insistir em quão pouco os axiomas exigem. Eles **não** dizem como calcular $P(A)$; dizem apenas com que regras qualquer proposta de cálculo tem de ser coerente. Essa parcimônia é uma virtude: o mesmo arcabouço serve para o dado honesto, para o retorno de um ativo e para o grau de crença de um analista, porque todos eles, se quiserem ser consistentes, obedecem às mesmas três regras.

Dois eventos merecem nome próprio. O **evento certo** é o próprio S, que ocorre sempre; o **evento impossível** é o conjunto vazio $\emptyset$, que nunca ocorre. Do segundo e do terceiro axioma sai de graça que

$$\boxed{P(\emptyset) = 0}$$

pois S e $\emptyset$ são disjuntos e sua união é S, de modo que $1 = P(S) = P(S) + P(\emptyset)$. Pelo mesmo raciocínio aplicado a A e A^c — disjuntos e com união S — obtemos a **regra do complementar**:

$$\boxed{P(A^c) = 1 - P(A)}$$

que é modesta na aparência e utilíssima na prática: sempre que o evento de interesse for do tipo "pelo menos um", calcule a probabilidade de "nenhum" e subtraia de um.

3.2.1 A abordagem clássica

Resta a questão de onde vêm os números. A resposta mais antiga é a **abordagem clássica**, válida quando o espaço amostral é finito e há razão para supor que todos os pontos amostrais sejam **igualmente prováveis**. Nesse caso,

$$\boxed{P(A) = \frac{\#A}{\#S}}$$

em que $\#A$ é o número de pontos amostrais favoráveis a A e $\#S$ o número total. Para o dado honesto, $P(A) = P(\{2, 4, 6\}) = 3/6 = 1/2$.

A hipótese de equiprobabilidade é uma **hipótese substantiva**, não uma definição — e a maior parte dos erros nessa abordagem vem de aplicá-la a um espaço amostral mal construído. Considere o lançamento de três moedas. Se listarmos os resultados pelo *número de caras*, teremos quatro possibilidades (0, 1, 2 ou 3 caras) que **não** são igualmente prováveis. O espaço amostral correto lista as sequências:

$$S = \{\text{CCC},\ \text{CCK},\ \text{CKC},\ \text{KCC},\ \text{CKK},\ \text{KCK},\ \text{KKC},\ \text{KKK}\},$$

com $\#S = 8$ resultados esses sim igualmente prováveis. O evento "sair exatamente duas caras" é $\{\text{CCK}, \text{CKC}, \text{KCC}\}$, com três pontos amostrais, e portanto

$$P(\text{duas caras}) = \frac{3}{8}.$$

A lição é geral: descreva o espaço amostral no nível de detalhe em que a simetria física do experimento se manifesta, e só então conte.

3.3 A Lei da Adição

O terceiro axioma cobre apenas o caso disjunto. E quando A e B podem ocorrer juntos? A intuição do problema é geométrica: ao somar $P(A)$ e $P(B)$, os pontos amostrais que estão *nos dois* eventos são contados **duas vezes**. Basta descontá-los uma vez:

$$\boxed{P(A \cup B) = P(A) + P(B) - P(A \cap B)}$$

que é a **Lei da Adição**. Quando $A \cap B = \emptyset$ o termo de correção some e recuperamos o terceiro axioma — ou seja, a Lei da Adição é a versão geral, e o axioma é seu caso particular.

Dois eventos com $A \cap B = \emptyset$ são chamados **mutuamente exclusivos** (ou excludentes): a ocorrência de um impede a do outro. "Sair face 2" e "sair face 5" são mutuamente exclusivos; "sair face par" e "sair face maior que 3" não são, pois $\{4, 6\}$ satisfaz ambos.

Exemplo (dois livros). Uma editora avalia dois títulos. A probabilidade de o primeiro ser adotado é $P(A) = 0{,}30$, a do segundo é $P(B) = 0{,}28$, e a de ambos serem adotados é $P(A \cap B) = 0{,}24$. A probabilidade de **pelo menos um** ser adotado é

$$P(A \cup B) = 0{,}30 + 0{,}28 - 0{,}24 = 0{,}34.$$

Sem o desconto, chegaríamos a 0,58 — quase o dobro do valor correto, porque os casos de adoção conjunta, que são a maior parte de cada evento isolado, entrariam duas vezes.

Exemplo (dois dados). Lançam-se dois dados honestos; $\#S = 36$. Qual a probabilidade de a soma ser 7 **ou** 11? Há seis maneiras de somar 7 e duas de somar 11, e os dois eventos são mutuamente exclusivos (uma soma não pode ser 7 e 11 ao mesmo tempo), de modo que

$$P = \frac{6}{36} + \frac{2}{36} = \frac{8}{36} = \frac{2}{9}.$$

Já a probabilidade de a soma ser **maior que 8 ou** as faces serem **iguais** exige a lei completa, porque $(5, 5)$ e $(6, 6)$ satisfazem as duas condições:

$$P = \frac{10}{36} + \frac{6}{36} - \frac{2}{36} = \frac{14}{36} = \frac{7}{18}.$$

3.4 Probabilidade condicional

Chegamos ao conceito central. Até aqui, toda probabilidade foi avaliada em relação ao espaço amostral completo. Mas a situação típica da vida — e da inferência — é outra: alguém nos informa que certo evento ocorreu, e queremos reavaliar a chance de outro à luz dessa informação.

Volte ao dado, com $A =$ "face par" $= \{2, 4, 6\}$ e $B =$ "face maior que 3" $= \{4, 5, 6\}$. Sem informação nenhuma, $P(A) = 3/6 = 1/2$. Suponha agora que alguém observe o resultado e nos diga apenas: *saiu face maior que 3*. O que muda? Os resultados 1, 2 e 3 estão descartados. O espaço amostral **encolheu** para $\{4, 5, 6\}$, e dentro desse novo espaço os resultados favoráveis a A são apenas $\{4, 6\}$. Logo, a probabilidade de A *dado que* B ocorreu é $2/3$, não $1/2$.

! A condicional é uma redução do espaço amostral

Condicionar em B significa **substituir** S **por** B como universo de referência. Todos os resultados fora de B passam a ter probabilidade zero, e os que estão em B têm suas probabilidades reescalonadas — divididas por $P(B)$ — para que voltem a somar 1. É por isso que $P(\cdot \mid B)$ é, ela própria, uma probabilidade legítima: satisfaz os três axiomas, com B no papel de S.

Formalizando essa contagem sobre o espaço reduzido, chegamos à definição:

$$\boxed{P(A \mid B) = \frac{P(A \cap B)}{P(B)}, \qquad P(B) > 0}$$

No exemplo do dado, $P(A \cap B) = P(\{4, 6\}) = 2/6$ e $P(B) = 3/6$, de modo que $P(A \mid B) = (2/6)/(3/6) = 2/3$, confirmando a contagem direta.

! Não confunda $P(A \mid B)$ com $P(A \cap B)$

$P(A \cap B)$ é a probabilidade de que **os dois** eventos ocorram, avaliada *antes* de qualquer informação — o denominador é S. $P(A \mid B)$ é a probabilidade de A **supondo já sabido** que B ocorreu — o denominador é B. Como $P(B) \leq 1$, dividir por $P(B)$ só pode aumentar o número: $P(A \mid B) \geq P(A \cap B)$ sempre.
A confusão é a fonte de erro mais frequente do capítulo, e ela também acontece na direção inversa, entre $P(A \mid B)$ e $P(B \mid A)$ — que são grandezas **diferentes**, ligadas justamente pelo Teorema de Bayes. Confundi-las é a "falácia da taxa base", de que trataremos adiante.

Exemplo (promoção e sexo). Uma empresa registra, entre seus 310 funcionários elegíveis:

	Promovidos	Não promovidos	Total
Masculino	46	184	230
Feminino	8	72	80
Total	**54**	**256**	**310**

Sejam $A =$ "ser promovido" e $B =$ "ser do sexo masculino". A probabilidade **conjunta** de um funcionário sorteado ao acaso ser homem *e* ter sido promovido é

$$P(A \cap B) = \frac{46}{310} = 0{,}1483,$$

enquanto a probabilidade **condicional** de ser promovido *dado que* é homem é

$$P(A \mid B) = \frac{46}{230} = 0{,}20.$$

O contraste entre 0,1483 e 0,20 é a distinção da caixa acima, em números: no primeiro caso o denominador é a empresa inteira; no segundo, apenas a coluna dos homens. Uma tabela de

dupla entrada é, aliás, o melhor instrumento pedagógico para a probabilidade condicional — condicionar é escolher uma linha (ou coluna) e recalcular as proporções dentro dela. Comparando com $P(A \mid B^c) = 8/80 = 0{,}10$, aliás, vê-se que a taxa de promoção entre os homens é o dobro da das mulheres.

3.5 Independência

Há um caso especial em que a informação **não muda nada**: saber que B ocorreu deixa a avaliação de A exatamente onde estava. Dizemos então que A e B são **independentes**:

$$P(A \mid B) = P(A).$$

Substituindo na definição de condicional, $P(A \cap B)/P(B) = P(A)$, e portanto a caracterização equivalente — em geral mais útil, por ser simétrica em A e B e por dispensar $P(B) > 0$:

$$\boxed{A \text{ e } B \text{ independentes } \iff P(A \cap B) = P(A)\,P(B)}$$

! Independentes $\neq$ mutuamente exclusivos

São conceitos **opostos**, não sinônimos. Se A e B são mutuamente exclusivos, então saber que B ocorreu me diz que A **certamente não** ocorreu: $P(A \mid B) = 0$. Isso é informação máxima, o extremo oposto da independência. De fato, para eventos com probabilidade positiva, mutuamente exclusivos implica $P(A \cap B) = 0 \neq P(A)P(B)$, ou seja, **dependentes**.
Mutuamente exclusivos é uma afirmação sobre a *estrutura conjuntista* dos eventos (não se sobrepõem); independência é uma afirmação sobre a *medida de probabilidade* (a proporção de B ocupada por A é a mesma que a de S). Uma se lê no diagrama de Venn, a outra não.

No exemplo do dado, $P(A)P(B) = (1/2)(1/2) = 1/4$ ao passo que $P(A \cap B) = 2/6 = 1/3$. Como $1/3 \neq 1/4$, "face par" e "face maior que 3" **não** são independentes — coerente com o fato de que $P(A \mid B) = 2/3$ difere de $P(A) = 1/2$.

Exemplo (álgebra e literatura). Numa turma, 75% dos alunos foram aprovados em álgebra (A), 84% em literatura (B) e 63% em ambas. Pela Lei da Adição, a probabilidade de um aluno ter sido aprovado em pelo menos uma das duas é

$$P(A \cup B) = 0{,}75 + 0{,}84 - 0{,}63 = 0{,}96.$$

E a probabilidade de aprovação em álgebra entre os aprovados em literatura é

$$P(A \mid B) = \frac{0{,}63}{0{,}84} = 0{,}75 = P(A).$$

Os eventos são **independentes**: ter passado em literatura não altera em nada a chance de ter passado em álgebra. Equivalentemente, $P(A)P(B) = 0{,}75 \times 0{,}84 = 0{,}63 = P(A \cap B)$. Note que a independência aqui é uma *constatação numérica*, não uma hipótese confortável que se assume por conveniência — e é assim que ela deveria ser tratada sempre.

3.6 A Lei da Multiplicação e o diagrama de árvore

Reescrevendo a definição de probabilidade condicional, obtemos a **Lei da Multiplicação**:

$$\boxed{P(A \cap B) = P(A \mid B)\,P(B) = P(B \mid A)\,P(A)}$$

Ela é a ferramenta de cálculo para experimentos que se desenrolam em **etapas**: a probabilidade de um caminho completo é o produto das probabilidades ao longo dele, cada uma condicionada ao que já aconteceu. No caso independente, $P(A \mid B) = P(A)$ e o produto vira $P(A)P(B)$ — mas esse é o caso particular, não a regra.

Exemplo (urna, sem reposição). Uma urna contém 8 bolas azuis e 4 vermelhas. Retiram-se duas bolas **sem reposição**. Seja $A =$ "a primeira é azul" e $B =$ "a segunda é vermelha". Na primeira extração, todas as 12 bolas estão disponíveis, logo

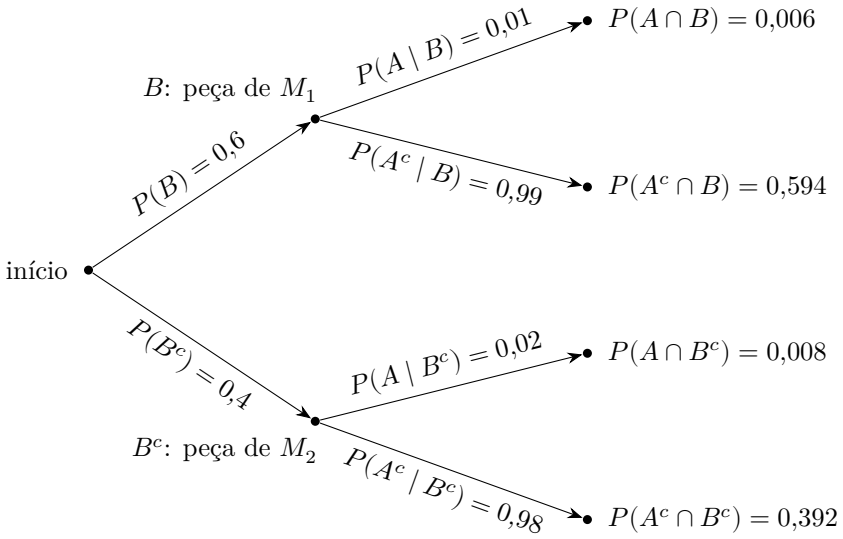
$$P(A) = \frac{8}{12} = \frac{2}{3}.$$

Dado que a primeira foi azul, restam 11 bolas na urna, das quais 4 vermelhas, e portanto $P(B \mid A) = 4/11$. Pela Lei da Multiplicação,

$$P(A \cap B) = P(B \mid A)\,P(A) = \frac{4}{11} \cdot \frac{8}{12} = \frac{8}{33} \approx 0{,}2424.$$

A ausência de reposição é exatamente o que torna as extrações **dependentes**: o resultado da primeira altera a composição da urna e, com ela, as probabilidades da segunda. Com reposição, teríamos $P(B \mid A) = P(B) = 4/12$ e as extrações seriam independentes.

O instrumento gráfico que organiza esses cálculos é o **diagrama de árvore**: cada nó é um estágio do experimento, cada ramo carrega a probabilidade condicional de segui-lo dado o caminho percorrido, e a probabilidade de cada folha é o produto dos ramos que a alcançam. Eis a árvore do exemplo das máquinas que resolveremos adiante:



Duas propriedades da árvore merecem registro, porque servem de verificação: as probabilidades dos ramos que saem de um mesmo nó somam 1, e as probabilidades das folhas — que são os eventos conjuntos, mutuamente exclusivos e exaustivos — também somam 1. No diagrama acima, $0{,}006 + 0{,}594 + 0{,}008 + 0{,}392 = 1$.

3.7 A Lei da Probabilidade Total

Olhe de novo para a árvore. As folhas em que a peça é defeituosa (A) são **duas**: a peça pode ser defeituosa vindo de M_1 ou vindo de M_2. Como essas duas maneiras são mutuamente exclusivas e esgotam todas as possibilidades, a probabilidade de A é a soma delas. Escrevendo cada uma pela Lei da Multiplicação, obtemos a **Lei da Probabilidade Total** (LPT):

$$\boxed{P(A) = P(A \mid B)\,P(B) + P(A \mid B^c)\,P(B^c)}$$

A leitura é iluminadora: a probabilidade não condicional de A é uma **média ponderada** das probabilidades condicionais de A em cada cenário, com pesos iguais às probabilidades dos cenários. Ela responde à pergunta "e se eu não souber em que cenário estou?" — a resposta é: use todos, pesados por sua plausibilidade.

A construção não se limita a dois cenários. Se $B_1, B_2, \ldots, B_k$ formam uma **partição** de S — são mutuamente exclusivos dois a dois e sua união é S —, então

$$\boxed{P(A) = \sum_{i=1}^{k} P(A \mid B_i)\,P(B_i)}$$

Exemplo (rota do caminhão). Um caminhão faz um trajeto por uma de duas rotas. Seja B o evento "toma a rota 1", com $P(B) = 0{,}5$, e A o evento "consome mais diesel que o previsto", com $P(A \mid B) = 0{,}3$ e $P(A \mid B^c) = 0{,}6$. Então

$$P(A \cap B) = P(A \mid B)P(B) = 0{,}3 \times 0{,}5 = 0{,}15,$$

$$P(A) = 0{,}3 \times 0{,}5 + 0{,}6 \times 0{,}5 = 0{,}45,$$

e, pela Lei da Adição, $P(A \cup B) = 0{,}45 + 0{,}50 - 0{,}15 = 0{,}80$. Note que a rota 2 é a consumidora ineficiente (0,6 contra 0,3), e que o 0,45 agregado é apenas a média das duas — um número que não descreve rota nenhuma, mas descreve corretamente a frota.

3.8 O Teorema de Bayes

A LPT percorre a árvore da **esquerda para a direita**: das causas para os efeitos, do cenário para o resultado. Mas a pergunta científica é quase sempre a inversa. O que observamos é o efeito — a peça saiu defeituosa, o teste deu positivo, o candidato foi aprovado — e o que queremos saber é a causa: de que máquina veio, o paciente está doente, o candidato tinha feito o curso? Precisamos percorrer a árvore **da direita para a esquerda**.

O instrumento é o **Teorema de Bayes**, que sai em duas linhas da Lei da Multiplicação. Escrevendo $P(A \cap B)$ das duas maneiras possíveis,

$$P(B \mid A)\,P(A) = P(A \cap B) = P(A \mid B)\,P(B),$$

e isolando o termo de interesse:

$$\boxed{P(B \mid A) = \frac{P(A \mid B)\,P(B)}{P(A)} = \frac{P(A \mid B)\,P(B)}{P(A \mid B)P(B) + P(A \mid B^c)P(B^c)}}$$

> **!** O denominador de Bayes é a LPT
>
> A segunda igualdade é o ponto operacional do teorema: o denominador $P(A)$ quase nunca é dado no enunciado — ele precisa ser **construído pela Lei da Probabilidade Total**, a partir das mesmas condicionais que aparecem no numerador. Na prática, todo exercício de Bayes é um exercício de LPT seguido de uma divisão. Se você souber montar a árvore, o teorema é aritmética.

A estrutura é a da aprendizagem: $P(B)$ é a probabilidade **a priori** do cenário B, antes de observar qualquer coisa; $P(A \mid B)$ é a **verossimilhança**, o quanto o cenário B tornaria provável o que observamos; e $P(B \mid A)$ é a probabilidade **a posteriori**, a crença revisada. Toda a inferência bayesiana é essa fórmula levada a sério.

Exemplo (duas máquinas). Uma fábrica produz peças em duas máquinas. A máquina M_1 responde por 60% da produção ($P(B) = 0{,}6$) e produz 1% de peças defeituosas ($P(A \mid B) = 0{,}01$); a M_2 responde pelos 40% restantes e produz 2% de defeituosas ($P(A \mid B^c) = 0{,}02$). Uma peça é sorteada ao acaso.

Pela LPT, a proporção global de defeituosas é

$$P(A) = 0{,}01 \times 0{,}6 + 0{,}02 \times 0{,}4 = 0{,}006 + 0{,}008 = 0{,}014,$$

isto é, 1,4% — entre 1% e 2%, mais perto do primeiro porque M_1 produz mais. Agora suponha que a peça sorteada **seja defeituosa**. Qual a probabilidade de ter vindo de M_1?

$$P(B \mid A) = \frac{0{,}006}{0{,}014} = 0{,}4286.$$

Repare no movimento: a priori, a peça tinha 60% de chance de vir de M_1; sabendo que é defeituosa, essa chance **cai para 42,9%**. A evidência "defeituosa" aponta para M_2, porque M_2 é a máquina mais

propensa a produzi-la — mas não aponta de modo esmagador, porque M_1 produz muito mais peças. É a tensão entre verossimilhança e taxa base, e é essa tensão que Bayes resolve quantitativamente. Ignorar o $P(B) = 0{,}6$ e concluir "veio de M_2" é cometer a **falácia da taxa base**.

Exemplo (candidato ao MFEE). Seja B = "o candidato fez o curso preparatório", com $P(B) = 0{,}7$, e A = "foi aprovado", com $P(A \mid B) = 0{,}9$ e $P(A \mid B^c) = 0{,}3$. A taxa global de aprovação é

$$P(A) = 0{,}9 \times 0{,}7 + 0{,}3 \times 0{,}3 = 0{,}63 + 0{,}09 = 0{,}72,$$

e, entre os aprovados, a proporção dos que fizeram o curso é

$$P(B \mid A) = \frac{0{,}63}{0{,}72} = 0{,}875.$$

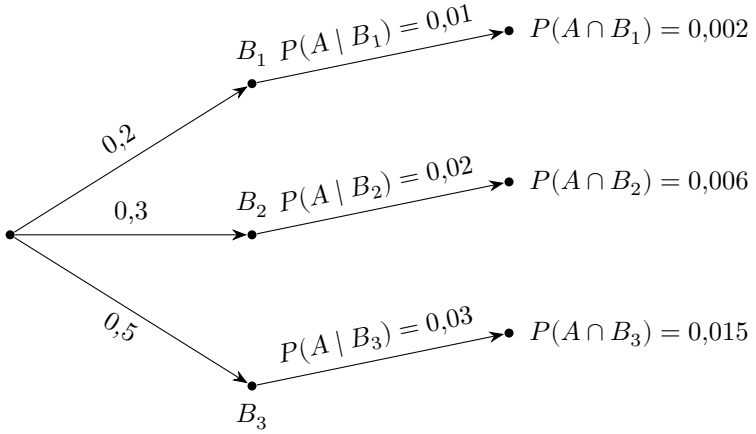
Aqui a evidência confirma: de 70% a priori para 87,5% a posteriori. Vale, contudo, o alerta do capítulo anterior — nada disso é causal. Os 0,9 contra 0,3 podem refletir o efeito do curso ou apenas o fato de que candidatos mais dedicados tanto fazem o curso quanto passam.

3.8.1 Bayes com mais de dois cenários

Quando a partição tem k elementos, a fórmula se generaliza sem novidade conceitual: o denominador é a LPT com k parcelas, e o numerador é a parcela correspondente ao cenário de interesse:

$$\boxed{P(B_j \mid A) = \frac{P(A \mid B_j)\, P(B_j)}{\sum_{i=1}^{k} P(A \mid B_i)\, P(B_i)}}$$

Exemplo (três fornecedores). Três fornecedores respondem por $P(B_1) = 0{,}2$, $P(B_2) = 0{,}3$ e $P(B_3) = 0{,}5$ dos insumos de uma linha de montagem, com taxas de defeito $P(A \mid B_1) = 0{,}01$, $P(A \mid B_2) = 0{,}02$ e $P(A \mid B_3) = 0{,}03$. A árvore tem agora três ramos no primeiro estágio:



Somando os três caminhos que levam a A,

$$P(A) = 0{,}002 + 0{,}006 + 0{,}015 = 0{,}023,$$

e a probabilidade de uma peça defeituosa ter vindo do segundo fornecedor é

$$P(B_2 \mid A) = \frac{0{,}006}{0{,}023} = 0{,}2609.$$

De novo o mesmo padrão: a priori B_2 tinha 30%; a posteriori, 26,1%. Caiu porque B_2 tem taxa de defeito intermediária, e a evidência favorece relativamente o fornecedor mais descuidado, B_3 (cuja posteriori sobe de 50% para $0{,}015/0{,}023 = 65{,}2\%$).

3.9 A independência é condicional: a armadilha do seguro

Encerramos a teoria com o ponto mais sutil do capítulo — aquele que, na minha experiência, separa quem entendeu probabilidade condicional de quem apenas memorizou a fórmula.

Uma seguradora divide seus segurados em dois grupos de risco. O grupo X, que reúne 30% da carteira, tem probabilidade 0,4 de sofrer um sinistro em um dado ano; o grupo Y, com os 70% restantes, tem probabilidade 0,2. Sorteia-se um segurado ao acaso.

(a) Qual a probabilidade de ele sofrer sinistro em 2026? É LPT pura:

$$P(S_{26}) = 0{,}4 \times 0{,}30 + 0{,}2 \times 0{,}70 = 0{,}12 + 0{,}14 = 0{,}26.$$

(b) Dado que sofreu sinistro em 2026, qual a probabilidade de ser do grupo X? É Bayes:

$$P(X \mid S_{26}) = \frac{0{,}12}{0{,}26} = 0{,}4615.$$

O sinistro é evidência de que o segurado pertence ao grupo de risco alto: de 30% a priori para 46,2% a posteriori.

(c) Qual a probabilidade de ele sofrer sinistro em 2026 *e* em 2027? Aqui está a armadilha. A resposta tentadora é multiplicar: $0{,}26 \times 0{,}26 = 0{,}0676$. **Está errada.** A resposta correta é

$$P(S_{26} \cap S_{27}) = 0{,}4^2 \times 0{,}30 + 0{,}2^2 \times 0{,}70 = 0{,}048 + 0{,}028 = 0{,}076.$$

> **!** Independência condicional não implica independência incondicional
>
> A hipótese do problema é que os sinistros de anos distintos são independentes **dentro de cada grupo** — condicionalmente ao grupo. Ela **não** implica que sejam independentes incondicionalmente, e de fato não são.
> A razão é que o grupo é **desconhecido**. Saber que o segurado sofreu sinistro em 2026 é informação sobre *que tipo de segurado ele é* — pelo item (b), eleva a chance de ser do grupo

X de 30% para 46,2% — e um segurado com maior chance de ser do grupo X tem maior chance de sofrer sinistro também em 2027. A dependência entre os anos é **induzida pela heterogeneidade não observada**.
Em números: $P(S_{27} \mid S_{26}) = 0{,}076/0{,}26 = 0{,}2923$, contra $P(S_{27}) = 0{,}26$ incondicional. O primeiro sinistro aumenta a probabilidade do segundo em três pontos percentuais, e é exatamente por isso que $0{,}076 > 0{,}0676$.

Vale enunciar o mecanismo em geral, porque ele reaparece o curso inteiro. Condicionando no grupo $G \in \{X, Y\}$,

$$P(S_{26} \cap S_{27}) = \sum_g P(S_{26} \cap S_{27} \mid G = g)\, P(G = g) = \sum_g p_g^2\, P(G = g) = \mathbb{E}[p_G^2],$$

ao passo que $P(S_{26})P(S_{27}) = (\mathbb{E}[p_G])^2$. A diferença entre os dois é exatamente $\mathbb{E}[p_G^2] - (\mathbb{E}[p_G])^2 = \mathrm{V}(p_G)$, a **variância dos riscos entre grupos** — que aqui vale $0{,}076 - 0{,}0676 = 0{,}0084$. Ou seja: a dependência induzida é nula se e somente se todos os grupos tiverem o mesmo risco. Heterogeneidade *é* dependência.

O leitor reconhecerá, nessa identidade, a fórmula de cálculo da variância de Seção 2 — e ela reaparecerá formalizada como decomposição da variância em Seção 4. Economicamente, é o fenômeno que a literatura de seguros e de mercado de trabalho chama de **dependência espúria de estado**: uma correlação temporal que parece indicar que o passado *causa* o futuro, mas que decorre apenas de diferenças persistentes e não observadas entre indivíduos.

3.10 Laboratório computacional em R

Começamos pelo que o computador faz melhor que nós: **enumerar** o espaço amostral. Para dois dados, S tem 36 pontos, e todas as probabilidades do capítulo viram contagens. Verificamos aqui a Lei da Adição no exemplo "soma maior que 8 ou faces iguais".

```
S <- expand.grid(d1 = 1:6, d2 = 1:6)
S$soma <- S$d1 + S$d2
A <- S$soma > 8      # evento A: soma maior que 8
B <- S$d1 == S$d2    # evento B: faces iguais
round(c(card_S = nrow(S), P_A = mean(A), P_B = mean(B), P_AB = mean(A & B),
        P_AuB_pela_lei = mean(A) + mean(B) - mean(A & B),
        P_AuB_direta   = mean(A | B),
        valor_teorico  = 7 / 18), 4)
```

```
       card_S           P_A           P_B          P_AB P_AuB_pela_lei
      36.0000        0.2778        0.1667        0.0556         0.3889
P_AuB_direta valor_teorico
      0.3889        0.3889
```

As duas formas coincidem, como devem: a Lei da Adição não é uma aproximação, é uma identidade contábil. Repare que `mean()` sobre um vetor lógico é precisamente $\#A/\#S$ — a abordagem clássica em uma função.

A condicional é igualmente direta: basta restringir o denominador ao evento condicionante. Retomamos o dado com $A =$ "par" e $B =$ "maior que 3".

```
Sd <- 1:6
Ad <- Sd %in% c(2, 4, 6)   # face par
Bd <- Sd > 3               # face maior que 3
round(c(P_A = mean(Ad), P_B = mean(Bd), P_AB = mean(Ad & Bd),
        P_A_dado_B = mean(Ad & Bd) / mean(Bd),
        produto_P_A_P_B = mean(Ad) * mean(Bd)), 4)
```

```
          P_A           P_B          P_AB    P_A_dado_B produto_P_A_P_B
       0.5000        0.5000        0.3333        0.6667          0.2500
```

Confirmam-se os 2/3 e a **dependência**: $P(A \cap B) = 0{,}3333$ difere de $P(A)P(B) = 0{,}25$.

Passamos agora à simulação de Monte Carlo, que é a verificação mais convincente de um resultado probabilístico: em vez de manipular fórmulas, **fabricamos** um grande número de realizações do experimento e contamos. Simulamos a fábrica das duas máquinas.

```
set.seed(202601)
n <- 500000
maquina <- ifelse(runif(n) < 0.6, "M1", "M2")              # 60% de M1
defeito <- runif(n) < ifelse(maquina == "M1", 0.01, 0.02) # 1% e 2%
round(c(P_A_simulado = mean(defeito), P_A_teorico = 0.014,
        Bayes_simulado = mean(maquina[defeito] == "M1"),
        Bayes_teorico  = 0.006 / 0.014,
        n_defeituosas  = sum(defeito)), 4)
```

```
  P_A_simulado    P_A_teorico Bayes_simulado  Bayes_teorico  n_defeituosas
        0.0143         0.0140         0.4323         0.4286      7138.0000
```

O `mean(maquina[defeito] == "M1")` merece uma pausa: ele **seleciona** apenas as peças defeituosas e, dentro desse subconjunto, calcula a proporção vinda de M_1. É a redução do espaço amostral, implementada como um filtro. O resultado ronda 0,43, contra os 0,4286 teóricos — e note que a precisão é limitada não pelas 500 mil peças simuladas, mas pelas poucas milhares que saíram defeituosas, que são as únicas que informam sobre $P(B \mid A)$. Condicionar em evento raro custa amostra.

O gráfico a seguir mostra a convergência: a estimativa oscila muito no começo e se aproxima da linha teórica à medida que peças defeituosas se acumulam.

```r
idx  <- which(defeito)
conv <- data.frame(k = seq_along(idx),
                   est = cumsum(maquina[idx] == "M1") / seq_along(idx))
g1 <- ggplot(subset(conv, k >= 10), aes(x = k, y = est)) +
  geom_line(colour = az, linewidth = 0.4) +
  geom_hline(yintercept = 0.006 / 0.014, colour = vm, linetype = "dashed") +
  scale_x_log10() +
  labs(x = "pecas defeituosas acumuladas (escala log)", y = "proporcao vinda de M1")

set.seed(202602)
m      <- 200000
grupo  <- ifelse(runif(m) < 0.30, "X", "Y")
p      <- ifelse(grupo == "X", 0.4, 0.2)
ambos  <- (runif(m) < p) & (runif(m) < p)   # sinistro em 2026 E em 2027
conv2  <- data.frame(k = seq_len(m), est = cumsum(ambos) / seq_len(m))
g2 <- ggplot(subset(conv2, k >= 100), aes(x = k, y = est)) +
  geom_line(colour = vd, linewidth = 0.4) +
  geom_hline(yintercept = 0.076,  colour = vm, linewidth = 0.6) +
  geom_hline(yintercept = 0.0676, colour = cz, linetype = "dotted", linewidth = 0.6) +
  scale_x_log10() + coord_cartesian(ylim = c(0.05, 0.11)) +
  labs(x = "segurados simulados (escala log)", y = "proporcao com sinistro nos 2 anos")
library(patchwork)
g1 + g2
```

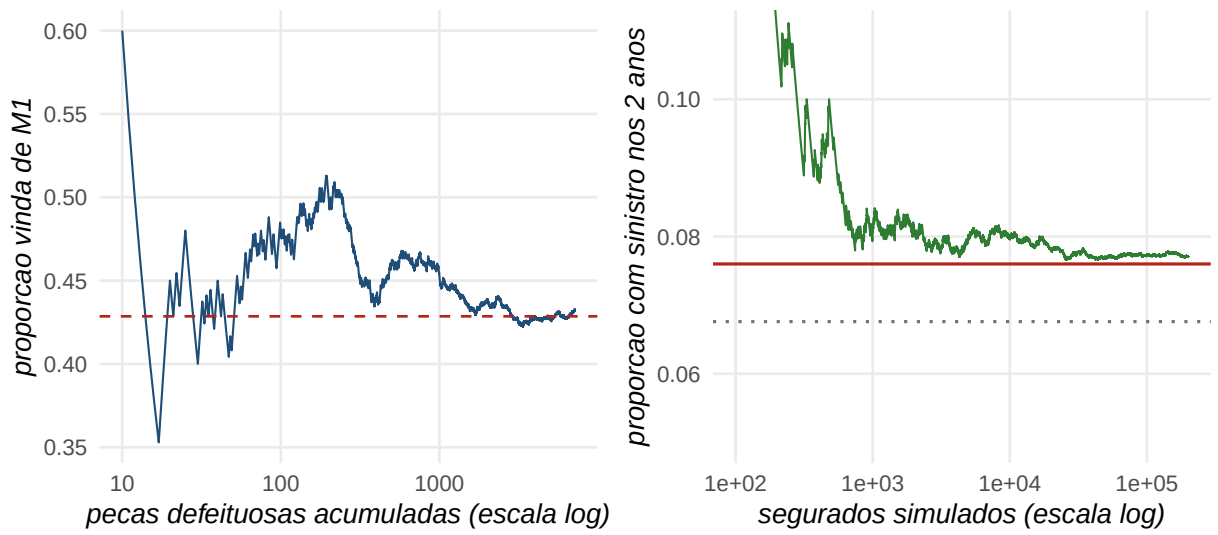


Figura 2: Convergência de Monte Carlo. À esquerda, a estimativa de $P(M_1 \mid \text{defeito})$ conforme se acumulam peças defeituosas, com a linha tracejada em 0,4286. À direita, o exemplo do seguro: a proporção de segurados com sinistro nos dois anos converge para 0,076 (linha cheia), não para $0{,}26^2 = 0{,}0676$ (linha pontilhada).

Por fim, a armadilha do seguro em números. Simulamos 200 mil segurados, cada um sorteado para

um grupo e depois exposto a dois anos de risco, com a probabilidade **do seu grupo** em ambos os anos — que é exatamente a hipótese de independência *condicional*.

```
set.seed(202603)
m     <- 200000
grupo <- ifelse(runif(m) < 0.30, "X", "Y")   # 30% no grupo de risco alto
p     <- ifelse(grupo == "X", 0.4, 0.2)      # risco anual do grupo
s26   <- runif(m) < p                        # sinistro em 2026
s27   <- runif(m) < p                        # sinistro em 2027
round(c(a_P_sinistro_2026     = mean(s26),
        b_P_grupoX_dado_sin   = mean(grupo[s26] == "X"),
        c_P_ambos_os_anos     = mean(s26 & s27),
        se_fossem_independent = mean(s26) * mean(s27),
        P_2027_dado_2026      = mean(s27[s26])), 4)
```

```
    a_P_sinistro_2026   b_P_grupoX_dado_sin     c_P_ambos_os_anos
               0.2617                 0.4662                0.0778
se_fossem_independent      P_2027_dado_2026
               0.0680                 0.2973
```

Os três itens do exercício aparecem nas três primeiras linhas: 0,26, 0,4615 e 0,076. E as duas últimas linhas fecham o argumento — o produto das marginais dá 0,0676, abaixo do valor simulado, e $P(S_{27} \mid S_{26}) \approx 0{,}29$ excede $P(S_{27}) = 0{,}26$. A simulação não *supôs* dependência em lugar nenhum do código: ela emergiu, sozinha, do fato de que o grupo é sorteado uma única vez e vale para os dois anos.

3.11 Exercícios

1. **Espaço amostral e contagem.** Lançam-se três moedas honestas. (a) Escreva S explicitamente. (b) Calcule a probabilidade de sair exatamente duas caras, pelo menos duas caras e nenhuma cara. (c) Por que o espaço amostral $\{0, 1, 2, 3\}$ caras, embora correto como lista de resultados, **não** admite a abordagem clássica $\#A/\#S$?

2. **Axiomas e complementar.** Usando apenas os três axiomas, demonstre que (a) $P(A^c) = 1 - P(A)$; (b) se $A \subseteq B$ então $P(A) \leq P(B)$; (c) $P(A \cup B) \leq P(A) + P(B)$ para quaisquer A, B (desigualdade de Boole). Em que passo de

   (c) a Lei da Adição é usada?

3. **Lei da Adição.** Numa carteira de crédito, 40% dos tomadores têm renda acima da mediana (A), 25% têm garantia real (B) e 15% têm as duas coisas. (a) Calcule $P(A \cup B)$. (b) Calcule $P(A^c \cap B^c)$ usando De Morgan. (c) Os eventos são independentes? E mutuamente exclusivos? Justifique cada resposta separadamente.

4. **Condicional versus conjunta.** Retome a tabela promovidos × sexo. Calcule $P(A \cap B)$, $P(A \mid B)$, $P(B \mid A)$ e $P(A \mid B^c)$, e explique em uma frase o que muda no *denominador* de cada uma. Por que $P(A \mid B) \neq P(B \mid A)$ aqui, e o que seria necessário para que fossem iguais?

5. **Sem reposição.** Da urna com 8 bolas azuis e 4 vermelhas, retiram-se **três** bolas sem reposição. Calcule a probabilidade de (a) as três serem azuis; (b) exatamente uma ser vermelha; (c) pelo menos uma ser vermelha — use o complementar. (d) Refaça (a) supondo extrações **com** reposição e explique a direção da diferença.

6. **Bayes e a taxa base.** Um teste para uma doença que atinge 0,5% da população tem sensibilidade 99% (detecta o doente) e especificidade 95% (não acusa o sadio). (a) Qual a probabilidade de um teste sorteado ao acaso dar positivo? (b) Dado um resultado positivo, qual a probabilidade de a pessoa estar doente? (c) O resultado de (b) surpreende? Explique por que a resposta é baixa mesmo com um teste tão acurado, apoiando-se no exemplo das duas máquinas.

7. **Bayes com três cenários.** Refaça o exemplo dos três fornecedores calculando as três posterioris $P(B_1 \mid A)$, $P(B_2 \mid A)$ e $P(B_3 \mid A)$. (a) Verifique que somam 1 — por que isso *tem* de acontecer? (b) Qual fornecedor teve a maior *variação relativa* entre a priori e a posteriori, e o que na estrutura do problema explica isso?

8. **A armadilha da independência condicional.** Generalize o exemplo do seguro: sejam k grupos com riscos anuais $p_1, \dots, p_k$ e proporções $w_1, \dots, w_k$ na carteira, com os sinistros independentes **dado o grupo**. (a) Mostre que $P(S_{26} \cap S_{27}) - P(S_{26})P(S_{27}) = \mathrm{V}(p_G)$, a variância dos riscos entre grupos.

   (b) Conclua que os sinistros são incondicionalmente independentes **se e somente se** todos os p_i forem iguais. (c) Um atuário que estime o risco agregado em 0,26 e precifique dois anos de cobertura como $0{,}26^2$ está subestimando ou superestimando a exposição da seguradora? Discuta a implicação econômica.

4 Variáveis Aleatórias

Os dois capítulos anteriores contam histórias que ainda não se encontraram. Em Seção 2 trabalhamos com **números** — trinta faturamentos, cinco salários, prazos de entrega — e os resumimos por média, variância, desvio padrão, coeficiente de variação. Em Seção 3 trabalhamos com **eventos** — "sair cara", "o cliente ser inadimplente", "a peça estar defeituosa" — e lhes atribuímos probabilidades. Um mundo é numérico e determinístico; o outro é conjuntista e incerto. Falta a ponte.

A **variável aleatória** é essa ponte. Ela é, em uma frase, a representação *numérica* dos resultados de um experimento aleatório: uma regra que pega cada ponto do espaço amostral e lhe atribui um número. Feito isso, todo o vocabulário do primeiro capítulo passa a estar disponível para o segundo. A média descritiva $\mu = \sum x_i/n$ vira o **valor esperado** $\mathbb{E}(X) = \sum x\,P(X = x)$ — a mesma média ponderada, com as probabilidades no lugar das frequências relativas. A variância descritiva $\sigma^2 = \sum(x_i - \mu)^2/n$ vira $\mathrm{V}(X) = \mathbb{E}[(X - \mathbb{E}(X))^2]$. Cada medida-resumo dos dados observados ganha uma contraparte teórica, definida sobre o modelo probabilístico e não sobre uma tabela de números.

Esse é o fio condutor do capítulo, e vale mantê-lo à vista: **não estamos aprendendo conceitos novos, estamos reescrevendo os antigos em um mundo onde os pesos são probabilidades**. O ganho é enorme. Uma média descritiva descreve o que aconteceu; uma esperança descreve o que o modelo prevê — e portanto permite decidir *antes* de observar, que é o que um economista precisa fazer.

4.1 O que é uma variável aleatória

Um experimento aleatório produz resultados que, em geral, não são números: "cara", "coroa", "cliente pagou", "peça defeituosa". Para calcular médias e variâncias precisamos de números. A variável aleatória faz essa tradução.

> **!** Definição — variável aleatória
>
> Seja S o espaço amostral de um experimento aleatório. Uma **variável aleatória** (v.a.) é uma função
>
> $$X : S \longrightarrow \mathbb{R}$$
>
> que associa a cada resultado $s \in S$ um número real $X(s)$. É, portanto, uma **representação numérica dos resultados** do experimento.
> O adjetivo *aleatória* não se refere à função, que é perfeitamente determinística, mas ao seu argumento: não sabemos qual s ocorrerá, e por isso não sabemos qual número observaremos.

Exemplo (três moedas). Lançamos uma moeda honesta três vezes. O espaço amostral tem oito pontos equiprováveis,

$$S = \{ccc,\ cck,\ ckc,\ kcc,\ ckk,\ kck,\ kkc,\ kkk\},$$

com c para cara e k para coroa. Defina X = número de caras. Então $X(ccc) = 3$, $X(cck) = X(ckc) = X(kcc) = 2$, e assim por diante. A função X colapsa oito resultados qualitativos em quatro números: $0, 1, 2, 3$.

Note o que aconteceu: **eventos viraram valores**. O evento "sair exatamente duas caras" — um subconjunto de S com três elementos — passa a ser escrito simplesmente como $\{X = 2\}$. Probabilidades de eventos viram probabilidades de valores, e é sobre esses valores que faremos aritmética.

> **!** Convenção de notação (usada em todo o curso)
>
> Letras **maiúsculas** — X, Y, Z — designam a **variável aleatória**, isto é, a função $S \to \mathbb{R}$, antes de o experimento ser realizado. Letras **minúsculas** — x, y, z — designam os **valores** que ela pode assumir, isto é, números concretos.
> Assim, $P(X = x)$ lê-se "a probabilidade de que a variável aleatória X assuma o valor x". Escrever $P(x = x)$ ou $P(X = X)$ é erro de notação, e um erro que atrapalha o raciocínio.

4.1.1 Discretas e contínuas

A distinção fundamental é sobre o **conjunto de valores possíveis** de X, chamado seu *suporte*:

- X é **discreta** quando esse conjunto é finito ou infinito enumerável — tipicamente resultados de **contagens**: número de caras, número de clientes inadimplentes, número de defeitos por lote, número de sinistros no mês;
- X é **contínua** quando esse conjunto é um intervalo (ou uma união de intervalos) da reta — tipicamente resultados de **medições**: tempo de espera, consumo de energia, retorno de um ativo, altura, preço.

A distinção não é preciosismo: ela determina se a probabilidade se distribui em *massas pontuais* (discreto, e somamos) ou em *densidade sobre um contínuo* (contínuo, e integramos). Tudo o que vem a seguir aparece em duas versões, uma para cada caso, e o leitor que perceber que são a mesma ideia — com $\sum$ virando $\int$ — terá entendido o capítulo.

4.2 Distribuição de probabilidade

Conhecer a variável aleatória é conhecer não apenas quais valores ela pode assumir, mas **com que probabilidade** cada um deles ocorre. Essa lista completa — valores mais probabilidades — é a **distribuição de probabilidade** de X. Ela é o análogo exato da tabela de frequências de Seção 2, com uma diferença: as frequências relativas observadas dão lugar a probabilidades teóricas.

4.2.1 O caso discreto

! Definição — distribuição de probabilidade discreta

Se X é discreta com valores possíveis $x_1, x_2, ...$, sua distribuição de probabilidade é a função $P(X = x)$, que deve satisfazer duas propriedades:

$$\boxed{\text{(i)} \quad P(X = x) \geq 0 \ \text{ para todo } x; \qquad \text{(ii)} \quad \sum_x P(X = x) = 1}$$

A primeira é a não negatividade dos axiomas de Kolmogorov; a segunda diz que a variável *tem* de assumir algum de seus valores possíveis.

Retomando as três moedas: como os oito pontos de S são equiprováveis, basta contar quantos levam a cada valor de X. Um único resultado dá zero caras (kkk), três dão uma cara, três dão duas e um dá três. Logo

$$P(X = 0) = \tfrac{1}{8}, \qquad P(X = 1) = \tfrac{3}{8}, \qquad P(X = 2) = \tfrac{3}{8}, \qquad P(X = 3) = \tfrac{1}{8},$$

e de fato $\frac{1}{8} + \frac{3}{8} + \frac{3}{8} + \frac{1}{8} = 1$. Trocamos um espaço amostral de oito elementos por uma tabela de quatro linhas sem perder nada do que interessa sobre o número de caras — é a economia de descrição que a variável aleatória proporciona.

4.2.2 O caso contínuo

No contínuo, algo muda de forma decisiva. Se X é o consumo de energia de uma residência em MWh, há uma infinidade não enumerável de valores possíveis. Atribuir probabilidade positiva a cada um deles faria a soma explodir. A saída é atribuir probabilidade a **intervalos**, e descrever como ela se distribui por meio de uma **função densidade de probabilidade** $f(x)$.

! Definição — função densidade de probabilidade

Uma função $f : \mathbb{R} \to \mathbb{R}$ é uma **densidade** da v.a. contínua X quando satisfaz as três propriedades:

$$\boxed{\begin{aligned} &\text{(i)} \quad f(x) \geq 0 \ \text{ para todo } x \in \mathbb{R}; \\ &\text{(ii)} \quad \int_{-\infty}^{\infty} f(x)\,dx = 1; \\ &\text{(iii)} \quad P(X = x) = 0 \ \text{ para todo } x \in \mathbb{R}. \end{aligned}}$$

E a probabilidade de um intervalo é a **área sob a densidade** naquele intervalo:

$$\boxed{P(a \leq X \leq b) = \int_a^b f(x)\,dx}$$

A propriedade (iii) merece parar um instante, porque é contraintuitiva. Ela diz que a probabilidade de X assumir *exatamente* um valor específico é zero — a probabilidade de o consumo ser exatamente 1,0000... MWh, com infinitas casas decimais nulas, é nula. Isso não significa que o evento seja impossível; significa que, num contínuo, a probabilidade só tem substância quando espalhada sobre um intervalo de comprimento positivo. Formalmente, $\int_a^a f(x)\,dx = 0$: um intervalo degenerado tem área nula.

Uma consequência prática, que usaremos o tempo todo: **no contínuo, incluir ou excluir os extremos não altera a probabilidade**,

$$P(a < X < b) = P(a \le X < b) = P(a < X \le b) = P(a \le X \le b).$$

No discreto isso é falso, e a distinção entre $<$ e $\le$ é vital.

> **i** Ferramenta matemática — a integral definida como área
>
> Todo o tratamento do caso contínuo repousa sobre a leitura da integral definida $\int_a^b f(x)\,dx$ como a **área sob o gráfico** de f entre a e b, e sobre o Teorema Fundamental do Cálculo, que a conecta à antiderivada: se $F' = f$, então $\int_a^b f = F(b) - F(a)$. Esse mesmo teorema reaparece adiante como a relação entre a função de distribuição acumulada e a densidade. *(Integral definida, Teorema Fundamental do Cálculo e técnicas de integração: capítulos de* Cálculo Integral *das Notas de Matemática, Profa. Silvia Matos.)*

Exemplo (consumo de energia). Suponha que o consumo mensal X, em MWh, de certa residência tenha densidade

$$f(x) = c\,x^2, \qquad 0 < x < 2,$$

e $f(x) = 0$ fora desse intervalo. Antes de qualquer conta, precisamos determinar c — e a propriedade (ii) o determina sozinha:

$$\int_0^2 c\,x^2\,dx = c\left[\frac{x^3}{3}\right]_0^2 = c\cdot\frac{8}{3} = 1 \quad\Longrightarrow\quad c = \frac{3}{8}.$$

A densidade é portanto $f(x) = \frac{3}{8}x^2$ em $(0, 2)$. Repare que $f(2) = 3/2 > 1$: uma densidade **pode** exceder 1, porque ela não é uma probabilidade — é uma probabilidade *por unidade de* x. O que não pode exceder 1 é a área.

Qual a probabilidade de o consumo superar 1 MWh?

$$P(X > 1) = \int_1^2 \frac{3}{8}x^2\,dx = \frac{3}{8}\left[\frac{x^3}{3}\right]_1^2 = \frac{1}{8}\,(8-1) = \frac{7}{8}.$$

Ou seja, 87,5% dos meses. A densidade crescente x^2 concentra massa à direita, o que torna esse resultado esperado: consumos altos são mais prováveis que baixos nesta residência.

4.3 Valor esperado

Chegamos ao conceito central. A pergunta é a mesma de Seção 2 — em torno de que valor a distribuição se concentra? — mas agora feita sobre um modelo, não sobre dados.

! Definição — valor esperado (ou esperança, ou média)

$$\mathbb{E}(X) = \sum_x x\,P(X = x) \quad \text{(discreto)}, \qquad \mathbb{E}(X) = \int_{-\infty}^{\infty} x\,f(x)\,dx \quad \text{(contínuo)}.$$

Escreve-se também μ ou μ_X. É uma **média ponderada** dos valores possíveis, em que os pesos são as probabilidades.

A conexão com o capítulo anterior é literal. Lá, a média ponderada era $\mu_p = \sum \omega_j x_j / \sum \omega_j$; aqui os pesos são as probabilidades, que já somam 1, de modo que o denominador desaparece. **A esperança é a média ponderada da distribuição.**

Duas observações do professor merecem destaque, porque corrigem o mal-entendido mais comum do curso.

! O que o valor esperado NÃO é

$\mathbb{E}(X)$ **não é um valor que se espera que ocorra**. Em geral, ele sequer é um valor possível da variável. Nas três moedas, $\mathbb{E}(X) = 1{,}5$ caras — e ninguém jamais observará uma cara e meia. No exemplo dos faturamentos de Seção 2, a média 10,3 não constava da tabela.
A leitura correta é mecânica: $\mathbb{E}(X)$ é o **ponto de equilíbrio da distribuição**. Imagine os valores possíveis dispostos sobre uma régua sem peso e, sobre cada um, um peso igual à sua probabilidade. O ponto onde a régua se equilibra é $\mathbb{E}(X)$. É por isso que valores extremos com probabilidade pequena ainda deslocam a esperança: eles têm braço de alavanca longo.

Exemplo (v.a. discreta). Seja X com $P(X = 0) = \frac{1}{2}$, $P(X = 1) = \frac{1}{3}$, $P(X = 2) = \frac{1}{6}$. As três probabilidades somam 1, como deve ser. Então

$$\mathbb{E}(X) = 0 \cdot \tfrac{1}{2} + 1 \cdot \tfrac{1}{3} + 2 \cdot \tfrac{1}{6} = \tfrac{1}{3} + \tfrac{1}{3} = \frac{2}{3}.$$

Novamente, 2/3 não é um valor possível de X — é onde a distribuição se equilibra.

Exemplo (consumo de energia, continuação). Com $f(x) = \frac{3}{8}x^2$ em $(0, 2)$:

$$\mathbb{E}(X) = \int_0^2 x \cdot \frac{3}{8}x^2\,dx = \frac{3}{8}\int_0^2 x^3\,dx = \frac{3}{8}\left[\frac{x^4}{4}\right]_0^2 = \frac{3}{8} \cdot 4 = \frac{3}{2}.$$

O consumo esperado é de 1,5 MWh. Como a densidade é crescente, o ponto de equilíbrio fica à direita do meio do intervalo — coerente com a assimetria da distribuição.

4.3.1 Moda e mediana de uma variável aleatória

As outras duas medidas de posição de Seção 2 também se transportam.

A **moda** Mo é o valor de maior probabilidade (discreto) ou o ponto de máximo da densidade (contínuo). Nas três moedas há empate entre 1 e 2, e a distribuição é bimodal. No consumo de energia, $f(x) = \frac{3}{8}x^2$ é crescente em todo o intervalo, de modo que o máximo está na fronteira: $Mo = 2$.

A **mediana** Md é o valor que parte a distribuição ao meio. No contínuo, é o número k tal que

$$\boxed{\int_{-\infty}^{k} f(x)\,dx = 0{,}5}$$

Para o consumo de energia,

$$\int_0^k \frac{3}{8}x^2\,dx = \frac{k^3}{8} = \frac{1}{2} \quad\Longrightarrow\quad k^3 = 4 \quad\Longrightarrow\quad Md = \sqrt[3]{4} \approx 1{,}5874.$$

Temos aqui, em um único exemplo, as três medidas distintas: $\mathbb{E}(X) = 1{,}5 < Md \approx 1{,}59 < Mo = 2$. Elas só coincidem em distribuições simétricas e unimodais; a assimetria as separa, exatamente como no capítulo descritivo. A Figura 4 mostra as três sobre o gráfico da densidade.

4.4 Variância, desvio padrão e coeficiente de variação

A esperança sozinha é um resumo perigosamente incompleto — a lição do forno e do freezer, transposta. Nada ilustra melhor essa insuficiência do que um problema de decisão de investimento.

Exemplo (dois projetos). Uma empresa avalia dois projetos mutuamente exclusivos, cujos retornos dependem do cenário macroeconômico do próximo ano:

Cenário	Probabilidade	Projeto X	Projeto Y
Recessão	0,2	50.000	0
Estabilidade	0,5	150.000	100.000
Crescimento	0,3	200.000	300.000

Os retornos esperados são

$$\mathbb{E}(X) = 0{,}2(50.000) + 0{,}5(150.000) + 0{,}3(200.000) = 10.000 + 75.000 + 60.000 = 145.000,$$

$$\mathbb{E}(Y) = 0{,}2(0) + 0{,}5(100.000) + 0{,}3(300.000) = 0 + 50.000 + 90.000 = 140.000.$$

Pelo critério do valor esperado, escolhe-se o **projeto** X: R$ 145.000 contra R$ 140.000. A decisão parece resolvida.

Agora a provocação. Suponha que o projeto X tivesse, em vez daqueles, os seguintes retornos:

Cenário	Probabilidade	Projeto X (alternativo)
Recessão	0,2	-500.000
Estabilidade	0,5	220.000
Crescimento	0,3	450.000

O valor esperado é

$$\mathbb{E}(X) = 0{,}2(-500.000) + 0{,}5(220.000) + 0{,}3(450.000) = -100.000 + 110.000 + 135.000 = 145.000,$$

exatamente o mesmo de antes. Pelo critério da esperança, os dois projetos X são indistinguíveis — e ambos continuam batendo Y. Mas o segundo embute uma perda de meio milhão com 20% de chance, o que pode simplesmente quebrar a empresa. Qualquer decisor real distingue os dois casos.

A moral é a mais importante do capítulo: **o valor esperado ignora o risco**. Ele é o primeiro momento da distribuição e nada diz sobre a dispersão em torno dele. Precisamos de uma segunda medida — e ela é, transposta ao mundo probabilístico, exatamente a variância de Seção 2. O tratamento completo da fronteira risco-retorno, com carteiras e diversificação, é o objeto de Seção 7; aqui construímos a ferramenta.

! Definição — variância, desvio padrão e coeficiente de variação

$$\boxed{\mathrm{V}(X) = \mathbb{E}\Big[(X - \mathbb{E}(X))^2\Big]}$$

isto é, o valor esperado do quadrado do desvio em relação à média. Explicitamente,

$$\mathrm{V}(X) = \sum_x (x - \mathbb{E}(X))^2 P(X = x) \quad \text{ou} \quad \mathrm{V}(X) = \int_{-\infty}^{\infty} (x - \mathbb{E}(X))^2 f(x)\, dx.$$

O **desvio padrão** e o **coeficiente de variação** são então

$$\boxed{\mathrm{DP}(X) = \sqrt{\mathrm{V}(X)}} \qquad \boxed{\mathrm{CV}(X) = \frac{\mathrm{DP}(X)}{\mathbb{E}(X)}}$$

As justificativas são as mesmas do capítulo descritivo, e não vale repeti-las: elevamos ao quadrado para que desvios positivos e negativos não se cancelem; tomamos a raiz para recuperar a unidade original; dividimos pela média para comparar dispersões entre variáveis de magnitudes ou unidades diferentes.

4.4.1 A forma de cálculo

Aplicar a definição exige conhecer $\mathbb{E}(X)$ antes de calcular os desvios — duas passagens pela distribuição. Há um atalho algébrico, e é por ele que a variância **costuma ser obtida** na prática. Expandindo o quadrado e usando a linearidade da esperança (que provaremos na próxima seção):

$$\begin{aligned}\mathrm{V}(X) &= \mathbb{E}\big[(X-\mathbb{E}(X))^2\big] = \mathbb{E}\big[X^2 - 2X\,\mathbb{E}(X) + \mathbb{E}^2(X)\big] \\ &= \mathbb{E}(X^2) - 2\,\mathbb{E}(X)\,\mathbb{E}(X) + \mathbb{E}^2(X) = \mathbb{E}(X^2) - \mathbb{E}^2(X).\end{aligned}$$

$$\boxed{\mathrm{V}(X) = \mathbb{E}(X^2) - \mathbb{E}^2(X)}$$

Seguindo a notação do professor, $\mathbb{E}^2(X)$ denota $[\mathbb{E}(X)]^2$ — o *quadrado da média* — e não deve ser confundido com $\mathbb{E}(X^2)$, o *valor esperado do quadrado*. A distinção entre os dois é o conteúdo inteiro da fórmula: a variância é exatamente a diferença entre a média dos quadrados e o quadrado da média, e o fato de essa diferença ser sempre não negativa é a desigualdade de Jensen em sua forma mais simples.

Para calcular $\mathbb{E}(X^2)$ usa-se a mesma regra, aplicada aos quadrados dos valores:

$$\mathbb{E}(X^2) = \sum_{x} x^2\,P(X=x) \qquad \text{ou} \qquad \mathbb{E}(X^2) = \int_{-\infty}^{\infty} x^2 f(x)\,dx.$$

Exemplo (a v.a. discreta). Com $P(X=0)=\frac{1}{2}$, $P(X=1)=\frac{1}{3}$, $P(X=2)=\frac{1}{6}$ e $\mathbb{E}(X)=\frac{2}{3}$:

$$\mathbb{E}(X^2) = 0^2\cdot\tfrac{1}{2} + 1^2\cdot\tfrac{1}{3} + 2^2\cdot\tfrac{1}{6} = \tfrac{1}{3} + \tfrac{4}{6} = 1,$$

$$\mathrm{V}(X) = \mathbb{E}(X^2) - \mathbb{E}^2(X) = 1 - \left(\frac{2}{3}\right)^2 = 1 - \frac{4}{9} = \frac{5}{9} \approx 0{,}5556,$$

e portanto $\mathrm{DP}(X) = \sqrt{5/9} \approx 0{,}745$.

Exemplo (consumo de energia). Com $f(x) = \frac{3}{8}x^2$ em $(0,2)$ e $\mathbb{E}(X)=\frac{3}{2}$:

$$\mathbb{E}(X^2) = \int_0^2 x^2\cdot\frac{3}{8}x^2\,dx = \frac{3}{8}\left[\frac{x^5}{5}\right]_0^2 = \frac{3}{8}\cdot\frac{32}{5} = \frac{12}{5},$$

$$\mathrm{V}(X) = \frac{12}{5} - \left(\frac{3}{2}\right)^2 = \frac{12}{5} - \frac{9}{4} = \frac{48-45}{20} = \frac{3}{20} = 0{,}15,$$

de modo que $\mathrm{DP}(X) = \sqrt{0{,}15} \approx 0{,}387$ MWh e $\mathrm{CV}(X) = 0{,}387/1{,}5 \approx 25{,}8\%$.

4.4.2 Uma decisão de investimento

Exemplo (investimento de risco). Uma aplicação rende 10% com probabilidade 0,4 e perde 4% com probabilidade 0,6. Vale a pena? O retorno esperado é

$$\mathbb{E}(R) = 0{,}4(10\%) + 0{,}6(-4\%) = 4\% - 2{,}4\% = 1{,}6\%.$$

O número é positivo, e a tentação é aprovar o investimento. Mas 1,6% *contra o quê*? Um retorno só se avalia contra a alternativa disponível — o **custo de oportunidade**. Se a poupança rende, digamos, algo em torno de 6% no mesmo período, com risco praticamente nulo, então a aplicação oferece menos retorno esperado **e** carrega o risco de uma perda de 4% com probabilidade 0,6, que é o cenário mais provável dos dois. A conclusão é direta: **não compensa**.

Duas lições ficam. Primeira, a esperança é o insumo de uma comparação, nunca a resposta isolada. Segunda, a comparação relevante em finanças é sempre par a par, retorno *e* risco — o que exige as duas medidas juntas, e é o assunto de Seção 7.

4.5 Propriedades do valor esperado e da variância

O que acontece com $\mathbb{E}$ e V quando transformamos a variável? A pergunta é onipresente: converter dólares em reais, acrescentar um custo fixo, aplicar uma alíquota, padronizar para comparar. As três propriedades a seguir respondem a praticamente tudo que se precisa no caso linear.

! Propriedades de $\mathbb{E}$ e V sob transformações lineares

Sejam a e b constantes reais e X uma variável aleatória.
(1) Constante. Se $Y = b$, então

$$\boxed{\mathbb{E}(Y) = b \qquad \text{e} \qquad \mathrm{V}(Y) = 0}$$

(2) Multiplicação por constante. Se $Y = aX$, então

$$\boxed{\mathbb{E}(Y) = a\,\mathbb{E}(X) \qquad \text{e} \qquad \mathrm{V}(Y) = a^2\,\mathrm{V}(X)}$$

(3) Transformação afim. Se $Y = aX + b$, então

$$\boxed{\mathbb{E}(Y) = a\,\mathbb{E}(X) + b \qquad \text{e} \qquad \mathrm{V}(Y) = a^2\,\mathrm{V}(X)}$$

As demonstrações são de uma linha cada. Para a esperança, tudo decorre da linearidade do somatório (ou da integral): $\sum (ax+b)P(X=x) = a\sum xP(X=x) + b\sum P(X=x) = a\mathbb{E}(X) + b$, usando $\sum P(X=x) = 1$ no último termo. Para a variância,

$$\mathrm{V}(aX+b) = \mathbb{E}\big[(aX + b - a\mathbb{E}(X) - b)^2\big] = \mathbb{E}\big[a^2(X - \mathbb{E}(X))^2\big] = a^2\mathrm{V}(X),$$

porque o b se cancela dentro do parêntese.

E é justamente esse cancelamento que carrega a intuição: **a constante aditiva desloca a distribuição inteira sem alterar sua forma**, e dispersão é uma propriedade da forma, não da posição. Somar R$ 100 a todos os salários aumenta a média em 100 e não muda o desvio padrão em nada. Já a **constante multiplicativa estica a distribuição**, e por isso entra na variância **ao quadrado** — o que faz sentido dimensional, já que a variância está em unidades ao quadrado. No desvio padrão, ela volta a entrar linearmente: $\mathrm{DP}(aX) = |a|\mathrm{DP}(X)$.

Exemplo (produto importado). Uma empresa importa um produto cujo preço em dólares tem $\mathbb{E}(X) = \mu = 80$ e $\mathrm{DP}(X) = \sigma = 8$ — portanto $\mathrm{V}(X) = 64$ e $\mathrm{CV}(X) = 8/80 = 10\%$.

(a) Conversão cambial. Com a taxa de câmbio fixada em 2 R$/US$, o preço em reais é $Y = 2X$. Então

$$\mathbb{E}(Y) = 2(80) = 160, \qquad \mathrm{V}(Y) = 2^2(64) = 256, \qquad \mathrm{DP}(Y) = \sqrt{256} = 16, \qquad \mathrm{CV}(Y) = \frac{16}{160} = 10\%.$$

O preço médio e o desvio padrão dobram, a variância quadruplica — e o coeficiente de variação **não se altera**. Isso é o esperado: CV é adimensional e invariante a mudanças de escala, exatamente como a correlação de Seção 2. A dispersão *relativa* do preço é a mesma, medida em dólares ou em reais.

(b) Aumento fixo. Suponha, alternativamente, que o fornecedor acrescente 10 dólares ao preço: $Z = X + 10$. Então

$$\mathbb{E}(Z) = 80 + 10 = 90, \qquad \mathrm{V}(Z) = 1^2(64) = 64, \qquad \mathrm{DP}(Z) = 8, \qquad \mathrm{CV}(Z) = \frac{8}{90} \approx 8{,}89\%.$$

A média sobe, a variância e o desvio padrão **ficam intactos**, e o CV cai — porque o denominador cresceu enquanto o numerador não. A incerteza absoluta sobre o preço é a mesma de antes; ela apenas pesa relativamente menos sobre um preço maior.

A lição, em uma frase: **a constante multiplicativa entra ao quadrado na variância; a aditiva não a afeta**.

4.5.1 Padronização

Um caso particular de (3) é tão útil que ganha nome. Dada X com $\mathbb{E}(X) = \mu$ e $\mathrm{DP}(X) = \sigma > 0$, defina a variável **padronizada**

$$\boxed{Z = \frac{X - \mu}{\sigma}}$$

que é uma transformação afim com $a = 1/\sigma$ e $b = -\mu/\sigma$. Aplicando as propriedades:

$$\mathbb{E}(Z) = \frac{1}{\sigma}\mathbb{E}(X) - \frac{\mu}{\sigma} = \frac{\mu - \mu}{\sigma} = 0, \qquad \mathrm{V}(Z) = \frac{1}{\sigma^2}\,\mathrm{V}(X) = \frac{\sigma^2}{\sigma^2} = 1.$$

$$\boxed{\mathbb{E}(Z) = 0 \qquad \text{e} \qquad \mathrm{V}(Z) = 1}$$

Padronizar é subtrair a média (centrar em zero) e dividir pelo desvio padrão (colocar em unidades de desvio padrão). O valor z resultante responde à pergunta "a quantos desvios padrão da média está esta observação?", que é *comparável entre variáveis distintas*: um aluno com $z = 1{,}8$ em matemática foi relativamente melhor do que um com $z = 1{,}2$ em história, ainda que as notas brutas digam outra coisa.

Guarde esta transformação. Ela é o mecanismo que permitirá usar uma única tabela — a da Normal padrão — para qualquer distribuição Normal em Seção 6, e é o que dá forma à estatística de teste em Seção 10.

4.6 A função de distribuição acumulada

Até aqui, para calcular $P(a \le X \le b)$ no caso contínuo, resolvemos uma integral. Isso é trabalhoso e, para densidades como a Normal, impossível em forma fechada. Existe um objeto que armazena essa informação de uma vez por todas.

! Definição — função de distribuição acumulada (FDA)

A **função de distribuição acumulada** de uma v.a. X é

$$\boxed{F(x) = P(X \le x), \qquad x \in \mathbb{R}}$$

isto é, a probabilidade acumulada até o ponto x. Ela existe e é definida em toda a reta, para variáveis discretas e contínuas indistintamente — é o objeto mais geral do capítulo.

Suas propriedades decorrem diretamente dos axiomas de probabilidade:

1. $0 \le F(x) \le 1$ para todo x, pois F é uma probabilidade;
2. F é **não decrescente**: se $x_1 < x_2$, então $F(x_1) \le F(x_2)$, porque o evento $\{X \le x_1\}$ está contido em $\{X \le x_2\}$;
3. $\lim_{x \to -\infty} F(x) = 0$ e $\lim_{x \to +\infty} F(x) = 1$;
4. no caso contínuo, F é contínua; no discreto, é uma **função em degraus**, contínua à direita, com um salto de altura $P(X = x)$ em cada valor possível.

Exemplo (FDA discreta). Para $P(X = 0) = \frac{1}{2}$, $P(X = 1) = \frac{1}{3}$, $P(X = 2) = \frac{1}{6}$, a acumulada é

$$F(x) = \begin{cases} 0, & x < 0, \\ \frac{1}{2}, & 0 \le x < 1, \\ \frac{1}{2} + \frac{1}{3} = \frac{5}{6}, & 1 \le x < 2, \\ 1, & x \ge 2. \end{cases}$$

Os degraus estão em 0, $\frac{1}{2}$, $\frac{5}{6}$ e 1, e as alturas dos saltos — $\frac{1}{2}$, $\frac{1}{3}$, $\frac{1}{6}$ — são exatamente as probabilidades pontuais. A distribuição pode ser lida a partir da FDA, e vice-versa: as duas descrições são equivalentes.

4.6.1 A FDA evita integrais

No caso contínuo, o Teorema Fundamental do Cálculo estabelece a ponte entre F e f:

$$F(x) = \int_{-\infty}^{x} f(t)\,dt \qquad \Longleftrightarrow \qquad \boxed{f(x) = \frac{dF(x)}{dx}}$$

a densidade é a derivada da acumulada. E daí segue o resultado que torna a FDA tão prática:

$$\boxed{P(a < X < b) = F(b) - F(a)}$$

Uma vez conhecida F, calcular a probabilidade de qualquer intervalo é uma **subtração**, e não uma integral. É por isso que as tabelas estatísticas (Normal, t, qui-quadrado) tabulam sempre a acumulada, e por que as funções `pnorm`, `pt`, `pexp` do R retornam $F(x)$: a integral já foi resolvida uma vez, numericamente, e o usuário só subtrai. E como no contínuo $P(X = a) = P(X = b) = 0$, **as quatro desigualdades coincidem**:

$$P(a < X < b) = P(a \le X < b) = P(a < X \le b) = P(a \le X \le b) = F(b) - F(a).$$

> **Ferramenta matemática — derivada como inversa da integral**
>
> A relação $f = dF/dx$ é a segunda metade do Teorema Fundamental do Cálculo: derivar a função área devolve o integrando. Na prática, isso significa que se pode navegar livremente entre densidade e acumulada — integrando em um sentido, derivando no outro —, e que basta conhecer uma delas. *(Derivadas, regras de derivação e o Teorema Fundamental do Cálculo: capítulos de* Cálculo Diferencial *e* Cálculo Integral *das Notas de Matemática, Profa. Silvia Matos.)*

Exemplo (da densidade à acumulada). Seja $f(x) = 2x$ em $(0, 1)$. Então, para $0 \le x \le 1$,

$$F(x) = \int_{0}^{x} 2t\,dt = \left[t^2\right]_0^x = x^2,$$

com $F(x) = 0$ para $x < 0$ e $F(x) = 1$ para $x > 1$. Verificação rápida: $F(1) = 1$, como exige a propriedade (3). E $P(0{,}2 < X < 0{,}5) = F(0{,}5) - F(0{,}2) = 0{,}25 - 0{,}04 = 0{,}21$, sem integrar nada.

Exemplo (da acumulada à densidade). Seja $F(x) = 1 - e^{-x}$ para $x \ge 0$ (e $F(x) = 0$ para $x < 0$). Derivando,

$$f(x) = \frac{d}{dx}\left(1 - e^{-x}\right) = e^{-x}, \qquad x \ge 0.$$

Esta é a distribuição **exponencial** de parâmetro 1, que reencontraremos em Seção 6 como o modelo canônico de tempos de espera. Note como F satisfaz as propriedades: é não decrescente, vale 0 em $x = 0$ e tende a 1 quando $x \to \infty$.

4.7 Laboratório computacional em R

O R traz a função `integrate`, que resolve integrais definidas numericamente — o que nos permite verificar todos os resultados do exemplo do consumo de energia sem confiar na conta à mão. A sintaxe é a do professor: define-se a função a integrar e chamam-se os limites.

```
dens <- function(x) (3/8) * x^2        # candidata a densidade em (0,2)
f    <- function(x) x * (3/8) * x^2    # x f(x), para E(X)
f2   <- function(x) x^2 * (3/8) * x^2  # x^2 f(x), para E(X^2)

area <- integrate(dens, 0, 2)$value
EX   <- integrate(f,    0, 2)$value
EX2  <- integrate(f2,   0, 2)$value
c(area = area, EX = EX, EX2 = EX2, VX = EX2 - EX^2,
  DP = sqrt(EX2 - EX^2), CV = sqrt(EX2 - EX^2) / EX)
```

```
     area        EX       EX2        VX        DP        CV
1.0000000 1.5000000 2.4000000 0.1500000 0.3872983 0.2581989
```

A área vale 1, confirmando que $c = 3/8$ é a constante correta; $\mathbb{E}(X) = 1{,}5$, $\mathbb{E}(X^2) = 2{,}4$ e $\mathrm{V}(X) = 0{,}15 = 3/20$, exatamente os valores obtidos analiticamente. A probabilidade da cauda e a mediana também saem por integração — esta última resolvendo $\int_0^k f = 0{,}5$ numericamente com `uniroot`:

```
PX1 <- integrate(dens, 1, 2)$value
med <- uniroot(function(k) integrate(dens, 0, k)$value - 0.5, c(0, 2))$root
c(P_X_maior_1 = PX1, exato_7_8 = 7/8, mediana = med, raiz_cubica_4 = 4^(1/3))
```

```
  P_X_maior_1     exato_7_8       mediana raiz_cubica_4
     0.875000      0.875000      1.587391      1.587401
```

Confere: $P(X > 1) = 7/8$ e $Md = \sqrt[3]{4}$.

Simulação e a lei dos grandes números. A esperança é uma média ponderada teórica. Faz sentido perguntar se ela tem alguma contraparte empírica — e tem: se sorteamos repetidamente valores da distribuição, a **média amostral converge para** $\mathbb{E}(X)$. Vejamos com a v.a. discreta do capítulo.

```
set.seed(20261)
valores <- c(0, 1, 2); probs <- c(1/2, 1/3, 1/6)
EX_d  <- sum(valores * probs)          # E(X)  = 2/3
EX2_d <- sum(valores^2 * probs)        # E(X^2) = 1
amostra <- sample(valores, size = 10000, replace = TRUE, prob = probs)
c(EX_teorico = EX_d, media_amostral = mean(amostra),
  VX_teorico = EX2_d - EX_d^2, var_amostral = mean((amostra - mean(amostra))^2))
```

```
 EX_teorico media_amostral     VX_teorico   var_amostral
  0.6666667      0.6843000      0.5555556      0.5610335
```

A média de dez mil sorteios fica próxima de $2/3 \approx 0{,}6667$, e a variância amostral, de $5/9 \approx 0{,}5556$. O gráfico a seguir mostra a convergência acontecendo:

```
d_lgn <- data.frame(n = 1:10000, media = cumsum(amostra) / (1:10000))
ggplot(d_lgn, aes(n, media)) +
  geom_hline(yintercept = 2/3, colour = vm, linewidth = 0.7) +
  geom_line(colour = az, linewidth = 0.35) +
  scale_x_log10() +
  coord_cartesian(ylim = c(0.3, 1.1)) +
  labs(x = "n (escala log)", y = "media dos n primeiros sorteios")
```

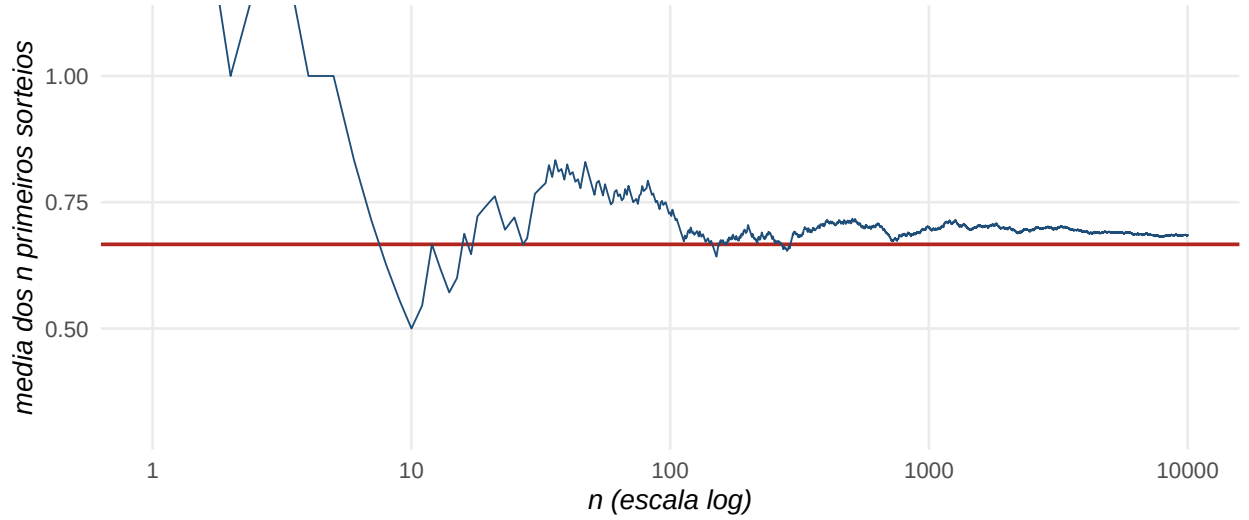


Figura 3: Lei dos grandes números: média dos n primeiros sorteios (azul) convergindo para $\mathbb{E}(X) = 2/3$ (linha vermelha). Eixo horizontal em escala logarítmica.

As oscilações violentas do início cedem lugar a uma trajetória cada vez mais colada ao valor teórico. Esse é o conteúdo da **lei dos grandes números**, que formalizaremos em Seção 7 e que é, no fundo, a razão pela qual a inferência estatística funciona: médias de amostras grandes informam sobre esperanças populacionais.

As propriedades de $\mathbb{E}$ e V, numericamente. Verifiquemos a propriedade (3) com $a = 2$ e $b = 10$, calculando diretamente sobre a distribuição de $Y = 2X + 10$ e comparando com a fórmula:

```
a <- 2; b <- 10
EY <- sum((a * valores + b) * probs)
VY <- sum((a * valores + b - EY)^2 * probs)
rbind(direto  = c(E = EY,           V = VY),
      formula = c(E = a * EX_d + b, V = a^2 * (EX2_d - EX_d^2)))
```

```
               E        V
direto  11.33333 2.222222
formula 11.33333 2.222222
```

As duas linhas coincidem. O caso do produto importado, com $\mu = 80$ e $\sigma = 8$, fica:

```r
mu <- 80; sigma <- 8
rbind("Y = 2X    " = c(E = 2*mu,  V = 4*sigma^2, DP = 2*sigma, CV = (2*sigma)/(2*mu)),
      "Z = X + 10" = c(E = mu+10, V = sigma^2,   DP = sigma,   CV = sigma/(mu+10)))
```

```
             E   V DP         CV
Y = 2X     160 256 16 0.10000000
Z = X + 10  90  64  8 0.08888889
```

Exatamente os números do texto: CV invariante à multiplicação (10% nos dois casos de X e Y) e reduzido a 8,89% pelo acréscimo aditivo, que aumenta a média sem tocar na dispersão.

Média, mediana e moda numa distribuição assimétrica. Por fim, a figura que resume a seção de posição — as três medidas sobre a densidade $f(x) = 3x^2/8$:

```r
gx  <- seq(0, 2, length.out = 400)
d_f <- data.frame(x = gx, f = dens(gx))
marcas <- data.frame(
  medida = factor(c("moda", "mediana", "media"), levels = c("moda", "mediana", "media")),
  x = c(2, 4^(1/3), 3/2))
ggplot(d_f, aes(x, f)) +
  geom_area(fill = az, alpha = 0.18) +
  geom_line(colour = az, linewidth = 0.8) +
  geom_vline(data = marcas, aes(xintercept = x, colour = medida, linetype = medida),
             linewidth = 0.8) +
  scale_colour_manual(values = c(moda = lr, mediana = vd, media = vm)) +
  scale_linetype_manual(values = c(moda = "dotted", mediana = "dashed", media = "solid")) +
  labs(x = "x", y = "f(x) = 3x^2/8")
```

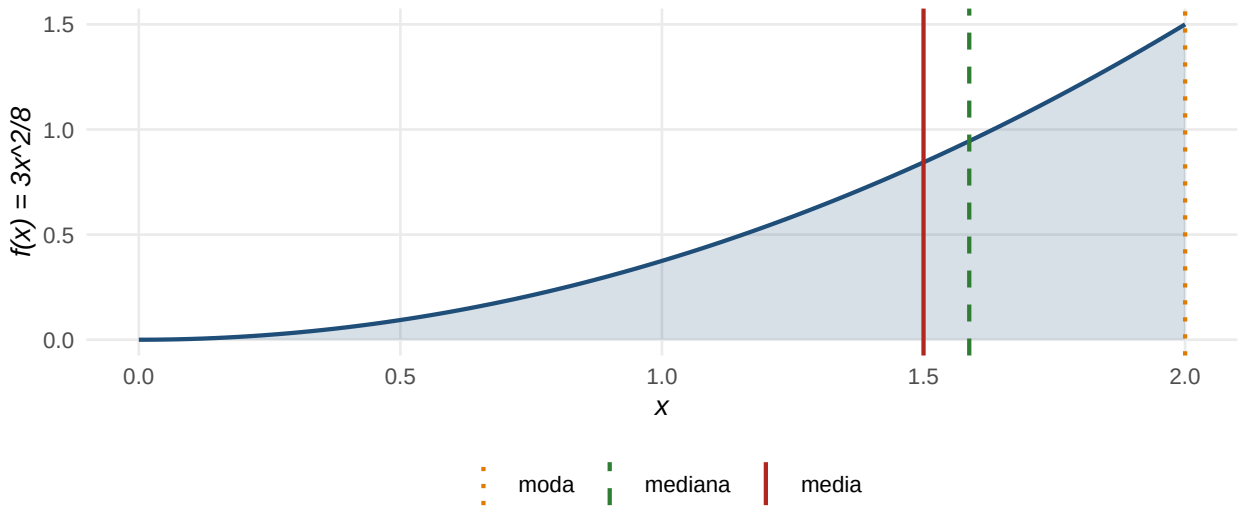


Figura 4: Densidade $f(x) = 3x^2/8$ em $(0, 2)$, com média $3/2$, mediana $\sqrt[3]{4}$ e moda 2. Numa distribuição assimétrica as três medidas de posição diferem.

A densidade cresce em todo o suporte, empurrando a moda para a fronteira direita; a mediana fica entre média e moda. Compare com o histograma dos faturamentos em Seção 2, onde média e mediana quase coincidiam: lá a distribuição era aproximadamente simétrica, aqui não é — e a distância entre as três medidas *é* a assimetria.

4.8 Exercícios

1. **Distribuição de uma contagem.** Lançam-se quatro moedas honestas e define-se X como o número de caras. (a) Construa a distribuição de probabilidade de X e verifique as duas propriedades. (b) Calcule $\mathbb{E}(X)$, $\mathbb{E}(X^2)$ e $\mathrm{V}(X)$. (c) Compare $\mathbb{E}(X)$ com o resultado do caso de três moedas e conjecture a fórmula geral para n lançamentos.

2. **Determinando a constante.** Seja $f(x) = k(1 - x^2)$ para $-1 < x < 1$ e zero fora.

   (a) Determine k. (b) Calcule $\mathbb{E}(X)$ e $\mathrm{V}(X)$. (c) Obtenha a mediana e a moda e explique, apelando à forma da densidade, por que as três medidas coincidem aqui e não coincidem no exemplo do consumo de energia.

3. **Da densidade à acumulada e de volta.** Para $f(x) = \frac{3}{8}x^2$ em $(0, 2)$: (a) obtenha $F(x)$ em toda a reta; (b) recalcule $P(X > 1)$ usando F, sem integrar; (c) derive F e confirme que retorna f; (d) resolva $F(k) = 0{,}5$ e recupere a mediana $\sqrt[3]{4}$.

4. **Por que $P(X = x) = 0$.** Explique, sem apelar à fórmula, por que uma variável aleatória contínua atribui probabilidade nula a qualquer valor pontual. Em seguida, argumente por que isso *não* torna tais valores impossíveis, e discuta o que muda no caso discreto ao trocar $P(a < X < b)$ por $P(a \le X \le b)$.

5. **O valor esperado ignora o risco.** Retome os dois projetos do texto e a tabela alternativa para X. (a) Calcule $\mathrm{DP}(X)$, $\mathrm{DP}(Y)$ e DP do X alternativo.

(b) Ordene as três alternativas por retorno esperado e por risco. (c) Construa uma quarta distribuição, com $\mathbb{E} = 145.000$, que seja preferível a todas as demais para um decisor avesso ao risco. (d) Que critério de decisão você proporia que use as duas medidas?

6. **Propriedades sob transformação.** Uma empresa tem custo total $C = 40.000 + 3Q$, onde a quantidade vendida Q tem $\mathbb{E}(Q) = 10.000$ e $\mathrm{DP}(Q) = 800$. (a) Calcule $\mathbb{E}(C)$, $\mathrm{V}(C)$, $\mathrm{DP}(C)$ e $\mathrm{CV}(C)$. (b) Compare $\mathrm{CV}(C)$ com $\mathrm{CV}(Q)$ e explique a diferença.

   (c) Mostre em geral que $\mathrm{CV}(aX) = \mathrm{CV}(X)$ para $a > 0$, mas $\mathrm{CV}(X + b) \neq \mathrm{CV}(X)$ para $b \neq 0$.

7. **Padronização.** Seja X com $\mu = 50$ e $\sigma = 5$. (a) Verifique diretamente que $Z = (X - 50)/5$ tem $\mathbb{E}(Z) = 0$ e $\mathrm{V}(Z) = 1$. (b) A que valor de X corresponde $z = -1{,}5$?

   (c) Duas provas têm médias 6,0 e 70 e desvios 1,0 e 8. Um aluno tirou 7,5 na primeira e 82 na segunda. Em qual foi relativamente melhor?

8. **FDA em degraus.** Seja X discreta com $P(X = -1) = 0{,}3$, $P(X = 0) = 0{,}2$, $P(X = 1) = 0{,}5$. (a) Escreva e esboce $F(x)$. (b) Calcule $P(X \leq 0)$, $P(X < 0)$ e $P(-1 < X \leq 1)$, atentando às desigualdades. (c) Recupere a distribuição a partir das alturas dos saltos de F. (d) Explique por que essa recuperação não é possível, do mesmo modo, no caso contínuo — e o que a substitui.

5 Distribuições Discretas

O capítulo anterior nos deu a gramática: variável aleatória, distribuição de probabilidade, valor esperado, variância. Com ela, sabemos o que fazer diante de *uma* variável aleatória concreta — listar seus valores, atribuir probabilidades, somar $x\,P(X = x)$. O problema é que esse procedimento recomeça do zero a cada novo experimento, e isso é insustentável. Ninguém constrói a tabela de probabilidades do número de caras em vinte lançamentos ponto a ponto.

A saída é uma mudança de perspectiva que dá à estatística boa parte de seu poder. Em vez de descrever cada variável aleatória individualmente, catalogamos **modelos** — famílias de distribuições que recorrem, indexadas por um punhado de parâmetros. E o critério do catálogo não é matemático, é **estrutural**: cada modelo corresponde a uma *estrutura de experimento*. Se o experimento que temos diante dos olhos tem a estrutura X, sua distribuição é a distribuição X, e todas as fórmulas — probabilidade pontual, esperança, variância — vêm de graça.

O ganho é grande. Reconhecer que "número de peças defeituosas em uma amostra de 20, com sorteio independente e probabilidade fixa de defeito" e "número de acertos de um atirador em 5 tiros" são o *mesmo* experimento, com nomes diferentes, é o que permite resolver os dois com uma única fórmula deduzida uma única vez. Esse é o trabalho intelectual central deste capítulo, e ele é mais de **leitura de enunciado** do que de álgebra: a pergunta a fazer diante de qualquer problema não é "que conta eu faço?", mas **"que estrutura de experimento é esta?"**.

Vamos catalogar cinco modelos discretos. Os quatro primeiros descrevem experimentos com resultado dicotômico — sucesso ou fracasso — e se distinguem pelo que se conta e pelo modo como as repetições se relacionam. O quinto descreve contagens de eventos raros ao longo de um contínuo de tempo ou espaço. Ao final, veremos que dois deles são, em condições apropriadas, boas aproximações de outros dois — o que reduz ainda mais o catálogo, e revela que as famílias não são cinco ilhas, mas um sistema com ligações internas.

> **!** A pergunta que organiza o capítulo
>
> Diante de um problema, percorra este roteiro antes de escrever qualquer fórmula:
>
> 1. **O que estou contando?** Número de sucessos em um número fixo de tentativas? Número de tentativas até o primeiro sucesso? Número de ocorrências em um intervalo?
> 2. **As tentativas são independentes?** Ou o resultado de uma altera as probabilidades das seguintes?
> 3. **Há reposição?** Extrair de uma população finita **sem** reposição destrói a independência — e troca o modelo.
>
> As respostas a essas três perguntas determinam o modelo. As fórmulas vêm depois.

5.1 O ensaio de Bernoulli

Comecemos pelo tijolo elementar. Um **ensaio de Bernoulli** é um experimento aleatório com exatamente dois resultados possíveis, convencionalmente chamados *sucesso* e *fracasso*. A moeda que dá cara ou coroa; o cliente que paga ou não paga; a peça que sai boa ou defeituosa; o eleitor que vota ou não vota em determinado candidato.

Os nomes são neutros — "sucesso" não carrega juízo de valor algum. Numa aplicação de controle de qualidade, o "sucesso" costuma ser justamente a peça defeituosa, porque é o que se está contando. A convenção é: sucesso é o evento de interesse.

! Definição — variável aleatória de Bernoulli

Seja um experimento com dois resultados, sucesso (com probabilidade p) e fracasso (com probabilidade $1-p$). A variável aleatória

$$X = \begin{cases} 1, & \text{se ocorre sucesso,} \\ 0, & \text{se ocorre fracasso,} \end{cases}$$

tem **distribuição de Bernoulli** de parâmetro p, e escrevemos $X \sim \text{Bernoulli}(p)$. Sua distribuição de probabilidade é

$$\boxed{P(X = x) = p^x (1-p)^{1-x}, \qquad x = 0, 1}$$

A fórmula parece uma complicação gratuita de algo trivial, e em certo sentido é: ela apenas escreve $P(X = 1) = p$ e $P(X = 0) = 1 - p$ numa única linha. Verifique substituindo. Com $x = 1$, $p^1(1-p)^0 = p$; com $x = 0$, $p^0(1-p)^1 = 1 - p$. O expoente funciona como um interruptor que liga um fator e desliga o outro.

Essa forma compacta não é adorno. Ela é o que torna possível escrever a verossimilhança de uma amostra de n ensaios como um produto $\prod p^{x_i}(1-p)^{1-x_i}$, cujo logaritmo é uma soma — e daí sai o estimador de máxima verossimilhança de p em Seção 8. Guarde a fórmula por essa razão, não pela conta em si.

5.1.1 Esperança e variância

Os slides não trazem $\mathbb{E}(X)$ e $\mathrm{V}(X)$ da Bernoulli, mas a dedução cabe em três linhas e vale fazê-la, porque todas as demais esperanças do capítulo se apoiam nela. Aplicando a definição de Seção 4 sobre os dois únicos valores possíveis:

$$\mathbb{E}(X) = \sum_{x=0}^{1} x\, P(X = x) = 0 \cdot (1-p) + 1 \cdot p = p.$$

Para a variância, use a forma de cálculo $\mathrm{V}(X) = \mathbb{E}(X^2) - \mathbb{E}^2(X)$. Note que, como X só assume os valores 0 e 1, tem-se $X^2 = X$ — uma variável indicadora é idempotente —, e portanto $\mathbb{E}(X^2) = \mathbb{E}(X) = p$. Logo

$$\mathrm{V}(X) = \mathbb{E}(X^2) - \mathbb{E}^2(X) = p - p^2 = p(1-p).$$

$$\boxed{X \sim \text{Bernoulli}(p) \implies \mathbb{E}(X) = p \qquad \text{e} \qquad \mathrm{V}(X) = p(1-p)}$$

Vale olhar para $\mathrm{V}(X) = p(1-p)$ com atenção, porque essa função reaparece no curso inteiro — no erro padrão de uma proporção em Seção 8, na largura do intervalo de confiança para p, na estatística

de teste. É uma parábola invertida em p, nula em $p = 0$ e em $p = 1$ e máxima em $p = 1/2$, onde vale $1/4$.

A leitura é intuitiva: quando $p = 0$ ou $p = 1$ o resultado é **certo** e não há incerteza alguma — variância zero, como manda a propriedade da constante. A incerteza é máxima quando os dois resultados são igualmente prováveis, porque é aí que menos se sabe sobre o que vai acontecer. Sucesso quase certo e fracasso quase certo são, ambos, situações informativas; meio a meio é o estado de ignorância máxima.

5.2 A distribuição Binomial

Um ensaio de Bernoulli isolado é pouco interessante. O que interessa é repeti-lo. E aqui a estrutura do experimento precisa ser especificada com cuidado, porque é ela — não a fórmula — que define o modelo.

> **!** A estrutura do experimento binomial
>
> Um experimento é **binomial** quando satisfaz as quatro condições:
>
> 1. consiste em uma sequência de n ensaios **idênticos**, com n **fixado de antemão**;
> 2. cada ensaio tem apenas dois resultados possíveis, sucesso e fracasso;
> 3. a probabilidade de sucesso p é **constante** de ensaio para ensaio;
> 4. os ensaios são **independentes**: o resultado de um não afeta os demais.
>
> A variável de interesse é $X =$ **número de sucessos nos** n **ensaios**, e escrevemos $X \sim \text{Bin}(n, p)$.

As condições 3 e 4 são as que se violam na prática, e a violação típica tem uma causa só: **extração sem reposição de uma população finita**. Se tiro uma peça de um lote de dez e não a devolvo, a probabilidade de a segunda ser defeituosa depende do que aconteceu na primeira. Esse caso tem modelo próprio, a hipergeométrica, e é a próxima seção. A condição 1 também merece nota: n é fixo *antes* de observar. Se em vez disso eu decido parar quando obtiver o primeiro sucesso, o número de ensaios vira aleatório — e o modelo vira geométrico.

5.2.1 Deduzindo a fórmula

Vale deduzir a distribuição binomial em vez de apenas exibi-la, porque o raciocínio é o mesmo que resolverá metade dos problemas do capítulo.

Queremos $P(X = x)$: a probabilidade de exatamente x sucessos em n ensaios. Pense primeiro em **uma** sequência específica que produza esse resultado — digamos, os x primeiros ensaios dão sucesso e os $n - x$ restantes dão fracasso. Como os ensaios são independentes, a probabilidade dessa sequência particular é o produto das probabilidades individuais:

$$\underbrace{p \cdot p \cdots p}_{x \text{ vezes}} \cdot \underbrace{(1-p) \cdots (1-p)}_{n-x \text{ vezes}} = p^x (1-p)^{n-x}.$$

Mas essa não é a única sequência com x sucessos. Qualquer **rearranjo** das posições dos sucessos serve, e todos os rearranjos têm exatamente a mesma probabilidade — porque o produto não depende

da ordem dos fatores. Falta então contar de quantas maneiras se pode escolher quais x dos n ensaios serão os bem-sucedidos. Essa é uma combinação de n elementos tomados x a x, e multiplicá-la pela probabilidade comum encerra a dedução.

Ferramenta matemática — análise combinatória

O número de maneiras de escolher x objetos dentre n **sem levar em conta a ordem** é a **combinação**

$$\binom{n}{x} = \frac{n!}{x!\,(n-x)!},$$

com a convenção $0! = 1$. Duas propriedades economizam contas: a simetria $\binom{n}{x} = \binom{n}{n-x}$ — escolher quem entra é o mesmo que escolher quem fica de fora — e os casos triviais $\binom{n}{0} = \binom{n}{n} = 1$, $\binom{n}{1} = n$.

É a *ordem não importar* que justifica usar combinação e não arranjo: as sequências SSF e SFS têm dois sucessos cada, e o que nos interessa é o total, não a posição. *(Princípio fundamental da contagem, permutações, arranjos e combinações: capítulo* Análise Combinatória *das Notas de Matemática, Profa. Silvia Matos.)*

Definição — distribuição Binomial

Se $X \sim \text{Bin}(n, p)$ conta o número de sucessos em n ensaios de Bernoulli independentes com probabilidade de sucesso p, então

$$\boxed{P(X = x) = \binom{n}{x} p^x (1-p)^{n-x}, \qquad x = 0, 1, 2, \ldots, n}$$

com esperança e variância

$$\boxed{\mathbb{E}(X) = np \qquad \text{e} \qquad \text{V}(X) = np(1-p)}$$

A esperança e a variância dispensam demonstração formal aqui — elas seguem de X ser a soma de n Bernoullis independentes, $X = X_1 + \cdots + X_n$, e das propriedades de soma de variáveis aleatórias que estabeleceremos em Seção 7. Mas a intuição é imediata: se cada ensaio contribui com p sucessos em média, n ensaios contribuem com np; se cada um contribui com $p(1-p)$ de variância e eles não interferem entre si, as variâncias se somam.

Repare no papel da independência. Ela é o que permite multiplicar as probabilidades na dedução da fórmula **e** o que permite somar as variâncias. É a hipótese que carrega o modelo inteiro.

5.2.2 Exemplos

Três moedas, de volta. Em Seção 3 calculamos $P(\text{duas caras}) = 3/8$ enumerando o espaço amostral de oito sequências e contando as três favoráveis. Com o modelo binomial, o mesmo resultado sai sem enumerar nada: $X \sim \text{Bin}(3;\, 1/2)$ e

$$P(X=2) = \binom{3}{2}\left(\frac{1}{2}\right)^2\left(\frac{1}{2}\right)^1 = 3\cdot\frac{1}{4}\cdot\frac{1}{2} = \frac{3}{8}.$$

O $\binom{3}{2} = 3$ é literalmente a contagem que fizemos à mão — as três sequências CCK, CKC, KCC —, e $(1/2)^2(1/2)^1 = 1/8$ é a probabilidade comum a cada uma delas. A fórmula não faz mágica; ela apenas automatiza o que já sabíamos fazer. A diferença é que com $n = 20$ a enumeração é impraticável e a fórmula continua funcionando.

O atirador. Um atirador acerta o alvo com probabilidade $p = 2/3$ em cada tiro, e os tiros são independentes. Em 5 tentativas, qual a probabilidade de exatamente 3 acertos? Aqui $X \sim \text{Bin}(5;\, 2/3)$ e

$$P(X=3) = \binom{5}{3}\left(\frac{2}{3}\right)^3\left(\frac{1}{3}\right)^2 = 10\cdot\frac{8}{27}\cdot\frac{1}{9} = \frac{80}{243} \approx 0{,}3292.$$

Curioso notar que $\mathbb{E}(X) = 5\cdot 2/3 \approx 3{,}33$: três acertos é o resultado mais próximo da média, e ainda assim sua probabilidade é de apenas 33%. Não há contradição — a distribuição está espalhada, e nenhum valor isolado concentra muita massa. É a mesma lição de Seção 4: a esperança é o ponto de equilíbrio, não o resultado a esperar.

A prova de múltipla escolha. Uma prova tem 20 questões, cada uma com 5 alternativas e apenas uma correta. Um aluno responde tudo no chute. Como os chutes são independentes e a probabilidade de acerto é $1/5 = 0{,}2$ em cada questão, o número de acertos é $X \sim \text{Bin}(20;\, 0{,}2)$.

A probabilidade de acertar exatamente 6 questões é

$$P(X=6) = \binom{20}{6}(0{,}2)^6(0{,}8)^{14} = 38.760 \times (0{,}2)^6 \times (0{,}8)^{14} \approx 0{,}1091.$$

Se a aprovação exige acertar 6 ou mais, a probabilidade de o chutador passar é uma soma de quinze termos — de $x = 6$ até $x = 20$. Ninguém faz isso: usa-se o complementar, que troca quinze parcelas por seis,

$$P(X \geq 6) = 1 - P(X \leq 5) = 1 - \sum_{x=0}^{5}\binom{20}{x}(0{,}2)^x(0{,}8)^{20-x} \approx 1 - 0{,}8042 = 0{,}1958.$$

É a regra do complementar de Seção 3 em ação, e ela é *a* técnica de cálculo com distribuições discretas: sempre que o evento for do tipo "pelo menos k", passe ao complementar "no máximo $k-1$". No R, isso é `1 - pbinom(5, 20, 0.2)`, e não por acaso o argumento é 5 e não 6 — voltaremos a esse detalhe.

Por fim, o número esperado de acertos é $\mathbb{E}(X) = 20 \times 0{,}2 = 4$, com $\text{V}(X) = 20(0{,}2)(0{,}8) = 3{,}2$ e $\text{DP}(X) \approx 1{,}79$. Quatro acertos em vinte é o que o chute rende em média; a chance de o acaso levar alguém de 4 a 6 acertos existe, mas é de menos de 20%.

i Nota computacional — Binomial no R e no Excel

No R, `dbinom(x, n, p)` devolve $P(X = x)$ e `pbinom(x, n, p)` devolve a acumulada $P(X \leq x)$. A letra inicial segue uma convenção que vale para todas as distribuições: `d` de *density* (probabilidade pontual, no caso discreto), `p` de *probability* (acumulada), `q` de *quantile* (a inversa da acumulada) e `r` de *random* (sorteio).
No Excel, a função correspondente é `DISTR.BINOM(x; n; p; cumulativo)`, em que o quarto argumento é `FALSO` para a pontual e `VERDADEIRO` para a acumulada — o análogo exato do par `dbinom`/`pbinom`.

5.3 A distribuição Hipergeométrica

Retome a condição que mais se viola no modelo binomial: a probabilidade de sucesso constante, garantida pela independência entre ensaios. Ela vale quando se sorteia **com reposição**, ou quando a população é tão grande que retirar alguns elementos não a altera de modo apreciável. Não vale quando se extrai **sem reposição** de uma população pequena.

O exemplo mínimo esclarece. Uma urna tem 3 bolas azuis e 2 vermelhas. A probabilidade de a primeira extraída ser azul é 3/5. E a segunda? Depende: se a primeira foi azul, restam 2 azuis em 4 bolas e a probabilidade cai a 2/4; se foi vermelha, restam 3 azuis em 4 e a probabilidade sobe a 3/4. A probabilidade de sucesso **não é constante** e os ensaios **não são independentes**. O modelo binomial está fora.

! A estrutura do experimento hipergeométrico

Considere uma população **finita** de N elementos, dos quais r são de um tipo (os "sucessos") e $N - r$ são do outro. Extrai-se uma amostra de n elementos **sem reposição**. A variável $X =$ número de sucessos na amostra tem distribuição **hipergeométrica**, e escrevemos $X \sim \text{Hiper}(N, r, n)$.
O que caracteriza o modelo é a **amostragem sem reposição de população finita** — as três palavras importam. Sem reposição, porque com reposição voltaríamos à binomial; finita, porque é a finitude que faz cada extração alterar a composição do que resta.

A dedução é puro raciocínio combinatório, e mais direta que a da binomial. Todas as amostras de tamanho n são igualmente prováveis — estamos na abordagem clássica de Seção 3 —, de modo que basta contar. O total de amostras possíveis é $\binom{N}{n}$. Uma amostra favorável tem x sucessos, escolhidos dentre os r disponíveis de $\binom{r}{x}$ maneiras, e $n - x$ fracassos, escolhidos dentre os $N - r$ de $\binom{N-r}{n-x}$ maneiras. Pelo princípio fundamental da contagem, o número de amostras favoráveis é o produto.

! Definição — distribuição Hipergeométrica

Se $X \sim \text{Hiper}(N, r, n)$,

$$\boxed{P(X = x) = \frac{\binom{r}{x}\binom{N-r}{n-x}}{\binom{N}{n}}}$$

para os valores de x que respeitam $\max(0,\, n-(N-r)) \le x \le \min(n, r)$ — não se pode obter mais sucessos do que existem na população, nem do que o tamanho da amostra permite. A esperança e a variância são

$$\boxed{\mathbb{E}(X) = n\frac{r}{N} \qquad \text{e} \qquad \mathrm{V}(X) = n\frac{r}{N}\left(1 - \frac{r}{N}\right)\frac{N-n}{N-1}}$$

Compare essas duas expressões com as da binomial e a estrutura salta aos olhos. Escrevendo $p = r/N$ para a **proporção de sucessos na população**, a esperança é $\mathbb{E}(X) = np$ — exatamente a da binomial. A amostragem sem reposição **não altera o número esperado de sucessos**: em média, a amostra reproduz a composição da população, com ou sem reposição.

A variância, essa sim, muda. Ela é $np(1-p)$ — a variância binomial — multiplicada por um fator adicional.

! O fator de correção de população finita

O termo

$$\boxed{\frac{N-n}{N-1}}$$

é o **fator de correção de população finita**. Como $n \ge 1$, ele é sempre menor ou igual a 1: **a amostragem sem reposição produz menos variabilidade** que a amostragem com reposição.

A razão é informacional. Sem reposição, cada elemento extraído é informação que não se repete: à medida que a amostra cresce, resta menos incerteza sobre o que sobrou na população. No caso extremo $n = N$ — quando a amostra é a população inteira —, o fator vale $(N-N)/(N-1) = 0$ e a variância se anula, como deve ser: examinando todos os elementos, sabe-se exatamente quantos sucessos há, sem erro nenhum.

No outro extremo, quando n é pequeno diante de N, o fator se aproxima de 1 e a correção some. É esse limite que fundamenta a aproximação da hipergeométrica pela binomial, adiante. O mesmo fator reaparecerá em Seção 8, corrigindo o erro padrão da média em amostragem de populações finitas.

5.3.1 Exemplos

A urna. Uma urna contém 8 bolas azuis e 5 vermelhas, num total de $N = 13$. Extraem-se 4 bolas, sem reposição. Qual a probabilidade de exatamente 3 serem azuis? Com $r = 8$, $n = 4$, $x = 3$:

$$P(X=3)=\frac{\binom{8}{3}\binom{5}{1}}{\binom{13}{4}}=\frac{56\times 5}{715}=\frac{280}{715}\approx 0{,}3916.$$

No R, `dhyper(3, 8, 5, 4)`. Note a ordem dos argumentos, que **não** é a da notação do professor: `dhyper(x, m, n, k)` recebe $m=r$ (o número de sucessos na população), $n=N-r$ (o número de fracassos) e k (o tamanho da amostra). Ou seja, informa-se a decomposição $8+5=13$, e não o total 13 — um erro clássico é passar N onde o R espera $N-r$.

O lote de peças. Um lote tem 10 peças, das quais 4 são defeituosas. Inspecionam-se 5 peças, sem reposição. A probabilidade de encontrar exatamente 2 defeituosas é, com $N=10$, $r=4$, $n=5$:

$$P(X=2)=\frac{\binom{4}{2}\binom{6}{3}}{\binom{10}{5}}=\frac{6\times 20}{252}=\frac{120}{252}\approx 0{,}4762.$$

Este é o cenário canônico do controle de qualidade por amostragem, e ilustra bem por que a binomial não serviria: com $N=10$ e $n=5$, metade do lote é inspecionada, e a composição do que resta muda drasticamente a cada peça retirada.

O traficante. Um traficante transporta 15 comprimidos, dos quais 5 são narcóticos e 10 são aspirinas. A polícia apreende a bagagem e leva 4 comprimidos à análise. Ele será preso se **pelo menos um** dos analisados for narcótico. Qual a probabilidade de ele ser preso?

O evento "pelo menos um" pede o complementar. Com $N=15$, $r=5$, $n=4$:

$$P(X=0)=\frac{\binom{5}{0}\binom{10}{4}}{\binom{15}{4}}=\frac{1\times 210}{1365}=\frac{210}{1365}\approx 0{,}1539,$$

e portanto

$$P(\text{preso})=P(X\geq 1)=1-P(X=0)\approx 1-0{,}1539=0{,}8461.$$

Cerca de 85%. Repare que quatro comprimidos em quinze — pouco mais de um quarto da bagagem — já bastam para uma probabilidade alta de detecção, porque a fração de narcóticos (1/3) é substancial. É a mesma lógica que sustenta a inspeção por amostragem: não é preciso examinar tudo para detectar um problema com alta probabilidade.

5.4 A distribuição Geométrica

Volte à estrutura binomial e altere uma coisa: em vez de fixar o número de ensaios e contar sucessos, fixamos o **número de sucessos em um** e contamos ensaios. A pergunta passa a ser *quantas tentativas até o primeiro sucesso?*

A inversão parece pequena e muda tudo. Agora n não é mais parâmetro — é a própria variável aleatória, e ela não tem limite superior: pode-se, em princípio, tentar indefinidamente sem acertar. O suporte é infinito enumerável.

! A estrutura do experimento geométrico

Repetem-se ensaios de Bernoulli independentes, todos com probabilidade de sucesso p, **até que ocorra o primeiro sucesso**. A variável

$$X = \text{número do ensaio em que ocorre o primeiro sucesso}$$

tem distribuição **geométrica** de parâmetro p, e escrevemos $X \sim \text{Geom}(p)$.
Os ensaios são os mesmos da binomial — independentes e com p constante. O que muda é a **regra de parada**: lá o número de ensaios era fixo e a contagem de sucessos era aleatória; aqui a contagem de sucessos é fixa (um) e o número de ensaios é aleatório.

A fórmula sai sem esforço. Para que o primeiro sucesso ocorra exatamente no ensaio x, é necessário e suficiente que os $x-1$ primeiros ensaios sejam fracassos e o x-ésimo seja sucesso. Essa é *uma única* sequência — não há o que rearranjar, e por isso não aparece coeficiente combinatório —, e a independência permite multiplicar:

$$\underbrace{(1-p)(1-p)\cdots(1-p)}_{x-1 \text{ fracassos}} \cdot p.$$

! Definição — distribuição Geométrica

Se $X \sim \text{Geom}(p)$,

$$\boxed{P(X = x) = (1-p)^{x-1}p, \qquad x = 1, 2, 3, ...}$$

com

$$\boxed{\mathbb{E}(X) = \frac{1}{p} \qquad \text{e} \qquad \text{V}(X) = \frac{1-p}{p^2}}$$

Que $\mathbb{E}(X) = 1/p$ é um dos resultados mais memoráveis da estatística elementar, porque é exatamente o que a intuição sugere: se acerto um em cada dez, espero dez tentativas até acertar; se a probabilidade é 2/3, espero 1,5 tentativa. **A espera média é o inverso da chance.**

i Ferramenta matemática — série geométrica

O modelo tem infinitos valores possíveis, e é preciso verificar que as probabilidades ainda somam 1. Usando a soma da série geométrica de razão $q = 1-p$, com $|q| < 1$,

$$\sum_{k=0}^{\infty} q^k = \frac{1}{1-q},$$

obtemos

$$\sum_{x=1}^{\infty}(1-p)^{x-1}p = p\sum_{k=0}^{\infty}(1-p)^k = p \cdot \frac{1}{1-(1-p)} = \frac{p}{p} = 1,$$

como exige a propriedade (ii) de Seção 4. A mesma série, derivada termo a termo —

$\sum_{k\geq 1} kq^{k-1} = 1/(1-q)^2$ —, produz $\mathbb{E}(X) = 1/p$, e a acumulada tem forma fechada:

$$P(X \leq x) = 1 - (1-p)^x,$$

que se lê diretamente: falhar nas x primeiras tentativas tem probabilidade $(1-p)^x$. *(Séries geométricas, convergência e soma: capítulo* Sequências e Séries *das Notas de Matemática, Profa. Silvia Matos.)*

! Cuidado — o deslocamento de índice em `dgeom`

Existem **duas convenções** para a geométrica, e o R adota a que *não* é a do curso. Nestas notas, X conta o **número de ensaios até o primeiro sucesso**, com suporte $\{1, 2, 3, \dots\}$. O R, na função `dgeom`, define a variável como o **número de fracassos antes do primeiro sucesso**, com suporte $\{0, 1, 2, \dots\}$. As duas diferem por uma unidade: $X_{\mathrm{R}} = X - 1$. Logo, para calcular $P(X = x)$ na convenção do curso é preciso escrever

```
dgeom(x - 1, p)
```

e, para a acumulada, `pgeom(x - 1, p)`. Esquecer o -1 é o erro mais comum do capítulo em laboratório, e ele é silencioso: o R devolve um número perfeitamente plausível, apenas errado. As esperanças também diferem — $\mathbb{E}(X) = 1/p$ contra $\mathbb{E}(X_{\mathrm{R}}) = (1-p)/p$ —, embora as variâncias coincidam, porque somar uma constante não altera a dispersão.

5.4.1 Exemplos

O atirador, de novo. Aquele atirador com $p = 2/3$ atira até acertar. Qual a probabilidade de o primeiro acerto ocorrer somente no quinto tiro? Isso exige quatro erros seguidos e, no quinto tiro, um acerto:

$$P(X = 5) = \left(\frac{1}{3}\right)^4 \cdot \frac{2}{3} = \frac{1}{81} \cdot \frac{2}{3} = \frac{2}{243} \approx 0{,}0082.$$

Menos de 1%, e a razão é clara: com $p = 2/3$ o atirador é bom, $\mathbb{E}(X) = 1{,}5$ tiros, e errar quatro seguidos é raro. A geométrica decai geometricamente — daí o nome —, e o decaimento é tão mais rápido quanto maior for p.

Os pênaltis. Um jogador desperdiça pênaltis com probabilidade 0,1, e as cobranças são independentes. Duas perguntas parecidas que exigem modelos diferentes:

(a) Em 5 cobranças, qual a probabilidade de desperdiçar exatamente uma? Aqui o número de tentativas é fixo e contamos desperdícios: é binomial, $X \sim \mathrm{Bin}(5;\, 0{,}1)$,

$$P(X = 1) = \binom{5}{1} (0{,}1)^1 (0{,}9)^4 = 5 \times 0{,}1 \times 0{,}6561 = 0{,}32805.$$

(b) Qual a probabilidade de o primeiro desperdício ocorrer na quinta cobrança? Agora contamos tentativas até o primeiro sucesso: é geométrica, $X \sim \mathrm{Geom}(0{,}1)$,

$$P(X = 5) = (0{,}9)^4(0{,}1) = 0{,}6561 \times 0{,}1 = 0{,}06561.$$

O contraste é instrutivo, e é a razão de o professor pôr os dois itens no mesmo enunciado. Os dois cálculos usam os mesmos ingredientes, $(0{,}9)^4$ e 0,1; o que os separa é o fator $\binom{5}{1} = 5$. Em (a), o desperdício pode ocorrer em qualquer uma das cinco cobranças, e há cinco arranjos favoráveis. Em (b), o desperdício tem de ocorrer na quinta, e apenas nela — a posição está fixada, e o coeficiente desaparece. Exatamente por isso $0{,}32805 = 5 \times 0{,}06561$.

5.4.2 Uma condicional que engana

O exemplo a seguir é uma armadilha, e o professor a assinala explicitamente em aula. Ele casa este capítulo com Seção 3 e vale por vários exercícios.

Enunciado. Com o mesmo jogador de pênaltis, $X \sim \text{Geom}(0{,}1)$: sabendo que ele já cobrou três pênaltis **sem desperdiçar nenhum**, qual a probabilidade de o primeiro desperdício ocorrer até a quinta cobrança?

A informação "já cobrou três sem desperdiçar" é o evento $\{X > 3\}$, e o que se pede é a probabilidade **condicional**

$$P(X \le 5 \mid X > 3) = \frac{P(\{X \le 5\} \cap \{X > 3\})}{P(X > 3)} = \frac{P(X = 4) + P(X = 5)}{1 - P(X \le 3)}.$$

A interseção merece um segundo de atenção: os valores de X que são simultaneamente ≤ 5 e > 3 são apenas 4 e 5. O numerador é, portanto,

$$P(X = 4) + P(X = 5) = (0{,}9)^3(0{,}1) + (0{,}9)^4(0{,}1) = 0{,}0729 + 0{,}06561 = 0{,}13851.$$

O denominador é a probabilidade de sobreviver às três primeiras cobranças, que a forma fechada da acumulada dá de imediato:

$$P(X > 3) = (0{,}9)^3 = 0{,}729.$$

Logo

$$P(X \le 5 \mid X > 3) = \frac{0{,}13851}{0{,}729} = 0{,}19.$$

> **!** A pegadinha
>
> "Um olhar desatento calcularia $P(X = 4) + P(X = 5) = 0{,}1385$" — e pararia aí, ignorando a condicionante. O erro é esquecer que a informação recebida **encolheu o espaço amostral**, no sentido preciso de Seção 3: já não estamos no universo de todas as sequências possíveis, mas apenas naquelas em que as três primeiras cobranças foram convertidas. Nesse universo menor, os casos favoráveis pesam mais — e por isso a resposta correta, 0,19, é **maior** que a ingênua, 0,1385.

Dividir por $P(X > 3) = 0{,}729$ é exatamente o reescalonamento que devolve a soma 1 dentro do novo universo. Sempre que o enunciado disser "sabendo que", "dado que" ou "já tendo ocorrido", há um denominador a escrever.

Vale registrar um fato que este exemplo quase revela. Calculando na mesma linha $P(X \leq 5 \mid X > 3)$ e $P(X \leq 2) = 1 - (0{,}9)^2 = 0{,}19$, obtemos o **mesmo número**. Não é coincidência: a geométrica é a única distribuição discreta **sem memória** — dado que ainda não houve sucesso após k tentativas, a distribuição do tempo *adicional* de espera é idêntica à distribuição original. O jogador que já converteu três pênaltis não está "mais perto" nem "mais longe" de desperdiçar do que estava no começo. Sua contraparte contínua, a exponencial, tem a mesma propriedade e aparece em Seção 6.

5.5 A distribuição de Poisson

Os quatro modelos anteriores contam sucessos em **ensaios discretos** identificáveis: tiros, questões, cobranças, peças. Há uma classe de fenômenos em que isso não faz sentido. Quantos acidentes ocorrem em uma rodovia por hora? Quantos clientes chegam a uma agência bancária em dez minutos? Quantos defeitos há por metro quadrado de tecido? Quantos sinistros uma seguradora registra por mês?

Nesses casos não há um "n" — não existe um número identificável de oportunidades de acidente dentro de uma hora. O que existe é uma **taxa média de ocorrência** por unidade de tempo ou de espaço, e as ocorrências se distribuem ao longo desse contínuo de modo independente.

! A estrutura do experimento de Poisson

Conta-se o número de ocorrências de um evento em um **intervalo** de tempo, de espaço, de área ou de volume, sob as hipóteses de que:

1. a probabilidade de ocorrência é **proporcional ao tamanho do intervalo** — dobrar o intervalo dobra o número esperado de eventos;
2. ocorrências em intervalos **disjuntos são independentes**;
3. a probabilidade de **duas ou mais** ocorrências em um intervalo muito pequeno é desprezível — os eventos não se acumulam em um mesmo instante.

Sob essas hipóteses, $X =$ número de ocorrências no intervalo tem distribuição de **Poisson** de parâmetro λ, e escrevemos $X \sim \text{Poi}(\lambda)$. O parâmetro λ é o **número médio de ocorrências no intervalo considerado**.

! Definição — distribuição de Poisson

Se $X \sim \text{Poi}(\lambda)$, com $\lambda > 0$,

$$\boxed{P(X = x) = \frac{\lambda^x e^{-\lambda}}{x!}, \qquad x = 0, 1, 2, \ldots}$$

e, o que é uma peculiaridade do modelo,

$$\boxed{\mathbb{E}(X) = \lambda \qquad \text{e} \qquad \mathrm{V}(X) = \lambda}$$

isto é, **média e variância coincidem**.

A igualdade $\mathbb{E}(X) = \mathrm{V}(X) = \lambda$ é a assinatura da Poisson, e tem consequência prática imediata: o modelo tem um único parâmetro, de modo que fixar a média já fixa toda a dispersão. Não há liberdade para ajustar as duas separadamente. Quando dados de contagem exibem variância bem maior que a média — situação chamada **superdispersão**, comum em econometria de contagem —, isso é evidência empírica *contra* a Poisson, e o diagnóstico é uma simples comparação de dois números.

i Ferramenta matemática — a expansão de e^x

Que as probabilidades de Poisson somam 1 decorre da série de Maclaurin da exponencial,

$$e^{\lambda} = \sum_{x=0}^{\infty} \frac{\lambda^x}{x!} = 1 + \lambda + \frac{\lambda^2}{2!} + \frac{\lambda^3}{3!} + \cdots,$$

válida para todo $\lambda \in \mathbb{R}$. De fato,

$$\sum_{x=0}^{\infty} \frac{\lambda^x e^{-\lambda}}{x!} = e^{-\lambda} \sum_{x=0}^{\infty} \frac{\lambda^x}{x!} = e^{-\lambda} e^{\lambda} = 1.$$

O fator $e^{-\lambda}$ é, portanto, exatamente a **constante de normalização** que faz a série somar 1 — não é um enfeite da fórmula, é o que a torna uma distribuição de probabilidade. Derivando a mesma série obtém-se $\mathbb{E}(X) = \lambda$. *(Série de Taylor e Maclaurin, função exponencial: capítulos de* Sequências e Séries *e* Funções Elementares *das Notas de Matemática, Profa. Silvia Matos.)*

5.5.1 A reescala de λ

Antes dos exemplos, um ponto que merece destaque próprio porque é onde a maior parte dos erros acontece.

! λ é uma taxa — e taxas se reescalam

O parâmetro λ **não é uma característica fixa do fenômeno**: é o número médio de ocorrências **no intervalo que se está considerando**. Trocar o intervalo troca λ.
Pela hipótese 1 da estrutura de Poisson, a proporcionalidade é exata. Se a taxa é λ_0 por unidade de tempo e o intervalo de interesse tem t unidades, então

$$\boxed{\lambda = \lambda_0\, t}$$

Uma taxa de 3 acidentes por hora significa $\lambda = 3$ para uma hora, $\lambda = 6$ para duas horas, $\lambda = 1{,}5$ para meia hora e $\lambda = 1$ para vinte minutos. **Antes de aplicar a fórmula, verifique se a unidade de λ e a unidade do intervalo da pergunta são a mesma.** É onde os alunos erram, e é um erro que a fórmula não denuncia.

5.5.2 Exemplos

A rodovia. Em determinado trecho de rodovia ocorrem, em média, 3 acidentes por hora.

(a) Qual a probabilidade de ocorrerem exatamente 2 acidentes em uma hora? O intervalo da pergunta é uma hora, o mesmo em que a taxa está expressa. Logo $\lambda = 3$ e

$$P(X = 2) = \frac{3^2 e^{-3}}{2!} = \frac{9e^{-3}}{2} = 4{,}5\, e^{-3} \approx 0{,}2240.$$

(b) Qual a probabilidade de ocorrerem 2 ou mais acidentes em vinte minutos? Agora o intervalo mudou. Vinte minutos são um terço de hora, de modo que

$$\lambda = 3 \times \frac{20}{60} = 3 \times \frac{1}{3} = 1.$$

Usar $\lambda = 3$ aqui seria responder à pergunta errada — a probabilidade de 2 ou mais acidentes em uma *hora*, que é bem maior. Com $\lambda = 1$ e a regra do complementar:

$$P(X \geq 2) = 1 - P(X = 0) - P(X = 1) = 1 - \frac{1^0 e^{-1}}{0!} - \frac{1^1 e^{-1}}{1!} = 1 - e^{-1} - e^{-1} = 1 - 2e^{-1} \approx 0{,}2642.$$

Cerca de 26%. Repare que a conta usou duas ferramentas de uma vez: a reescala de λ e o complementar. Os dois passos são quase automáticos depois de vistos, e quase sempre esquecidos antes.

> **Nota computacional — Poisson no R**
>
> `dpois(x, lambda)` devolve $P(X = x)$ e `ppois(x, lambda)` a acumulada $P(X \leq x)$. O item (b) acima é `1 - ppois(1, 1)` — e não `1 - ppois(2, 1)`. A regra é a mesma para toda função `p` do R: `p*(k, ...)` inclui k, de modo que $P(X \geq k) = 1 - P(X \leq k-1)$ se escreve com $k-1$ como argumento. Nas distribuições discretas, esse deslocamento de uma unidade é a fonte de erro mais frequente, e a razão pela qual convém sempre traduzir o enunciado para uma acumulada antes de digitar.

5.6 Aproximações entre modelos

Os cinco modelos não são cinco ilhas. Sob certas condições um deles se torna computacionalmente conveniente no lugar de outro, com erro desprezível — e as relações entre eles são, além de práticas, esclarecedoras sobre a natureza de cada um.

5.6.1 Hipergeométrica pela Binomial

A hipergeométrica difere da binomial porque a extração sem reposição altera as probabilidades ao longo da amostragem. Mas *quanto* ela altera depende do tamanho relativo da amostra. Se a população tem um milhão de elementos e retiramos dez, a composição do que resta praticamente não muda: a probabilidade de sucesso é, para todos os efeitos, constante, e os ensaios são quase independentes.

! Aproximação Hipergeométrica → Binomial

Quando a população é grande relativamente à amostra, especificamente quando

$$\boxed{N \geq 20\,n}$$

isto é, quando a amostra não excede 5% da população, pode-se aproximar

$$\text{Hiper}(N, r, n) \;\approx\; \text{Bin}\Big(n,\; p = \frac{r}{N}\Big)\,.$$

A justificativa formal está no fator de correção: com $n \leq N/20$, tem-se $(N-n)/(N-1) \geq 0{,}95$, de modo que as variâncias dos dois modelos diferem em menos de 5%. As esperanças, como vimos, já são exatamente iguais.

Exemplo (a pesquisa eleitoral). Uma cidade tem 1.000 habitantes, dos quais 300 são eleitores de certo candidato. Sorteia-se uma amostra de 10 pessoas, sem reposição. Qual a probabilidade de exatamente 5 serem eleitores desse candidato?

O modelo exato é Hiper$(1000, 300, 10)$:

$$P(X = 5) = \frac{\binom{300}{5}\binom{700}{5}}{\binom{1000}{10}} \approx 0{,}1026.$$

Como $N = 1000 \geq 20 \times 10 = 200$, a condição está satisfeita com folga, e podemos usar Bin$(10;\, 0{,}3)$ com $p = 300/1000 = 0{,}3$:

$$P(X = 5) \approx \binom{10}{5}(0{,}3)^5(0{,}7)^5 = 252 \times 0{,}00243 \times 0{,}16807 \approx 0{,}1029.$$

A diferença aparece na quarta casa decimal — irrelevante para qualquer uso prático. E a economia de esforço é considerável: os coeficientes $\binom{1000}{10}$ e $\binom{300}{5}$ são números astronômicos, ao passo que $\binom{10}{5} = 252$ se calcula de cabeça.

Essa aproximação é, aliás, a razão pela qual quase toda a teoria de amostragem do curso usa a binomial. Pesquisas eleitorais reais amostram alguns milhares de pessoas de um eleitorado de dezenas de milhões: a fração amostrada é ínfima, o fator de correção é indistinguível de 1, e a distinção entre com e sem reposição perde qualquer relevância prática.

5.6.2 Binomial pela Poisson

A segunda aproximação vai em outra direção. Considere uma binomial com n muito grande e p muito pequeno — eventos raros observados em muitas oportunidades. Cada tentativa isolada quase nunca dá sucesso, mas há tantas tentativas que alguns sucessos acabam ocorrendo. Isso é precisamente a descrição informal de um processo de Poisson: ocorrências raras espalhadas por um contínuo grande.

! Aproximação Binomial → Poisson

Quando n é grande e p é pequeno, de modo que o produto np permaneça moderado,

$$\boxed{\text{Bin}(n,p) \;\approx\; \text{Poi}(\lambda), \qquad \lambda = np}$$

Formalmente, $\text{Bin}(n,p) \to \text{Poi}(\lambda)$ quando $n \to \infty$ e $p \to 0$ com $np \to \lambda$ fixo. Note a coerência das duas primeiras exigências: a esperança é preservada, pois $\mathbb{E} = np = \lambda$ nos dois modelos; e a variância $np(1-p)$ tende a $np = \lambda$ justamente porque $p \to 0$ faz $(1-p) \to 1$. A Poisson é o limite da binomial em que p é tão pequeno que a distinção entre "com" e "sem" o fator $(1-p)$ desaparece.

Exemplo (a seguradora). Uma seguradora tem 20.000 apólices de determinado tipo. A probabilidade de um segurado sofrer sinistro no ano é $p = 0{,}00005$. Qual a probabilidade de ocorrerem exatamente 3 sinistros no ano?

O modelo exato é Bin(20.000; 0,00005), cuja probabilidade pontual envolve $\binom{20000}{3}$ — um número da ordem de 10^{11} — multiplicado por potências minúsculas. À mão, é impraticável; mesmo em ponto flutuante, é um cálculo com risco de perda de precisão.

Como n é enorme e p é diminuto, aplicamos a aproximação com

$$\lambda = np = 20.000 \times 0{,}00005 = 1,$$

e portanto

$$P(X = 3) \approx \frac{1^3 e^{-1}}{3!} = \frac{e^{-1}}{6} \approx 0{,}0613.$$

O valor exato pela binomial é 0,06131, contra 0,06131 pela Poisson: coincidem até a quinta casa decimal. O trabalho passou de um coeficiente binomial gigantesco para uma divisão por $3! = 6$.

Este exemplo também explica por que a Poisson é chamada de **lei dos eventos raros** e por que ela é o modelo padrão em atuária, em confiabilidade e em economia de seguros. Todos esses campos observam um número enorme de unidades expostas, cada uma com risco individual minúsculo, e se interessam pelo agregado — que é exatamente a estrutura limite acima.

5.7 Um catálogo dos cinco modelos

Vale reunir o capítulo em uma única página. A tabela a seguir é para consulta, mas a coluna mais importante não é a das fórmulas: é a última, a da estrutura do experimento. É por ela que se escolhe o modelo.

Modelo	**Suporte**	$\mathbb{E}(X)$ / $\mathrm{V}(X)$	**Quando usar**
Bernoulli(p)	$\{0, 1\}$	p / $p(1-p)$	Um único ensaio dicotômico.
Bin(n, p)	$\{0, \dots, n\}$	np / $np(1-p)$	Sucessos em n ensaios independentes, com n fixo e p constante (equivale a sortear com reposição).
Hiper(N, r, n)	$\{\max(0, n-N+r), \dots, \min(n, r)\}$	np / $np(1-p)\dfrac{N-n}{N-1}$ com $p = r/N$	Sucessos em amostra de tamanho n retirada *sem reposição* de população finita de tamanho N.
Geom(p)	$\{1, 2, 3, \dots\}$	$1/p$ / $(1-p)/p^2$	Ensaios *até* o primeiro sucesso; o número de ensaios é aleatório.
Poi(λ)	$\{0, 1, 2, \dots\}$	λ / λ	Ocorrências em um intervalo de tempo ou espaço, à taxa média λ; eventos raros, sem n identificável.

Três observações de leitura fecham o catálogo.

A primeira: **Bernoulli, Binomial e Geométrica compartilham o mesmo ensaio** e diferem apenas no que se conta e em quando se para. Um ensaio, contando o resultado: Bernoulli. n ensaios, contando sucessos: Binomial. Ensaios até o primeiro sucesso, contando ensaios: Geométrica.

A segunda: **Binomial e Hipergeométrica respondem à mesma pergunta** — quantos sucessos na amostra — e diferem apenas no mecanismo de sorteio, com ou sem reposição. Têm a mesma esperança np e variâncias que diferem pelo fator de correção, que tende a 1 quando a amostra é pequena diante da população.

A terceira: **a Poisson é a única que não tem** n. Ela não conta sucessos em tentativas; conta ocorrências em um contínuo. É por isso que seu parâmetro é uma taxa, e é por isso que precisa ser reescalado quando o intervalo da pergunta muda.

5.8 Laboratório computacional em R

O R implementa as cinco distribuições com a convenção de prefixos já mencionada: `d` para a probabilidade pontual, `p` para a acumulada, `q` para o quantil e `r` para a geração de valores aleatórios. Começamos reproduzindo todos os exemplos numéricos do capítulo.

```r
c(moedas_P2      = dbinom(2, 3, 1/2),
  atirador_P3    = dbinom(3, 5, 2/3),
  prova_P6       = dbinom(6, 20, 0.2),
  prova_Pmaior6  = 1 - pbinom(5, 20, 0.2),
  urna_P3azuis   = dhyper(3, 8, 5, 4),
  lote_P2defeito = dhyper(2, 4, 6, 5),
  traficante     = 1 - dhyper(0, 5, 10, 4))
```

```
     moedas_P2    atirador_P3       prova_P6  prova_Pmaior6   urna_P3azuis
     0.3750000      0.3292181      0.1090997      0.1957922      0.3916084
lote_P2defeito     traficante
     0.4761905      0.8461538
```

Confere com o texto: 3/8, 0,3292, 0,1091, 0,1958, 0,3916, 0,4762 e 0,8461. Observe em `dhyper(2, 4, 6, 5)` que os argumentos são $r = 4$ **defeituosas** e $N - r = 6$ **boas** — não o total $N = 10$.

Agora a geométrica e a Poisson, com atenção ao deslocamento de índice:

```r
c(geom_atirador  = dgeom(5 - 1, 2/3),          # note o "- 1": convencao do R
  geom_a_mao     = (1 - 2/3)^(5 - 1) * (2/3),
  penalti_binom  = dbinom(1, 5, 0.1),
  penalti_geom   = dgeom(5 - 1, 0.1),
  rodovia_1h_P2  = dpois(2, lambda = 3),
  rodovia_exato  = 4.5 * exp(-3),
  rodovia_20min  = 1 - ppois(1, lambda = 1),   # lambda reescalado: 3 * (20/60)
  rodovia_exato2 = 1 - 2 * exp(-1))
```

```
 geom_atirador     geom_a_mao  penalti_binom   penalti_geom  rodovia_1h_P2
   0.008230453    0.008230453    0.328050000    0.065610000    0.224041808
 rodovia_exato  rodovia_20min rodovia_exato2
   0.224041808    0.264241118    0.264241118
```

As linhas `geom_atirador`/`geom_a_mao`, `rodovia_1h_P2`/`rodovia_exato` e `rodovia_20min`/`rodovia_exato2` coincidem par a par — a fórmula à mão e a função do R dão o mesmo número, o que valida tanto o uso do `- 1` quanto a reescala de λ.

O erro do deslocamento de índice merece ser visto uma vez, para nunca mais ser cometido:

```r
p <- 0.1
c(correto_P_X_igual_5 = dgeom(5 - 1, p),   # convencao do curso: 5 ensaios
  errado_sem_o_menos1 = dgeom(5,     p),   # o R entende: 5 fracassos, 6 ensaios
  E_do_curso          = 1 / p,
  E_do_R              = (1 - p) / p)
```

```
correto_P_X_igual_5 errado_sem_o_menos1          E_do_curso             E_do_R
           0.065610            0.059049           10.000000           9.000000
```

O valor errado, 0,059, é perfeitamente plausível — está na mesma ordem de grandeza do correto, 0,0656 — e é justamente por isso que o erro passa despercebido. As esperanças, 10 contra 9, deixam a diferença de convenção explícita.

5.8.1 O catálogo em código

A tabela-resumo da seção anterior, construída programaticamente a partir das fórmulas, com um caso numérico concreto para cada modelo:

```
catalogo <- data.frame(
  modelo   = c("Bernoulli(0,3)", "Binomial(10; 0,3)", "Hiper(50, 15, 10)",
               "Geometrica(0,3)", "Poisson(3)"),
  suporte  = c("{0,1}", "{0,...,10}", "{0,...,10}", "{1,2,3,...}", "{0,1,2,...}"),
  E        = c(0.3, 10 * 0.3, 10 * (15/50), 1 / 0.3, 3),
  V        = c(0.3 * 0.7, 10 * 0.3 * 0.7,
               10 * (15/50) * (1 - 15/50) * (50 - 10) / (50 - 1),
               (1 - 0.3) / 0.3^2, 3))
catalogo$DP <- sqrt(catalogo$V)
catalogo
```

```
             modelo     suporte        E        V        DP
1    Bernoulli(0,3)       {0,1} 0.300000 0.210000 0.4582576
2 Binomial(10; 0,3)  {0,...,10} 3.000000 2.100000 1.4491377
3 Hiper(50, 15, 10)  {0,...,10} 3.000000 1.714286 1.3093073
4   Geometrica(0,3) {1,2,3,...} 3.333333 7.777778 2.7888668
5        Poisson(3) {0,1,2,...} 3.000000 3.000000 1.7320508
```

Compare a segunda e a terceira linhas: mesma esperança $\mathbb{E}(X) = 3$, porque $p = r/N = 0{,}3$ nos dois casos, mas a variância da hipergeométrica é menor — 1,714 contra 2,1 —, por efeito do fator de correção $(50-10)/(50-1) \approx 0{,}816$. É o resultado teórico da seção sobre população finita, verificado numericamente.

5.8.2 Verificando $\mathbb{E}$ e V por simulação

Fórmulas se demonstram, mas convém vê-las funcionar. Geramos 200 mil valores de cada distribuição e comparamos as médias e variâncias amostrais com as teóricas. Pela lei dos grandes números de Seção 4, elas devem praticamente coincidir.

```
set.seed(20264)
M <- 200000

sim <- list(
  Bernoulli  = rbinom(M, size = 1,  prob = 0.3),
  Binomial   = rbinom(M, size = 10, prob = 0.3),
  Hiper      = rhyper(M, m = 15, n = 35, k = 10),
```

```
  Geometrica = rgeom(M, prob = 0.3) + 1,        # "+ 1": convencao do curso
  Poisson    = rpois(M, lambda = 3))

teorico <- data.frame(
  E_teor = c(0.3, 3, 3, 1/0.3, 3),
  V_teor = c(0.21, 2.1, 10*0.3*0.7*(50-10)/(50-1), 0.7/0.09, 3))

resultado <- data.frame(
  modelo    = names(sim),
  E_simul   = sapply(sim, mean),
  E_teorico = teorico$E_teor,
  V_simul   = sapply(sim, var),
  V_teorico = teorico$V_teor)
round(resultado[, -1], 4)
```

```
           E_simul E_teorico V_simul V_teorico
Bernoulli   0.3000    0.3000  0.2100    0.2100
Binomial    2.9993    3.0000  2.0970    2.1000
Hiper       2.9946    3.0000  1.7165    1.7143
Geometrica  3.3376    3.3333  7.8497    7.7778
Poisson     3.0011    3.0000  3.0073    3.0000
```

As colunas simuladas e teóricas batem até a segunda ou terceira casa decimal em todos os cinco casos. Note o `+ 1` na geração da geométrica, o mesmo ajuste de convenção de antes: sem ele, a média simulada seria $(1-p)/p \approx 2{,}33$ em vez de $1/p \approx 3{,}33$.

5.8.3 As duas aproximações, graficamente

Hipergeométrica e Binomial. A aproximação melhora conforme N cresce com n e $p = r/N$ fixos. Fixamos $n = 10$ e $p = 0{,}3$ e deixamos N variar, medindo a discrepância pelo **erro máximo absoluto** entre as duas distribuições, $\max_x |P_{\text{Hiper}}(x) - P_{\text{Bin}}(x)|$:

```
n_am <- 10; p_pop <- 0.3
Ns   <- seq(50, 2000, by = 10)
erro <- sapply(Ns, function(N) {
  r <- round(p_pop * N)
  max(abs(dhyper(0:n_am, r, N - r, n_am) - dbinom(0:n_am, n_am, p_pop)))
})
d_err <- data.frame(razao = Ns / n_am, erro = erro)

g1 <- ggplot(d_err, aes(razao, erro)) +
  geom_vline(xintercept = 20, linetype = "dashed", colour = vm) +
  geom_line(colour = az, linewidth = 0.8) +
  labs(x = "N / n  (tracejada: criterio N >= 20n)", y = "erro maximo absoluto")

N1 <- 1000; r1 <- 300
```

```r
d_cmp <- data.frame(
  x = rep(0:n_am, 2),
  prob = c(dhyper(0:n_am, r1, N1 - r1, n_am), dbinom(0:n_am, n_am, p_pop)),
  modelo = rep(c("Hipergeometrica", "Binomial"), each = n_am + 1))

g2 <- ggplot(d_cmp, aes(factor(x), prob, fill = modelo)) +
  geom_col(position = position_dodge(width = 0.75), width = 0.7, alpha = 0.85) +
  scale_fill_manual(values = c(Hipergeometrica = az, Binomial = lr)) +
  labs(x = "x  (N = 1000, r = 300, n = 10)", y = "P(X = x)")

library(patchwork)
g1 + g2 + plot_layout(widths = c(1, 1.3))
```

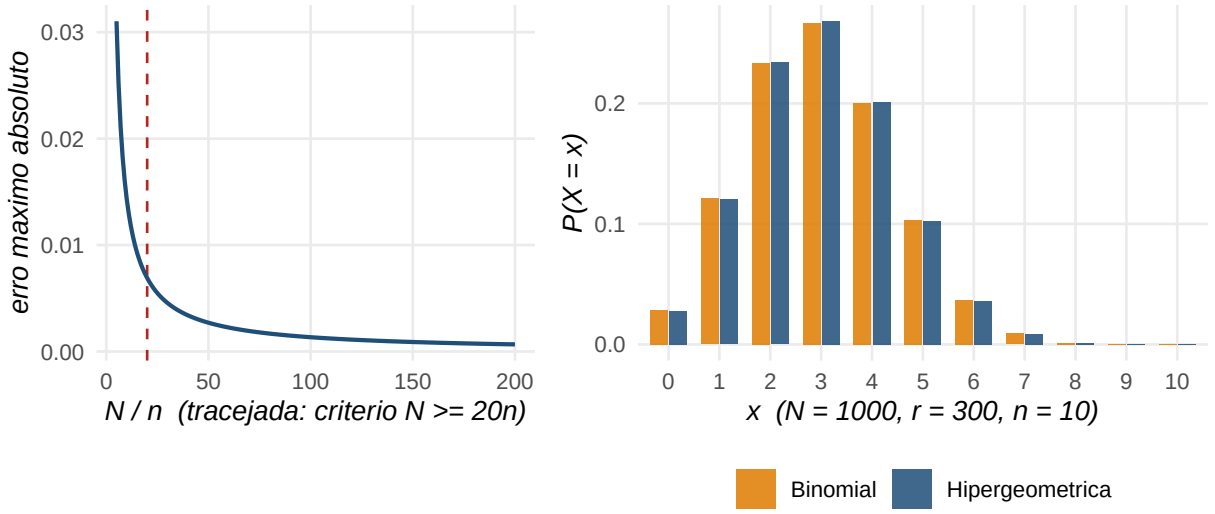


Figura 5: Aproximação da Hipergeométrica pela Binomial. À esquerda, o erro máximo absoluto entre as duas distribuições cai conforme N/n cresce, com $n = 10$ e $p = 0{,}3$ fixos; a linha tracejada marca o critério $N \geq 20n$. À direita, as duas distribuições em $N = 1000$: visualmente indistinguíveis.

O erro máximo decai monotonicamente e já é da ordem de 10^{-3} na fronteira do critério $N \geq 20n$, tornando-se invisível bem antes disso. As barras à direita, no caso da pesquisa eleitoral, praticamente se sobrepõem — a diferença de 0,1026 para 0,1029 em $x = 5$ é a maior de todas, e não se vê no gráfico.

Binomial e Poisson. Agora fixamos $\lambda = np = 3$ e fazemos n crescer, forçando $p = 3/n$ a encolher. A aproximação deve melhorar na mesma medida:

```r
lam <- 3
ns  <- seq(10, 600, by = 5)
erro2 <- sapply(ns, function(n) {
  max(abs(dbinom(0:20, n, lam / n) - dpois(0:20, lam)))
```

```r
})
d_err2 <- data.frame(n = ns, erro = erro2)

g3 <- ggplot(d_err2, aes(n, erro)) +
  geom_line(colour = az, linewidth = 0.8) +
  labs(x = "n  (com p = 3/n, de modo que np = 3)", y = "erro maximo absoluto")

n2 <- 500
d_cmp2 <- data.frame(
  x = rep(0:12, 2),
  prob = c(dbinom(0:12, n2, lam / n2), dpois(0:12, lam)),
  modelo = rep(c("Binomial(500; 0,006)", "Poisson(3)"), each = 13))

g4 <- ggplot(d_cmp2, aes(x, prob, colour = modelo, shape = modelo)) +
  geom_point(size = 2.2) +
  geom_line(linewidth = 0.4, alpha = 0.6) +
  scale_colour_manual(values = c("Binomial(500; 0,006)" = az, "Poisson(3)" = vm)) +
  scale_x_continuous(breaks = 0:12) +
  labs(x = "x", y = "P(X = x)")

g3 + g4 + plot_layout(widths = c(1, 1.3))
```

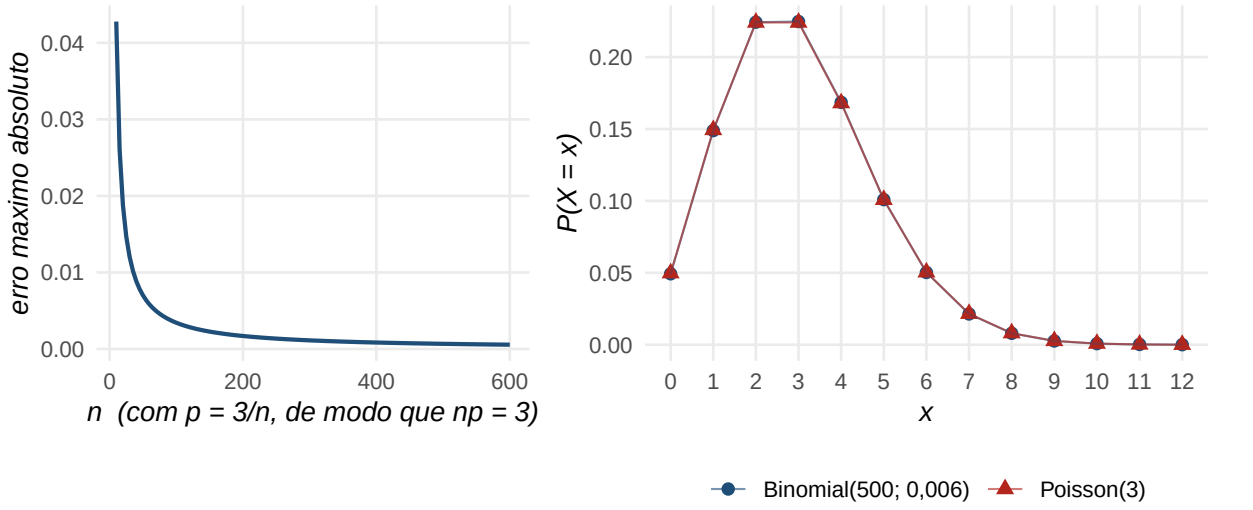


Figura 6: Aproximação da Binomial pela Poisson com $\lambda = np = 3$ fixo. À esquerda, o erro máximo absoluto cai conforme n cresce e $p = 3/n$ encolhe. À direita, Bin(500; 0,006) contra Poi(3): as duas curvas se confundem.

O padrão é o mesmo: erro decrescente em n, e sobreposição visual completa já em algumas centenas de tentativas. Vale conferir também o caso da seguradora, com n na casa das dezenas de milhares:

```r
n_ap <- 20000; p_ap <- 0.00005
c(binomial_exata = dbinom(3, n_ap, p_ap),
  poisson_aprox  = dpois(3, lambda = n_ap * p_ap),
  formula_mao    = exp(-1) / 6,
  erro_absoluto  = abs(dbinom(3, n_ap, p_ap) - dpois(3, n_ap * p_ap)))
```

```
binomial_exata  poisson_aprox    formula_mao  erro_absoluto
  6.131171e-02   6.131324e-02   6.131324e-02   1.533016e-06
```

O erro absoluto é da ordem de 10^{-6} — a aproximação é, para todos os efeitos práticos, exata.

5.8.4 A pegadinha da condicional, por simulação

Por fim, verificamos a condicional geométrica sem usar a fórmula. Simulamos meio milhão de sequências de cobranças, **descartamos** aquelas em que o desperdício ocorreu nas três primeiras — que é literalmente o que "condicionar em $X > 3$" significa — e calculamos a proporção das restantes em que $X \leq 5$:

```r
set.seed(20265)
p  <- 0.1
X  <- rgeom(500000, p) + 1            # convencao do curso: 1, 2, 3, ...

sobreviventes <- X[X > 3]             # o espaco amostral reduzido
prop_condicional <- mean(sobreviventes <= 5)

num <- dgeom(4 - 1, p) + dgeom(5 - 1, p)   # P(4) + P(5)
den <- 1 - pgeom(3 - 1, p)                 # P(X > 3) = 0,9^3

c(simulada        = prop_condicional,
  formula         = num / den,
  numerador       = num,
  denominador     = den,
  resposta_ingenua = num,
  sem_memoria     = 1 - (1 - p)^2)          # P(X <= 2)
```

```
        simulada          formula        numerador      denominador
       0.1898719        0.1900000        0.1385100        0.7290000
resposta_ingenua      sem_memoria
       0.1385100        0.1900000
```

A simulação devolve algo muito próximo de 0,19, confirmando a conta do texto — e mostrando que a resposta ingênua 0,1385, que é apenas o numerador, está longe. Repare que a fração de sequências descartadas é justamente $1 - 0{,}729 = 27{,}1\%$; é essa remoção que infla a proporção de 0,1385 para 0,19.

A última linha traz o fato registrado no texto: $P(X \leq 2) = 1 - 0{,}9^2 = 0{,}19$, **idêntico** à condicional. É a falta de memória da geométrica — ter cobrado três pênaltis sem desperdício deixa o jogador exatamente na mesma situação em que estava antes de começar. Um resultado que reencontraremos, na versão contínua, com a distribuição exponencial em Seção 6.

5.9 Exercícios

1. **Reconhecendo a estrutura.** Para cada situação, identifique o modelo, seus parâmetros e escreva $P(X = x)$ sem calcular: (a) número de caras em 12 lançamentos de uma moeda viciada com $P(\text{cara}) = 0{,}4$; (b) número de cartas de copas em 5 cartas retiradas de um baralho de 52 sem reposição; (c) número de lançamentos de um dado até sair o primeiro 6;

   (d) número de chamadas recebidas por um call center entre 14h e 14h30, sabendo que a média é de 20 chamadas por hora; (e) número de sorteios com reposição, dentre 8, em que se retira bola azul de uma urna com 3 azuis e 7 brancas. Em cada caso, justifique explicitamente por que os modelos rivais foram descartados.

2. **Bernoulli e o máximo da variância.** (a) Derive $\mathbb{E}(X)$ e $\mathrm{V}(X)$ para $X \sim \text{Bernoulli}(p)$ sem usar a identidade $X^2 = X$, isto é, calculando $\mathbb{E}(X^2)$ pela definição. (b) Mostre, por cálculo diferencial, que $\mathrm{V}(X) = p(1-p)$ é máxima em $p = 1/2$ e vale $1/4$ nesse ponto.

   (c) Interprete economicamente: em que sentido $p = 1/2$ é a situação de maior incerteza?
   (d) Que consequência isso tem para o tamanho de amostra necessário em uma pesquisa eleitoral com disputa apertada, em comparação com uma eleição decidida?

3. **Binomial na prova de chute.** Retome a prova de 20 questões com 5 alternativas.

   (a) Calcule $P(X = 0)$ e interprete. (b) Se a aprovação exigisse 10 acertos, qual seria a probabilidade de aprovação no chute? (c) Quantas questões o aluno acerta em média, e com que desvio padrão? (d) Se as alternativas passassem de 5 para 4, o que aconteceria com $\mathbb{E}(X)$, $\mathrm{V}(X)$ e $P(X \geq 6)$? Resolva os itens com `pbinom` e confira ao menos um deles pela fórmula.

4. **Hipergeométrica e o fator de correção.** Uma comissão de 5 pessoas será sorteada de um departamento com 12 professores, dos quais 4 são economistas. (a) Qual a probabilidade de a comissão ter exatamente 2 economistas? (b) E pelo menos 1? (c) Calcule $\mathbb{E}(X)$ e $\mathrm{V}(X)$ e compare a variância com a de uma $\text{Bin}(5;\, 4/12)$. (d) Qual o valor do fator de correção, e por que ele é tão distante de 1 aqui? (e) A aproximação binomial seria aceitável neste caso? Justifique pelo critério $N \geq 20n$.

5. **Geométrica e a convenção do R.** Um vendedor fecha negócio com probabilidade 0,2 por visita, de forma independente. (a) Qual a probabilidade de fechar o primeiro negócio na quarta visita? (b) Quantas visitas ele faz em média até o primeiro negócio, e qual o desvio padrão? (c) Qual a probabilidade de precisar de mais de 10 visitas? Use a forma fechada $P(X > x) = (1-p)^x$. (d) Escreva as três respostas em R, com e sem o ajuste `- 1`, e explique em palavras o que a versão sem ajuste está de fato calculando.

6. **A reescala de λ.** Uma agência bancária recebe, em média, 12 clientes por hora.

(a) Qual a probabilidade de receber exatamente 3 clientes em 15 minutos? (b) E nenhum cliente em 5 minutos? (c) E mais de 20 clientes em uma hora? (d) Um analista respondeu ao item (a) usando $\lambda = 12$; que pergunta ele respondeu, e por que o erro não é detectável olhando apenas para o resultado? (e) Mostre que $\mathbb{E}(X)$ e $\mathrm{V}(X)$ escalam ambas linearmente com o tamanho do intervalo, e explique por que isso implica que o **desvio padrão** escala com a raiz do intervalo.

7. **As duas aproximações.** (a) Uma fábrica produz 5.000 peças, das quais 250 são defeituosas; inspecionam-se 20. Calcule $P(X = 2)$ pela hipergeométrica e pela binomial, verifique o critério $N \geq 20n$ e reporte o erro relativo. (b) A probabilidade de um determinado tipo de falha em um servidor é 0,0002 por dia, e uma empresa opera 5.000 servidores; calcule a probabilidade de haver exatamente 1 falha em um dia pela binomial e pela Poisson. (c) Em (b), a Poisson subestima ou superestima a variância da binomial? Em quanto? (d) Construa um exemplo em que a aproximação Poisson falha visivelmente e explique qual das hipóteses foi violada.

8. **A condicional geométrica.** Seja $X \sim \mathrm{Geom}(p)$ com $p = 0{,}25$. (a) Calcule $P(X \leq 6 \mid X > 2)$ pela definição de probabilidade condicional de Seção 3.

   (b) Calcule $P(X \leq 4)$ e compare com o item (a) — o que você observa? (c) Demonstre em geral a **falta de memória**, isto é, que $P(X > s + t \mid X > s) = P(X > t)$ para quaisquer inteiros $s, t \geq 0$, usando a forma fechada $P(X > x) = (1 - p)^x$. (d) Um jogador de cassino argumenta que, tendo perdido dez vezes seguidas, "está na hora de ganhar". Use o item (c) para refutá-lo, e discuta por que a falta de memória é uma propriedade tão contraintuitiva.

6 Distribuições Contínuas

Em Seção 5 percorremos os modelos que descrevem **contagens**: quantas peças defeituosas num lote, quantos clientes inadimplentes na carteira, quantos sinistros no mês. Toda a aritmética de lá era de somatórios, porque a probabilidade estava distribuída em massas pontuais e bastava somá-las.

Este capítulo faz o mesmo percurso do outro lado da fronteira traçada em Seção 4. Aqui as variáveis resultam de **medições** — tempo de espera na fila, consumo de energia, duração de uma lâmpada, altura de uma pessoa, preço de um ativo — e assumem qualquer valor de um intervalo da reta. A probabilidade deixa de ser massa e passa a ser **densidade**; o $\sum$ vira $\int$; e a pergunta "qual a probabilidade de X valer exatamente isto?" perde o sentido, substituída por "qual a probabilidade de X cair *neste intervalo*?".

Estudaremos quatro modelos, em ordem crescente de sofisticação e de importância econômica:

- a **uniforme**, o modelo da ignorância total dentro de um intervalo, e o mais simples de todos — sua densidade é uma reta horizontal e todas as contas saem por áreas de retângulos;
- a **exponencial**, o modelo canônico de **tempos de espera** e de **durações**, irmã contínua da Poisson, e portadora de uma propriedade tão elegante quanto perigosa: a falta de memória;
- a **Normal**, a distribuição mais importante da estatística — a que dá forma à regra empírica, à padronização de Seção 4 e a praticamente todo intervalo de confiança de Seção 9;
- a **lognormal**, obtida da Normal por exponenciação, que é o pressuposto padrão para **preços de ativos** e o modelo natural para variáveis positivas e assimétricas à direita, como a renda.

O fio condutor é sempre o mesmo. Para cada modelo, três perguntas: qual a **densidade** $f(x)$; qual a **acumulada** $F(x)$, que dispensa integrar a cada cálculo; e quanto valem $\mathbb{E}(X)$ e $\mathrm{V}(X)$. Respondidas essas três, o modelo está dominado.

6.1 A distribuição Uniforme

Comecemos pelo caso mais simples que existe. Suponha que tudo o que sabemos sobre uma grandeza é que ela cai em algum ponto do intervalo (α, β), e que **nenhuma região desse intervalo é mais provável que outra de mesmo comprimento**. Essa é a descrição verbal da distribuição uniforme, e ela determina a densidade sozinha: se f tem de ser constante e integrar 1 sobre um intervalo de comprimento $\beta - \alpha$, então a constante só pode ser o recíproco desse comprimento.

! Definição — distribuição Uniforme

A variável aleatória X tem **distribuição uniforme** no intervalo (α, β), e escrevemos $X \sim \mathrm{Unif}(\alpha, \beta)$, quando sua densidade é

$$f(x) = \begin{cases} \dfrac{1}{\beta - \alpha}, & \alpha < x < \beta, \\ 0, & \text{caso contrário.} \end{cases}$$

A verificação de que se trata mesmo de uma densidade é imediata: $f \geq 0$ porque $\beta > \alpha$, e

$$\int_\alpha^\beta \frac{1}{\beta-\alpha}\,dx = \frac{1}{\beta-\alpha}\Big[x\Big]_\alpha^\beta = \frac{\beta-\alpha}{\beta-\alpha} = 1.$$

6.1.1 Probabilidades: áreas de retângulos

Como a densidade é constante, a área sob ela em qualquer subintervalo é simplesmente base × altura. Para $\alpha \le a \le b \le \beta$,

$$\boxed{P(a \le X \le b) = \int_a^b \frac{dx}{\beta-\alpha} = \frac{b-a}{\beta-\alpha}}$$

isto é, a probabilidade é a **fração do intervalo total** que o subintervalo ocupa. É a tradução exata da ideia de equiprobabilidade: intervalos de mesmo comprimento têm a mesma probabilidade, onde quer que estejam.

A acumulada segue da mesma conta, com $a = \alpha$:

$$F(x) = P(X \le x) = \begin{cases} 0, & x \le \alpha, \\ \dfrac{x-\alpha}{\beta-\alpha}, & \alpha < x < \beta, \\ 1, & x \ge \beta, \end{cases}$$

uma reta que sobe de 0 a 1 ao longo do suporte — a acumulada mais simples que existe.

6.1.2 Esperança e variância

! Momentos da Uniforme

Se $X \sim \text{Unif}(\alpha, \beta)$, então

$$\boxed{\mathbb{E}(X) = \frac{\alpha+\beta}{2} \qquad \text{V}(X) = \frac{(\beta-\alpha)^2}{12}}$$

A esperança é o **ponto médio do intervalo**, o que a intuição do ponto de equilíbrio de Seção 4 já antecipava: uma barra de densidade homogênea equilibra-se no meio. Formalmente,

$$\mathbb{E}(X) = \int_\alpha^\beta \frac{x}{\beta-\alpha}\,dx = \frac{1}{\beta-\alpha}\left[\frac{x^2}{2}\right]_\alpha^\beta = \frac{\beta^2-\alpha^2}{2(\beta-\alpha)} = \frac{(\beta-\alpha)(\beta+\alpha)}{2(\beta-\alpha)} = \frac{\alpha+\beta}{2},$$

onde a fatoração da diferença de quadrados faz todo o trabalho. Para a variância, usamos a forma de cálculo $\text{V}(X) = \mathbb{E}(X^2) - \mathbb{E}^2(X)$:

$$\mathbb{E}(X^2) = \int_\alpha^\beta \frac{x^2}{\beta-\alpha}\,dx = \frac{1}{\beta-\alpha}\left[\frac{x^3}{3}\right]_\alpha^\beta = \frac{\beta^3-\alpha^3}{3(\beta-\alpha)} = \frac{\alpha^2+\alpha\beta+\beta^2}{3},$$

usando agora a fatoração $\beta^3-\alpha^3 = (\beta-\alpha)(\beta^2+\alpha\beta+\alpha^2)$. Subtraindo o quadrado da média,

$$\mathrm{V}(X) = \frac{\alpha^2+\alpha\beta+\beta^2}{3} - \frac{(\alpha+\beta)^2}{4} = \frac{4\alpha^2+4\alpha\beta+4\beta^2-3\alpha^2-6\alpha\beta-3\beta^2}{12} = \frac{\alpha^2-2\alpha\beta+\beta^2}{12} = \frac{(\beta-\alpha)^2}{12}.$$

Repare no que a fórmula diz: a variância depende **apenas do comprimento** do intervalo, não de sua posição. Deslocar (α,β) para a direita não altera a dispersão — é a propriedade $\mathrm{V}(X+b)=\mathrm{V}(X)$ de Seção 4, aqui visível a olho nu.

Exemplo (consumo de energia). O consumo mensal de energia de uma planta industrial, medido em GWh, distribui-se uniformemente entre 1,5 e 2 GWh. Qual a probabilidade de o consumo superar 1,7 GWh?

Temos $X \sim \mathrm{Unif}(1{,}5;\,2)$, de modo que a densidade vale $1/(2-1{,}5) = 2$ em todo o suporte. Então

$$P(X > 1{,}7) = \frac{2-1{,}7}{2-1{,}5} = \frac{0{,}3}{0{,}5} = 0{,}6.$$

Em 60% dos meses o consumo supera 1,7 GWh. Note que a densidade vale 2, maior que 1 — o que é perfeitamente legítimo, pela razão discutida em Seção 4: densidade não é probabilidade, é probabilidade por unidade de x. O que não pode passar de 1 é a área, e ela vale exatamente $0{,}5 \times 2 = 1$.

Completando a caracterização,

$$\mathbb{E}(X) = \frac{1{,}5+2}{2} = 1{,}75 \text{ GWh}, \qquad \mathrm{V}(X) = \frac{(2-1{,}5)^2}{12} = \frac{0{,}25}{12} \approx 0{,}0208,$$

com $\mathrm{DP}(X) \approx 0{,}144$ GWh.

6.2 A distribuição Exponencial

Passamos ao primeiro modelo com conteúdo econômico substantivo. A pergunta que ele responde é: **quanto tempo até que algo aconteça?** Quanto tempo até o próximo cliente chegar à fila, até a máquina quebrar, até a lâmpada queimar, até o próximo sinistro ser comunicado à seguradora.

! Definição — distribuição Exponencial

A variável aleatória X tem **distribuição exponencial** de parâmetro $\lambda > 0$, e escrevemos $X \sim \mathrm{Expo}(\lambda)$, quando sua densidade é

$$f(x) = \begin{cases} \lambda e^{-\lambda x}, & x > 0, \\ 0, & x \le 0. \end{cases}$$

O parâmetro λ é a **taxa de ocorrência** — eventos por unidade de tempo — e tem dimensão inversa à de X.

O suporte é $\mathbb{R}_+$, o que já é adequado ao objeto modelado: tempos são positivos. A densidade é **decrescente** em todo o suporte, com máximo em $x \to 0^+$, onde vale λ. Isso diz algo forte sobre o fenômeno: esperas curtas são sempre mais prováveis que esperas longas, e a moda de um tempo exponencial é zero.

Há um parentesco direto com Seção 5 que vale registrar. Se as ocorrências de um evento seguem um processo de Poisson com taxa λ por unidade de tempo — de modo que o **número** de ocorrências em um intervalo unitário é $\text{Poi}(\lambda)$ —, então o **tempo entre ocorrências consecutivas** é $\text{Expo}(\lambda)$. São duas leituras do mesmo processo: a Poisson conta eventos com o tempo fixo; a exponencial mede tempo com a contagem fixa em um.

6.2.1 A função de distribuição acumulada

Aqui a exponencial é generosa: sua acumulada tem forma fechada simples, e por isso raramente precisamos integrar de novo.

! FDA da Exponencial

Se $X \sim \text{Expo}(\lambda)$, então

$$\boxed{F(x) = P(X \leq x) = 1 - e^{-\lambda x}, \qquad x > 0}$$

e, complementarmente, a **função de sobrevivência**

$$\boxed{P(X > x) = e^{-\lambda x}}$$

Demonstração (integração direta). Para $x > 0$,

$$F(x) = \int_0^x \lambda e^{-\lambda t}\, dt.$$

A antiderivada de $\lambda e^{-\lambda t}$ em relação a t é $-e^{-\lambda t}$, pois derivando esta última pela regra da cadeia recuperamos $\lambda e^{-\lambda t}$. Logo

$$F(x) = \Big[-e^{-\lambda t} \Big]_0^x = -e^{-\lambda x} - (-e^0) = 1 - e^{-\lambda x}. \qquad \blacksquare$$

Verificação das propriedades da FDA: $F(0) = 1 - e^0 = 0$; F é crescente, pois sua derivada é $\lambda e^{-\lambda x} > 0$; e $F(x) \to 1$ quando $x \to \infty$, porque $e^{-\lambda x} \to 0$. Tudo como manda Seção 4. Derivando de volta, $F'(x) = \lambda e^{-\lambda x} = f(x)$, o que fecha o círculo do Teorema Fundamental do Cálculo.

A função de sobrevivência $P(X > x) = 1 - F(x) = e^{-\lambda x}$ é, na prática, a forma mais usada: "qual a probabilidade de a máquina durar mais de x horas?".

6.2.2 Esperança e variância

! Momentos da Exponencial

Se $X \sim \text{Expo}(\lambda)$, então

$$\boxed{\mathbb{E}(X) = \frac{1}{\lambda} \qquad \text{V}(X) = \frac{1}{\lambda^2}}$$

e portanto $\text{DP}(X) = 1/\lambda = \mathbb{E}(X)$: na exponencial, **desvio padrão e média coincidem**, de modo que $\text{CV}(X) = 1$ sempre, qualquer que seja λ.

O resultado $\mathbb{E}(X) = 1/\lambda$ é o que torna o parâmetro interpretável na prática. Se chegam em média 6 clientes por hora ($\lambda = 6$ por hora), o tempo médio entre chegadas é 1/6 de hora, ou 10 minutos. Taxa e tempo médio são recíprocos.

Demonstração de $\mathbb{E}(X)$ por integração por partes. Por definição,

$$\mathbb{E}(X) = \int_0^\infty x\,\lambda e^{-\lambda x}\,dx.$$

Escolhemos $u = x$ e $dv = \lambda e^{-\lambda x}dx$, de modo que $du = dx$ e $v = -e^{-\lambda x}$. A fórmula $\int u\,dv = uv - \int v\,du$ dá

$$\mathbb{E}(X) = \Big[-xe^{-\lambda x}\Big]_0^\infty + \int_0^\infty e^{-\lambda x}\,dx.$$

O termo de fronteira se anula nos dois extremos: em $x = 0$ porque o fator x é nulo; em $x \to \infty$ porque a exponencial decai mais rápido do que x cresce, de modo que $xe^{-\lambda x} \to 0$. Resta

$$\mathbb{E}(X) = \int_0^\infty e^{-\lambda x}\,dx = \Big[-\frac{1}{\lambda}e^{-\lambda x}\Big]_0^\infty = 0 - \Big(-\frac{1}{\lambda}\Big) = \frac{1}{\lambda}. \qquad \blacksquare$$

i Ferramenta matemática — integração por partes e integral imprópria

Duas ferramentas se combinam na demonstração acima.
A **integração por partes**, $\int u\,dv = uv - \int v\,du$, é a regra do produto lida ao contrário: derivando uv obtém-se $u\,dv + v\,du$, e integrar os dois lados e isolar $\int u\,dv$ produz a fórmula. Ela serve exatamente ao caso em que o integrando é o produto de um polinômio por uma exponencial: cada aplicação rebaixa em uma unidade o grau do polinômio, até restar só a exponencial. Aplicando-a **duas** vezes a $\int_0^\infty x^2\lambda e^{-\lambda x}dx$ obtém-se $\mathbb{E}(X^2) = 2/\lambda^2$, e daí a variância.
A **integral imprópria** $\int_0^\infty$ é definida como o limite $\lim_{b\to\infty}\int_0^b$, e só existe quando esse limite é finito. Aqui ele é, e a razão é o **decaimento exponencial dominar o crescimento polinomial**: para todo n fixo, $x^n e^{-\lambda x} \to 0$ quando $x \to \infty$. Esse fato — verificável por L'Hôpital aplicada n vezes a $x^n/e^{\lambda x}$ — é o que anula os termos de fronteira e faz a exponencial ter momentos de todas as ordens.

(Integração por partes, integrais impróprias, regra de L'Hôpital e limites no infinito: capítulos de Cálculo Integral *e* Cálculo Diferencial *das Notas de Matemática, Profa. Silvia Matos.)*

Para a variância, o mesmo caminho com $u = x^2$ leva a $\mathbb{E}(X^2) = 2/\lambda^2$ e portanto

$$\mathrm{V}(X) = \mathbb{E}(X^2) - \mathbb{E}^2(X) = \frac{2}{\lambda^2} - \frac{1}{\lambda^2} = \frac{1}{\lambda^2}.$$

6.2.3 Um exemplo completo: a fila

Exemplo (tempo de espera na fila). O tempo de espera de um cliente em uma fila de atendimento tem distribuição exponencial com média de 10 minutos. Como $\mathbb{E}(X) = 1/\lambda = 10$, segue que

$$\lambda = \frac{1}{10} = 0{,}1 \text{ cliente por minuto,}$$

e a acumulada é $F(x) = 1 - e^{-0{,}1x}$. Calculemos quatro probabilidades.

(a) Probabilidade de esperar no máximo 12 minutos.

$$P(X \leq 12) = F(12) = 1 - e^{-0{,}1(12)} = 1 - e^{-1{,}2} = 1 - 0{,}3012 = 0{,}6988.$$

(b) Probabilidade de esperar no máximo 7 minutos.

$$P(X \leq 7) = F(7) = 1 - e^{-0{,}7} = 1 - 0{,}4966 = 0{,}5034.$$

(c) Probabilidade de esperar entre 7 e 12 minutos. Basta subtrair as acumuladas, como manda Seção 4 — sem integrar nada:

$$P(7 \leq X \leq 12) = F(12) - F(7) = 0{,}6988 - 0{,}5034 = 0{,}1954.$$

(d) Probabilidade de esperar mais que a média.

$$P(X > 10) = 1 - F(10) = e^{-0{,}1(10)} = e^{-1} = 0{,}3679.$$

Esse último resultado merece destaque, porque é mais geral do que parece.

! A probabilidade de superar a média é sempre e^{-1}

Para **qualquer** $\lambda > 0$, a probabilidade de uma variável exponencial superar sua própria média é

$$\boxed{P\left(X > \tfrac{1}{\lambda}\right) = e^{-\lambda \cdot \frac{1}{\lambda}} = e^{-1} \approx 0{,}3679}$$

O parâmetro **cancela**: o resultado não depende de λ. Fila de banco, vida útil de lâmpada, intervalo entre sinistros — em todos os casos, apenas **36,79%** das observações superam a

média, e portanto **63,21%** ficam abaixo dela.
A consequência é imediata e importante: na exponencial, a **média é sempre maior que a mediana**. De fato, resolvendo $F(Md) = 0{,}5$,

$$1 - e^{-\lambda\, Md} = 0{,}5 \implies e^{-\lambda\, Md} = 0{,}5 \implies \boxed{Md = \frac{\ln 2}{\lambda} \approx \frac{0{,}6931}{\lambda}}$$

isto é, $Md \approx 0{,}693\, \mathbb{E}(X)$ — a mediana é sempre cerca de 69,3% da média, independentemente de λ. Essa é a assinatura da **assimetria positiva**: a cauda longa à direita puxa a média para além da mediana, exatamente o padrão descrito em Seção 2.

Exemplo (durabilidade de lâmpadas). A vida útil de certo tipo de lâmpada é exponencial com $\lambda = 0{,}01$ por hora, de modo que a vida média é $1/0{,}01 = 100$ horas. Qual a mediana?

$$Md = \frac{\ln 2}{0{,}01} = \frac{0{,}6931}{0{,}01} = 69{,}31 \text{ horas},$$

que no R se obtém diretamente com `qexp(0.5, 0.01)`. Metade das lâmpadas queima antes de 69,31 horas, embora a média seja 100 — a distância entre as duas medidas é toda a assimetria do modelo. Um gerente que planejasse a reposição pela média subestimaria gravemente a frequência de trocas na primeira metade do estoque.

6.2.4 A falta de memória

Chegamos à propriedade mais característica — e mais controversa — da exponencial.

! Propriedade da falta de memória

Se $X \sim \text{Expo}(\lambda)$, então para quaisquer $x, s > 0$

$$\boxed{P(X > x + s \mid X > x) = P(X > s)}$$

Em palavras: **dado que o evento ainda não ocorreu até o instante x, a distribuição do tempo adicional de espera é idêntica à distribuição original**. O passado é irrelevante; o processo "não se lembra" de quanto já esperou.
A exponencial é a **única** distribuição contínua com essa propriedade — sua contraparte discreta é a geométrica de Seção 5.

Demonstração. Partimos da definição de probabilidade condicional de Seção 3:

$$P(X > x + s \mid X > x) = \frac{P(\{X > x + s\} \cap \{X > x\})}{P(X > x)}.$$

O primeiro passo é notar que, como $s > 0$, temos $x + s > x$, de modo que o evento $\{X > x + s\}$ **está contido** em $\{X > x\}$: se a espera passou de $x + s$, certamente passou de x. A interseção é, portanto, o próprio $\{X > x + s\}$. Usando agora a função de sobrevivência $P(X > t) = e^{-\lambda t}$:

$$P(X > x+s \mid X > x) = \frac{P(X > x+s)}{P(X > x)} = \frac{e^{-\lambda(x+s)}}{e^{-\lambda x}} = e^{-\lambda(x+s)+\lambda x} = e^{-\lambda s} = P(X > s). \qquad \blacksquare$$

Toda a demonstração se apoia em um único fato: $e^{a+b} = e^a e^b$, de modo que a divisão de exponenciais **elimina** o x. A propriedade funcional da exponencial é a propriedade probabilística do modelo.

Exemplo (televisor de LED). A vida útil de um televisor de LED é exponencial com vida esperada de 500 dias, isto é, $\lambda = 1/500 = 0{,}002$ por dia. Um aparelho já funcionou 200 dias. Qual a probabilidade de que dure ao menos mais 300 dias?

Queremos $P(X > 500 \mid X > 200)$. Pela falta de memória, com $x = 200$ e $s = 300$,

$$P(X > 500 \mid X > 200) = P(X > 300) = e^{-0{,}002(300)} = e^{-0{,}6} = 0{,}5488.$$

Se, em vez disso, perguntássemos pela probabilidade de durar mais 100 dias:

$$P(X > 300 \mid X > 200) = P(X > 100) = e^{-0{,}002(100)} = e^{-0{,}2} = 0{,}8187.$$

Os 200 dias já vividos **não entram na conta**. Um televisor com 200 dias de uso e um recém saído da caixa têm exatamente a mesma distribuição de vida restante.

6.2.5 Uma crítica ao uso da exponencial para durabilidade

O parágrafo anterior deve ter provocado desconforto — e deve mesmo. Vale explicitar por quê.

> **!** A falta de memória é uma hipótese forte, não um teorema sobre o mundo
>
> Reformulada em termos concretos, a propriedade diz: *se a lâmpada já durou x horas, a probabilidade de durar mais s horas é a mesma que ela teria quando nova.* Ou seja, **não há desgaste**. Um equipamento com 200 dias de uso é, probabilisticamente, indistinguível de um equipamento zero-quilômetro.
> Para a maioria dos bens duráveis isso é **empiricamente falso**. Componentes se fadigam, rolamentos se desgastam, filamentos se afinam, baterias perdem capacidade. A taxa de falha de um equipamento real tipicamente cresce com a idade — a chamada "curva da banheira" da engenharia de confiabilidade tem uma fase inicial de mortalidade infantil, um platô, e uma fase final crescente de desgaste. A exponencial modela **apenas o platô**, e mesmo assim apenas aproximadamente.

Onde a hipótese *é* razoável? Precisamente nos fenômenos em que as ocorrências são causadas por **choques externos independentes**, e não por deterioração interna:

- o tempo até a próxima chegada de cliente a uma fila — o cliente que vai chegar não sabe há quanto tempo ninguém chega;
- o tempo até o próximo sinistro comunicado a uma seguradora, quando os sinistros decorrem de acidentes independentes;
- o tempo até a próxima falha causada por surto elétrico, e não por fadiga;

- o tempo até a próxima chegada de pedido em um sistema de estoques.

Em todos esses casos, o processo subjacente é o de Poisson, e a falta de memória é a tradução correta da hipótese de independência entre ocorrências.

O critério prático de diagnóstico é aquele $\mathrm{CV}(X) = 1$ observado acima. Se os dados de duração exibem desvio padrão sensivelmente **menor** que a média ($\mathrm{CV} < 1$), há regularidade incompatível com a exponencial e o desgaste está presente; se exibem $\mathrm{CV} > 1$, há mais heterogeneidade que a exponencial admite. Para durabilidade com desgaste, os modelos usuais são a **Weibull** e a **gama**, que generalizam a exponencial admitindo taxa de falha crescente — a exponencial é o caso particular de taxa constante. O ponto metodológico é geral e vale a pena guardar: **um modelo probabilístico embute afirmações substantivas sobre o mecanismo gerador**, e escolher a exponencial para vida útil de equipamentos é afirmar, ainda que involuntariamente, que eles não envelhecem.

6.3 A distribuição Normal

Chegamos à distribuição mais importante da estatística. Sua centralidade não é convenção nem tradição: ela decorre do Teorema Central do Limite (Seção 7), segundo o qual somas e médias de variáveis aleatórias tendem à Normal independentemente da distribuição das parcelas. Como a maior parte das grandezas observadas — alturas, erros de medição, retornos agregados, médias amostrais — resulta do acúmulo de muitos efeitos pequenos e independentes, a forma de sino aparece por toda parte.

! Definição — distribuição Normal

A variável aleatória X tem **distribuição Normal** com parâmetros μ e σ^2, e escrevemos $X \sim N(\mu, \sigma^2)$, quando sua densidade é

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

Os parâmetros são diretamente os momentos:

$$\mathbb{E}(X) = \mu \qquad \mathrm{V}(X) = \sigma^2$$

Vale um alerta de notação que causa erro recorrente: o **segundo argumento é a variância**, não o desvio padrão. Escrever $X \sim N(170, 25)$ significa $\mu = 170$ e $\sigma^2 = 25$, portanto $\sigma = 5$. No R, ao contrário, as funções `dnorm`, `pnorm`, `qnorm` e `rnorm` recebem o **desvio padrão** no argumento `sd`. A conversão entre as duas convenções é fonte constante de erro.

6.3.1 Forma e propriedades

A densidade tem quatro características que se leem diretamente da fórmula:

1. **Formato de sino.** O expoente é $-(x-\mu)^2/(2\sigma^2)$, sempre não positivo, com máximo (valor zero) em $x = \mu$. Logo f atinge seu máximo em $x = \mu$ e decai monotonicamente para os dois lados.

2. **Simetria em torno de μ.** A densidade depende de x apenas por $(x-\mu)^2$, que é invariante à troca de $x-\mu$ por $-(x-\mu)$. Assim $f(\mu-c)=f(\mu+c)$ para todo c, e a distribuição é perfeitamente simétrica. Consequência imediata, já conhecida de Seção 2: **média, mediana e moda coincidem**, todas iguais a μ. Em particular $P(X>\mu)=P(X<\mu)=0{,}5$.
3. **Caudas que nunca tocam o eixo.** Como $e^{-t}>0$ para todo t finito, $f(x)>0$ em toda a reta. A Normal atribui probabilidade positiva a qualquer intervalo, por mais distante que esteja de μ — embora essa probabilidade decaia extraordinariamente rápido.
4. **μ localiza, σ dispersa.** Alterar μ **desloca** a curva sem deformá-la; alterar σ **achata e alarga** (se cresce) ou **afina e eleva** (se diminui), mantendo a área em 1. Os dois painéis da Figura 8 mostram exatamente esses dois movimentos.

6.3.2 Por que existe uma tabela

Aqui aparece a dificuldade técnica que organiza toda a prática com a Normal. Calcular $P(a \le X \le b)$ exige

$$\int_a^b \frac{1}{\sigma\sqrt{2\pi}}\, e^{-\frac{(x-\mu)^2}{2\sigma^2}}\, dx,$$

e esta integral **não tem solução analítica**: não existe função elementar cuja derivada seja e^{-x^2}. Nenhuma combinação finita de polinômios, exponenciais, logaritmos e funções trigonométricas resolve o problema — o resultado é um teorema, não uma limitação de engenhosidade. A antiderivada existe (o integrando é contínuo), mas não é expressável em forma fechada; ela recebeu nome próprio, a função erro, e é calculada numericamente.

Daí a **tabela da Normal padrão**, que é simplesmente o valor numérico dessa integral, tabulado de uma vez por todas. E daí, também, a necessidade da padronização: seria impossível tabular uma curva diferente para cada par (μ, σ^2).

6.3.3 Padronização

! Padronização e a Normal padrão

Se $X \sim N(\mu, \sigma^2)$, então

$$\boxed{Z = \frac{X-\mu}{\sigma} \sim N(0,1)}$$

que se chama **Normal padrão**. Sua densidade é $\phi(z) = \frac{1}{\sqrt{2\pi}} e^{-z^2/2}$ e sua acumulada denota-se $\Phi(z) = P(Z \le z)$ — é ela que a tabela fornece.
Qualquer probabilidade sobre X converte-se em uma sobre Z:

$$\boxed{P(a \le X \le b) = P\left(\frac{a-\mu}{\sigma} \le Z \le \frac{b-\mu}{\sigma}\right) = \Phi\left(\frac{b-\mu}{\sigma}\right) - \Phi\left(\frac{a-\mu}{\sigma}\right)}$$

Que $\mathbb{E}(Z) = 0$ e $\mathrm{V}(Z) = 1$ já sabíamos de Seção 4, por ser Z uma transformação afim de X. O que é **novo e específico da Normal** é que Z continua Normal — a família normal é fechada sob transformações lineares, propriedade que nem todas as famílias têm.

i Ferramenta matemática — mudança de variável na integral

A afirmação "Z é Normal padrão" prova-se por **mudança de variável**. Queremos a densidade de $Z = (X - \mu)/\sigma$. Comece pela acumulada:

$$F_Z(z) = P(Z \le z) = P\left(\frac{X-\mu}{\sigma} \le z\right) = P(X \le \mu + \sigma z) = F_X(\mu + \sigma z),$$

usando $\sigma > 0$ para preservar o sentido da desigualdade. Derivando em z pela regra da cadeia,

$$f_Z(z) = \sigma f_X(\mu + \sigma z) = \sigma \cdot \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(\mu+\sigma z-\mu)^2}{2\sigma^2}} = \frac{1}{\sqrt{2\pi}} e^{-z^2/2},$$

que é exatamente $\phi(z)$. O fator σ vindo da regra da cadeia **cancela** o $1/\sigma$ da densidade — e esse cancelamento é o conteúdo do jacobiano da transformação.
O mesmo raciocínio, na forma integral, é a substituição $u = (x - \mu)/\sigma$, com $du = dx/\sigma$, que converte $\int f(x)dx$ em $\int \phi(u)du$ com os limites transformados. É a técnica que usaremos de novo, adiante, para obter a lognormal a partir da Normal.
(Mudança de variável, substituição em integrais e regra da cadeia: capítulos de Cálculo Integral *e* Cálculo Diferencial *das Notas de Matemática, Profa. Silvia Matos.)*

6.3.4 A regra empírica

A padronização produz um subproduto de enorme valor prático: como toda Normal vira a mesma $N(0,1)$, as probabilidades de intervalos medidos **em desvios padrão** são universais.

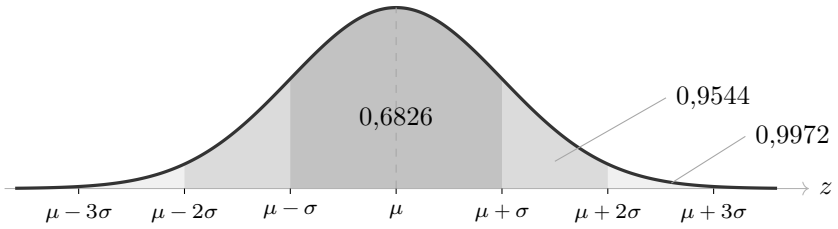
! Regra empírica (68–95–99,7)

Se $X \sim N(\mu, \sigma^2)$, então

$$\begin{array}{|rl|}\hline P(\mu - \sigma < X < \mu + \sigma) = 0{,}6826 & (\approx 68\%) \\ P(\mu - 2\sigma < X < \mu + 2\sigma) = 0{,}9544 & (\approx 95\%) \\ P(\mu - 3\sigma < X < \mu + 3\sigma) = 0{,}9972 & (\approx 99{,}72\%) \\ \hline \end{array}$$

Os valores **não dependem** de μ nem de σ: são propriedades da forma da curva.

É esta regra que dá conteúdo preciso à afirmação de Seção 2, feita então sem justificativa, de que o desvio padrão "ganha interpretação sob normalidade" e de que valores fora de $[\mu - 3\sigma, \mu + 3\sigma]$ são candidatos naturais a outlier: eles ocorrem em menos de 0,3% dos casos. A figura a seguir representa as três faixas.



Regra empírica. As três faixas centrais da densidade Normal, em desvios padrão em torno da média. As áreas valem 0,6826, 0,9544 e 0,9972 e independem de μ e σ.

6.3.5 Alturas: dez probabilidades

Exemplo (alturas). A altura dos indivíduos de certa população, em centímetros, é $X \sim N(170, 25)$ — portanto $\mu = 170$ e $\sigma = \sqrt{25} = 5$. Calculemos dez probabilidades, que juntas cobrem todos os casos que aparecem na prática.

(a) $P(170 < X < 172{,}3)$. Padronizando os dois limites: $z_1 = (170 - 170)/5 = 0$ e $z_2 = (172{,}3 - 170)/5 = 0{,}46$. Logo

$$P(170 < X < 172{,}3) = P(0 < Z < 0{,}46) = \Phi(0{,}46) - \Phi(0) = 0{,}6772 - 0{,}5 = 0{,}1772.$$

(b) $P(170 < X < 175)$. Agora $z_2 = (175 - 170)/5 = 1$:

$$P(0 < Z < 1) = \Phi(1) - 0{,}5 = 0{,}8413 - 0{,}5 = 0{,}3413.$$

(c) $P(165 < X < 170)$. Aqui $z_1 = (165 - 170)/5 = -1$ e $z_2 = 0$. Pela **simetria**, esta área é idêntica à do item (b), refletida em torno de zero:

$$P(-1 < Z < 0) = P(0 < Z < 1) = 0{,}3413.$$

(d) $P(165 < X < 175)$. É a união dos dois intervalos anteriores, ou seja, a faixa de um desvio padrão para cada lado:

$$P(-1 < Z < 1) = 2(0{,}3413) = 0{,}6826.$$

Reconhecemos a primeira linha da regra empírica.

(e) $P(170 < X < 180)$. Com $z_2 = (180 - 170)/5 = 2$:

$$P(0 < Z < 2) = \Phi(2) - 0{,}5 = 0{,}9772 - 0{,}5 = 0{,}4772.$$

(f) $P(160 < X < 180)$. Dois desvios de cada lado:

$$P(-2 < Z < 2) = 2(0{,}4772) = 0{,}9544.$$

Segunda linha da regra empírica.

(g) $P(X > 170)$. O ponto é a própria média, e a simetria resolve sem tabela:

$$P(X > 170) = P(Z > 0) = 0{,}5.$$

(h) $P(X > 175)$. Com $z = 1$, e usando o complementar:

$$P(Z > 1) = 1 - \Phi(1) = 1 - 0{,}8413 = 0{,}1587.$$

(i) $P(X < 175)$. É o complementar do anterior:

$$P(Z < 1) = \Phi(1) = 0{,}8413.$$

(j) $P(X < 165)$. Com $z = -1$, e de novo pela simetria — a cauda esquerda além de -1 tem a mesma área da cauda direita além de $+1$:

$$P(Z < -1) = P(Z > 1) = 0{,}1587.$$

Vale extrair o método destes dez itens, porque ele se repete em todo o curso: **padronize, desenhe a área, e reduza-a a somas e diferenças de** Φ, explorando simetria ($\Phi(-z) = 1 - \Phi(z)$) e complementaridade ($P(Z > z) = 1 - \Phi(z)$) para evitar tabelas de duas caudas. Note também a coerência interna: os itens (h) e (j) dão o mesmo número, e $0{,}1587 + 0{,}6826 + 0{,}1587 = 1$, como deve ser.

Exemplo (valor presente líquido). O VPL de um projeto de investimento, em milhões de reais, é $X \sim N(80, 16)$, isto é, $\mu = 80$ e $\sigma = 4$.

(a) $P(80 < X < 83)$. Padronizando, $z_1 = 0$ e $z_2 = (83 - 80)/4 = 0{,}75$:

$$P(0 < Z < 0{,}75) = \Phi(0{,}75) - 0{,}5 = 0{,}7734 - 0{,}5 = 0{,}2734.$$

(b) $P(79 < X < 82)$. Agora $z_1 = (79 - 80)/4 = -0{,}25$ e $z_2 = (82 - 80)/4 = 0{,}50$. Como os limites estão em lados opostos da média, **somamos** as duas áreas parciais:

$$P(-0{,}25 < Z < 0{,}50) = \Phi(0{,}50) - \Phi(-0{,}25) = 0{,}6915 - 0{,}4013 = 0{,}2902.$$

Equivalentemente, $P(-0{,}25 < Z < 0) + P(0 < Z < 0{,}50) = 0{,}0987 + 0{,}1915 = 0{,}2902$. As duas leituras coincidem, e escolher entre elas é questão do formato da tabela disponível.

6.3.6 Fazendo isso no R

O R dispensa a tabela, e admite os dois caminhos. Podemos passar μ e σ diretamente para `pnorm`, ou padronizar à mão e usar a Normal padrão:

```r
pnorm(175, mean = 170, sd = 5)      # sem padronizar
```

```
[1] 0.8413447
```

```r
pnorm((175 - 170) / 5)              # padronizando: mean = 0, sd = 1 por default
```

```
[1] 0.8413447
```

Os dois resultados são idênticos, como devem ser — a padronização é uma identidade, não uma aproximação. Repare que o R usa `sd` (**desvio padrão**), enquanto a notação $N(\mu, \sigma^2)$ traz a **variância**: com $N(170, 25)$, passa-se `sd = 5` e nunca `sd = 25`.

> **!** Muita atenção com o separador decimal!
>
> O aviso é do professor e vale repetir. Em português escrevemos 172,3; o R — como toda linguagem de programação — exige **ponto**: `172.3`. Escrever `pnorm(172,3, 170, 5)` não gera erro de sintaxe: o R interpreta como três argumentos, `q = 172`, `mean = 3`, `sd = 170`, e devolve um número perfeitamente plausível e **completamente errado**. É a pior espécie de bug — o que não avisa. Confira sempre os separadores antes de confiar em um resultado.

6.4 A distribuição Lognormal

O último modelo do capítulo nasce de uma pergunta natural: e se, em vez de a *variável* ser Normal, for o seu **logaritmo**?

A pergunta é econômica antes de ser matemática. Muitas grandezas de interesse são **estritamente positivas** — preços, rendas, faturamentos, valores de imóveis — o que já as torna incompatíveis com a Normal, cujo suporte é toda a reta. Além disso, elas costumam ser **assimétricas à direita**: muitas observações moderadas e uma cauda longa de valores altos, exatamente o padrão de assimetria positiva descrito em Seção 2. E, sobretudo, elas variam de forma **multiplicativa**: um ativo rende 2% ao dia, não R$ 2 ao dia; um salário sobe 8%, não R$ 8. Ora, o logaritmo converte produtos em somas — e é a somas que o Teorema Central do Limite de Seção 7 se aplica.

> **!** Definição — distribuição Lognormal
>
> Se $X \sim N(\mu, \sigma^2)$ e definimos
>
> $$\boxed{Y = e^X}$$
>
> então Y tem **distribuição lognormal** de parâmetros μ e σ^2, com densidade
>
> $$\boxed{f(y) = \frac{1}{y\,\sigma\sqrt{2\pi}}\, e^{-\frac{(\ln y - \mu)^2}{2\sigma^2}}, \qquad y > 0.}$$
>
> Equivalentemente, e é a forma mais útil na prática: Y **é lognormal se e somente se** $\ln Y$ **é Normal**.

O nome é, portanto, literal e às vezes confunde: lognormal não é "o log de uma normal", é a variável **cujo log** é normal. Note dois detalhes da densidade. O suporte é $y > 0$, como convém, porque $e^X > 0$ sempre. E aparece um fator $1/y$ que não estava na Normal — ele é o jacobiano da transformação, e sua origem fica clara na dedução.

Dedução da densidade (mudança de variável). Como $y \mapsto e^y$ é estritamente crescente, podemos passar pela acumulada:

$$F_Y(y) = P(Y \le y) = P(e^X \le y) = P(X \le \ln y) = F_X(\ln y), \qquad y > 0,$$

onde a passagem usa que o logaritmo é crescente e preserva a desigualdade. Derivando em y pela regra da cadeia, com $\frac{d}{dy} \ln y = 1/y$:

$$f_Y(y) = f_X(\ln y) \cdot \frac{1}{y} = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(\ln y - \mu)^2}{2\sigma^2}} \cdot \frac{1}{y},$$

que é exatamente a densidade enunciada. O fator $1/y$ é a derivada da transformação inversa — o jacobiano —, e é ele que produz a assimetria: como $1/y$ decresce, a densidade é comprimida perto da origem e esticada à direita.

6.4.1 Momentos

! Momentos da Lognormal

Se Y é lognormal de parâmetros μ e σ^2, então

$$\boxed{\mathbb{E}(Y) = e^{\mu + \frac{\sigma^2}{2}} \qquad \mathrm{V}(Y) = e^{2\mu+\sigma^2}\left(e^{\sigma^2} - 1\right)}$$

Estas fórmulas contêm a lição mais importante da seção, e ela é uma armadilha clássica:

$$\mathbb{E}(Y) = \mathbb{E}(e^X) = e^{\mu + \frac{\sigma^2}{2}} \;\neq\; e^{\mathbb{E}(X)} = e^{\mu}.$$

A esperança da exponencial não é a exponencial da esperança. O fator extra $e^{\sigma^2/2}$ é sempre maior que 1 (para $\sigma > 0$), de modo que $\mathbb{E}(Y) > e^{\mu}$ estritamente. Isso é a desigualdade de Jensen aplicada à função convexa e^x, e não é uma sutileza acadêmica: exponenciar a média de log-retornos para obter o retorno médio esperado **subestima sistematicamente** o resultado, erro corriqueiro em análise financeira.

As três medidas de posição da lognormal, aliás, separam-se de forma ordenada e permanente:

$$\boxed{Mo = e^{\mu-\sigma^2} \;<\; Md = e^{\mu} \;<\; \mathbb{E}(Y) = e^{\mu+\frac{\sigma^2}{2}}}$$

A ordem moda $<$ mediana $<$ média é a assinatura canônica da **assimetria positiva** de Seção 2, e aqui ela vale sempre, para qualquer $\sigma > 0$ — não é acidente de parametrização. A mediana é particularmente fácil de justificar: como ln é monótono, $P(Y \le e^{\mu}) = P(\ln Y \le \mu) = P(X \le \mu) = 0{,}5$, porque μ é a mediana da Normal.

6.4.2 Calculando probabilidades

Não é preciso integrar a densidade lognormal nem tabelar nada de novo. A monotonicidade do logaritmo transporta qualquer pergunta sobre Y para uma pergunta sobre a Normal:

$$\boxed{P(Y \le y) = P(\ln Y \le \ln y) = P\left(Z \le \frac{\ln y - \mu}{\sigma}\right) = \Phi\left(\frac{\ln y - \mu}{\sigma}\right)}$$

Ou seja: **tome o logaritmo do valor, padronize, e consulte a Normal**. Três passos, sempre os mesmos.

Exemplo (preço de um ativo). O preço Y de um ativo daqui a um período é lognormal com parâmetros $\mu = 5{,}11$ e $\sigma^2 = 1$ (logo $\sigma = 1$), isto é, $\ln Y \sim N(5{,}11; 1)$. Qual a probabilidade de o preço ficar abaixo de 164?

Primeiro, o logaritmo do valor de corte: $\ln 164 \approx 5{,}10$. Depois, a padronização:

$$z = \frac{\ln 164 - \mu}{\sigma} = \frac{5{,}10 - 5{,}11}{1} = -0{,}01.$$

Por fim, a tabela:

$$P(Y < 164) = P(Z < -0{,}01) = \Phi(-0{,}01) = 0{,}4960 \approx 0{,}496.$$

Há quase exatamente 50% de chance de o preço ficar abaixo de 164 — o que era previsível antes de qualquer conta, já que a mediana do ativo é $e^{5{,}11} = 165{,}67$, muito próxima de 164. Note, porém, o **preço esperado**:

$$\mathbb{E}(Y) = e^{5{,}11 + 0{,}5} = e^{5{,}61} = 273{,}14,$$

bastante acima da mediana de 165,67 e da moda de $e^{5{,}11 - 1} = 60{,}95$. A distância entre esses três números é enorme, e é toda ela assimetria: a possibilidade de ganhos muito grandes puxa a média para cima sem afetar a mediana. Um relatório que informasse apenas o "preço esperado de 273" daria uma impressão profundamente enganosa de um ativo cuja probabilidade de superar esse valor é de apenas 31%.

6.4.3 Por que preços de ativos e por que renda

Duas aplicações justificam o lugar da lognormal no curso.

Preços de ativos. O pressuposto padrão em finanças — o de Black-Scholes e de praticamente todo modelo de tempo contínuo — é que os preços são lognormais, isto é, que os **log-retornos** são normais. A justificativa encadeia tudo o que vimos. Escreva o preço em T como o preço inicial multiplicado pelos retornos brutos de cada período:

$$P_T = P_0 \prod_{t=1}^{T} (1 + R_t) \qquad \Longrightarrow \qquad \ln P_T = \ln P_0 + \sum_{t=1}^{T} \ln(1 + R_t).$$

O logaritmo transformou um **produto** de muitos fatores em uma **soma** de muitas parcelas. Se os log-retornos $\ln(1+R_t)$ são aproximadamente independentes e identicamente distribuídos, o Teorema Central do Limite de Seção 7 garante que a soma é aproximadamente Normal — e portanto P_T, sendo a exponencial de uma Normal, é aproximadamente lognormal. Dois subprodutos vêm de graça e são economicamente essenciais: o preço nunca fica negativo (responsabilidade limitada), e a distribuição é assimétrica à direita (perdas limitadas a 100%, ganhos ilimitados). É a mesma estrutura multiplicativa das carteiras de Seção 7, vista agora ao longo do tempo em vez de entre ativos.

Distribuição de renda. A renda do trabalho exibe, em praticamente toda economia observada, forte assimetria positiva: a média fica bem acima da mediana, e uma cauda longa de rendas altas domina os agregados. O modelo lognormal reproduz esse padrão naturalmente e tem uma justificativa geradora análoga à dos preços — a chamada *lei do efeito proporcional*: se a renda de um indivíduo evolui por choques **multiplicativos** independentes (aumentos percentuais, promoções, valorizações), então seu logaritmo acumula choques **aditivos** e tende à Normal. Daí a prática, universal em economia do trabalho, de rodar regressões com ln(renda) como variável dependente: além de normalizar os resíduos, os coeficientes passam a se ler diretamente como variações percentuais.

6.5 Resumo dos quatro modelos

A tabela a seguir consolida o capítulo. Vale lê-la como um mapa: dadas a densidade e a acumulada, tudo o mais é aplicação das definições de Seção 4.

	Unif(α,β)	Expo(λ)	$N(\mu,\sigma^2)$	Lognormal(μ,σ^2)
Suporte	(α,β)	$(0,\infty)$	$(-\infty,\infty)$	$(0,\infty)$
Densidade $f(x)$	$\frac{1}{\beta-\alpha}$	$\lambda e^{-\lambda x}$	$\frac{1}{\sigma\sqrt{2\pi}}e^{-\frac{(x-\mu)^2}{2\sigma^2}}$	$\frac{1}{x\sigma\sqrt{2\pi}}e^{-\frac{(\ln x-\mu)^2}{2\sigma^2}}$
Acumulada $F(x)$	$\frac{x-\alpha}{\beta-\alpha}$	$1-e^{-\lambda x}$	sem forma fechada	$\Phi\left(\frac{\ln x-\mu}{\sigma}\right)$
$\mathbb{E}(X)$	$\frac{\alpha+\beta}{2}$	$\frac{1}{\lambda}$	μ	$e^{\mu+\sigma^2/2}$
$\mathrm{V}(X)$	$\frac{(\beta-\alpha)^2}{12}$	$\frac{1}{\lambda^2}$	σ^2	$e^{2\mu+\sigma^2}(e^{\sigma^2}-1)$
Mediana	$\frac{\alpha+\beta}{2}$	$\frac{\ln 2}{\lambda}$	μ	e^{μ}
Assimetria	nula (simétrica)	positiva (fixa)	nula (simétrica)	positiva (cresce com σ)
Modela	ignorância no intervalo	tempos de espera	somas e médias	preços, renda
No R	`unif`	`exp`	`norm`	`lnorm`

Três leituras transversais merecem registro. Primeira, **duas famílias são simétricas e duas são assimétricas**, e nas simétricas média, mediana e moda coincidem, enquanto nas assimétricas positivas vale sempre moda $<$ mediana $<$ média. Segunda, **só a Normal não tem acumulada fechada** — é a razão de existir tabela para ela e não para as outras. Terceira, os prefixos do R são universais: `d` para densidade, `p` para acumulada, `q` para quantil (a inversa da acumulada) e `r` para gerar amostras, combinados com o sufixo da última linha — `dnorm`, `pexp`, `qlnorm`, `runif`.

6.6 Laboratório computacional em R

Verificamos numericamente cada resultado do capítulo. O padrão é o de sempre: nunca aceitar uma fórmula sem confrontá-la com integração numérica ou simulação.

A uniforme. Comecemos pelo exemplo do consumo de energia, checando de uma vez a área total, a probabilidade pedida e os dois momentos — cada um por dois caminhos independentes, a integral e a fórmula fechada.

```
alpha <- 1.5; beta <- 2
f_unif <- function(x) dunif(x, alpha, beta)

c(area       = integrate(f_unif, alpha, beta)$value,
  P_maior1.7 = integrate(f_unif, 1.7, beta)$value,
  P_punif    = 1 - punif(1.7, alpha, beta),
  E_integral = integrate(function(x) x * f_unif(x), alpha, beta)$value,
  E_formula  = (alpha + beta) / 2,
  V_integral = integrate(function(x) x^2 * f_unif(x), alpha, beta)$value -
               ((alpha + beta) / 2)^2,
  V_formula  = (beta - alpha)^2 / 12)
```

```
      area P_maior1.7    P_punif E_integral  E_formula V_integral  V_formula
1.00000000 0.60000000 0.60000000 1.75000000 1.75000000 0.02083333 0.02083333
```

Área 1, $P(X > 1{,}7) = 0{,}6$ pelos dois métodos, $\mathbb{E}(X) = 1{,}75$ e $\mathrm{V}(X) = 0{,}0208$: todos os números do texto confirmados.

A exponencial da fila. Os quatro itens do exemplo, mais a mediana, em uma chamada:

```
lambda <- 0.1
round(c(a_ate12        = pexp(12, lambda),
        b_ate7         = pexp(7, lambda),
        c_entre_7_12   = pexp(12, lambda) - pexp(7, lambda),
        d_acima_media  = 1 - pexp(1 / lambda, lambda),
        e_menos_um     = exp(-1),
        mediana        = qexp(0.5, lambda)), 4)
```

```
      a_ate12        b_ate7  c_entre_7_12 d_acima_media    e_menos_um
       0.6988        0.5034        0.1954        0.3679        0.3679
      mediana
       6.9315
```

Exatamente 0,6988, 0,5034, 0,1954 e 0,3679 — e o item (d) coincide com e^{-1} até a última casa, como a álgebra previa. A mediana de 6,93 minutos está bem abaixo da média de 10.

A invariância em λ. A afirmação de que $P(X > \mathbb{E}(X)) = e^{-1}$ e $Md/\mathbb{E}(X) = \ln 2$ valem para *qualquer* taxa é forte o bastante para merecer verificação sobre uma grade deliberadamente heterogênea de valores de λ:

```r
lambdas <- c(0.01, 0.1, 0.5, 1, 7.3)
data.frame(lambda          = lambdas,
           media           = 1 / lambdas,
           mediana         = qexp(0.5, lambdas),
           P_X_maior_media = 1 - pexp(1 / lambdas, lambdas),
           razao_md_media  = qexp(0.5, lambdas) / (1 / lambdas))
```

```
  lambda       media     mediana P_X_maior_media razao_md_media
1   0.01 100.0000000 69.31471806       0.3678794      0.6931472
2   0.10  10.0000000  6.93147181       0.3678794      0.6931472
3   0.50   2.0000000  1.38629436       0.3678794      0.6931472
4   1.00   1.0000000  0.69314718       0.3678794      0.6931472
5   7.30   0.1369863  0.09495167       0.3678794      0.6931472
```

As duas últimas colunas são **constantes**: 0,3679 e 0,6931 em todas as linhas, de $\lambda = 0{,}01$ a $\lambda = 7{,}3$. A média e a mediana mudam por duas ordens de grandeza; sua **razão**, não. A segunda linha reproduz a fila (média 10, mediana 6,93) e a primeira, as lâmpadas (média 100, mediana 69,31).

A falta de memória, por simulação. A demonstração algébrica foi de três linhas, mas a propriedade é contraintuitiva o bastante para valer uma verificação empírica. A estratégia é direta: simulamos duzentos mil televisores, **descartamos** os que falharam antes de 200 dias, e olhamos a vida **residual** dos sobreviventes. Se a exponencial não tem memória, essa vida residual deve ter a mesma distribuição da vida original.

```r
set.seed(20265)
lam_tv <- 1 / 500
X      <- rexp(200000, lam_tv)
resid  <- X[X > 200] - 200          # vida ADICIONAL dos sobreviventes aos 200 dias

round(c(n_sobreviventes             = length(resid),
        sim_P_resid_maior_100       = mean(resid > 100),
        teo_P_X_maior_100           = 1 - pexp(100, lam_tv),
        sim_media_residual          = mean(resid),
        teo_media_original          = 1 / lam_tv,
        sim_P_X_maior_300_dado_200  = mean(X[X > 200] > 300)), 4)
```

```
           n_sobreviventes       sim_P_resid_maior_100
                 133964.0000                      0.8198
          teo_P_X_maior_100          sim_media_residual
                    0.8187                    499.9630
        teo_media_original sim_P_X_maior_300_dado_200
                  500.0000                      0.8198
```

O resultado é o esperado e ainda assim surpreende: a vida residual média dos aparelhos que já duraram 200 dias é de **500 dias** — a mesma vida média de um aparelho novo. Os 200 dias vividos não descontaram nada. E a probabilidade simulada de durar mais 100 dias, 0,8198, bate com o valor

teórico $e^{-0{,}2} = 0{,}8187$ do exemplo do televisor. A figura compara as duas distribuições inteiras, não apenas suas médias:

```
d_mem <- rbind(
  data.frame(t = X[1:60000],     grupo = "original: X"),
  data.frame(t = resid[1:60000], grupo = "residual: X - 200 | X > 200"))
ggplot(d_mem, aes(t, colour = grupo, linetype = grupo)) +
  geom_density(linewidth = 0.8, bw = 40, bounds = c(0, Inf)) +   # suporte positivo
  scale_colour_manual(values = c(az, vm)) +
  coord_cartesian(xlim = c(0, 2000)) +
  labs(x = "tempo de vida (dias)", y = "densidade")
```

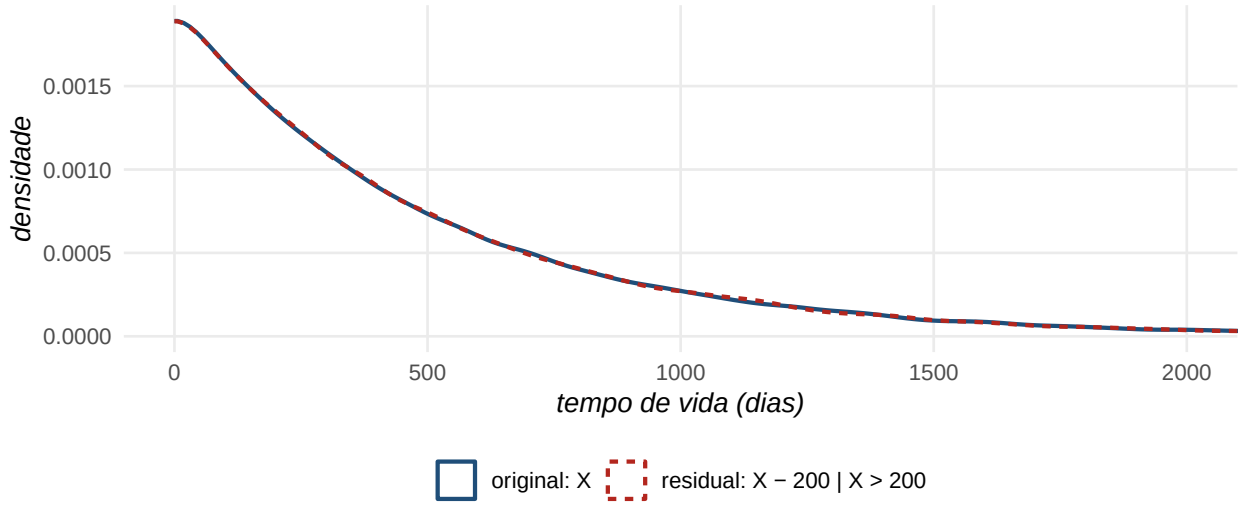


Figura 7: Falta de memória por simulação: a distribuição da vida residual dos televisores que sobreviveram a 200 dias (vermelho tracejado) é indistinguível da distribuição da vida original (azul). Condicionar em $X > 200$ e subtrair 200 devolve a mesma exponencial.

As duas curvas se sobrepõem em toda a extensão. Este é o conteúdo gráfico da falta de memória — e, como argumentamos no texto, a razão para desconfiar do modelo quando aplicado a equipamentos que de fato se desgastam.

A regra empírica, por integração e por simulação. Verifiquemos as três linhas por três caminhos independentes: integrando a densidade Normal com `integrate`, usando a acumulada `pnorm`, e simulando um milhão de alturas.

```
mu <- 170; sigma <- 5
phi <- function(x) (1 / (sigma * sqrt(2 * pi))) * exp(-(x - mu)^2 / (2 * sigma^2))

k         <- 1:3
integral  <- sapply(k, function(j) integrate(phi, mu - j * sigma, mu + j * sigma)$value)
acumulada <- pnorm(k) - pnorm(-k)
```

```r
set.seed(20266)
amostra  <- rnorm(1e6, mu, sigma)
simulado <- sapply(k, function(j) mean(abs(amostra - mu) < j * sigma))

round(data.frame(k, integral, acumulada, simulado,
                 tabela = c(0.6826, 0.9544, 0.9972)), 6)
```

```
  k integral acumulada simulado tabela
1 1 0.682689  0.682689 0.681852 0.6826
2 2 0.954500  0.954500 0.954683 0.9544
3 3 0.997300  0.997300 0.997228 0.9972
```

As três primeiras colunas coincidem até a sexta casa decimal, e a simulação as reproduz até a terceira. A última coluna traz os valores que o professor apresenta em sala, arredondados para quatro casas a partir da tabela — a pequena diferença na segunda e terceira linhas (0,9544 contra 0,954500; 0,9972 contra 0,997300) é apenas o truncamento da tabela, que fornece $\Phi(2) = 0{,}9772$ e dobra a distância à média.

As dez alturas. Reproduzamos de uma vez os dez itens do exemplo:

```r
round(c(a = pnorm(172.3, 170, 5) - pnorm(170, 170, 5),
        b = pnorm(175,   170, 5) - pnorm(170, 170, 5),
        c = pnorm(170,   170, 5) - pnorm(165, 170, 5),
        d = pnorm(175,   170, 5) - pnorm(165, 170, 5),
        e = pnorm(180,   170, 5) - pnorm(170, 170, 5),
        f = pnorm(180,   170, 5) - pnorm(160, 170, 5),
        g = 1 - pnorm(170, 170, 5),
        h = 1 - pnorm(175, 170, 5),
        i =     pnorm(175, 170, 5),
        j =     pnorm(165, 170, 5)), 4)
```

```
     a      b      c      d      e      f      g      h      i      j
0.1772 0.3413 0.3413 0.6827 0.4772 0.9545 0.5000 0.1587 0.8413 0.1587
```

Todos os dez batem com o texto. Os itens (d) e (f) saem como 0,6827 e 0,9545, contra os 0,6826 e 0,9544 obtidos ao dobrar os valores tabelados — de novo, o arredondamento da tabela, não um erro. E, como observado no texto, (h) e (j) coincidem por simetria.

Normal: o papel de μ e de σ. Os dois painéis reproduzem os gráficos dos slides. À esquerda, variamos μ mantendo σ fixo; à direita, o contrário.

```r
gx <- seq(-9, 12, length.out = 500)
d_mu <- rbind(
  data.frame(x = gx, f = dnorm(gx, -2, 2), p = "mu = -2"),
  data.frame(x = gx, f = dnorm(gx,  0, 2), p = "mu = 0"),
  data.frame(x = gx, f = dnorm(gx,  3, 2), p = "mu = 3"))
```

```r
d_sg <- rbind(
  data.frame(x = gx, f = dnorm(gx, 1, 1),   p = "sigma = 1"),
  data.frame(x = gx, f = dnorm(gx, 1, 2),   p = "sigma = 2"),
  data.frame(x = gx, f = dnorm(gx, 1, 3.5), p = "sigma = 3,5"))

p1 <- ggplot(d_mu, aes(x, f, colour = p, linetype = p)) +
  geom_line(linewidth = 0.8) + scale_colour_manual(values = c(az, vd, vm)) +
  labs(x = "x  |  sigma = 2 fixo", y = "f(x)")
p2 <- ggplot(d_sg, aes(x, f, colour = p, linetype = p)) +
  geom_line(linewidth = 0.8) + scale_colour_manual(values = c(az, vd, vm)) +
  labs(x = "x  |  mu = 1 fixo", y = "f(x)")

library(patchwork)
p1 + p2
```

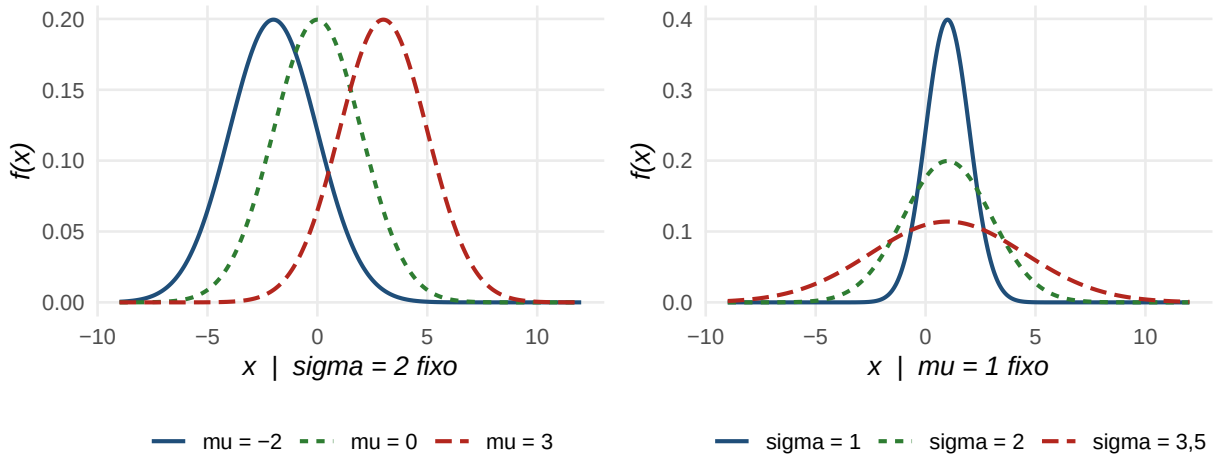


Figura 8: Os dois parâmetros da Normal. Esquerda: variar μ desloca a curva sem deformá-la. Direita: variar σ altera a dispersão — curvas mais largas são necessariamente mais baixas, porque a área permanece igual a 1.

O painel da direita torna visível uma restrição que a fórmula esconde: a densidade $N(1;\,3{,}5^2)$ é tão mais baixa quanto mais larga, porque a área sob qualquer densidade vale 1. Dispersão maior significa não só cauda mais espalhada, mas **menos massa perto da média**.

O VPL. Rápida verificação do segundo exemplo normal:

```r
round(c(a = pnorm(83, 80, 4) - pnorm(80, 80, 4),
        b = pnorm(82, 80, 4) - pnorm(79, 80, 4)), 4)
```

```
     a      b
0.2734 0.2902
```

0,2734 e 0,2902, como no texto.

A lognormal. Primeiro o exemplo do ativo, pelos dois caminhos — `plnorm` direto, e a padronização manual sobre `pnorm`:

```
mu_l <- 5.11; sg_l <- 1
round(c(P_menor164_plnorm = plnorm(164, mu_l, sg_l),
        P_menor164_pnorm  = pnorm((log(164) - mu_l) / sg_l),
        ln_164            = log(164),
        z                 = (log(164) - mu_l) / sg_l,
        aprox_tabela      = pnorm(-0.01)), 4)
```

```
P_menor164_plnorm  P_menor164_pnorm            ln_164                 z
           0.4960            0.4960            5.0999           -0.0101
     aprox_tabela
           0.4960
```

Os dois caminhos dão 0,4960, e a aproximação de sala — arredondar $\ln 164$ para 5,10, obter $z = -0{,}01$ e consultar a tabela — reproduz o mesmo número. Agora os momentos, confrontando as fórmulas fechadas com meio milhão de sorteios:

```
media_t <- exp(mu_l + sg_l^2 / 2)
var_t   <- exp(2 * mu_l + sg_l^2) * (exp(sg_l^2) - 1)

set.seed(20267)
Y <- rlnorm(500000, mu_l, sg_l)

round(rbind(
  teorico  = c(media = media_t, variancia = var_t, dp = sqrt(var_t),
               mediana = exp(mu_l), moda = exp(mu_l - sg_l^2)),
  simulado = c(mean(Y), var(Y), sd(Y), median(Y), NA)), 2)
```

```
          media variancia     dp mediana  moda
teorico  273.14  128197.2 358.05  165.67 60.95
simulado 273.28  128928.4 359.07  165.83    NA
```

Média e mediana simuladas reproduzem as teóricas com três dígitos significativos, e a variância — muito mais difícil de estimar, por depender da cauda — fica dentro de 1%. Repare na magnitude: o desvio padrão (358) é **maior que a média** (273), de modo que CV > 1. É uma distribuição de cauda pesada, e a moda de 61 mostra o quão longe da média fica o valor mais provável.

A assimetria, graficamente. A figura fecha o capítulo conectando-o a Seção 2:

```
gy   <- seq(0.1, 1200, length.out = 800)
d_ln <- data.frame(y = gy, f = dlnorm(gy, mu_l, sg_l))
marcas <- data.frame(
  medida = factor(c("moda", "mediana", "media"),
```

```
                  levels = c("moda", "mediana", "media")),
  y = c(exp(mu_l - sg_l^2), exp(mu_l), exp(mu_l + sg_l^2 / 2)))

ggplot(d_ln, aes(y, f)) +
  geom_area(fill = az, alpha = 0.18) +
  geom_line(colour = az, linewidth = 0.8) +
  geom_vline(data = marcas, aes(xintercept = y, colour = medida, linetype = medida),
             linewidth = 0.8) +
  scale_colour_manual(values = c(moda = lr, mediana = vd, media = vm)) +
  scale_linetype_manual(values = c(moda = "dotted", mediana = "dashed", media = "solid")) +
  labs(x = "y (preco do ativo)", y = "f(y)")
```

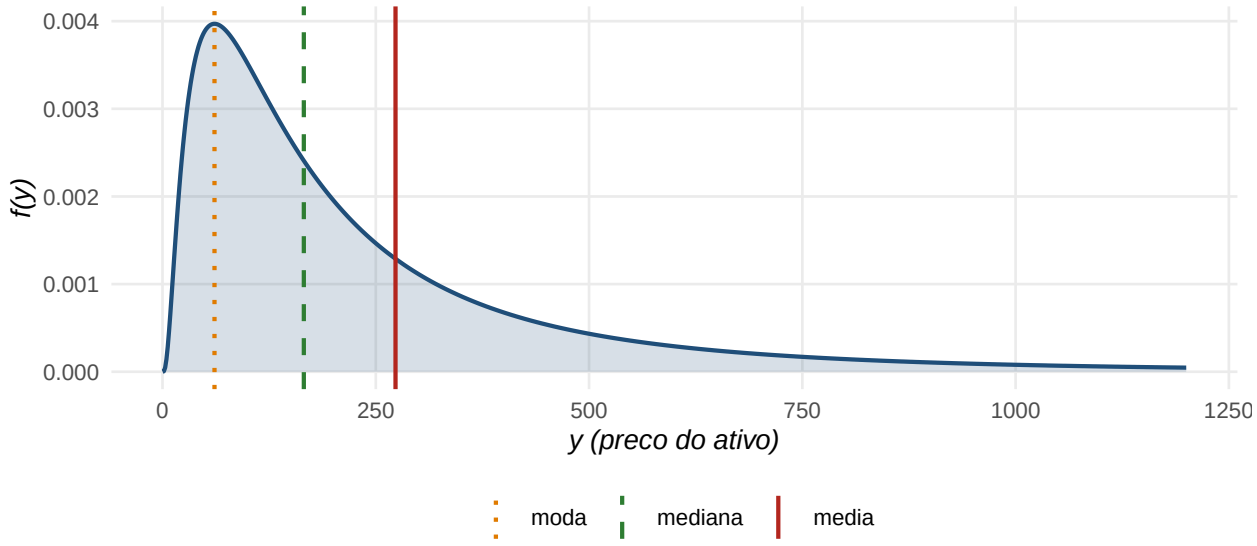


Figura 9: Densidade lognormal com $\mu = 5{,}11$ e $\sigma = 1$. A assimetria positiva separa as três medidas de posição na ordem canônica moda $<$ mediana $<$ média — aqui 60,95, 165,67 e 273,14. Compare com a densidade simétrica da Normal, em que as três coincidem.

As três linhas verticais estão nitidamente separadas, e na ordem prevista. É a mesma figura conceitual da densidade $3x^2/8$ de Seção 4 — só que ali a assimetria era negativa (moda na fronteira direita) e aqui é positiva. Vale contrastar com o histograma aproximadamente simétrico dos faturamentos de Seção 2, onde média e mediana quase coincidiam: **a distância entre as medidas de posição é a assimetria**, e a lognormal é o modelo que a produz por construção.

6.7 Exercícios

1. **Uniforme.** O tempo de duração de uma reunião distribui-se uniformemente entre 45 e 90 minutos. (a) Escreva $f(x)$ e $F(x)$. (b) Calcule $P(X > 60)$, $P(50 < X < 70)$ e $P(X = 60)$.

   (c) Obtenha $\mathbb{E}(X)$, $\mathrm{V}(X)$, $\mathrm{DP}(X)$ e $\mathrm{CV}(X)$. (d) Mostre, em geral, que para $X \sim \mathrm{Unif}(\alpha, \beta)$ a mediana coincide com a média, e explique por que isso decorre da simetria da densidade.

2. **Exponencial: os três objetos.** Uma central de atendimento recebe em média 12 chamados por hora. (a) Identifique λ na unidade "por minuto" e escreva f e F.

   (b) Calcule a probabilidade de o intervalo entre dois chamados consecutivos exceder 10 minutos. (c) Obtenha a mediana e verifique que $Md/\mathbb{E}(X) = \ln 2$. (d) Confirme que $P(X > \mathbb{E}(X)) = e^{-1}$ e explique por que esse número não depende de λ.

3. **Demonstrações da exponencial.** Seja $X \sim \text{Expo}(\lambda)$. (a) Obtenha $F(x)$ por integração direta. (b) Demonstre $\mathbb{E}(X) = 1/\lambda$ por integração por partes, justificando cuidadosamente por que o termo de fronteira se anula em $x \to \infty$. (c) Aplique integração por partes duas vezes para obter $\mathbb{E}(X^2) = 2/\lambda^2$ e conclua $\text{V}(X) = 1/\lambda^2$. (d) Mostre que $\text{CV}(X) = 1$ para todo λ e discuta como usar esse fato como teste diagnóstico informal em uma amostra de tempos observados.

4. **Falta de memória e sua crítica.** (a) Demonstre que $P(X > x + s \mid X > x) = P(X > s)$ para a exponencial, indicando em que passo se usa que $\{X > x + s\} \subset \{X > x\}$. (b) Um motor tem vida útil exponencial com média de 8 anos e já opera há 5; qual a probabilidade de operar ao menos mais 3? (c) Um engenheiro objeta que a resposta é implausível. Formule a objeção dele com precisão e diga que característica dos dados de durabilidade a exponencial é incapaz de reproduzir. (d) Que modelo alternativo você sugeriria e qual parâmetro adicional ele introduziria?

5. **Normal: o método completo.** O peso de pacotes de café tem distribuição $N(500, 64)$ gramas. (a) Identifique μ e σ — atenção ao segundo argumento.

   (b) Calcule $P(X < 490)$, $P(495 < X < 510)$ e $P(X > 516)$. (c) Determine o valor c tal que $P(X > c) = 0{,}05$. (d) Se o rótulo declara 500 g e a fiscalização autua pacotes abaixo de 485 g, que fração da produção está sujeita a autuação? (e) Que μ a empresa deveria calibrar para reduzir essa fração a 0,1%, mantendo $\sigma = 8$?

6. **Regra empírica e outliers.** (a) Deduza as três linhas da regra empírica a partir de $\Phi(1) = 0{,}8413$, $\Phi(2) = 0{,}9772$ e $\Phi(3) = 0{,}9987$. (b) Mostre que os valores não dependem de μ nem de σ, apelando à padronização. (c) Retome o critério de outlier de Seção 2 — valores fora de $[\mu - 3\sigma, \mu + 3\sigma]$ — e calcule quantas observações atípicas se esperaria, sob normalidade, em uma amostra de 10.000 pontos.

   (d) Compare com o critério de $1{,}5\,\Delta Q$ do box-plot: qual dos dois é mais rigoroso sob normalidade?

7. **Lognormal.** O faturamento anual das firmas de um setor, em milhões de reais, é lognormal com $\mu = 3$ e $\sigma^2 = 0{,}64$. (a) Calcule $\mathbb{E}(Y)$, Md e Mo e ordene-os.

   (b) Obtenha $P(Y < 20)$ e $P(Y > 40)$. (c) Que fração das firmas fatura acima da média do setor? Compare com os 50% que ficariam acima da mediana e interprete economicamente.
   (c) Mostre que $\mathbb{E}(Y) > e^{\mu}$ para todo $\sigma > 0$ e explique por que exponenciar a média dos logs subestima a média dos níveis.

8. **Da Normal à lognormal, e a escolha do modelo.** (a) Deduza a densidade lognormal a partir de $Y = e^X$ com $X \sim N(\mu, \sigma^2)$, explicitando de onde vem o fator $1/y$.

(b) Argumente por que a Normal é inadequada para modelar preços de ativos, apontando duas propriedades incompatíveis. (c) Explique, usando $\ln P_T = \ln P_0 + \sum_t \ln(1 + R_t)$ e Seção 7, por que a lognormalidade dos preços é uma consequência quase automática da hipótese de log-retornos i.i.d. (d) Renda e preços de ativos são ambos modelados por lognormais; aponte uma diferença econômica relevante entre os dois casos que a distribuição, sozinha, não captura.

7 Funções Lineares de Variáveis Aleatórias

Os cinco capítulos anteriores construíram, peça por peça, o aparato probabilístico: o que é uma variável aleatória, como calcular sua esperança e sua variância, e que modelos — Binomial, Poisson, Exponencial, Normal — descrevem os fenômenos típicos. Em todos eles, porém, olhamos **uma** variável de cada vez. Este capítulo remove essa restrição, e ao removê-la muda a natureza do curso.

A razão é simples de enunciar e difícil de exagerar. Ninguém observa uma variável aleatória isolada: observam-se n delas, $X_1, X_2, \ldots, X_n$ — os retornos de n dias, os pesos de n pacotes, as respostas de n entrevistados —, e a pergunta que se faz é sobre a **soma** ou sobre a **média** desses valores. A média amostral $\bar{X}$ não é um número: é uma função de n variáveis aleatórias e, portanto, **é ela mesma uma variável aleatória**. Descobrir qual é a distribuição de $\bar{X}$ é descobrir o que se pode e o que não se pode afirmar a partir de uma amostra. É por isso que este é o capítulo-charneira do curso: aqui a probabilidade vira inferência.

O capítulo tem duas metades. Na primeira, estudamos somas e médias de variáveis aleatórias e chegamos ao **Teorema Central do Limite**, o resultado mais importante da estatística. Na segunda, aplicamos o maquinário a economia e finanças: retorno e volatilidade, comparação de estratégias por probabilidade de perda e a teoria de carteiras — onde a covariância de Seção 2 reaparece, agora com a função precisa de explicar por que a diversificação funciona.

> **!** O ponto de chegada, antecipado
>
> Ao fim deste capítulo saberemos que, se $X_1, \ldots, X_n$ são independentes com média μ e variância σ^2 finitas, então para n grande
>
> $$\bar{X} \overset{\text{aprox.}}{\sim} N\left(\mu,\ \frac{\sigma^2}{n}\right)$$
>
> **qualquer que seja a distribuição das** X_i. Guarde esta linha. Tudo o que se fará em Seção 8, Seção 9 e Seção 10 é consequência dela.

7.1 Somas de variáveis aleatórias

Comecemos pelo objeto mais simples. Dadas n variáveis aleatórias $X_1, X_2, \ldots, X_n$, defina a **soma**

$$S = \sum_{i=1}^{n} X_i = X_1 + X_2 + \cdots + X_n.$$

Como cada X_i é uma função do espaço amostral em $\mathbb{R}$, a soma delas também é — logo S é uma variável aleatória, com sua própria distribuição, sua própria esperança e sua própria variância. Queremos as duas últimas.

7.1.1 A esperança da soma: sempre vale

! Esperança de uma soma

Para quaisquer variáveis aleatórias $X_1, \dots, X_n$ com esperanças finitas,

$$\boxed{\mathbb{E}(S) = \mathbb{E}\left(\sum_{i=1}^{n} X_i\right) = \sum_{i=1}^{n} \mathbb{E}(X_i) = \mathbb{E}(X_1) + \mathbb{E}(X_2) + \cdots + \mathbb{E}(X_n)}$$

Nenhuma hipótese sobre a relação entre as variáveis é necessária. Elas podem ser independentes, correlacionadas, funções umas das outras — a igualdade vale sempre.

Vale insistir nesse ponto, porque ele é fonte perene de confusão. A esperança é um operador **linear**, e linearidade é uma propriedade da operação de somar, não da estrutura de dependência. No caso de duas variáveis discretas, a demonstração é uma troca de ordem de somatório sobre a distribuição conjunta:

$$\mathbb{E}(X+Y) = \sum_x \sum_y (x+y) P(X=x, Y=y) = \sum_x x \sum_y P(x,y) + \sum_y y \sum_x P(x,y) = \mathbb{E}(X) + \mathbb{E}(Y),$$

porque $\sum_y P(x,y) = P(X=x)$ e $\sum_x P(x,y) = P(Y=y)$ são as distribuições marginais. Em nenhum passo se usou $P(x,y) = P(x)P(y)$. O caso geral segue por indução em n.

A intuição econômica é direta: o faturamento esperado de uma rede é a soma dos faturamentos esperados das filiais, mesmo que as filiais compitam entre si, mesmo que uma canibalize a outra, mesmo que todas respondam ao mesmo ciclo macroeconômico. A dependência afeta o **risco** do faturamento agregado, jamais o seu valor esperado.

7.1.2 A variância da soma: exige descorrelação

Com a variância a história é outra. Aqui a estrutura de dependência entra, e entra de forma essencial.

! Variância de uma soma

Se $X_1, \dots, X_n$ são **independentes** — ou, mais fracamente, apenas **descorrelacionadas duas a duas**, isto é, $\mathrm{Cov}(X_i, X_j) = 0$ para todo $i \neq j$ —, então

$$\boxed{\mathrm{V}(S) = \mathrm{V}\left(\sum_{i=1}^{n} X_i\right) = \sum_{i=1}^{n} \mathrm{V}(X_i) = \mathrm{V}(X_1) + \mathrm{V}(X_2) + \cdots + \mathrm{V}(X_n)}$$

Sem essa hipótese a igualdade **falha**, e a fórmula correta traz termos de covariância — como veremos na seção sobre combinações lineares.

Repare na assimetria entre os dois resultados: **médias somam sempre; variâncias somam só sob descorrelação**. Essa assimetria é o motor de tudo o que vem depois, inclusive da diversificação

de carteiras: se conseguimos combinar ativos cujas covariâncias sejam negativas, o retorno esperado da carteira continua sendo a média ponderada dos retornos esperados, mas o risco fica **abaixo** da soma dos riscos. É almoço grátis, e é o único que a teoria de finanças reconhece.

Note também que as variâncias **somam**, nunca se subtraem. Mesmo a variância de uma *diferença* de variáveis independentes é a **soma** das variâncias: $\mathrm{V}(X-Y) = \mathrm{V}(X) + \mathrm{V}(Y)$, pois $\mathrm{V}(-Y) = (-1)^2\mathrm{V}(Y) = \mathrm{V}(Y)$ pela propriedade (2) de Seção 4. Incerteza acumula; ela não se cancela por subtração.

7.1.3 Soma de Normais independentes

Conhecer $\mathbb{E}(S)$ e $\mathrm{V}(S)$ não é conhecer a distribuição de S. Há, porém, um caso em que a distribuição sai de graça — e é o caso que mais usaremos.

! Soma de Normais independentes

Se $X_1, \ldots, X_n$ são **independentes** e cada uma tem distribuição $N(\mu_i, \sigma_i^2)$, então

$$S = \sum_{i=1}^{n} X_i \sim N\left(\sum_{i=1}^{n} \mu_i,\ \sum_{i=1}^{n} \sigma_i^2\right).$$

No caso particular em que todas têm a **mesma** média μ e a **mesma** variância σ^2,

$$\boxed{S \sim N(n\mu,\ n\sigma^2)}$$

A propriedade é notável e não é genérica: somar duas exponenciais não dá uma exponencial, somar duas uniformes não dá uma uniforme. A Normal é **estável sob adição**, e é uma das razões pelas quais ela ocupa o lugar que ocupa. A demonstração formal exige a técnica das funções geradoras de momentos ou a convolução de densidades, que estão além do escopo do curso; o resultado, entretanto, é de uso imediato.

Observe o comportamento das duas escalas. A média da soma cresce com n; o desvio padrão cresce com $\sqrt{n}$, pois $\mathrm{DP}(S) = \sqrt{n\sigma^2} = \sigma\sqrt{n}$. A soma fica maior e mais dispersa em termos absolutos, mas *relativamente* menos dispersa: o coeficiente de variação de S é $\sigma\sqrt{n}/(n\mu) = \mathrm{CV}(X)/\sqrt{n}$, que tende a zero. Agregar reduz incerteza relativa — é o princípio de toda a indústria de seguros.

i Ferramenta matemática — linearidade do somatório e a raiz de n

Os dois resultados desta seção repousam sobre propriedades elementares do somatório: $\sum(a_i + b_i) = \sum a_i + \sum b_i$ e $\sum c\,a_i = c\sum a_i$. O fato de a variância envolver quadrados é o que produz o $\sqrt{n}$: somar n variâncias iguais dá $n\sigma^2$, e a raiz disso é $\sigma\sqrt{n}$, não $n\sigma$. A distinção entre crescimento linear (da média) e crescimento em $\sqrt{n}$ (do desvio padrão) é o conteúdo geométrico de todo este capítulo. *(Propriedades do somatório e funções potência: capítulo* Nivelamento *das Notas de Matemática.)*

7.1.4 Exemplo: o elevador

Um elevador tem capacidade declarada de 500 kg. O peso dos usuários é uma variável aleatória Normal com média $\mu = 70$ kg e variância $\sigma^2 = 100$ (portanto $\sigma = 10$ kg). Sete pessoas entram no elevador. Qual a probabilidade de a capacidade ser excedida?

O peso total é $S = X_1 + \cdots + X_7$, soma de sete Normais independentes com mesma média e mesma variância. Pelo resultado anterior,

$$\mathbb{E}(S) = 7 \times 70 = 490, \qquad \mathrm{V}(S) = 7 \times 100 = 700, \qquad S \sim N(490,\ 700),$$

com $\mathrm{DP}(S) = \sqrt{700} \approx 26{,}46$ kg. Padronizando na forma de Seção 6,

$$P(S > 500) = P\left(Z > \frac{500 - 490}{\sqrt{700}}\right) = P(Z > 0{,}38) = 1 - 0{,}6480 = 0{,}3520.$$

Há **35,2%** de chance de o elevador exceder a capacidade com sete ocupantes — um número desconfortavelmente alto, e o tipo de cálculo que sustenta normas de segurança. Note que a resposta não seria obtida comparando 500 com $7 \times 70 = 490$ e concluindo "cabe": o que importa é a distribuição inteira de S, não apenas seu centro.

7.1.5 Exemplo: os pacotes de café

Uma torrefadora envasa pacotes cujo peso, em gramas, segue $N(500, 16)$ — logo $\mu = 500$ e $\sigma = 4$. Uma caixa contém $n = 100$ pacotes, com pesos independentes.

(a) Qual a probabilidade de a caixa pesar mais de 49.960 g? O peso da caixa é a soma:

$$S \sim N(100 \times 500,\ 100 \times 16) = N(50.000,\ 1.600),$$

com $\mathrm{DP}(S) = \sqrt{1.600} = 40$ g. Então

$$P(S > 49.960) = P\left(Z > \frac{49.960 - 50.000}{40}\right) = P(Z > -1) = P(Z < 1) = 0{,}8413.$$

Com 84,13% de probabilidade a caixa supera 49.960 g. Voltaremos a este exemplo em um instante, na versão que interessa à inferência.

7.2 A média amostral é uma variável aleatória

Chegamos ao conceito que reorganiza o curso.

Considere n variáveis aleatórias $X_1, \dots, X_n$ e defina

$$\boxed{\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i = \frac{X_1 + X_2 + \cdots + X_n}{n}}$$

Formalmente, $\bar{X}$ é apenas S/n — a soma multiplicada pela constante $1/n$. Mas a mudança de perspectiva que ela introduz é radical, e vale enunciá-la sem economia de palavras.

> **!** O ponto crucial do capítulo
>
> Em Seção 2, a média $\mu = \sum x_i/n$ era um **número**: calculado sobre trinta faturamentos observados, dava 10,3 e pronto. Aqui, $\bar{X} = \sum X_i/n$ é uma **variável aleatória**: função de n variáveis aleatórias, ela tem valores possíveis, distribuição de probabilidade, esperança e variância próprias.
> A diferença é a diferença entre *antes* e *depois* de coletar a amostra. Depois de coletada, temos $\bar{x}$, um número. Antes de coletá-la — que é o momento em que se raciocina sobre o que a amostra pode dizer —, temos $\bar{X}$, uma variável aleatória cuja distribuição descreve *todas as amostras que poderíamos ter obtido*. Essa distribuição chama-se **distribuição amostral da média**, e conhecê-la é o que permite afirmar coisas como "com 95% de confiança, μ está entre tais limites".
> Sem esta observação não existe inferência estatística. Com ela, existe Seção 8 e existe Seção 9.

7.2.1 Esperança e variância de $\bar{X}$

As duas demonstrações a seguir são curtas e devem ser sabidas de cor.

Esperança. Usando $\mathbb{E}(aX) = a\mathbb{E}(X)$ e a linearidade da esperança, e supondo apenas que todas as X_i tenham a mesma média $\mathbb{E}(X_i) = \mu$ (nem sequer é preciso independência):

$$\mathbb{E}(\bar{X}) = \mathbb{E}\left(\frac{1}{n}\sum_{i=1}^{n} X_i\right) = \frac{1}{n}\,\mathbb{E}\left(\sum_{i=1}^{n} X_i\right) = \frac{1}{n}\sum_{i=1}^{n}\mathbb{E}(X_i) = \frac{1}{n}\cdot n\mu$$

$$\boxed{\mathbb{E}(\bar{X}) = \mu}$$

Variância. Agora usando $\mathrm{V}(aX) = a^2\mathrm{V}(X)$ e a aditividade da variância — que **exige** descorrelação —, com $\mathrm{V}(X_i) = \sigma^2$ para todo i:

$$\mathrm{V}(\bar{X}) = \mathrm{V}\left(\frac{1}{n}\sum_{i=1}^{n} X_i\right) = \frac{1}{n^2}\,\mathrm{V}\left(\sum_{i=1}^{n} X_i\right) = \frac{1}{n^2}\sum_{i=1}^{n}\mathrm{V}(X_i) = \frac{1}{n^2}\cdot n\sigma^2$$

$$\boxed{\mathrm{V}(\bar{X}) = \frac{\sigma^2}{n}}$$

Os dois resultados dizem coisas muito diferentes, e é a combinação delas que faz a estatística funcionar.

O primeiro diz que $\bar{X}$ está **centrada no parâmetro que se quer conhecer**. Em média, sobre todas as amostras possíveis, a média amostral acerta μ — nem superestima sistematicamente, nem subestima. Em Seção 8 chamaremos essa propriedade de **não viés**, e ela é a primeira coisa que se exige de um estimador.

O segundo diz que a **dispersão de $\bar{X}$ encolhe com n**. A variância da média amostral é a variância populacional dividida pelo tamanho da amostra: quanto mais observações, mais concentrada em torno de μ fica a distribuição de $\bar{X}$, e quando $n \to \infty$ a variância vai a zero. Isso é a **lei dos grandes números** que já havíamos visto operar por simulação em Seção 4, agora com uma taxa explícita.

7.2.2 O erro-padrão

Tomando a raiz da variância de $\bar{X}$:

$$\boxed{\mathrm{DP}(\bar{X}) = \frac{\sigma}{\sqrt{n}}}$$

Esta quantidade é tão central que tem nome próprio na literatura: chama-se **erro-padrão da média**.

> **!** Sobre o termo erro-padrão
>
> O professor não emprega o termo "erro-padrão" neste módulo, referindo-se sempre a $\sigma/\sqrt{n}$ como o desvio padrão de $\bar{X}$ — o que é rigorosamente correto e evita uma palavra que confunde iniciantes ("erro" sugere engano, e não há engano nenhum aqui). Mas o termo é padrão em toda a literatura e em todo *output* de software, e reaparecerá em Seção 9 e Seção 10. Vale fixar desde já: **erro-padrão de um estimador é o desvio padrão de sua distribuição amostral**. Para a média, é $\sigma/\sqrt{n}$.

O que merece atenção é o $\sqrt{n}$ **no denominador**, e não o n. A precisão da média amostral melhora, mas melhora devagar: para reduzir o erro-padrão à metade é preciso **quadruplicar** a amostra; para reduzi-lo a um décimo, multiplicá-la por cem. É a lei dos rendimentos decrescentes da informação, e é o motivo econômico pelo qual pesquisas de opinião param em mil ou dois mil entrevistados — dobrar esse número custa o dobro e melhora a precisão em apenas 41%.

7.2.3 Distribuição de $\bar{X}$ sob normalidade

Se as X_i são Normais independentes, sabemos que $S \sim N(n\mu, n\sigma^2)$, e $\bar{X}$ é uma transformação linear de S com $a = 1/n$. Como transformações lineares de Normais são Normais:

$$\boxed{X_i \sim N(\mu, \sigma^2) \text{ independentes} \quad \Longrightarrow \quad \bar{X} \sim N\left(\mu,\ \frac{\sigma^2}{n}\right)}$$

Não é apenas que $\bar{X}$ tenha média μ e variância σ^2/n — sabemos sua **distribuição inteira**, e portanto podemos calcular a probabilidade de qualquer evento envolvendo $\bar{X}$, padronizando como sempre:

$$Z = \frac{\bar{X} - \mu}{\sigma/\sqrt{n}} \sim N(0, 1).$$

Esta é a estatística que aparecerá em cada intervalo de confiança e em cada teste de hipótese do resto do curso. Repare no denominador: **é o erro-padrão, não** σ. Trocar um pelo outro é o erro de conta mais comum da disciplina.

7.2.4 Exemplo: os pacotes de café, versão média

Retomemos a torrefadora, com pacotes $N(500, 16)$ e $n = 100$.

(b) Qual a probabilidade de o peso médio dos 100 pacotes ser inferior a 500,7 g?

Agora a variável de interesse é $\bar{X}$, não S. Temos

$$\mathbb{E}(\bar{X}) = 500, \qquad \mathrm{V}(\bar{X}) = \frac{16}{100} = 0{,}16, \qquad \mathrm{DP}(\bar{X}) = \sqrt{0{,}16} = 0{,}4 \text{ g},$$

de modo que $\bar{X} \sim N(500;\ 0{,}16)$. Padronizando,

$$P(\bar{X} < 500{,}7) = P\left(Z < \frac{500{,}7 - 500}{0{,}4}\right) = P(Z < 1{,}75) = 0{,}9599.$$

Vale contrastar as duas perguntas do exemplo. A dispersão do peso de **um** pacote é $\sigma = 4$ g; a dispersão da **média** de cem pacotes é 0,4 g — dez vezes menor, porque $\sqrt{100} = 10$. Um desvio de 0,7 g é irrelevante para um pacote individual (menos de 0,2 desvio padrão) e é grande para a média de cem (1,75 desvios padrão). É exatamente esse efeito que torna possível detectar desregulagens pequenas de uma máquina inspecionando amostras: **a média é um instrumento muito mais preciso que a observação individual**.

7.2.5 Exemplo: as baterias

A vida útil de certa bateria automotiva tem média $\mu = 40$ meses e desvio padrão $\sigma = 8$ meses. Toma-se uma amostra de $n = 64$ baterias. Qual a probabilidade de a vida média da amostra ficar entre 38 e 42 meses?

O erro-padrão é

$$\mathrm{DP}(\bar{X}) = \frac{8}{\sqrt{64}} = \frac{8}{8} = 1 \text{ mês},$$

de modo que $\bar{X} \sim N(40, 1)$ e os limites 38 e 42 estão exatamente a **dois erros-padrão** da média. Pela regra empírica de Seção 6, ou padronizando,

$$P(38 < \bar{X} < 42) = P(-2 < Z < 2) = 0{,}9772 - 0{,}0228 = 0{,}9544.$$

Note dois pontos. Primeiro: o enunciado **não afirma que a vida útil das baterias seja Normal** — e não precisa, porque com $n = 64$ o Teorema Central do Limite, que enunciaremos a seguir, garante a normalidade aproximada de $\bar{X}$. Segundo: o intervalo $[38, 42]$ tem apenas ± 2 meses de largura em torno de 40, enquanto a vida de uma bateria individual tem desvio padrão de 8 meses. Não há contradição: baterias individuais variam muito, médias de 64 baterias variam pouco.

7.3 O Teorema Central do Limite

Tudo o que fizemos até aqui sobre a distribuição de S e de $\bar{X}$ pressupôs que as X_i fossem Normais. É uma hipótese forte e, em geral, falsa: pesos, tempos, retornos, contagens raramente são exatamente Normais. Se a normalidade da amostra fosse condição necessária, a inferência estatística seria uma técnica de aplicação restrita.

O Teorema Central do Limite (TCL) diz que não é.

! Teorema Central do Limite

A soma e a média de n variáveis aleatórias independentes, com médias e variâncias finitas — quaisquer que sejam suas distribuições —, seguem distribuição aproximadamente Normal se n é suficientemente grande $(n > 30)$**.**
Em símbolos, se $X_1, \ldots, X_n$ são independentes com $\mathbb{E}(X_i) = \mu$ e $\mathrm{V}(X_i) = \sigma^2$ finitas, então para n suficientemente grande

$$\boxed{S = \sum_{i=1}^{n} X_i \overset{\text{aprox.}}{\sim} N(n\mu,\ n\sigma^2) \qquad \text{e} \qquad \bar{X} \overset{\text{aprox.}}{\sim} N\left(\mu,\ \frac{\sigma^2}{n}\right)}$$

equivalentemente, em forma padronizada,

$$Z = \frac{\bar{X} - \mu}{\sigma/\sqrt{n}} \overset{\text{aprox.}}{\sim} N(0, 1).$$

7.3.1 Por que é o resultado mais importante do curso

Três observações explicam a reverência.

Primeira: a hipótese que ele dispensa. As fórmulas $\mathbb{E}(\bar{X}) = \mu$ e $\mathrm{V}(\bar{X}) = \sigma^2/n$ valem sempre (sob descorrelação), mas sozinhas não permitem calcular probabilidade nenhuma — para isso é preciso a distribuição. O TCL fornece a distribuição **sem exigir nada sobre a distribuição das parcelas**, além de médias e variâncias finitas. A cláusula "quaisquer que sejam suas distribuições" é a coisa notável do teorema: a população pode ser exponencial, uniforme, Bernoulli, bimodal, grotescamente assimétrica — a média amostral converge para a Normal do mesmo jeito. A Figura 10 mostra isso acontecendo com três populações escolhidas justamente por serem o mais distantes possível da Normal.

Segunda: por que a Normal é onipresente. O TCL explica por que tantas grandezas empíricas são aproximadamente Normais — alturas, erros de medida, notas, produtividade. Não é coincidência nem harmonia pré-estabelecida: é que essas grandezas resultam da **soma de muitos efeitos pequenos e razoavelmente independentes**. A altura de uma pessoa agrega centenas de contribuições genéticas e nutricionais; o erro de um instrumento agrega imperfeições diversas. Somas de muitas coisas tendem à Normal, e o mundo é feito de somas. (O mesmo argumento em escala multiplicativa produz a lognormal de Seção 6: se os *logaritmos* é que se somam, o nível é lognormal.)

Terceira: é a licença para toda a inferência. Em Seção 9 vamos afirmar que a média populacional μ pertence a $\bar{x} \pm 1{,}96\,\sigma/\sqrt{n}$ com 95% de confiança. O 1,96 vem da Normal padrão. O direito de usar a Normal padrão — sem saber se a população é Normal — vem do TCL, e só dele. Retire o teorema e todo o edifício da inferência clássica desaba.

7.3.2 O que "suficientemente grande" quer dizer

A regra $n > 30$ merece ser desmontada com cuidado, porque é ensinada como se fosse parte do teorema. **Não é.**

O TCL é um resultado **assintótico**: ele afirma que a distribuição de Z converge para a $N(0,1)$ quando $n \to \infty$. Rigorosamente, para n finito qualquer, a aproximação é apenas isso — uma aproximação, com erro não nulo. O teorema não diz quão grande n precisa ser para que o erro seja tolerável, porque isso **depende da distribuição das parcelas**:

- se a população já é **simétrica e sem caudas pesadas**, a convergência é rapidíssima: com uma uniforme, $n = 5$ já produz um histograma visualmente indistinguível de uma Normal, e mesmo $n = 2$ dá uma distribuição triangular razoavelmente comportada;
- se a população é **fortemente assimétrica** — uma exponencial, um retorno com caudas gordas —, a convergência é mais lenta, e $n = 30$ pode ser insuficiente para as caudas, ainda que já seja bom para o centro;
- o caso patológico é a **Bernoulli com p extremo**: com $p = 0{,}01$, uma amostra de 30 observações tem em média 0,3 sucessos, e a distribuição de $\bar{X}$ é essencialmente discreta e concentrada em zero — longe de qualquer Normal. A regra prática usual para proporções, que reencontraremos em Seção 9, é exigir $np \geq 5$ **e** $n(1-p) \geq 5$.

! A regra $n > 30$ é convenção, não teorema

$n > 30$ é uma **regra de bolso**, calibrada para funcionar razoavelmente bem em casos típicos. Não há nada de matematicamente especial no número 30, e nenhum enunciado rigoroso do TCL o menciona. Ele deve ser lido como "a partir de algumas dezenas de observações, e desde que a população não seja patologicamente assimétrica, a aproximação Normal costuma ser aceitável". Na prática: com população simétrica, muito menos que 30 basta; com assimetria severa ou caudas pesadas, 30 pode ser insuficiente. E há um caso em que n nenhum resolve — quando a variância populacional é infinita, o TCL simplesmente não se aplica. Distribuições de retornos de ativos em alta frequência flertam com esse regime, o que é um dos motivos pelos quais modelos financeiros baseados em normalidade falham justamente nas crises.

7.3.3 Exemplo: as latas na prateleira

Uma prateleira suporta 200 kg. Sobre ela serão colocadas 49 latas de tinta, cujo peso tem média 4 kg e desvio padrão 1 kg. Os pesos são independentes. Qual a probabilidade de a prateleira ceder?

Note que o enunciado **não diz qual é a distribuição do peso de uma lata** — e não precisa dizer. Como $n = 49 > 30$, o TCL garante que a soma é aproximadamente Normal com

$$\mathbb{E}(S) = 49 \times 4 = 196 \text{ kg}, \qquad \mathrm{V}(S) = 49 \times 1^2 = 49, \qquad \mathrm{DP}(S) = \sqrt{49} = 7 \text{ kg},$$

isto é, $S \overset{\text{aprox.}}{\sim} N(196, 49)$. Então

$$P(S > 200) = P\left(Z > \frac{200 - 196}{7}\right) = P(Z > 0{,}571) = 1 - 0{,}7157 = 0{,}2843.$$

Há **28,4%** de chance de a prateleira ceder. Duas leituras merecem registro. Primeira, do ponto de vista da engenharia: uma margem de 4 kg sobre 196 é ilusória, porque a incerteza agregada é de 7 kg — a folga é menor que um desvio padrão. Segunda, do ponto de vista estatístico: resolvemos um problema sem conhecer a distribuição dos pesos individuais. Bastaram a média, o desvio padrão, a independência e n grande. Essa é, em uma frase, a utilidade prática do TCL.

7.4 Combinações lineares e a covariância

Somas e médias são casos particulares de um objeto mais geral. Dadas variáveis aleatórias $X_1, \ldots, X_n$ e constantes $a_1, \ldots, a_n$, a **combinação linear**

$$\boxed{C = \sum_{i=1}^{n} a_i X_i = a_1 X_1 + a_2 X_2 + \cdots + a_n X_n}$$

recupera a soma quando $a_i = 1$ para todo i, e a média quando $a_i = 1/n$. Sua importância econômica, porém, vem de outro lugar: C é exatamente a estrutura do **retorno de uma carteira**, em que a_i é a fração do capital alocada no ativo i e X_i é o retorno desse ativo.

Concentremo-nos no caso $n = 2$, que é onde toda a intuição está. Sejam X e Y variáveis aleatórias e a, b constantes, e defina $C = aX + bY$.

! Esperança e variância de uma combinação linear

$$\boxed{\mathbb{E}(C) = \mathbb{E}(aX + bY) = a\,\mathbb{E}(X) + b\,\mathbb{E}(Y)}$$

sempre — sem nenhuma hipótese sobre a relação entre X e Y. E

$$\boxed{\mathrm{V}(C) = \mathrm{V}(aX + bY) = a^2\,\mathrm{V}(X) + b^2\,\mathrm{V}(Y) + 2ab\,\mathrm{Cov}(X, Y)}$$

em que $\mathrm{Cov}(X, Y) = \mathbb{E}\big[(X - \mathbb{E}(X))(Y - \mathbb{E}(Y))\big] = \mathbb{E}(XY) - \mathbb{E}(X)\mathbb{E}(Y)$ é a **covariância** entre X e Y. Se X e Y são independentes (ou apenas descorrelacionadas), então $\mathrm{Cov}(X, Y) = 0$ e o termo cruzado desaparece, recuperando $\mathrm{V}(C) = a^2\mathrm{V}(X) + b^2\mathrm{V}(Y)$.

A demonstração é o desenvolvimento de um quadrado. Escrevendo $\mu_X = \mathbb{E}(X)$, $\mu_Y = \mathbb{E}(Y)$ e usando $\mathbb{E}(C) = a\mu_X + b\mu_Y$:

$$\begin{aligned}
\mathrm{V}(C) &= \mathbb{E}\big[(aX + bY - a\mu_X - b\mu_Y)^2\big] = \mathbb{E}\big[\big(a(X - \mu_X) + b(Y - \mu_Y)\big)^2\big] \\
&= \mathbb{E}\big[a^2(X - \mu_X)^2\big] + \mathbb{E}\big[b^2(Y - \mu_Y)^2\big] + \mathbb{E}\big[2ab(X - \mu_X)(Y - \mu_Y)\big] \\
&= a^2\mathrm{V}(X) + b^2\mathrm{V}(Y) + 2ab\,\mathrm{Cov}(X, Y).
\end{aligned}$$

O termo cruzado, que no quadrado de um binômio é $2ab$, aqui carrega a covariância — e é onde toda a ação está.

! O único lugar do curso em que a covariância entra sem independência

Vale explicitar: em todo o restante da disciplina, as variáveis aleatórias são supostas independentes, e por isso as variâncias simplesmente somam. **Esta seção é a exceção.** A fórmula $\mathrm{V}(aX + bY) = a^2\mathrm{V}(X) + b^2\mathrm{V}(Y) + 2ab\mathrm{Cov}(X, Y)$ é o único ponto em que a estrutura de dependência entre variáveis aleatórias entra explicitamente nas contas.
E não é acidente que seja aqui: é precisamente esse termo $2ab\mathrm{Cov}(X, Y)$ que explica a diversificação de carteiras. Sem ele, o risco de uma carteira seria sempre uma média ponderada dos riscos dos ativos, e não haveria razão econômica para diversificar. Toda a teoria moderna de portfólio mora no termo cruzado.

Convém escrever a covariância em termos da correlação de Seção 2. Como $\rho_{XY} = \mathrm{Cov}(X, Y)/(\sigma_X\sigma_Y)$, temos

$$\boxed{\mathrm{Cov}(X, Y) = \rho_{XY}\,\sigma_X\,\sigma_Y}$$

e a fórmula da variância assume a forma que usaremos em finanças:

$$\mathrm{V}(C) = a^2\sigma_X^2 + b^2\sigma_Y^2 + 2ab\,\rho_{XY}\,\sigma_X\,\sigma_Y.$$

Como $-1 \le \rho \le 1$, o termo cruzado é limitado, e isso terá consequência precisa adiante.

7.5 Aplicações em economia e finanças

A segunda metade do capítulo aplica o maquinário a decisões financeiras. O vocabulário muda, a matemática é a mesma.

7.5.1 Retorno e volatilidade

O objeto básico é o **retorno** de um ativo entre dois instantes. Se P_t é o preço no instante t,

$$\boxed{R_t = \frac{P_t - P_{t-1}}{P_{t-1}}}$$

é o retorno (simples) do período. Como P_t não é conhecido em $t-1$, o retorno é uma **variável aleatória** — e é por isso que toda a análise de investimentos é uma aplicação de probabilidade.

Sobre a distribuição de R_t interessam duas medidas, exatamente as duas de Seção 4:

- o **retorno esperado** $\mathbb{E}(R_t)$, que resume a rentabilidade;
- o **desvio padrão** $\mathrm{DP}(R_t)$, que em finanças recebe o nome de **volatilidade** e é *a* medida convencional de **risco**.

> **!** Volatilidade e prêmio de risco
>
> **Volatilidade** é o desvio padrão dos retornos de um ativo. Ela mede a dispersão dos resultados possíveis em torno do retorno esperado — quanto maior, mais imprevisível o resultado, e portanto mais arriscado o ativo.
> Ativos mais voláteis tendem a oferecer retorno esperado maior, e a diferença entre o retorno esperado de um ativo com risco e o de um ativo sem risco chama-se **prêmio de risco**. O adjetivo importa: o prêmio de risco é uma rentabilidade adicional **esperada**, não garantida. Quem aceita mais risco em troca de mais retorno esperado pode perfeitamente terminar com menos dinheiro — se o resultado fosse garantido, não haveria risco algum, e não haveria prêmio a pagar.

Essa distinção entre esperado e garantido é a fonte de metade dos mal-entendidos sobre investimento, e o exemplo das duas estratégias, adiante, a torna concreta.

7.5.2 Uma árvore binomial de preços

O exemplo a seguir é uma bela síntese do curso: usa a Binomial de Seção 5, as propriedades de transformação linear de Seção 4, e produz uma distribuição de preços completa.

Uma ação vale R$ 100 no dia 1. A cada dia, seu preço **sobe** R$ 1 com probabilidade 0,6 ou **desce** R$ 1 com probabilidade 0,4, e as variações diárias são independentes. Qual é a distribuição do preço no dia 5?

O primeiro passo é contar as variações. Do dia 1 ao dia 5 há **quatro** variações diárias (1→2, 2→3, 3→4, 4→5), e não cinco. Defina

$$X = \text{número de altas nas quatro variações } \sim \text{ Bin}(4;\ 0{,}6),$$

pois estão satisfeitas as quatro condições do experimento binomial: $n = 4$ fixo, dois resultados por ensaio, $p = 0{,}6$ constante e independência.

O segundo passo é escrever o preço final em função de X. Se houve x altas, houve $4 - x$ baixas, e o preço é

$$Y = 100 + 1 \cdot x - 1 \cdot (4 - x) = 100 + x - 4 + x = 2X + 96.$$

Y é uma transformação afim de X — exatamente o objeto de Seção 4, com $a = 2$ e $b = 96$.

A distribuição de Y segue da de X por $P(Y = 2x + 96) = P(X = x)$, com $P(X = x) = \binom{4}{x} 0{,}6^x 0{,}4^{4-x}$:

x	0	1	2	3	4
$y = 2x + 96$	96	98	100	102	104
$P(Y = y)$	0,0256	0,1536	0,3456	0,3456	0,1296

As cinco probabilidades somam 1, como devem. Note que o preço final não pode ser ímpar: com quatro variações de R$ 1, a paridade do preço se conserva.

O preço esperado, por dois caminhos. O interesse pedagógico do exemplo está em calcular $\mathbb{E}(Y)$ de duas maneiras e obter o mesmo número.

Via 1 — pela definição, somando sobre a distribuição de Y:

$$\begin{aligned}
\mathbb{E}(Y) &= \sum_y y\, P(Y = y) \\
&= 96(0{,}0256) + 98(0{,}1536) + 100(0{,}3456) + 102(0{,}3456) + 104(0{,}1296) \\
&= 2{,}4576 + 15{,}0528 + 34{,}56 + 35{,}2512 + 13{,}4784 \\
&= 100{,}8.
\end{aligned}$$

Via 2 — pelas propriedades, sem construir a distribuição de Y: como $Y = 2X + 96$ e $X \sim \text{Bin}(4; 0{,}6)$ tem $\mathbb{E}(X) = np = 4(0{,}6) = 2{,}4$,

$$\mathbb{E}(Y) = \mathbb{E}(2X + 96) = 2\,\mathbb{E}(X) + 96 = 2(2{,}4) + 96 = 4{,}8 + 96 = 100{,}8.$$

Mesmo resultado, três linhas em vez de cinco parcelas — e sem precisar da tabela. A lição metodológica é que **as propriedades de $\mathbb{E}$ e V existem para evitar reconstruir distribuições**. A variância sai do mesmo modo: $\mathrm{V}(X) = np(1-p) = 4(0{,}6)(0{,}4) = 0{,}96$, logo $\mathrm{V}(Y) = 2^2(0{,}96) = 3{,}84$ e $\mathrm{DP}(Y) = \sqrt{3{,}84} \approx 1{,}96$.

Economicamente: com probabilidade de alta 0,6 em cada dia, a ação tem preço esperado de R$ 100,80 ao fim de quatro variações — um retorno esperado de 0,8% — mas há 2,56% de chance de terminar em R$ 96, ou seja, com perda de 4%. Retorno esperado positivo convive com probabilidade de perda relevante, o que nos leva ao próximo critério.

7.5.3 Comparação de estratégias pela probabilidade de perda

Duas estratégias de investimento têm retornos, em percentual, aproximadamente Normais:

$$\text{Estratégia 1}: \; X_1 \sim N(5;\ 9) \quad (\mu_1 = 5\%,\ \sigma_1 = 3\%),$$
$$\text{Estratégia 2}: \; X_2 \sim N(6;\ 16) \quad (\mu_2 = 6\%,\ \sigma_2 = 4\%).$$

A estratégia 2 tem retorno esperado maior. Mas também é mais volátil. Um critério possível para decidir entre elas é a **probabilidade de prejuízo**, isto é, de retorno negativo.

Para a estratégia 1:

$$P(X_1 < 0) = P\left(Z < \frac{0-5}{3}\right) = P(Z < -1{,}67) = 0{,}0475.$$

Para a estratégia 2:

$$P(X_2 < 0) = P\left(Z < \frac{0-6}{4}\right) = P(Z < -1{,}50) = 0{,}0668.$$

Conclusão: pelo critério da probabilidade de perda, a estratégia 1 é a mais vantajosa — 4,75% contra 6,68% de chance de prejuízo — apesar de ter retorno esperado *menor*. A razão é que o acréscimo de 1 ponto percentual no retorno esperado veio acompanhado de um acréscimo de 1 ponto percentual na volatilidade, e o segundo pesou mais: o que importa para o cálculo é a distância de zero **medida em desvios padrão**, e ela caiu de 1,67 para 1,50.

! Discussão do critério

A probabilidade de perda é um critério **legítimo e incompleto**, e vale entender ambos os adjetivos.

É legítimo porque incorpora as duas dimensões — retorno e risco — em um único número interpretável, e porque corresponde a uma preocupação real: muitos investidores têm aversão específica a resultados negativos, não apenas à dispersão. Na literatura de gestão de risco, critérios desse tipo (probabilidade de perda, *Value at Risk*, *shortfall*) são padrão justamente por isso.

É incompleto por duas razões. Primeira, ele ignora a **magnitude** das perdas: uma estratégia com 4% de chance de perder tudo é preferida, por este critério, a outra com 5% de chance de perder um centavo. Segunda, ele fixa arbitrariamente o **limiar** em zero — nada impede comparar a probabilidade de ficar abaixo da inflação, ou abaixo da taxa livre de risco, e a ordenação das estratégias pode se inverter conforme o limiar escolhido. (Verifique: com limiar em 3%, qual estratégia ganha?)

A moral geral é a de Seção 4: nenhum número isolado resolve uma decisão sob incerteza. O que a estatística oferece é a distribuição inteira; escolher o critério que a resume é uma decisão econômica, informada por preferências.

7.5.4 Carteiras e diversificação

Chegamos à aplicação central. Uma carteira aloca a fração a do capital no ativo X e a fração $b = 1 - a$ no ativo Y. Seu retorno é a combinação linear

$$C = aX + bY, \qquad a + b = 1,\ a, b \geq 0,$$

e as duas fórmulas da seção anterior dão tudo o que se precisa:

$$\mathbb{E}(C) = a\,\mathbb{E}(X) + b\,\mathbb{E}(Y), \qquad \mathrm{V}(C) = a^2\sigma_X^2 + b^2\sigma_Y^2 + 2ab\,\rho\,\sigma_X\sigma_Y.$$

A primeira diz que **o retorno esperado da carteira é a média ponderada dos retornos esperados** — é a mesma média ponderada do investidor de Seção 2, agora com esperanças no lugar de rentabilidades determinísticas. Nada de surpreendente. A segunda é onde mora a novidade.

Exemplo numérico. Considere dois ativos:

Ativo	Peso	Retorno esperado	Volatilidade
X	$a = 0{,}6$	3%	$\sigma_X = 7\%$
Y	$b = 0{,}4$	6%	$\sigma_Y = 10\%$

com correlação $\rho = -0{,}5$ entre os retornos.

O retorno esperado da carteira é

$$\mathbb{E}(C) = 0{,}6(3\%) + 0{,}4(6\%) = 1{,}8\% + 2{,}4\% = 4{,}2\%,$$

entre os 3% de X e os 6% de Y, como tem de ser. A covariância é

$$\mathrm{Cov}(X, Y) = \rho\,\sigma_X\sigma_Y = (-0{,}5)(0{,}07)(0{,}10) = -0{,}0035,$$

e a variância da carteira,

$$\begin{aligned}\mathrm{V}(C) &= (0{,}6)^2(0{,}07)^2 + (0{,}4)^2(0{,}10)^2 + 2(0{,}6)(0{,}4)(-0{,}0035)\\ &= 0{,}36(0{,}0049) + 0{,}16(0{,}01) - 0{,}00168\\ &= 0{,}001764 + 0{,}0016 - 0{,}00168 = 0{,}001684,\end{aligned}$$

de modo que

$$\sigma_C = \sqrt{0{,}001684} \approx 0{,}041 = 4{,}1\%.$$

Pare e observe o número. A volatilidade da carteira é **4,1%** — menor que a volatilidade de X (7%) e menor que a de Y (10%). **A carteira é menos arriscada que qualquer um dos ativos que a compõem.**

! O hedge, ou por que a diversificação funciona

O resultado parece paradoxal — combinar dois ativos arriscados produzindo algo menos arriscado que ambos — mas é aritmética elementar do termo cruzado. Com $\rho < 0$, os dois ativos tendem a se mover em **direções opostas**: quando X decepciona, Y tende a compensar, e vice-versa. As oscilações se cancelam parcialmente, e o que sobra é uma série de retornos mais estável.
Esse cancelamento deliberado de riscos por meio de posições que se movem em sentidos opostos é o que se chama **hedge**, ou proteção. E note o que ele **não** custa: o retorno esperado da carteira continua sendo exatamente a média ponderada, 4,2% — o termo cruzado não aparece em $\mathbb{E}(C)$. Reduziu-se risco sem sacrificar retorno esperado, o que é a coisa mais próxima de um almoço grátis que a teoria financeira admite.

7.5.5 A diversificação funciona mesmo com $\rho > 0$

O ponto sutil, e o mais importante da seção, é que **não é preciso correlação negativa para diversificar**. Basta que a correlação seja **menor que 1**.

Para ver isso, tome o caso de dois ativos com pesos a e $b = 1 - a$ e compare a volatilidade da carteira com a média ponderada das volatilidades individuais, $a\sigma_X + b\sigma_Y$ — que é o risco que se teria se os riscos simplesmente somassem proporcionalmente. Partindo da fórmula da variância e usando $\rho \leq 1$:

$$\begin{aligned} \mathrm{V}(C) &= a^2\sigma_X^2 + b^2\sigma_Y^2 + 2ab\,\rho\,\sigma_X\sigma_Y \\ &\leq a^2\sigma_X^2 + b^2\sigma_Y^2 + 2ab\,\sigma_X\sigma_Y \qquad (\text{pois } 2ab\,\sigma_X\sigma_Y > 0 \text{ e } \rho \leq 1) \\ &= (a\sigma_X + b\sigma_Y)^2 , \end{aligned}$$

de onde

$$\boxed{\sigma_C \;\leq\; a\sigma_X + b\sigma_Y, \qquad \text{com igualdade se e somente se } \rho = 1}$$

A desigualdade é **estrita sempre que** $\rho < 1$ (com $a, b > 0$ e $\sigma_X, \sigma_Y > 0$), porque nesse caso $2ab\sigma_X\sigma_Y(\rho - 1) < 0$ estritamente. Ou seja:

- se $\rho = 1$, os ativos são substitutos perfeitos em termos de risco: a volatilidade da carteira é exatamente a média ponderada, e **não há ganho algum em diversificar**;
- se $\rho < 1$ — inclusive $\rho = 0{,}9$, inclusive $\rho = 0{,}99$ — a volatilidade da carteira fica **estritamente abaixo** da média ponderada. Há ganho de diversificação;
- quanto menor ρ, maior o ganho, e no caso extremo $\rho = -1$ é possível escolher pesos que zeram completamente a variância da carteira.

A Figura 12 mostra as cinco curvas correspondentes a $\rho = -1;\ -0{,}5;\ 0;\ 0{,}5;\ 1$, e nela se vê que só a última é uma reta — a média ponderada. Todas as demais são côncavas para baixo em relação a ela, e o afastamento entre a curva e a reta *é* o ganho de diversificação.

Isso tem uma consequência econômica prática de peso: como é raríssimo encontrar dois ativos com correlação exatamente 1, **quase toda diversificação reduz risco**. Não é preciso procurar ativos que se movam ao contrário do mercado (o que é difícil e caro); basta que não se movam em perfeita sincronia.

> **Ferramenta matemática — desigualdade e o quadrado do binômio**
>
> A demonstração acima usa apenas o quadrado do binômio $(u+v)^2 = u^2 + 2uv + v^2$ com $u = a\sigma_X$ e $v = b\sigma_Y$, mais a monotonicidade da raiz quadrada. O caso de igualdade $\rho = 1$ tem, aliás, leitura geométrica: correlação 1 significa que Y é uma função afim crescente de X, de modo que os dois ativos são, a menos de escala, o **mesmo** ativo — e diversificar entre um ativo e ele mesmo obviamente não faz nada. *(Desigualdades e produtos notáveis: capítulo* Nivelamento *das Notas de Matemática.)*

> **Nota — a carteira de variância mínima**
>
> Determinar os **pesos ótimos** de uma carteira é assunto de finanças, não desta disciplina. Mas o caso de dois ativos sai em poucas linhas e vale registrar. Com $b = 1 - a$,
>
> $$\mathrm{V}(C)(a) = a^2\sigma_X^2 + (1-a)^2\sigma_Y^2 + 2a(1-a)\mathrm{Cov}(X,Y),$$
>
> que é uma parábola em a com coeficiente líder $\sigma_X^2 + \sigma_Y^2 - 2\mathrm{Cov}(X,Y) = \mathrm{V}(X-Y) \geq 0$, portanto convexa. Derivando e igualando a zero:
>
> $$\frac{d\mathrm{V}(C)}{da} = 2a\sigma_X^2 - 2(1-a)\sigma_Y^2 + 2(1-2a)\mathrm{Cov}(X,Y) = 0$$
>
> $$\boxed{a^* = \frac{\sigma_Y^2 - \mathrm{Cov}(X,Y)}{\sigma_X^2 + \sigma_Y^2 - 2\mathrm{Cov}(X,Y)}}$$
>
> No exemplo numérico ($\sigma_X^2 = 0{,}0049$, $\sigma_Y^2 = 0{,}01$, Cov $= -0{,}0035$), isso dá $a^* = 0{,}0135/0{,}0219 \approx 0{,}616$ — muito próximo do peso 0,6 usado, o que explica por que a volatilidade obtida (4,10%) está quase no mínimo possível (4,096%). Note que o peso ótimo **não depende dos retornos esperados**: ele minimiza risco, ignorando retorno. A carteira que otimiza a relação entre os dois — a fronteira eficiente de Markowitz — é o passo seguinte, e é matéria de um curso de finanças.

7.6 Laboratório computacional em R

O laboratório deste capítulo é o mais importante das notas, porque o Teorema Central do Limite é um resultado sobre o que acontece quando n cresce — e simulação é a única forma de *ver* isso acontecer.

7.6.1 O TCL em ação

A estratégia é a seguinte. Escolhemos três populações deliberadamente **muito distantes da Normal**: uma exponencial (fortemente assimétrica à direita), uma Bernoulli com $p = 0{,}1$ (discreta,

com apenas dois valores e assimétrica) e uma uniforme (simétrica, mas de suporte limitado e densidade plana). Para cada uma, sorteamos 20.000 amostras de tamanho n, calculamos a média de cada amostra, e olhamos o histograma dessas 20.000 médias. Se o TCL funciona, os histogramas devem se aproximar de uma Normal conforme n cresce — apesar de nenhuma das três populações ser remotamente Normal.

```
set.seed(20266)
M <- 20000                                  # numero de amostras simuladas
ns <- c(1, 2, 5, 30, 100)                   # tamanhos amostrais

# tres populacoes nao-normais, com suas medias e desvios teoricos
pops <- list(
  Exponencial = list(r = function(k) rexp(k, rate = 1),          mu = 1,   sd = 1),
  "Bernoulli(0,1)" = list(r = function(k) rbinom(k, 1, 0.1),     mu = 0.1,
                          sd = sqrt(0.1 * 0.9)),
  "Uniforme(0,1)" = list(r = function(k) runif(k),               mu = 0.5,
                         sd = sqrt(1/12))
)

# para cada populacao e cada n, M medias amostrais, ja padronizadas
simular <- function(nome, n) {
  p <- pops[[nome]]
  medias <- colMeans(matrix(p$r(n * M), nrow = n, ncol = M))
  data.frame(pop = nome, n = n,
             z = (medias - p$mu) / (p$sd / sqrt(n)))
}
sim <- do.call(rbind, lapply(names(pops), function(nm)
  do.call(rbind, lapply(ns, function(n) simular(nm, n)))))
sim$n <- factor(sim$n, levels = ns, labels = paste0("n = ", ns))
sim$pop <- factor(sim$pop, levels = names(pops))
dim(sim)
```

```
[1] 300000      3
```

Padronizamos as médias por $(\bar{X} - \mu)/(\sigma/\sqrt{n})$ para que todos os painéis fiquem na mesma escala: se o TCL vale, **todos** devem convergir para a mesma $N(0,1)$, independentemente da população e do n.

```
ggplot(sim, aes(x = z)) +
  geom_histogram(aes(y = after_stat(density)), bins = 45,
                 fill = az, alpha = 0.5, colour = NA) +
  stat_function(fun = dnorm, colour = vm, linewidth = 0.6) +
  facet_grid(pop ~ n, scales = "free_y") +
  coord_cartesian(xlim = c(-4, 4)) +
  labs(x = "media amostral padronizada", y = "densidade") +
  theme(strip.text = element_text(size = 8),
        axis.text.y = element_blank())
```

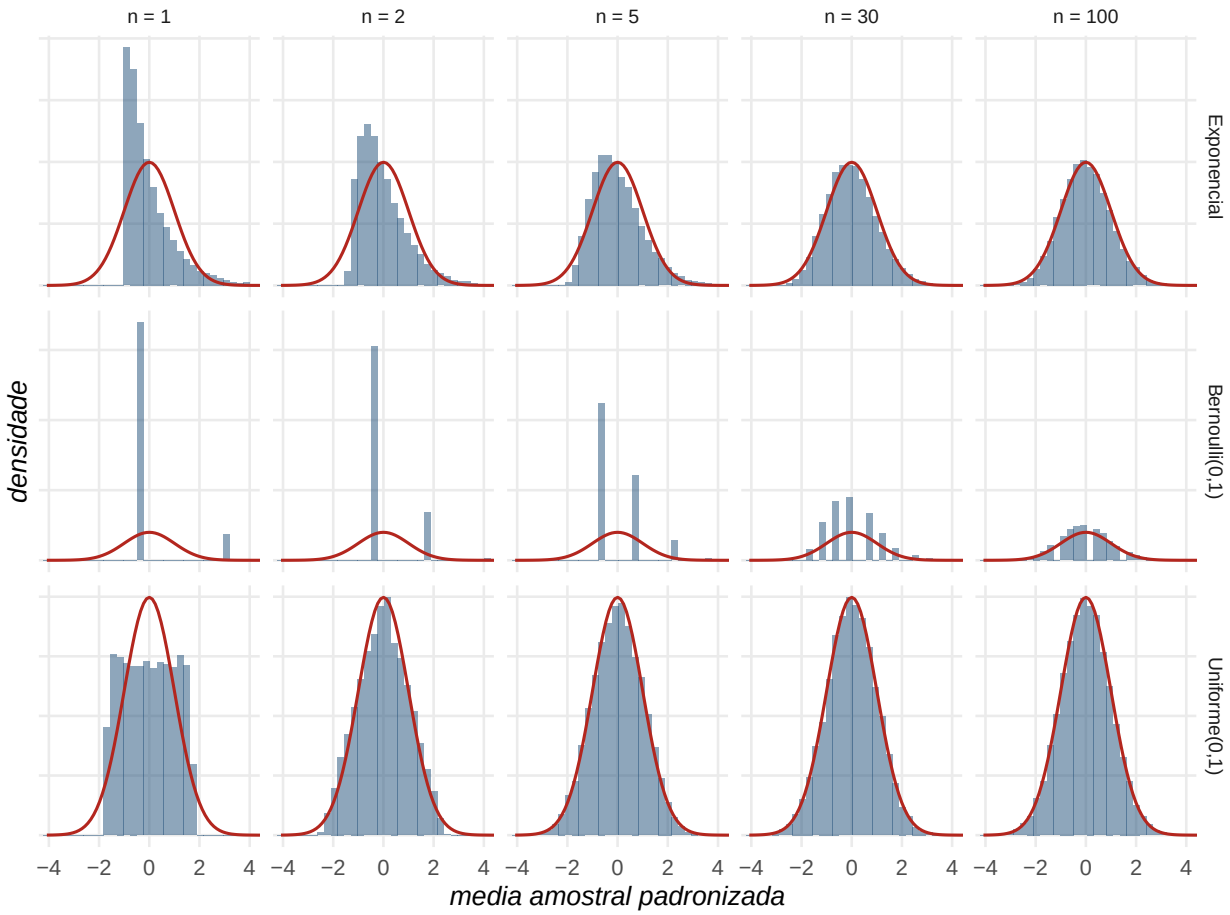


Figura 10: O Teorema Central do Limite em ação. Cada painel traz o histograma de 20.000 médias amostrais padronizadas, para três populações radicalmente não-normais (linhas) e cinco tamanhos amostrais (colunas). A curva vermelha é a densidade $N(0,1)$. Em $n = 1$ vê-se a população original — assimétrica, discreta ou plana; já em $n = 30$ os três histogramas são praticamente indistinguíveis da Normal.

A figura merece leitura demorada, coluna por coluna.

Em $n = 1$ não há média nenhuma: o histograma é a **própria população** padronizada. Vê-se a cauda longa da exponencial, os dois espetos da Bernoulli e o platô da uniforme — nada poderia estar mais longe da curva vermelha. Em $n = 2$ a exponencial já perdeu parte da cauda e a uniforme virou um triângulo. Em $n = 5$ a uniforme já é visualmente Normal e a exponencial ainda mostra assimetria residual. Em $n = 30$ — a fronteira convencional — os três estão essencialmente sobrepostos à Normal, com um resíduo de assimetria detectável apenas na exponencial. Em $n = 100$ o ajuste é perfeito para os três.

Duas conclusões, que são as do texto. Primeira: a convergência **acontece**, e acontece para populações escolhidas por serem o mais hostis possível. Segunda: a **velocidade** depende da população — a uniforme, simétrica, converge com $n = 5$; a exponencial, assimétrica, precisa de mais. Nenhum n é universalmente "suficiente", e a regra $n > 30$ é uma acomodação prática entre esses casos.

Vale medir a convergência, e não apenas olhá-la. O coeficiente de assimetria da média amostral deve tender a zero:

```
assimetria <- function(z) mean((z - mean(z))^3) / mean((z - mean(z))^2)^1.5
tab <- with(sim, tapply(z, list(pop, n), assimetria))
round(tab, 3)
```

```
               n = 1  n = 2  n = 5 n = 30 n = 100
Exponencial    1.975  1.381  0.891  0.372   0.211
Bernoulli(0,1) 2.670  1.868  1.172  0.518   0.262
Uniforme(0,1)  0.006 -0.001 -0.014 -0.009   0.035
```

A assimetria da exponencial cai de aproximadamente 2 (seu valor teórico) para perto de zero; a da Bernoulli(0,1), que parte de $\approx 2{,}67$, também. Nos dois casos a queda segue o padrão $1/\sqrt{n}$ previsto pela teoria: dividir a assimetria por dois exige quadruplicar n. A uniforme, simétrica desde o início, já parte de zero — é por isso que ela converge tão depressa.

7.6.2 Verificando $\mathbb{E}(\bar{X}) = \mu$, $\mathrm{V}(\bar{X}) = \sigma^2/n$ e o decaimento em $1/\sqrt{n}$

A simulação também permite verificar as duas fórmulas demonstradas no texto. Usemos a exponencial de taxa 1, que tem $\mu = 1$ e $\sigma = 1$:

```
set.seed(20267)
ns2 <- c(1, 2, 5, 10, 30, 100, 400, 1600)
res <- t(sapply(ns2, function(n) {
  medias <- colMeans(matrix(rexp(n * 20000, rate = 1), nrow = n))
  c(n = n,
    media_de_Xbar   = mean(medias),          # deve dar mu = 1
    var_de_Xbar     = var(medias),           # deve dar sigma^2/n = 1/n
    teorico_var     = 1 / n,
    dp_de_Xbar      = sd(medias),            # deve dar 1/sqrt(n)
    teorico_dp      = 1 / sqrt(n))
}))
round(res, 5)
```

```
        n media_de_Xbar var_de_Xbar teorico_var dp_de_Xbar teorico_dp
[1,]    1       1.00786     1.02539     1.00000    1.01262    1.00000
[2,]    2       1.00243     0.51367     0.50000    0.71671    0.70711
[3,]    5       0.99995     0.20146     0.20000    0.44884    0.44721
[4,]   10       0.99851     0.10070     0.10000    0.31733    0.31623
[5,]   30       1.00076     0.03370     0.03333    0.18359    0.18257
[6,]  100       0.99939     0.01014     0.01000    0.10067    0.10000
[7,]  400       1.00060     0.00249     0.00250    0.04991    0.05000
[8,] 1600       0.99978     0.00062     0.00062    0.02494    0.02500
```

As colunas confirmam as duas demonstrações. A média das médias fica sempre em torno de 1 — $\mathbb{E}(\bar{X}) = \mu$ **para todo** n, inclusive $n = 1$ —, enquanto a variância simulada acompanha $1/n$ e o desvio padrão acompanha $1/\sqrt{n}$. Note que o não viés vale desde a primeira linha, ao passo que a precisão só melhora com n: são propriedades independentes, e em Seção 8 elas receberão os nomes de **não viés** e **eficiência**.

O decaimento em $1/\sqrt{n}$ merece uma figura, porque é ele que governa o custo de obter precisão:

```
d_ep <- data.frame(n = res[, "n"], dp = res[, "dp_de_Xbar"])
curva <- data.frame(n = seq(1, 1600, length.out = 400))
curva$dp <- 1 / sqrt(curva$n)
g1 <- ggplot(d_ep, aes(n, dp)) +
  geom_line(data = curva, colour = vm, linewidth = 0.7) +
  geom_point(colour = az, size = 1.8) +
  labs(x = "n", y = "DP da media amostral")
g2 <- ggplot(d_ep, aes(n, dp)) +
  geom_line(data = curva, colour = vm, linewidth = 0.7) +
  geom_point(colour = az, size = 1.8) +
  scale_x_log10() + scale_y_log10() +
  labs(x = "n (escala log)", y = "DP (escala log)")
library(patchwork)
g1 + g2
```

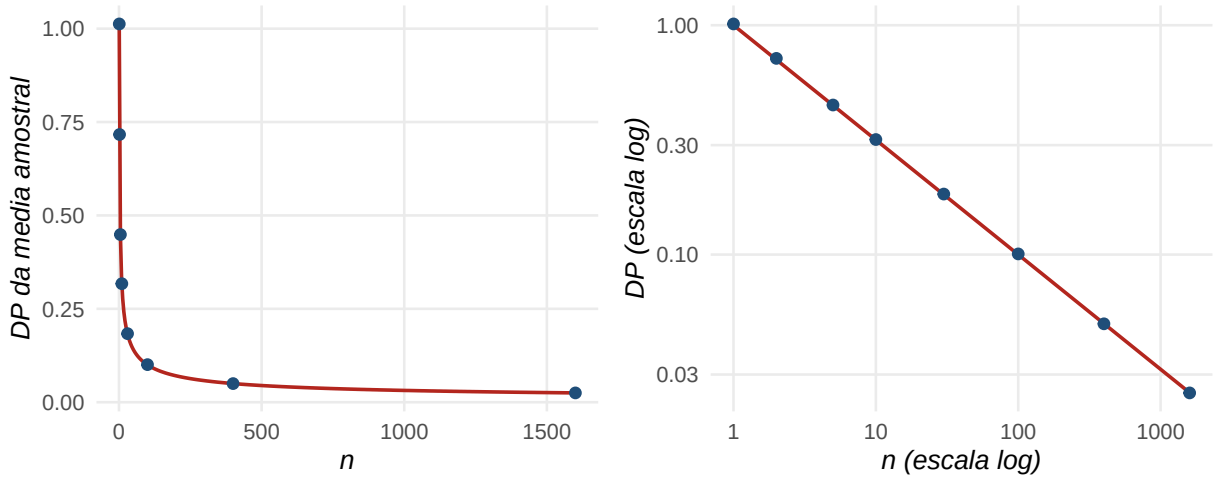


Figura 11: Decaimento do erro-padrão da média. Esquerda: desvio padrão simulado de $\bar{X}$ (pontos) contra a curva teórica $\sigma/\sqrt{n}$ (linha). Direita: os mesmos dados em escala log-log, onde a relação vira uma reta de inclinação $-1/2$. A precisão melhora com a raiz de n: quadruplicar a amostra reduz o erro-padrão à metade.

O painel da esquerda mostra a queda abrupta no início e o achatamento depois: os primeiros ganhos de precisão são baratos, os últimos são caríssimos. O painel da direita torna a lei exata: em log-log, $\log \mathrm{DP} = \log \sigma - \frac{1}{2} \log n$ é uma reta de inclinação $-1/2$, e os pontos simulados caem exatamente sobre ela.

7.6.3 Os exemplos numéricos do capítulo

Antes de prosseguir, convém confirmar as contas do texto. O R dispensa a tabela da Normal:

```
round(c(
  elevador_P_S_maior_500  = 1 - pnorm(500, mean = 7 * 70, sd = sqrt(7 * 100)),
  cafe_P_S_maior_49960    = 1 - pnorm(49960, mean = 50000, sd = sqrt(100 * 16)),
  cafe_P_Xbar_menor_500.7 = pnorm(500.7, mean = 500, sd = sqrt(16 / 100)),
  baterias_P_38_42        = pnorm(42, 40, 8 / sqrt(64)) - pnorm(38, 40, 8 / sqrt(64)),
  latas_P_S_maior_200     = 1 - pnorm(200, mean = 49 * 4, sd = sqrt(49 * 1)),
  estrategia1_P_perda     = pnorm(0, mean = 5, sd = 3),
  estrategia2_P_perda     = pnorm(0, mean = 6, sd = 4)
), 4)
```

```
 elevador_P_S_maior_500    cafe_P_S_maior_49960 cafe_P_Xbar_menor_500.7
                 0.3527                  0.8413                  0.9599
       baterias_P_38_42     latas_P_S_maior_200     estrategia1_P_perda
                 0.9545                  0.2839                  0.0478
    estrategia2_P_perda
                 0.0668
```

Todos batem com o texto: 0,3527 no elevador (contra 0,3520 pela tabela, arredondada a $z = 0{,}38$), 0,8413 e 0,9599 no café, 0,9545 nas baterias, 0,2839 nas latas, e 0,0478 contra 0,0668 na comparação de estratégias. As pequenas diferenças na terceira casa vêm do arredondamento de z ao consultar a tabela — o R usa o valor exato.

Vale ainda verificar o TCL diretamente no exemplo das latas, cuja distribuição individual nunca foi especificada. Suponhamos que o peso de cada lata seja, digamos, **uniforme** entre $4 - \sqrt{3}$ e $4 + \sqrt{3}$ (média 4, desvio padrão 1) — uma distribuição sem nenhuma semelhança com a Normal:

```
set.seed(20268)
somas <- colSums(matrix(runif(49 * 200000, 4 - sqrt(3), 4 + sqrt(3)), nrow = 49))
c(simulado = mean(somas > 200),
  aproximacao_normal = 1 - pnorm(200, 196, 7))
```

```
          simulado aproximacao_normal
         0.2844600          0.2838546
```

A aproximação Normal acerta a probabilidade com três casas decimais, embora as parcelas sejam uniformes. É o TCL trabalhando.

7.6.4 A fronteira de diversificação

Esta é a segunda figura essencial do capítulo. Fixamos as volatilidades dos dois ativos ($\sigma_X = 7\%$ e $\sigma_Y = 10\%$) e traçamos a volatilidade da carteira, $\sigma_C(a) = \sqrt{a^2\sigma_X^2 + (1-a)^2\sigma_Y^2 + 2a(1-a)\rho\sigma_X\sigma_Y}$, como função do peso a do ativo X, para cinco valores de ρ.

```
sX <- 0.07; sY <- 0.10
sigma_C <- function(a, rho) {
  sqrt(a^2 * sX^2 + (1 - a)^2 * sY^2 + 2 * a * (1 - a) * rho * sX * sY)
}
a_grid <- seq(0, 1, by = 0.005)
rhos   <- c(-1, -0.5, 0, 0.5, 1)
front <- do.call(rbind, lapply(rhos, function(r)
  data.frame(a = a_grid, rho = r, sigma = sigma_C(a_grid, r))))
front$rho_f <- factor(front$rho, levels = rhos,
                      labels = c("-1", "-0,5", "0", "0,5", "1"))

# peso de variancia minima, restrito a [0, 1] (sem venda a descoberto)
a_star <- function(rho) {
  cv <- rho * sX * sY
  pmin(pmax((sY^2 - cv) / (sX^2 + sY^2 - 2 * cv), 0), 1)
}
aa <- a_star(rhos)
round(data.frame(rho = rhos, a_otimo = aa,
                 sigma_min = sigma_C(aa, rhos),
                 media_ponderada = aa * sX + (1 - aa) * sY,
                 ganho = aa * sX + (1 - aa) * sY - sigma_C(aa, rhos)), 4)
```

```
   rho a_otimo sigma_min media_ponderada  ganho
1 -1.0  0.5882    0.0000          0.0824 0.0824
2 -0.5  0.6164    0.0410          0.0815 0.0405
3  0.0  0.6711    0.0573          0.0799 0.0225
4  0.5  0.8228    0.0682          0.0753 0.0071
5  1.0  1.0000    0.0700          0.0700 0.0000
```

A tabela já conta a história. Para cada ρ, a volatilidade mínima alcançável é **estritamente menor** que a média ponderada das volatilidades nos mesmos pesos — a coluna `ganho` mede exatamente essa diferença — exceto na última linha, $\rho = 1$, em que o ganho é nulo. E com $\rho = -1$ a volatilidade mínima é **exatamente zero**: correlação perfeitamente negativa permite eliminar o risco por completo.

Duas notas sobre a última linha. Primeira, com $\rho = 1$ a fórmula irrestrita devolveria $a^* = 3{,}33$, um peso fora de $[0, 1]$ — vender Y a descoberto para comprar mais X zeraria o risco, mas isso está fora do problema que colocamos, e por isso truncamos o peso em 1. Segunda, uma vez truncado, o mínimo é a carteira concentrada no ativo menos volátil, $\sigma_C = \sigma_X = 7\%$: quando os ativos são perfeitamente correlacionados, **a melhor estratégia de risco é não diversificar**, e sim concentrar no ativo de menor volatilidade.

```r
ponto <- data.frame(a = 0.6, sigma = sigma_C(0.6, -0.5))
ggplot(front, aes(a, sigma, colour = rho_f, linetype = rho_f)) +
  geom_line(linewidth = 0.75) +
  geom_point(data = ponto, aes(a, sigma), inherit.aes = FALSE,
             colour = vm, size = 2.4) +
  scale_colour_manual(values = c(vd, az, cz, lr, vm)) +
  scale_linetype_manual(values = c("solid", "solid", "solid", "solid", "22")) +
  scale_y_continuous(labels = function(x) paste0(round(100 * x, 1), "%")) +
  labs(x = "peso a do ativo X (o restante, 1 - a, vai para Y)",
       y = "volatilidade da carteira")
```

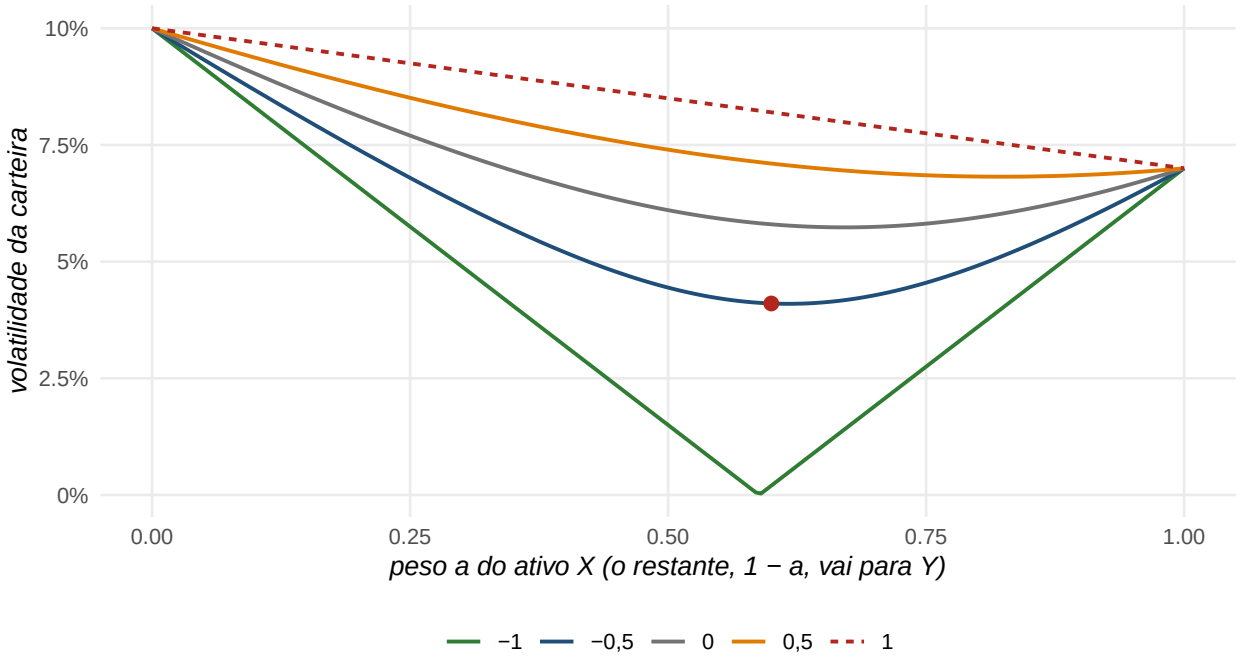


Figura 12: Fronteira de diversificação: volatilidade da carteira σ_C em função do peso a do ativo X ($\sigma_X = 7\%$), para cinco correlações. Os extremos $a = 0$ e $a = 1$ dão as volatilidades dos ativos isolados (10% e 7%) e são comuns a todas as curvas. Apenas $\rho = 1$ produz uma reta — a média ponderada, sem ganho de diversificação. Toda correlação menor que 1 gera uma curva abaixo dela, e o afastamento é o ganho. O ponto marca a carteira do exemplo do texto ($a = 0{,}6$, $\rho = -0{,}5$, $\sigma_C = 4{,}1\%$).

A figura é a demonstração gráfica do resultado da seção anterior. Todas as cinco curvas partem de $\sigma_Y = 10\%$ em $a = 0$ e chegam a $\sigma_X = 7\%$ em $a = 1$ — os extremos não dependem de ρ, porque uma carteira concentrada em um único ativo não tem termo cruzado. O que muda é o **caminho** entre os dois extremos:

- com $\rho = 1$ (tracejada vermelha) o caminho é a **reta** que liga 10% a 7%. Não há ganho: diversificar rende exatamente a média ponderada;
- com $\rho = 0{,}5$ (laranja) a curva já cai **abaixo** da reta — modestamente, mas cai. Este é o ponto sutil: correlação **positiva** e ainda assim há ganho;

- com $\rho = 0$ (cinza) e $\rho = -0{,}5$ (azul) o ganho cresce, e em ambos os casos existe uma faixa de pesos em que a carteira é menos volátil que **qualquer** dos ativos — o mínimo cai abaixo dos 7% de X;
- com $\rho = -1$ (verde) a curva toca o **eixo zero**: existe uma combinação que elimina totalmente o risco. É o caso-limite do hedge perfeito.

O ponto vermelho marca a carteira do texto. Está sobre a curva $\rho = -0{,}5$, em $a = 0{,}6$, a 4,1% — visivelmente abaixo dos 7% do ativo menos volátil, e praticamente no fundo da curva, como o cálculo de $a^* \approx 0{,}616$ havia antecipado.

7.6.5 A árvore binomial, pelas duas vias

Por fim, a verificação de que os dois caminhos para $\mathbb{E}(Y)$ dão o mesmo número:

```
x  <- 0:4
px <- dbinom(x, size = 4, prob = 0.6)      # X ~ Bin(4; 0,6)
y  <- 2 * x + 96                           # preco final
tab_arvore <- data.frame(x = x, y = y, P = round(px, 4))
tab_arvore
```

```
  x   y      P
1 0  96 0.0256
2 1  98 0.1536
3 2 100 0.3456
4 3 102 0.3456
5 4 104 0.1296
```

```
c(soma_das_probs = sum(px),
  via1_definicao = sum(y * px),            # E(Y) = soma de y P(y)
  via2_propriedade = 2 * (4 * 0.6) + 96,   # E(Y) = 2 E(X) + 96, com E(X) = np
  VY_definicao   = sum((y - sum(y * px))^2 * px),
  VY_propriedade = 2^2 * (4 * 0.6 * 0.4))  # V(Y) = a^2 V(X), com V(X) = np(1-p)
```

```
  soma_das_probs   via1_definicao via2_propriedade     VY_definicao
            1.00           100.80           100.80             3.84
  VY_propriedade
            3.84
```

As probabilidades reproduzem a tabela do texto — 0,0256; 0,1536; 0,3456; 0,3456; 0,1296 — e somam 1. As duas vias para $\mathbb{E}(Y)$ dão **100,8**, e as duas vias para $\mathrm{V}(Y)$ dão **3,84**. A coincidência não é numérica, é estrutural: a segunda via é a primeira, reorganizada pela linearidade da esperança.

Uma última verificação, por simulação, de que a construção da árvore está correta — sorteando um milhão de trajetórias de quatro dias:

```r
set.seed(20269)
variacoes <- matrix(sample(c(1, -1), 4 * 1e6, replace = TRUE, prob = c(0.6, 0.4)),
                    nrow = 4)
precos <- 100 + colSums(variacoes)
c(media_simulada = mean(precos), teorico = 100.8,
  dp_simulado = sd(precos), teorico_dp = sqrt(3.84),
  P_perda_simulada = mean(precos < 100), P_perda_teorica = sum(px[y < 100]))
```

```
  media_simulada          teorico      dp_simulado       teorico_dp
      100.802598       100.800000         1.958464         1.959592
P_perda_simulada  P_perda_teorica
        0.178865         0.179200
```

O preço médio simulado converge para R$ 100,80, o desvio padrão para $\sqrt{3{,}84} \approx 1{,}96$, e a probabilidade de terminar abaixo do preço inicial é de cerca de 17,9% (0,0256 + 0,1536). Retorno esperado positivo, com quase um em cada cinco cenários terminando no prejuízo: a distinção entre esperado e garantido, em números.

7.7 Exercícios

1. **Esperança e variância de somas.** Sejam X e Y variáveis aleatórias com $\mathbb{E}(X) = 10$, $\mathrm{V}(X) = 4$, $\mathbb{E}(Y) = 6$, $\mathrm{V}(Y) = 9$. (a) Calcule $\mathbb{E}(X + Y)$ — que hipótese foi necessária? (b) Calcule $\mathrm{V}(X + Y)$ supondo independência. (c) Refaça (b) com $\rho_{XY} = 0{,}5$ e com $\rho_{XY} = -0{,}8$. (d) Calcule $\mathrm{V}(X - Y)$ nos três casos e explique por que $\mathrm{V}(X - Y)$ pode ser maior que $\mathrm{V}(X + Y)$.

2. **O elevador, generalizado.** Com peso individual $N(70, 100)$: (a) refaça o cálculo para 6 e para 8 pessoas; (b) determine o número máximo de pessoas para que $P(S > 500) \leq 0{,}05$; (c) determine a capacidade que o elevador deveria ter para que, com 7 pessoas, essa probabilidade fosse 1%; (d) discuta por que a resposta de (b) não é simplesmente $\lfloor 500/70 \rfloor = 7$.

3. **Demonstrações de $\bar{X}$.** Sejam $X_1, \ldots, X_n$ com $\mathbb{E}(X_i) = \mu$ e $\mathrm{V}(X_i) = \sigma^2$.

   (a) Demonstre $\mathbb{E}(\bar{X}) = \mu$, indicando exatamente onde se usa a linearidade da esperança e por que nenhuma hipótese de independência é necessária. (b) Demonstre $\mathrm{V}(\bar{X}) = \sigma^2/n$, indicando em que passo a independência (ou descorrelação) é indispensável. (c) Suponha agora $\mathrm{Cov}(X_i, X_j) = c > 0$ para todo $i \neq j$; mostre que $\mathrm{V}(\bar{X}) = \sigma^2/n + c(n-1)/n$ e conclua que, sob correlação positiva constante, a variância de $\bar{X}$ **não** vai a zero. (d) Que implicação isso tem para amostras de séries temporais?

4. **Erro-padrão e tamanho de amostra.** Uma população tem $\sigma = 15$. (a) Calcule o erro-padrão de $\bar{X}$ para $n = 25$, 100, 400 e 1600. (b) Que n é necessário para que o erro-padrão seja 0,5? (c) Se o custo de coletar cada observação é constante, qual o custo relativo de reduzir o erro-padrão de 1,5 para 0,75? (d) Explique, com base em (c), por que institutos de pesquisa raramente ultrapassam 2.000 entrevistados.

5. **O TCL e seus limites.** (a) Enuncie o TCL com precisão, listando todas as hipóteses.

(b) Explique por que a regra $n > 30$ não faz parte do enunciado. (c) Para uma Bernoulli(0,02), calcule $\mathbb{E}(\bar{X})$ e $\mathrm{DP}(\bar{X})$ com $n = 30$ e argumente por que a aproximação Normal é ruim nesse caso — quantos sucessos se espera observar? (d) Que condição adicional a literatura impõe para proporções, e por quê? (e) Dê um exemplo de distribuição à qual o TCL, na forma enunciada, não se aplica.

6. **A árvore binomial, estendida.** Retome a ação a R$ 100 com alta de R$ 1 (prob. 0,6) ou baixa de R$ 1 (prob. 0,4) por dia. (a) Refaça a análise para o **dia 6** (cinco variações): escreva Y em função de $X \sim \mathrm{Bin}(5; 0{,}6)$, obtenha a distribuição de Y e calcule $\mathbb{E}(Y)$ pelas duas vias. (b) Calcule $\mathrm{V}(Y)$ e $\mathrm{DP}(Y)$. (c) Obtenha a probabilidade de o preço no dia 6 ser inferior ao preço inicial. (d) Mostre que, para k variações, $\mathbb{E}(Y) = 100 + k(2p - 1)$ e interprete a condição $p > 1/2$. (e) Use o TCL para aproximar a distribuição de Y após 100 variações e calcule $P(Y > 110)$.

7. **Comparação de estratégias.** Com $X_1 \sim N(5; 9)$ e $X_2 \sim N(6; 16)$ como no texto: (a) refaça a comparação usando como limiar a inflação de 3% em vez de zero — a ordenação se inverte? (b) Determine o limiar c para o qual as duas estratégias têm a mesma probabilidade de ficar abaixo dele. (c) Compare as duas estratégias pelo coeficiente de variação e discuta em que sentido esse critério difere do da probabilidade de perda.

   (d) Proponha uma terceira estratégia que domine ambas e explique o que "dominar" deve significar aqui.

8. **Carteiras e diversificação.** Dois ativos têm $\mathbb{E}(X) = 4\%$, $\sigma_X = 6\%$, $\mathbb{E}(Y) = 9\%$, $\sigma_Y = 12\%$. (a) Para $a = 0{,}5$, calcule $\mathbb{E}(C)$ e σ_C com $\rho = -1;\ -0{,}5;\ 0;\ 0{,}5;\ 1$, e verifique que $\mathbb{E}(C)$ não depende de ρ.

   (b) Demonstre que $\sigma_C \leq a\sigma_X + (1 - a)\sigma_Y$ com igualdade se e somente se $\rho = 1$. (c) Obtenha o peso de variância mínima a^* para cada ρ do item (a) e verifique que, com $\rho = -1$, ele zera a volatilidade. (d) Com $\rho = 0{,}9$, mostre numericamente que **ainda há** ganho de diversificação e quantifique-o. (e) Explique por que a^* não depende dos retornos esperados, e o que faltaria acrescentar ao problema para que dependesse.

8 Estimação Pontual

Até aqui, os parâmetros eram dados. Quando calculamos $P(X > 1)$ para o consumo de energia em Seção 4, a densidade $f(x) = \frac{3}{8}x^2$ estava posta. Quando computamos a probabilidade de três sinistros em um mês, o λ da Poisson estava posto (Seção 5). Quando padronizamos uma Normal para consultar a tabela, μ e σ estavam postos (Seção 6). Todo o edifício probabilístico dos capítulos anteriores foi construído sobre a hipótese de que **conhecemos o modelo e seus parâmetros**, e o que se pedia era deduzir consequências: dado o modelo, qual a probabilidade do dado?

Este capítulo inverte a seta. Na prática — e é sempre assim na prática — ninguém nos entrega o μ, o λ ou o p. Temos dados, uma amostra, e queremos recuperar deles o parâmetro desconhecido. Saímos da **dedução** probabilística e entramos na **indução** estatística; da pergunta "dado o modelo, o que esperar dos dados?" para "dados os dados, o que dizer do modelo?". É o salto lógico que a introdução anunciava, e ele abre a terça parte final destas notas: estimação pontual aqui, estimação intervalar em Seção 9, testes de hipóteses em Seção 10.

O programa do capítulo é curto e sua ordem importa. Primeiro fixamos o vocabulário — população, parâmetro, amostra, amostra aleatória simples — porque é nele que se esconde a maior parte das confusões. Depois separamos com cuidado o **estimador** da **estimativa**, distinção que parece pedante e é, na verdade, o coração do assunto: um é variável aleatória, o outro é número. Em seguida perguntamos o que faz de um estimador um bom estimador, e respondemos com a propriedade do **não-viés** — que nos obrigará, finalmente, a explicar o famoso $n - 1$. Por fim, os dois **métodos** de construção de estimadores que o curso cobre: a máxima verossimilhança e o método dos momentos.

8.1 População, parâmetro, amostra

> **!** Definição — população e parâmetro
>
> A **população** é o conjunto de todos os elementos ou resultados sob investigação — todo o universo de interesse. Um **parâmetro** é uma medida numérica que descreve uma característica dessa população.
> A **amostra** é um subconjunto da população, e uma medida numérica calculada a partir dela chama-se **estatística**.

A oposição a reter é essa: parâmetro é da população, estatística é da amostra. O parâmetro é um número **fixo e desconhecido** — a altura média de todos os brasileiros adultos existe, tem um valor determinado, e nós não o conhecemos. A estatística é um número **que calculamos** e, por isso mesmo, varia de amostra para amostra.

Por que trabalhar com amostras, afinal? Porque o censo — observar a população inteira — é quase sempre caro, demorado ou impossível. Impossível literalmente, em três situações frequentes: quando a população é infinita ou indefinidamente grande (todas as peças que uma máquina *virá a* produzir); quando o teste é destrutivo (para saber a vida útil média de uma lâmpada é preciso queimá-la, e ninguém queima o estoque inteiro); e quando a população é hipotética, definida por um processo gerador e não por uma lista de indivíduos — que é exatamente o caso em economia, onde a "população" por trás de uma série de retornos é o mecanismo que os gera.

A notação registra a distinção. Reservamos letras gregas para parâmetros populacionais e letras latinas para as estatísticas amostrais correspondentes:

Característica	Parâmetro (população)	Estatística (amostra)
Média	μ	$\bar{X}$
Variância	σ^2	S^2
Desvio padrão	σ	S
Proporção	p	$\hat{p}$

8.1.1 Amostra aleatória simples

Nem toda amostra serve. Uma amostra de conveniência — entrevistar quem passa em frente ao prédio da EPGE — pode ser sistematicamente enviesada de maneiras que nenhuma quantidade de matemática corrige depois. O requisito mínimo para que a inferência funcione é que a amostra seja **aleatória**.

! Definição — amostra aleatória simples (AAS)

Uma **amostra aleatória simples** de tamanho n é um conjunto de variáveis aleatórias $X_1, X_2, \dots, X_n$ tais que:

(i) as X_i são **independentes**;

(ii) todas têm a **mesma distribuição**, a da população.

Escreve-se, abreviadamente, $X_1, \dots, X_n \overset{\text{i.i.d.}}{\sim} F$ — independentes e identicamente distribuídas segundo a distribuição F da população.

Operacionalmente, uma AAS é obtida quando cada elemento da população tem a mesma probabilidade de ser sorteado e os sorteios não interferem uns nos outros — amostragem com reposição, ou de uma população suficientemente grande para que a reposição não faça diferença. A hipótese i.i.d. é forte e é frequentemente violada em economia (séries temporais são o contraexemplo óbvio: o PIB deste trimestre não é independente do anterior), mas é a hipótese sob a qual toda a teoria deste e dos dois próximos capítulos é construída.

Uma observação de notação, que vale a pena assentar agora porque será usada o tempo todo. A amostra, **antes de ser coletada**, é o vetor de variáveis aleatórias $(X_1, \dots, X_n)$ — maiúsculas. **Depois de coletada**, é o vetor de números $(x_1, \dots, x_n)$ — minúsculas. É a mesma convenção de Seção 4, e é dela que sai a distinção da próxima seção.

8.2 Estimador e estimativa

Suponha que o parâmetro de interesse seja a média populacional μ. A proposta natural — e correta, como veremos — é usar a média da amostra:

$$\hat{\mu} = \bar{X} = \frac{\sum_{i=1}^{n} X_i}{n}$$

O **chapéu** sobre o parâmetro é notação universal e significa uma coisa só: *estamos estimando*. Onde aparece $\hat{\theta}$, leia "nossa tentativa de recuperar θ a partir dos dados". O parâmetro θ é fixo e desconhecido; $\hat{\theta}$ é o que construímos para chegar perto dele.

Chegamos assim à distinção que dá nome à seção, e que é a mais importante do capítulo.

> **!** Estimador $\neq$ estimativa
>
> O **estimador** $\hat{\theta}$ é uma **função da amostra** $(X_1, \ldots, X_n)$ — e, como as X_i são variáveis aleatórias, o estimador **é ele próprio uma variável aleatória**. Tem distribuição, tem esperança, tem variância. Escreve-se com letras maiúsculas.
> A **estimativa** é o **valor numérico** que o estimador assume para uma amostra particular já observada, $(x_1, \ldots, x_n)$. É um número. Escreve-se com letras minúsculas.
> Em símbolos: o estimador é $\bar{X} = \sum X_i/n$; a estimativa é $\bar{x} = \sum x_i/n$.

A distinção não é preciosismo notacional — é o que torna a inferência possível. Se $\hat{\theta}$ é uma variável aleatória, então tem uma distribuição de probabilidade, e é sobre essa distribuição que faremos afirmações probabilísticas: "com 95% de confiança, o parâmetro está neste intervalo" (Seção 9), "rejeitamos a hipótese ao nível de 5%" (Seção 10). Nada disso faria sentido para um número fixo. É porque o estimador varia de amostra para amostra, e porque sabemos *como* ele varia, que podemos quantificar nossa incerteza sobre o parâmetro.

Exemplo (alturas). Sorteamos cinco estudantes e medimos suas alturas, em centímetros:

$$174 \qquad 186 \qquad 186 \qquad 180 \qquad 174$$

O estimador de μ é $\bar{X} = \sum_{i=1}^{5} X_i/5$ — uma variável aleatória, definida antes de vermos qualquer número, que descreve o que faríamos com *qualquer* amostra de cinco alturas. A estimativa é o número que resulta destes cinco valores concretos:

$$\bar{x} = \frac{174 + 186 + 186 + 180 + 174}{5} = \frac{900}{5} = 180.$$

Nossa estimativa da altura média da população é 180 cm. Se sorteássemos outros cinco estudantes, obteríamos outro número — 178, talvez 183. O estimador é o mesmo; a estimativa muda. E a pergunta que organiza o resto do capítulo é: **o que garante que esses números flutuem em torno do valor certo?**

8.3 Não-viés

A primeira e mais básica exigência que se faz a um estimador é que ele acerte **em média**. Não se pode pedir que $\hat{\theta}$ coincida com θ em cada amostra — isso seria impossível, porque $\hat{\theta}$ é aleatório e θ é fixo. O que se pode pedir é que não erre sistematicamente para cima nem para baixo.

! Definição — estimador não viesado e viés

Um estimador $\hat{\theta}$ é **não viesado** (ou *não tendencioso*, ou *centrado*) para o parâmetro θ quando

$$\boxed{\mathbb{E}(\hat{\theta}) = \theta}$$

De modo geral, o **viés** (ou *vício*, ou *tendenciosidade*) de $\hat{\theta}$ é

$$\boxed{B(\hat{\theta}) = \mathbb{E}(\hat{\theta}) - \theta}$$

de modo que o estimador é não viesado se e somente se $B(\hat{\theta}) = 0$.

A leitura é a do ponto de equilíbrio de Seção 4: a distribuição do estimador — a chamada **distribuição amostral** — se equilibra exatamente sobre o valor verdadeiro do parâmetro. Amostras individuais erram, umas para mais, outras para menos, mas os erros se cancelam na média. Um estimador viesado, ao contrário, erra na mesma direção sistematicamente, e nenhuma quantidade de dados corrige isso se o viés não desaparecer com n.

8.3.1 A média amostral é não viesada

O caso de $\bar{X}$ já está resolvido, e resolvido em Seção 7. Como $X_1, \ldots, X_n$ são i.i.d. com $\mathbb{E}(X_i) = \mu$, a linearidade da esperança dá

$$\mathbb{E}(\bar{X}) = \mathbb{E}\left(\frac{1}{n}\sum_{i=1}^{n} X_i\right) = \frac{1}{n}\sum_{i=1}^{n}\mathbb{E}(X_i) = \frac{1}{n}\cdot n\mu = \mu.$$

$$\boxed{\mathbb{E}(\bar{X}) = \mu \quad \Longrightarrow \quad \bar{X} \text{ é não viesado para } \mu}$$

Note que a demonstração **não usa independência** — apenas a linearidade de $\mathbb{E}$, que vale sempre. A independência entra na variância, $\mathrm{V}(\bar{X}) = \sigma^2/n$, que é o resultado que governa a precisão do estimador e que sustenta os intervalos de confiança do próximo capítulo. Mas para o não-viés basta que cada X_i tenha média μ.

8.3.2 Por que $n-1$: a variância amostral

Passemos à variância. O análogo natural de $\sigma^2 = \mathbb{E}[(X-\mu)^2]$ seria substituir a esperança pela média amostral e μ por $\bar{X}$, obtendo $\sum(X_i - \bar{X})^2/n$ — que é exatamente a variância descritiva de Seção 2. Ocorre que esse estimador **é viesado**. O estimador não viesado de σ^2 é

$$\boxed{S^2 = \frac{\sum_{i=1}^{n}(X_i - \bar{X})^2}{n-1}}$$

com divisor $n-1$, e a estatística $S = \sqrt{S^2}$ é o **desvio padrão amostral**. Este é o divisor que o R usa em `var()` e `sd()`, e é a convenção padrão de toda a inferência.

A pergunta que todo aluno faz — e que os slides deixam sem resposta — é *por quê*. A demonstração é clássica e cabe em poucas linhas; vale fazê-la, porque ela ilumina o mecanismo inteiro.

> **Ferramenta matemática — a decomposição do desvio**
>
> O truque é somar e subtrair μ dentro do desvio, escrevendo $X_i - \bar{X} = (X_i - \mu) - (\bar{X} - \mu)$, e expandir o quadrado. O termo cruzado colapsa porque $\sum(X_i - \mu) = n(\bar{X} - \mu)$ — a mesma manipulação de somatório usada em Seção 2 para mostrar que os desvios somam zero. Vale a pena reter o padrão: *decompor um desvio em torno de uma estatística num desvio em torno do parâmetro mais um desvio da estatística ao parâmetro* é o movimento algébrico mais recorrente da estatística matemática.

Comecemos pela soma de quadrados $SQ = \sum_{i=1}^{n}(X_i - \bar{X})^2$. Escrevendo $X_i - \bar{X} = (X_i - \mu) - (\bar{X} - \mu)$ e expandindo:

$$\begin{aligned}\sum_{i=1}^{n}(X_i - \bar{X})^2 &= \sum_{i=1}^{n}\Big[(X_i - \mu) - (\bar{X} - \mu)\Big]^2 \\ &= \sum_{i=1}^{n}(X_i - \mu)^2 - 2(\bar{X} - \mu)\sum_{i=1}^{n}(X_i - \mu) + n(\bar{X} - \mu)^2.\end{aligned}$$

Mas $\sum_{i=1}^{n}(X_i - \mu) = n\bar{X} - n\mu = n(\bar{X} - \mu)$, de modo que o termo do meio vale $-2n(\bar{X} - \mu)^2$ e cancela parcialmente o último:

$$\boxed{\sum_{i=1}^{n}(X_i - \bar{X})^2 = \sum_{i=1}^{n}(X_i - \mu)^2 - n(\bar{X} - \mu)^2}$$

Esta identidade já é instrutiva por si só: a soma de quadrados em torno de $\bar{X}$ é **sempre menor** que a soma em torno de μ — a média amostral é o ponto que minimiza a soma de quadrados dos desvios. E é aí que nasce o viés: ao centrar em $\bar{X}$ em vez de μ, subestimamos a dispersão.

Tomando esperanças e usando $\mathbb{E}[(X_i - \mu)^2] = \sigma^2$ e $\mathbb{E}[(\bar{X} - \mu)^2] = \mathrm{V}(\bar{X}) = \sigma^2/n$:

$$\mathbb{E}\left[\sum_{i=1}^{n}(X_i - \bar{X})^2\right] = n\sigma^2 - n \cdot \frac{\sigma^2}{n} = n\sigma^2 - \sigma^2 = (n-1)\sigma^2.$$

$$\boxed{\mathbb{E}\left[\sum_{i=1}^{n}(X_i - \bar{X})^2\right] = (n-1)\sigma^2}$$

O resultado responde tudo de uma vez. Dividindo por $n - 1$:

$$\mathbb{E}(S^2) = \frac{1}{n-1}\,\mathbb{E}\Big[\sum(X_i - \bar{X})^2\Big] = \frac{(n-1)\sigma^2}{n-1} = \sigma^2,$$

e S^2 é não viesado. Já o estimador com divisor n, que chamaremos $\hat{\sigma}^2$, tem

$$\mathbb{E}(\hat{\sigma}^2) = \mathbb{E}\left[\frac{\sum(X_i - \bar{X})^2}{n}\right] = \frac{n-1}{n}\,\sigma^2 \qquad \Longrightarrow \qquad B(\hat{\sigma}^2) = \frac{n-1}{n}\sigma^2 - \sigma^2 = -\frac{\sigma^2}{n}.$$

O viés é **negativo**: o estimador com divisor n subestima a variância, sistematicamente, em σ^2/n. A intuição que a identidade fornece é precisa — usamos os próprios dados para estimar o centro, e o centro estimado fica, por construção, mais próximo dos dados do que o centro verdadeiro. A correção $n-1$ compensa exatamente esse encolhimento. Dizemos que "perdemos um grau de liberdade" ao estimar μ: dos n desvios $(X_i - \bar{X})$, apenas $n-1$ são livres, já que sua soma é forçosamente nula e o n-ésimo fica determinado pelos demais.

Note também que $B(\hat{\sigma}^2) \to 0$ quando $n \to \infty$. O viés existe, mas some com o tamanho da amostra; com $n = 5$ ele custa 20% da variância, com $n = 1000$ custa 0,1%. É por isso que a distinção entre os dois divisores é dramática em amostras pequenas e irrelevante em amostras grandes — e a Figura 13 mostra isso acontecendo.

> **!** Resolvendo a aparente contradição com Seção 2
>
> Em Seção 2 definimos a variância com divisor n, e agora dizemos que o divisor correto é $n-1$. Não há contradição: **são objetos distintos, com finalidades distintas**.
> Lá, os trinta faturamentos eram tratados como *a população de interesse* — descrevíamos aquele conjunto, não inferíamos sobre nada maior. A variância com divisor n é a **definição do parâmetro** σ^2, e definições não têm viés: não estamos estimando coisa alguma.
> Aqui, os dados são uma *amostra* de uma população maior, e queremos recuperar o σ^2 dessa população. Aí sim a pergunta "o estimador acerta em média?" faz sentido, e a resposta exige $n-1$.
> A regra prática: divisor n quando os dados **são** a população (ou quando se conhece μ e se calculam desvios em torno dele); divisor $n-1$ quando os dados são uma **amostra** e μ foi estimado por $\bar{X}$. Do capítulo de inferência em diante, é sempre o segundo caso.

Uma advertência final, e ela será relevante daqui a duas seções: os estimadores de σ^2 produzidos tanto pela **máxima verossimilhança** quanto pelo **método dos momentos** usam divisor n e são, portanto, **viesados**. Os dois métodos que veremos a seguir são procedimentos mecânicos de construção de estimadores, e nada nos seus mecanismos garante não-viés. O $n-1$ é uma correção aplicada *depois*, à mão, e não uma saída natural de nenhum dos dois métodos.

8.4 Como construir estimadores

Até agora propusemos estimadores por analogia: para estimar a média populacional, use a média amostral. A analogia funciona bem para μ, mas não oferece orientação quando o parâmetro não tem contraparte amostral óbvia — qual é a "média amostral" do θ de uma densidade $f(x) = \theta x^{\theta-1}$? Precisamos de **métodos**: procedimentos sistemáticos que, dado um modelo e uma amostra, entreguem um estimador.

O curso apresenta dois:

1. **máxima verossimilhança** (MV), que escolhe o valor do parâmetro que torna a amostra observada a mais provável possível;
2. **método dos momentos**, que iguala momentos populacionais a momentos amostrais.

Há um terceiro, os **mínimos quadrados**, que escolhe o parâmetro minimizando a soma dos quadrados dos erros. Ele não é objeto deste curso — é o método central da **econometria**, onde reaparecerá como a base da estimação de modelos de regressão. Vale registrar a conexão: sob normalidade dos erros, o estimador de mínimos quadrados de um modelo de regressão **coincide** com o de máxima verossimilhança, o que dá à econometria clássica uma dupla justificativa.

8.5 Máxima verossimilhança

A ideia é de uma elegância notável, e se resume a uma pergunta invertida. A probabilidade responde: *dado o parâmetro, qual a chance de observar estes dados?* A verossimilhança fixa os dados — que já foram observados, afinal — e faz a pergunta ao contrário: *qual valor do parâmetro tornaria estes dados os mais prováveis de terem ocorrido?* Esse valor é o **estimador de máxima verossimilhança** (EMV).

! Definição — função de verossimilhança

Seja $X_1, \dots, X_n$ uma AAS de uma população com parâmetro θ desconhecido. A **função de verossimilhança** é a probabilidade conjunta da amostra, vista como função de θ:

$$\boxed{L(\theta) = \prod_{i=1}^{n} P(X_i = x_i) \quad \text{(discreto)}, \qquad L(\theta) = \prod_{i=1}^{n} f(x_i) \quad \text{(contínuo)}.}$$

O produtório vale porque a amostra é **independente**, e cada fator tem a mesma forma funcional porque é **identicamente distribuída** — as duas metades da hipótese de AAS.
A **log-verossimilhança** é

$$\boxed{\ell(\theta) = \ln L(\theta) = \sum_{i=1}^{n} \ln f(x_i)}$$

e o **estimador de máxima verossimilhança** de θ é o valor $\hat{\theta}$ que maximiza $L(\theta)$ — equivalentemente, $\ell(\theta)$.

Duas observações antes dos exemplos.

A primeira é sobre a mudança de ponto de vista. Em $L(\theta)$, os x_i são **constantes** — números já observados — e θ é a **variável**. É o oposto exato do que fizemos nos capítulos anteriores, onde θ era fixo e x variava. A função de verossimilhança não é uma distribuição de probabilidade em θ: ela não integra a 1 e θ não é aleatório. É apenas uma medida de quão bem cada valor candidato de θ "explica" os dados.

A segunda é sobre o logaritmo, e merece uma caixa.

i Ferramenta matemática — otimização, log e as condições de primeira e segunda ordem

Maximizar uma função derivável $g(\theta)$ em um intervalo aberto é resolver a **condição de primeira ordem** (CPO) $g'(\theta) = 0$ e verificar a **condição de segunda ordem** (CSO) $g''(\theta) < 0$, que distingue máximo de mínimo ou ponto de inflexão.
Aplicamos isso a ℓ, e não a L, por dois motivos. O primeiro é **algébrico**: o logaritmo transforma

produto em soma, $\ln \prod a_i = \sum \ln a_i$, e derivar uma soma de n termos é trivial enquanto derivar um produto de n fatores exige a regra do produto iterada. O segundo é **numérico**: o produto de centenas de probabilidades, cada uma menor que 1, subborda para zero em qualquer computador, ao passo que a soma dos logaritmos é perfeitamente representável.
E a troca é **legítima** porque ln é **estritamente crescente** — uma transformação monótona. Se $\ln u < \ln v$ sempre que $u < v$, então o ponto que maximiza L é exatamente o mesmo que maximiza $\ln L$: os valores máximos diferem, mas o **argmax** é idêntico. A Figura 14 mostra as duas curvas lado a lado, com o pico na mesma abscissa.
(Derivadas, regras de derivação, condições de primeira e segunda ordem e otimização de funções de uma variável: capítulos de Cálculo Diferencial *e* Otimização *das Notas de Matemática, Profa. Silvia Matos.)*

O procedimento se organiza em uma receita fixa, que o professor enuncia em cinco passos e que vale seguir literalmente.

! Os 5 passos da máxima verossimilhança

1. Escrever a função de verossimilhança
2. Escrever a função de log-verossimilhança
3. Derivar a função de log-verossimilhança
4. Igualar a derivada do passo 3 a zero, e resolver para o parâmetro de interesse
5. Aplicar a função encontrada em (4) a $\{X_1, \dots, X_n\}$, obtendo o EMV

O passo 5 é o que fecha o círculo conceitual do capítulo, e é fácil passar por ele sem perceber sua importância. Os passos 1 a 4 operam sobre a amostra observada $(x_1, \dots, x_n)$ e produzem uma **estimativa** — um número, ou melhor, uma *fórmula* aplicada a números. O passo 5 substitui as minúsculas por maiúsculas e converte essa fórmula em uma função da amostra aleatória: o **estimador**. É essa passagem que nos autoriza a perguntar, depois, se $\mathbb{E}(\hat{\theta}) = \theta$.

8.5.1 Exemplo 1: Poisson

Seja $X_1, \dots, X_n \overset{\text{i.i.d.}}{\sim} \text{Poi}(\lambda)$, com $P(X = x) = \dfrac{e^{-\lambda}\lambda^x}{x!}$, $x = 0, 1, 2, \dots$ (Seção 5).

Passo 1 — verossimilhança.

$$L(\lambda) = \prod_{i=1}^{n} \frac{e^{-\lambda}\lambda^{x_i}}{x_i!} = \frac{e^{-n\lambda}\,\lambda^{\sum x_i}}{\prod_{i=1}^{n} x_i!}.$$

Passo 2 — log-verossimilhança. Aplicando ln e usando $\ln(ab) = \ln a + \ln b$:

$$\ell(\lambda) = -n\lambda + \left(\sum_{i=1}^{n} x_i\right) \ln \lambda - \sum_{i=1}^{n} \ln(x_i!).$$

Repare que o último termo **não depende de** λ — é uma constante aditiva, e vai morrer na derivada. Isso é típico: boa parte da verossimilhança costuma ser irrelevante para a maximização, e identificá-la cedo poupa trabalho.

Passo 3 — derivar.

$$\ell'(\lambda) = \frac{d\ell}{d\lambda} = -n + \frac{\sum_{i=1}^{n} x_i}{\lambda}.$$

Passo 4 — igualar a zero e resolver.

$$-n + \frac{\sum x_i}{\lambda} = 0 \quad \Longrightarrow \quad \lambda = \frac{\sum_{i=1}^{n} x_i}{n} = \bar{x}.$$

A CSO confirma que é máximo: $\ell''(\lambda) = -\sum x_i/\lambda^2 < 0$ para todo $\lambda > 0$ (supondo ao menos uma observação positiva), de modo que ℓ é estritamente côncava.

Passo 5 — o estimador.

$$\boxed{\hat{\lambda}_{MV} = \bar{X} = \frac{\sum_{i=1}^{n} X_i}{n}}$$

O resultado é reconfortante: como $\mathbb{E}(X) = \lambda$ na Poisson, o EMV do parâmetro é simplesmente a média amostral, e por $\mathbb{E}(\bar{X}) = \mu = \lambda$ ele é não viesado.

8.5.2 Exemplo 2: exponencial

Seja $X_1, \ldots, X_n \overset{\text{i.i.d.}}{\sim} \text{Expo}(\lambda)$, com $f(x) = \lambda e^{-\lambda x}$ para $x > 0$ (Seção 6).

$$L(\lambda) = \prod_{i=1}^{n} \lambda e^{-\lambda x_i} = \lambda^n e^{-\lambda \sum x_i}, \qquad \ell(\lambda) = n \ln \lambda - \lambda \sum_{i=1}^{n} x_i.$$

Derivando e igualando a zero:

$$\ell'(\lambda) = \frac{n}{\lambda} - \sum_{i=1}^{n} x_i = 0 \quad \Longrightarrow \quad \lambda = \frac{n}{\sum x_i} = \frac{1}{\bar{x}}.$$

Com $\ell''(\lambda) = -n/\lambda^2 < 0$, confirmando o máximo. Logo

$$\boxed{\hat{\lambda}_{MV} = \frac{1}{\bar{X}}}$$

Também aqui o resultado é intuitivo: na exponencial $\mathbb{E}(X) = 1/\lambda$, isto é, λ é o inverso do tempo médio de espera. Estimamos o tempo médio por $\bar{X}$ e invertemos. Note, porém, que a invariância do EMV a transformações — o EMV de $g(\theta)$ é $g(\hat{\theta})$ — **não** preserva o não-viés: como $\mathbb{E}(1/\bar{X}) \neq 1/\mathbb{E}(\bar{X})$ pela desigualdade de Jensen, $1/\bar{X}$ é um estimador **viesado** de λ, ainda que seja o de máxima verossimilhança. É um bom lembrete de que os dois conceitos — ser EMV e ser não viesado — são independentes.

8.5.3 Exemplo 3: Bernoulli

Seja $X_1, \ldots, X_n \overset{\text{i.i.d.}}{\sim} \text{Bernoulli}(p)$, com $P(X = x) = p^x(1-p)^{1-x}$ para $x \in \{0, 1\}$. Escrevendo $k = \sum x_i$ para o número de sucessos observados:

$$L(p) = \prod_{i=1}^{n} p^{x_i}(1-p)^{1-x_i} = p^k(1-p)^{n-k}, \qquad \ell(p) = k \ln p + (n-k)\ln(1-p).$$

Derivando — atenção à regra da cadeia no segundo termo, cuja derivada interna é -1:

$$\ell'(p) = \frac{k}{p} - \frac{n-k}{1-p} = 0 \quad \Longrightarrow \quad k(1-p) = (n-k)p \quad \Longrightarrow \quad k = np,$$

de onde $p = k/n = \bar{x}$. Portanto

$$\boxed{\hat{p}_{MV} = \bar{X} = \frac{\sum_{i=1}^{n} X_i}{n}}$$

A proporção amostral de sucessos. Aqui o resultado é literalmente o que qualquer pessoa faria sem teoria alguma — se 12 de 40 clientes são inadimplentes, estime a taxa em $12/40 = 0{,}30$ — e é reconfortante que o método formal o confirme. Como $\mathbb{E}(X_i) = p$ na Bernoulli, $\hat{p}$ é não viesado. Este estimador é a base de toda a inferência sobre proporções em Seção 9.

8.5.4 Exemplo 4: uma densidade sem contraparte óbvia

Os três exemplos anteriores devolveram $\bar{X}$ ou seu inverso, o que pode dar a falsa impressão de que a MV é uma maneira complicada de justificar a média amostral. Este exemplo desfaz a impressão. Seja

$$f(x) = \theta x^{\theta-1}, \qquad 0 < x < 1, \qquad \theta > 0.$$

(Verifique que é densidade: $\int_0^1 \theta x^{\theta-1} dx = [x^\theta]_0^1 = 1$.) Não há nenhuma "média amostral de θ" a que apelar. Seguindo os cinco passos:

$$L(\theta) = \prod_{i=1}^{n} \theta x_i^{\theta-1} = \theta^n \left(\prod_{i=1}^{n} x_i \right)^{\theta-1},$$

$$\ell(\theta) = n \ln \theta + (\theta - 1) \sum_{i=1}^{n} \ln x_i,$$

$$\ell'(\theta) = \frac{n}{\theta} + \sum_{i=1}^{n} \ln x_i = 0 \quad \Longrightarrow \quad \theta = -\frac{n}{\sum_{i=1}^{n} \ln x_i}.$$

$$\boxed{\hat{\theta}_{MV} = -\frac{n}{\sum_{i=1}^{n} \ln X_i}}$$

O sinal negativo não é erro: como $0 < x_i < 1$, temos $\ln x_i < 0$ e a soma é negativa, de modo que $\hat{\theta} > 0$ como exigido. A CSO também confere: $\ell''(\theta) = -n/\theta^2 < 0$. Note que o estimador **não** é função de $\bar{X}$, mas da média dos logaritmos — o método encontrou sozinho a estatística relevante, que nenhuma analogia teria sugerido. É esse o valor da máxima verossimilhança: ela funciona onde a intuição não alcança.

8.5.5 Os EMV das distribuições do curso

Reunimos os resultados, incluindo os que ficam como exercício:

Distribuição	Parâmetro	EMV
Bernoulli(p)	p	$\hat{p} = \bar{X}$
Poi(λ)	λ	$\hat{\lambda} = \bar{X}$
Expo(λ)	λ	$\hat{\lambda} = 1/\bar{X}$
Geom(p)	p	$\hat{p} = 1/\bar{X}$
$N(\mu, \sigma^2)$	μ	$\hat{\mu} = \bar{X}$
$N(\mu, \sigma^2)$	σ^2	$\hat{\sigma}^2 = \dfrac{\sum (X_i - \bar{X})^2}{n}$

A última linha merece um comentário, porque é o ponto que a seção sobre não-viés antecipou. O EMV da variância normal tem divisor n, e é portanto **viesado**, com $\mathbb{E}(\hat{\sigma}^2) = \frac{n-1}{n}\sigma^2$. A máxima verossimilhança não promete não-viés — promete outra coisa (maximizar a plausibilidade dos dados observados), e as duas promessas não coincidem. Na prática se usa S^2, com $n-1$, aceitando abandonar o EMV em favor do estimador centrado.

8.6 Método dos momentos

O segundo método é bem mais simples, e o mais antigo dos dois. A ideia: a distribuição populacional tem **momentos** — $\mathbb{E}(X)$, $\mathbb{E}(X^2)$, $\mathbb{E}(X^3)$, … — que são funções dos parâmetros. A amostra tem momentos correspondentes. Iguale uns aos outros e resolva.

! Definição — momentos e o método

O k-ésimo **momento populacional** de X é $\mathbb{E}(X^k)$, e o k-ésimo **momento amostral** é

$$m_k = \frac{\sum_{i=1}^{n} X_i^k}{n}.$$

O **método dos momentos** estima os parâmetros resolvendo o sistema obtido ao igualar tantos momentos populacionais a amostrais quantos forem os parâmetros a estimar:

$$\mathbb{E}(X) = \frac{\sum X_i}{n}, \qquad \mathbb{E}(X^2) = \frac{\sum X_i^2}{n}, \qquad \ldots$$

A justificativa é a lei dos grandes números, que vimos operar em Seção 4 e que Seção 7 formaliza: momentos amostrais convergem para momentos populacionais quando n cresce. Se $m_k \approx \mathbb{E}(X^k)$, é razoável escolher o parâmetro que faz a igualdade valer exatamente na amostra que temos.

Um parâmetro. Basta a primeira equação, $\mathbb{E}(X) = \bar{X}$.

- **Poisson:** $\mathbb{E}(X) = \lambda$, logo $\lambda = \bar{X}$ e $\hat{\lambda}_{MM} = \bar{X}$.
- **Exponencial:** $\mathbb{E}(X) = 1/\lambda$, logo $1/\lambda = \bar{X}$ e $\hat{\lambda}_{MM} = 1/\bar{X}$.
- **Bernoulli:** $\mathbb{E}(X) = p$, logo $\hat{p}_{MM} = \bar{X}$.
- **Geométrica:** $\mathbb{E}(X) = 1/p$, logo $\hat{p}_{MM} = 1/\bar{X}$.

Nos quatro casos, **os mesmos estimadores da máxima verossimilhança**, obtidos em uma linha cada, sem derivar coisa alguma.

Dois parâmetros: a Normal. Precisamos de duas equações. Da Normal sabemos $\mathbb{E}(X) = \mu$ e, de $\mathrm{V}(X) = \mathbb{E}(X^2) - \mathbb{E}^2(X)$ (Seção 4), $\mathbb{E}(X^2) = \sigma^2 + \mu^2$. O sistema é

$$\begin{cases} \mu = \dfrac{\sum X_i}{n} = \bar{X}, \\ \sigma^2 + \mu^2 = \dfrac{\sum X_i^2}{n}. \end{cases}$$

A primeira equação dá $\hat{\mu}_{MM} = \bar{X}$ imediatamente. Substituindo na segunda:

$$\hat{\sigma}^2_{MM} = \frac{\sum X_i^2}{n} - \bar{X}^2 = \frac{\sum (X_i - \bar{X})^2}{n},$$

usando a forma de cálculo da variância de Seção 2. Ou seja:

$$\hat{\mu}_{MM} = \bar{X} \qquad \text{e} \qquad \hat{\sigma}^2_{MM} = \frac{\sum_{i=1}^{n} (X_i - \bar{X})^2}{n}$$

Exatamente os mesmos estimadores da máxima verossimilhança — inclusive o divisor n, e portanto inclusive o viés em $\hat{\sigma}^2$.

A vantagem do método dos momentos é a que os exemplos evidenciam: ele é **mais simples** que a máxima verossimilhança — não requer escrever verossimilhanças, derivar nem verificar condições de segunda ordem — e, na maior parte dos casos práticos, **leva ao mesmo resultado**. Sua desvantagem aparece nos casos em que os dois divergem: o EMV tem propriedades assintóticas ótimas que o estimador de momentos em geral não tem, e há situações em que o método dos momentos produz estimativas absurdas (fora do espaço de parâmetros, por exemplo uma variância negativa). Quando os dois coincidem — que é o caso de todas as distribuições deste curso — não há o que escolher.

8.7 Além do não-viés: eficiência e consistência

> **Ferramenta matemática — nota sobre o escopo**
>
> Esta seção vai **além do programa da disciplina**: os slides do curso não tratam de eficiência nem de consistência, e nada aqui será cobrado. Incluo-a porque é o vocabulário padrão com que a literatura discute estimadores, e o leitor de doutorado o encontrará na primeira página de qualquer livro de econometria. Trate-a como um mapa do que vem depois.

O não-viés é um critério fraco: ele diz que o estimador acerta em média, mas não diz **quão disperso** ele é em torno do alvo. Entre dois estimadores não viesados, prefere-se o de menor variância — o mais **eficiente**. Se $\hat{\theta}_1$ e $\hat{\theta}_2$ são ambos não viesados e $\mathrm{V}(\hat{\theta}_1) < \mathrm{V}(\hat{\theta}_2)$, então $\hat{\theta}_1$ é mais eficiente: suas estimativas se concentram mais perto do valor verdadeiro.

Mas há um trade-off entre as duas dimensões, e a medida que o resume é o **erro quadrático médio**:

$$\boxed{\mathrm{EQM}(\hat{\theta}) = \mathbb{E}\left[(\hat{\theta} - \theta)^2\right] = \mathrm{V}(\hat{\theta}) + \left[B(\hat{\theta})\right]^2}$$

A decomposição sai pela mesma manipulação da seção sobre o $n-1$: somando e subtraindo $\mathbb{E}(\hat{\theta})$ dentro do quadrado, o termo cruzado se anula e sobram a variância e o quadrado do viés. Ela diz que um estimador pode errar por duas razões — por estar descentrado (viés) ou por oscilar demais (variância) —, e que ambas custam o mesmo na métrica quadrática. A consequência é importante e contraintuitiva: **um estimador viesado pode ser preferível a um não viesado**, se seu viés for pequeno e sua variância suficientemente menor. É exatamente o argumento por trás dos estimadores de encolhimento e da regularização em aprendizado estatístico.

Por fim, a **consistência** é uma propriedade assintótica: $\hat{\theta}$ é consistente para θ quando converge em probabilidade para θ à medida que $n \to \infty$,

$$\hat{\theta}_n \xrightarrow{p} \theta \qquad \text{isto é} \qquad \lim_{n\to\infty} P(|\hat{\theta}_n - \theta| > \varepsilon) = 0 \quad \text{para todo } \varepsilon > 0.$$

Uma condição suficiente conveniente é $\mathrm{EQM}(\hat{\theta}_n) \to 0$, ou seja, viés e variância desaparecendo simultaneamente. É esse o sentido em que $\widehat{\sigma}^2$ com divisor n, embora viesado para todo n finito, é perfeitamente aceitável assintoticamente: seu viés $-\sigma^2/n$ tende a zero. E é o sentido em que $\bar{X}$ é consistente para μ, já que é não viesado e $\mathrm{V}(\bar{X}) = \sigma^2/n \to 0$ — que é, no fundo, a lei dos grandes números de Seção 7 reescrita como propriedade de um estimador.

Guardadas as três propriedades — não-viés, eficiência, consistência —, o resultado que justifica a máxima verossimilhança como método de eleição é assintótico: sob condições de regularidade, o EMV é consistente, assintoticamente não viesado, assintoticamente Normal e atinge a menor variância possível entre todos os estimadores consistentes. Nada disso está no programa, mas explica por que a MV domina a prática econométrica.

8.8 Laboratório computacional em R

8.8.1 O viés do divisor n, visto por simulação

Esta é a figura central do capítulo. A demonstração algébrica mostrou que $\mathbb{E}[\sum (X_i - \bar{X})^2] = (n-1)\sigma^2$; vamos vê-la acontecer. O experimento é direto: geramos vinte mil amostras de tamanho $n = 5$ de uma $N(10, 4)$ — de modo que $\sigma^2 = 4$ é conhecido —, calculamos as duas versões da variância em cada amostra e comparamos as médias.

```
set.seed(20268)
mu_v <- 10; sigma2_v <- 4; n_v <- 5; M <- 20000
amostras <- matrix(rnorm(M * n_v, mean = mu_v, sd = sqrt(sigma2_v)),
                   nrow = M, ncol = n_v)
xbar_m <- rowMeans(amostras)
SQ     <- rowSums((amostras - xbar_m)^2)   # soma de quadrados de cada amostra
v_n    <- SQ / n_v                         # estimador de MV / momentos
v_n1   <- SQ / (n_v - 1)                   # S^2, o nao viesado
round(c(sigma2         = sigma2_v,
        media_div_n    = mean(v_n),
        media_div_n1   = mean(v_n1),
        teorico_div_n  = (n_v - 1) / n_v * sigma2_v,
        vies_div_n     = mean(v_n) - sigma2_v,
        vies_div_n1    = mean(v_n1) - sigma2_v), 4)
```

```
      sigma2   media_div_n  media_div_n1 teorico_div_n    vies_div_n
      4.0000        3.2019        4.0024        3.2000       -0.7981
  vies_div_n1
      0.0024
```

Os números confirmam a álgebra com precisão notável. O estimador com divisor n tem média 3,20 contra o valor verdadeiro 4 — exatamente o previsto $\frac{n-1}{n}\sigma^2 = \frac{4}{5}\cdot 4 = 3{,}2$ —, com viés de $-0{,}80$, que é $-\sigma^2/n = -4/5$. O estimador com divisor $n-1$ tem média 4,0024: o desvio em relação a 4 é ruído de simulação, não viés.

O ponto crucial é que a subestimação é **sistemática**, não um acidente de amostras particulares. Isso fica visível na figura, que mostra as duas distribuições amostrais e como o viés se comporta à medida que n cresce.

```
d_vies <- data.frame(
  valor   = c(v_n, v_n1),
  divisor = rep(c("divisor n", "divisor n - 1"), each = M))
d_vies$divisor <- factor(d_vies$divisor, levels = c("divisor n", "divisor n - 1"))
medias <- d_vies |> group_by(divisor) |> summarise(m = mean(valor), .groups = "drop")

g1 <- ggplot(d_vies, aes(valor, fill = divisor, colour = divisor)) +
  geom_density(alpha = 0.18, linewidth = 0.7) +
  geom_vline(xintercept = sigma2_v, colour = cz, linetype = "dashed") +
```

```
  geom_vline(data = medias, aes(xintercept = m, colour = divisor),
             linewidth = 0.8, linetype = "dotted") +
  scale_fill_manual(values = c(vm, az)) +
  scale_colour_manual(values = c(vm, az)) +
  coord_cartesian(xlim = c(0, 16)) +
  labs(x = "estimativa da variancia  (verdadeiro: 4)", y = "densidade")

ns <- 2:30
conv <- do.call(rbind, lapply(ns, function(k) {
  A <- matrix(rnorm(4000 * k, mean = mu_v, sd = sqrt(sigma2_v)), nrow = 4000)
  S <- rowSums((A - rowMeans(A))^2)
  data.frame(n = k, div_n = mean(S / k), div_n1 = mean(S / (k - 1)))
}))
d_conv <- data.frame(n = rep(conv$n, 2), valor = c(conv$div_n, conv$div_n1),
                     divisor = rep(c("divisor n", "divisor n - 1"), each = nrow(conv)))
d_conv$divisor <- factor(d_conv$divisor, levels = c("divisor n", "divisor n - 1"))

g2 <- ggplot(d_conv, aes(n, valor, colour = divisor)) +
  geom_hline(yintercept = sigma2_v, colour = cz, linetype = "dashed") +
  geom_line(linewidth = 0.7) + geom_point(size = 0.8) +
  scale_colour_manual(values = c(vm, az)) +
  labs(x = "n", y = "media das estimativas")

g1 / g2
```

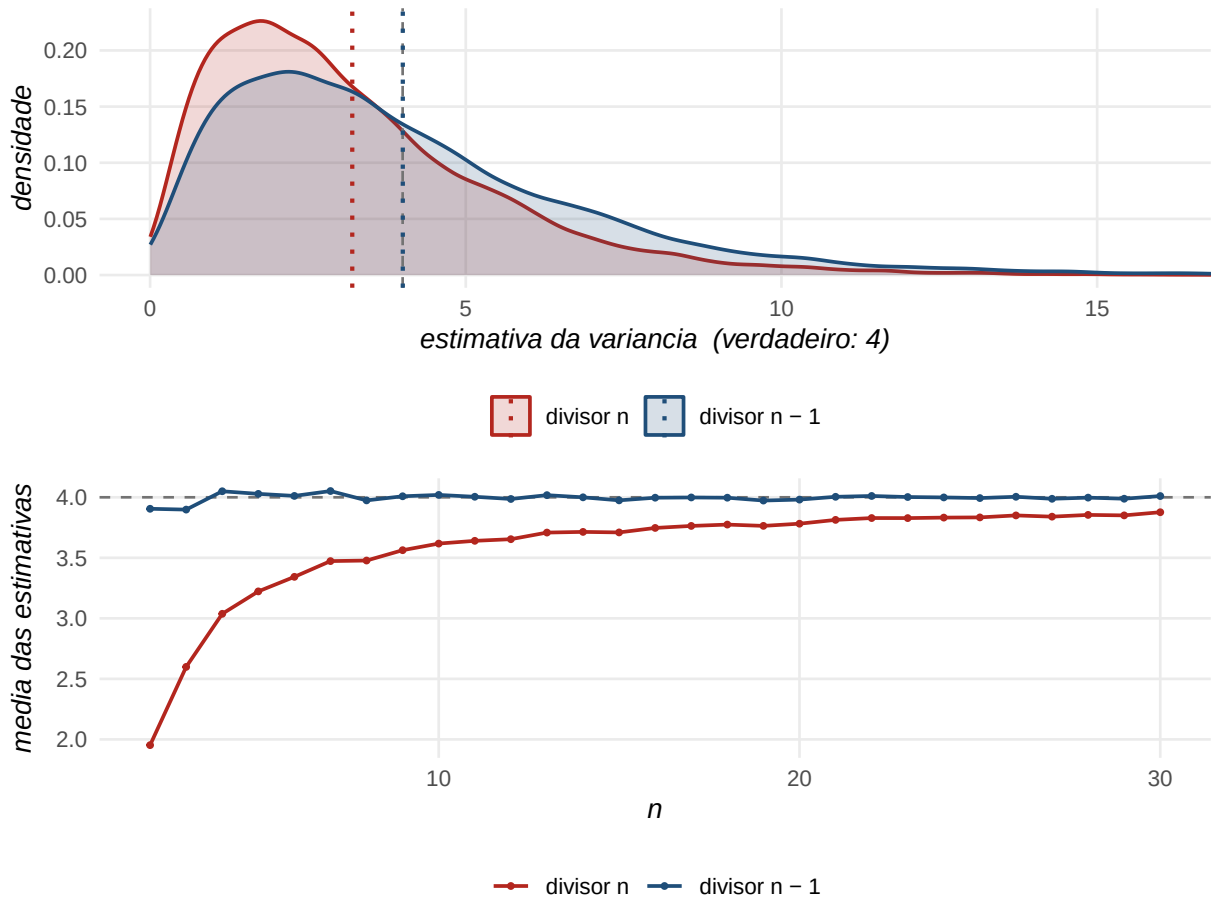


Figura 13: O viés do divisor n. Acima: distribuição das 20.000 estimativas de σ^2 com cada divisor ($n = 5$); a linha cinza tracejada é o valor verdadeiro $\sigma^2 = 4$ e as pontilhadas são as médias de cada estimador. Abaixo: média das estimativas em função de n — o divisor n (vermelho) converge ao valor verdadeiro apenas quando $n \to \infty$, enquanto o divisor $n-1$ (azul) o acerta para todo n.

O painel superior mostra duas distribuições com a **mesma forma** — uma é a outra multiplicada por $n/(n-1)$ —, mas centradas em pontos diferentes: a vermelha à esquerda do alvo, a azul sobre ele. O painel inferior mostra o custo do viés desaparecendo com n: em $n = 2$ o divisor n subestima σ^2 pela metade; em $n = 30$ o erro já é de 3%. A lição prática é a que a álgebra antecipou — **o $n-1$ importa muito em amostras pequenas e quase nada em amostras grandes** —, e a lição conceitual é que "quase nada" não é "nada", e não há razão para usar o estimador errado quando o certo custa uma subtração.

8.8.2 A verossimilhança visualizada

Vejamos agora a função de verossimilhança como objeto concreto. Tomamos uma amostra de dez contagens que suporemos Poisson e traçamos $L(\lambda)$ e $\ell(\lambda)$.

```r
x_poi <- c(2, 0, 3, 1, 4, 2, 1, 3, 2, 2)
c(n = length(x_poi), soma = sum(x_poi), xbar = mean(x_poi))
```

```
   n soma xbar
  10   20    2
```

```r
L_poi <- function(lam) sapply(lam, function(l) prod(dpois(x_poi, l)))
l_poi <- function(lam) sapply(lam, function(l) sum(dpois(x_poi, l, log = TRUE)))
c(L_no_max = L_poi(mean(x_poi)), l_no_max = l_poi(mean(x_poi)))
```

```
     L_no_max      l_no_max
1.563423e-07 -1.567122e+01
```

Note a ordem de grandeza de L no máximo: cerca de 10^{-6}. Com dez observações a verossimilhança já é minúscula, porque é um produto de dez probabilidades menores que 1; com mil observações ela seria numericamente zero em qualquer computador. É a razão numérica — além da algébrica — para trabalhar sempre com ℓ.

```r
grid_lam <- seq(0.5, 5, length.out = 500)
lam_hat  <- mean(x_poi)

gA <- ggplot(data.frame(lam = grid_lam, L = L_poi(grid_lam)), aes(lam, L)) +
  geom_line(colour = az, linewidth = 0.8) +
  geom_vline(xintercept = lam_hat, colour = vm, linetype = "dashed") +
  labs(x = "lambda", y = "L(lambda)")

gB <- ggplot(data.frame(lam = grid_lam, l = l_poi(grid_lam)), aes(lam, l)) +
  geom_line(colour = vd, linewidth = 0.8) +
  geom_vline(xintercept = lam_hat, colour = vm, linetype = "dashed") +
  labs(x = "lambda", y = "log L(lambda)")

gA + gB
```

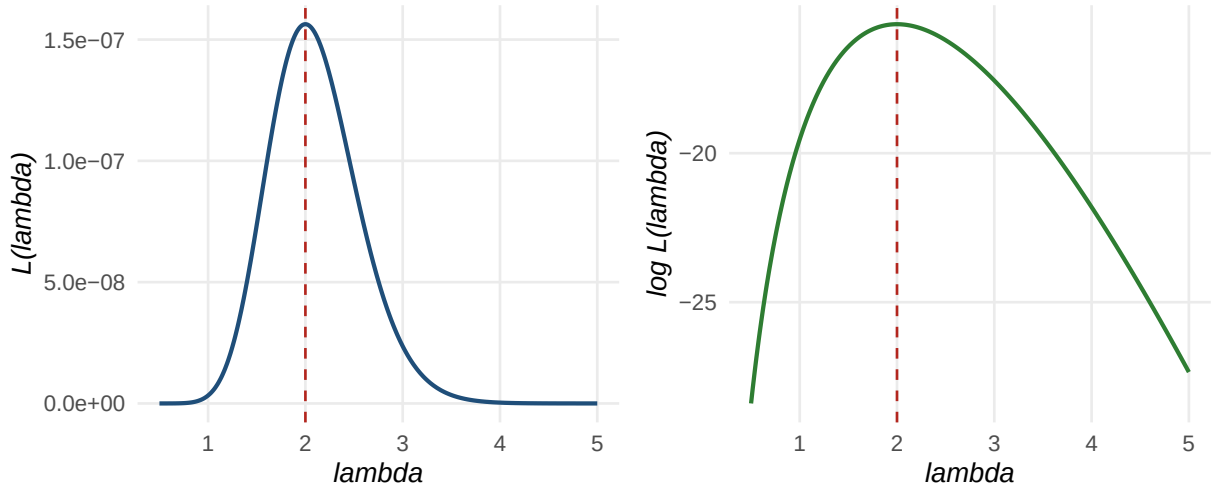


Figura 14: Verossimilhança $L(\lambda)$ (esquerda) e log-verossimilhança $\ell(\lambda)$ (direita) para a amostra Poisson. A linha vermelha tracejada marca $\hat{\lambda} = \bar{x} = 2$. As duas curvas têm formas muito diferentes, mas o **mesmo** ponto de máximo — o logaritmo é monótono crescente e preserva o argmax.

A figura é a tradução gráfica da caixa sobre o logaritmo. À esquerda, uma curva positiva com um pico agudo; à direita, uma curva côncava que vai a $-\infty$ nas bordas. Escalas incomparáveis, formas distintas — e o pico exatamente sobre $\lambda = 2$ nas duas. Maximizar uma é maximizar a outra.

Vale também observar a **curvatura** de ℓ perto do máximo, porque ela reaparecerá em Seção 9: quanto mais pontiaguda a log-verossimilhança, mais os dados discriminam entre valores candidatos do parâmetro, e mais precisa é a estimativa. Uma log-verossimilhança achatada significa que muitos valores de λ explicam os dados quase igualmente bem — pouca informação. Essa curvatura, $-\ell''(\hat{\theta})$, é a *informação de Fisher*.

8.8.3 Máxima verossimilhança numérica contra a fórmula fechada

Nos exemplos do capítulo conseguimos resolver a CPO analiticamente. Em modelos realistas isso raramente é possível, e a maximização é feita numericamente. Vale conferir que os dois caminhos concordam onde ambos existem. Usamos `optimize` para um parâmetro — lembrando que os otimizadores do R **minimizam**, de modo que passamos $-\ell$.

```
x_exp <- c(0.8, 2.1, 0.3, 1.4, 3.2, 0.9, 1.1, 0.5)

neg_l_poi <- function(lam) -sum(dpois(x_poi, lam, log = TRUE))
neg_l_exp <- function(lam) -sum(dexp(x_exp, rate = lam, log = TRUE))

opt_poi <- optimize(neg_l_poi, interval = c(1e-6, 20))
opt_exp <- optimize(neg_l_exp, interval = c(1e-6, 20))

round(rbind(Poisson     = c(numerico = opt_poi$minimum, analitico = mean(x_poi)),
            Exponencial = c(numerico = opt_exp$minimum, analitico = 1 / mean(x_exp))), 6)
```

```
            numerico analitico
Poisson     1.999982  2.000000
Exponencial 0.776708  0.776699
```

As duas colunas coincidem até a sexta casa: $\hat{\lambda} = \bar{x} = 2$ na Poisson e $\hat{\lambda} = 1/\bar{x} \approx 0{,}7767$ na exponencial. A pequena discrepância residual é a tolerância do otimizador, não erro de fórmula.

Para dois parâmetros usa-se `optim`. Testemos o caso da Normal, cujo EMV a tabela dá como $\hat{\mu} = \bar{X}$ e $\hat{\sigma}^2 = \sum(x_i - \bar{x})^2/n$ — a variância com divisor n, que é justamente a `var_p()` definida no preâmbulo destas notas:

```
x_nor <- c(9.2, 11.4, 10.1, 8.7, 12.3, 10.8, 9.9, 11.0)
neg_l_nor <- function(th) -sum(dnorm(x_nor, mean = th[1], sd = sqrt(th[2]), log = TRUE))
opt_nor <- optim(c(mean(x_nor), var_p(x_nor)), neg_l_nor,
                 method = "L-BFGS-B", lower = c(-Inf, 1e-8))
round(rbind(numerico  = opt_nor$par,
            analitico = c(mean(x_nor), var_p(x_nor)),
            S2_nao_viesado = c(mean(x_nor), var(x_nor))), 6)
```

```
                [,1]     [,2]
numerico       10.425 1.224375
analitico      10.425 1.224375
S2_nao_viesado 10.425 1.399286
```

A maximização numérica reproduz exatamente $\hat{\mu} = \bar{x}$ e $\hat{\sigma}^2$ com divisor n. A terceira linha mostra o S^2 não viesado, com divisor $n-1$: é **maior**, como tem de ser, e é ele que usaríamos na prática — mas **não** é o EMV. Os dois critérios divergem, e a tabela torna a divergência tangível.

Por fim, o exemplo em que a MV não devolve a média amostral. Simulamos de $f(x) = \theta x^{\theta-1}$ com $\theta = 3$ — o que se faz notando que se $U \sim \text{Unif}(0,1)$ então $U^{1/\theta}$ tem essa densidade — e comparamos a fórmula $\hat{\theta} = -n/\sum \ln x_i$ com a maximização numérica:

```
set.seed(11)
theta0 <- 3
x_th   <- runif(400)^(1 / theta0)
neg_l_th <- function(th) -sum(log(th) + (th - 1) * log(x_th))
round(c(analitico  = -length(x_th) / sum(log(x_th)),
        numerico   = optimize(neg_l_th, c(1e-6, 50))$minimum,
        verdadeiro = theta0), 6)
```

```
 analitico   numerico verdadeiro
  3.019211   3.019197   3.000000
```

Fórmula e otimizador concordam em 3,019, próximo do $\theta = 3$ verdadeiro — a diferença é erro amostral, que encolhe com n. Repare que aqui $\bar{x}$ não teria ajudado em nada: o estimador depende de $\sum \ln x_i$.

8.8.4 Máxima verossimilhança contra método dos momentos

Verifiquemos, enfim, a afirmação de que os dois métodos coincidem nas distribuições do curso. Geramos uma amostra de cada e calculamos os dois estimadores pelas respectivas fórmulas.

```
set.seed(20269)
x_ber <- rbinom(40, 1, 0.35)        # Bernoulli(p),  p = 0,35
x_geo <- rgeom(30, 0.25) + 1        # Geometrica(p), numero de tentativas, p = 0,25

comparacao <- rbind(
  Bernoulli   = c(EMV = mean(x_ber),      momentos = mean(x_ber)),
  Poisson     = c(mean(x_poi),            mean(x_poi)),
  Exponencial = c(1 / mean(x_exp),        1 / mean(x_exp)),
  Geometrica  = c(1 / mean(x_geo),        1 / mean(x_geo)),
  Normal_mu   = c(mean(x_nor),            mean(x_nor)),
  Normal_s2   = c(var_p(x_nor),           mean(x_nor^2) - mean(x_nor)^2))
round(comparacao, 6)
```

```
                EMV  momentos
Bernoulli    0.300000  0.300000
Poisson      2.000000  2.000000
Exponencial  0.776699  0.776699
Geometrica   0.225564  0.225564
Normal_mu   10.425000 10.425000
Normal_s2    1.224375  1.224375
```

Todas as linhas têm colunas idênticas. As duas últimas merecem atenção: para σ^2, o EMV foi calculado como $\sum(x_i - \bar{x})^2/n$ e o de momentos como $m_2 - m_1^2$, expressões algebricamente distintas que produzem o mesmo número — é a forma de cálculo da variância de Seção 2, verificada numericamente. E ambos usam divisor n; ambos são viesados.

Para fechar, uma verificação independente: os estimadores de momentos das duas distribuições simuladas contra os valores verdadeiros que usamos para gerar os dados.

```
round(c(p_bernoulli_est = mean(x_ber),  p_bernoulli_true = 0.35,
        p_geom_est      = 1 / mean(x_geo), p_geom_true   = 0.25), 4)
```

```
 p_bernoulli_est p_bernoulli_true       p_geom_est      p_geom_true
          0.3000           0.3500           0.2256           0.2500
```

Estimamos $\hat{p} = 0{,}30$ contra $p = 0{,}35$ verdadeiro na Bernoulli, e $\hat{p} = 0{,}2256$ contra 0,25 na geométrica. Os estimadores acertam a ordem de grandeza, erram por acaso amostral, e a pergunta natural — *de quanto pode ser esse erro?* — é precisamente a que a estimação pontual **não** responde. Um único número não carrega consigo nenhuma indicação de sua própria precisão. É essa lacuna que o próximo capítulo preenche, substituindo o ponto por um intervalo (Seção 9).

8.9 Exercícios

1. **Estimador e estimativa.** Uma amostra de seis salários iniciais, em milhares de reais, registrou 4,2; 3,8; 5,1; 4,0; 6,3; 4,4. (a) Escreva o estimador de μ e calcule a estimativa. (b) Calcule s^2 e $\hat{\sigma}^2$ (divisores $n-1$ e n) e diga qual é o EMV e qual é o não viesado. (c) Explique por que faz sentido perguntar "qual a variância de $\bar{X}$?" e não faz sentido perguntar "qual a variância de $\bar{x}$?".

2. **Não-viés de uma média ponderada.** Considere o estimador $\hat{\mu}_w = \sum_{i=1}^{n} w_i X_i$ com pesos constantes w_i. (a) Mostre que $\hat{\mu}_w$ é não viesado para μ se e somente se $\sum w_i = 1$. (b) Sob a hipótese i.i.d., calcule $\mathrm{V}(\hat{\mu}_w)$. (c) Mostre que, entre todos os pesos que somam 1, a variância é minimizada por $w_i = 1/n$ — ou seja, $\bar{X}$ é o mais eficiente da classe.

3. **O viés do divisor n, à mão.** Seja uma AAS de tamanho $n = 3$ de uma população com $\mu = 0$ e $\sigma^2 = 1$. (a) Use a identidade $\sum (X_i - \bar{X})^2 = \sum (X_i - \mu)^2 - n(\bar{X} - \mu)^2$ para obter $\mathbb{E}[\sum (X_i - \bar{X})^2]$. (b) Calcule o viés de $\hat{\sigma}^2$ e de S^2. (c) Suponha agora que μ seja **conhecido** e igual a 0: mostre que $\sum X_i^2/n$ é não viesado para σ^2, com divisor n. Por que a correção deixa de ser necessária?

4. **MV na geométrica.** Seja X o número de tentativas até o primeiro sucesso, com $P(X = x) = (1-p)^{x-1}p$ para $x = 1, 2, \ldots$ (Seção 5). (a) Escreva $L(p)$ e $\ell(p)$ para uma AAS de tamanho n. (b) Obtenha a CPO e resolva, mostrando que $\hat{p}_{MV} = 1/\bar{X}$. (c) Verifique a CSO. (d) Obtenha o mesmo resultado pelo método dos momentos em uma linha.

5. **MV na Normal, os dois parâmetros.** Para $X_1, \ldots, X_n \overset{\text{i.i.d.}}{\sim} N(\mu, \sigma^2)$, com $f(x) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\!\big[-\frac{(x-\mu)^2}{2\sigma^2}\big]$: (a) escreva $\ell(\mu, \sigma^2)$;

   (b) derive em relação a μ e obtenha $\hat{\mu} = \bar{X}$; (c) derive em relação a σ^2 (tratando-o como um único parâmetro, não σ) e obtenha $\hat{\sigma}^2 = \sum (X_i - \bar{X})^2/n$; (d) calcule $\mathbb{E}(\hat{\sigma}^2)$ e conclua que o EMV é viesado. (e) Por que a ordem importa — por que substituímos $\hat{\mu}$ antes de resolver para σ^2?

6. **Uma verossimilhança com fronteira.** Seja $X_1, \ldots, X_n \overset{\text{i.i.d.}}{\sim} \mathrm{Unif}(0, \theta)$, com $f(x) = 1/\theta$ para $0 < x < \theta$. (a) Escreva $L(\theta)$ atentando ao suporte: a verossimilhança é nula se algum $x_i > \theta$. (b) Mostre que L é **decrescente** em θ no domínio onde é positiva e conclua que $\hat{\theta}_{MV} = \max\{X_1, \ldots, X_n\}$. (c) Por que os cinco passos falham aqui — o que dá errado no passo 3? (d) Obtenha o estimador de momentos e mostre que é $2\bar{X}$; qual dos dois pode produzir uma estimativa **impossível** (menor que algum x_i observado)?

7. **Viés, variância e EQM.** Retome $\hat{\sigma}^2$ (divisor n) e S^2 (divisor $n-1$) numa população Normal, onde se sabe que $\mathrm{V}(S^2) = 2\sigma^4/(n-1)$. (a) Obtenha $\mathrm{V}(\hat{\sigma}^2)$ usando $\hat{\sigma}^2 = \frac{n-1}{n} S^2$ e a propriedade $\mathrm{V}(aX) = a^2 \mathrm{V}(X)$. (b) Calcule o EQM de cada um. (c) Mostre que $\mathrm{EQM}(\hat{\sigma}^2) < \mathrm{EQM}(S^2)$ para todo n. (d) Discuta: se o estimador viesado tem menor erro quadrático médio, por que a prática usa S^2?

8. **Simulação.** Adapte o código do laboratório para investigar o viés de $\hat{\lambda} = 1/\bar{X}$ como estimador do parâmetro de uma exponencial. (a) Gere 10.000 amostras de tamanho $n = 5$ de uma Expo(2) e calcule a média de $1/\bar{x}$; ela é próxima de 2? (b) Repita com $n = 50$ e $n = 500$ e descreva o comportamento do viés. (c) O que esse exercício ilustra sobre a relação entre ser EMV, ser não viesado e ser consistente?

(d) Verifique numericamente a desigualdade de Jensen comparando $\mathbb{E}(1/\bar{X})$ com $1/\mathbb{E}(\bar{X})$.

9 Intervalos de Confiança

O capítulo anterior terminou com uma lacuna. Estimamos $\hat{p} = 0{,}30$ contra um $p = 0{,}35$ verdadeiro, $\hat{\lambda} = 0{,}2256$ contra $0{,}25$ — e a pergunta natural, *de quanto pode ser esse erro?*, ficou sem resposta. Não foi acidente: a estimação pontual **não pode** respondê-la. Um número solto não carrega consigo nenhuma indicação de sua própria precisão.

E o problema é pior do que parece. Se a população é contínua, o estimador $\bar{X}$ é uma variável aleatória contínua, e vimos em Seção 6 que uma contínua atribui probabilidade zero a qualquer ponto isolado. Logo

$$P(\bar{X} = \mu) = 0.$$

A estimativa pontual, por melhor que seja o estimador, **quase certamente erra**. Erra pouco, esperamos — mas erra. O **erro de estimação** é

$$\bar{X} - \mu,$$

e ele é desconhecido, porque μ é desconhecido. Estamos numa posição curiosa: sabemos que erramos, não sabemos por quanto, e o próprio objeto que nos permitiria medir o erro é o que buscamos.

A saída é abandonar a pretensão de acertar o ponto e passar a apostar num **intervalo**. Em vez de dizer "a altura média é 180 cm" — afirmação que é quase certamente falsa —, diremos "a altura média está entre 174,74 e 185,26 cm, com 95% de confiança" — afirmação que tem uma chance razoável de ser verdadeira, e cuja margem de erro é explícita. Esse é o objeto deste capítulo: a **estimação intervalar**, ou **intervalo de confiança**.

O plano é o seguinte. Construímos primeiro, com todo o detalhe algébrico, o intervalo para μ quando σ é conhecido — é o caso-mãe, e todos os demais são variações dele. Depois paramos longamente na **interpretação**, que é o ponto conceitualmente mais difícil de todo o curso e a fonte de mais erros em prova. Em seguida relaxamos a hipótese de σ conhecido, o que nos obriga a apresentar a **distribuição t de Student**. Daí em diante é aplicação do mesmo molde a outros parâmetros: proporção, variância (via **qui-quadrado**), diferença de médias, diferença de proporções e razão de variâncias (via F). O capítulo fecha com uma tabela-resumo de todos os intervalos e um laboratório em R cuja primeira figura resolve, visualmente, o mal-entendido conceitual.

9.1 A ideia da estimação intervalar

! Definição — estimativa intervalar

Uma **estimativa intervalar** de um parâmetro θ é um intervalo $[L, U]$, construído a partir da amostra, ao qual se associa um **grau de confiança** $1 - \alpha$: a proporção de intervalos assim construídos que conteriam θ, se o procedimento fosse repetido indefinidamente.
O número α é o **nível de significância**; $1 - \alpha$, o **nível** ou **coeficiente de confiança**. Valores usuais são $\alpha = 0{,}05$ (confiança de 95%), $\alpha = 0{,}10$ (90%) e $\alpha = 0{,}01$ (99%).

A construção segue sempre a mesma receita, e vale enunciá-la antes de executá-la, porque ela se repetirá seis vezes ao longo do capítulo:

1. encontrar uma **quantidade pivotal** — uma função da amostra e do parâmetro cuja distribuição seja **conhecida e não dependa do parâmetro**;
2. escrever uma afirmação probabilística sobre essa quantidade, usando os quantis da sua distribuição;
3. **isolar o parâmetro** no meio da desigualdade, por manipulação algébrica;
4. substituir as variáveis aleatórias pelos valores observados.

O passo 1 é o único que exige teoria; os demais são álgebra. E o passo 4 — aparentemente inócuo — é onde mora toda a dificuldade de interpretação.

9.2 O intervalo para a média com σ conhecido

Suponha $X_1, \ldots, X_n$ uma AAS de uma população $N(\mu, \sigma^2)$, com μ desconhecido e σ^2 **conhecido**. A hipótese é artificial — quem conhece σ mas não μ? —, mas é o ponto de partida pedagógico correto, porque isola a lógica da construção sem a complicação extra de estimar a variância.

9.2.1 Passo 1: a quantidade pivotal

Sabemos de Seção 7 que a média amostral de uma população Normal é ela própria Normal, com

$$\bar{X} \sim N\left(\mu,\ \frac{\sigma^2}{n}\right),$$

pois $\mathbb{E}(\bar{X}) = \mu$ e $\mathrm{V}(\bar{X}) = \sigma^2/n$. Padronizando à maneira de Seção 6 — subtrair a média, dividir pelo desvio padrão —, obtemos

$$\boxed{Z = \frac{\bar{X} - \mu}{\sigma/\sqrt{n}} \sim N(0,1)}$$

Esta é a quantidade pivotal. Repare no que a torna útil: ela **contém** o parâmetro desconhecido μ, mas sua **distribuição não depende dele** — é uma Normal padrão qualquer que seja μ. É essa dupla propriedade que permitirá, adiante, transferir a informação probabilística sobre Z para uma afirmação sobre μ.

Note também que o denominador $\sigma/\sqrt{n}$ é o **erro-padrão** de $\bar{X}$, discutido em Seção 7. Ele é a unidade natural de medida do erro de estimação, e reaparecerá em todas as fórmulas do capítulo.

9.2.2 Passos 2 a 5: isolando μ

Seja $z_{\alpha/2}$ o valor da Normal padrão que deixa área $\alpha/2$ **à direita**, de modo que o intervalo $[-z_{\alpha/2},\, z_{\alpha/2}]$ concentra probabilidade $1-\alpha$, com $\alpha/2$ em cada cauda. Por simetria da Normal, esse é o intervalo **mais curto** com essa probabilidade. Partimos então de

$$P\big(-z_{\alpha/2} \le Z \le z_{\alpha/2}\big) = 1-\alpha.$$

Substituindo Z pela sua expressão e isolando μ em cinco movimentos algébricos:

$$P\left(-z_{\alpha/2} \le \frac{\bar{X}-\mu}{\sigma/\sqrt{n}} \le z_{\alpha/2}\right) = 1-\alpha \quad \text{(1) substituir a pivotal}$$

$$P\left(-z_{\alpha/2}\frac{\sigma}{\sqrt{n}} \le \bar{X}-\mu \le z_{\alpha/2}\frac{\sigma}{\sqrt{n}}\right) = 1-\alpha \quad \text{(2) multiplicar por } \sigma/\sqrt{n} > 0$$

$$P\left(-\bar{X}-z_{\alpha/2}\frac{\sigma}{\sqrt{n}} \le -\mu \le -\bar{X}+z_{\alpha/2}\frac{\sigma}{\sqrt{n}}\right) = 1-\alpha \quad \text{(3) subtrair } \bar{X}$$

$$P\left(\bar{X}+z_{\alpha/2}\frac{\sigma}{\sqrt{n}} \ge \mu \ge \bar{X}-z_{\alpha/2}\frac{\sigma}{\sqrt{n}}\right) = 1-\alpha \quad \text{(4) multiplicar por } -1$$

$$P\left(\bar{X}-z_{\alpha/2}\frac{\sigma}{\sqrt{n}} \le \mu \le \bar{X}+z_{\alpha/2}\frac{\sigma}{\sqrt{n}}\right) = 1-\alpha \quad \text{(5) reordenar}$$

O passo (4) é o único que exige atenção: multiplicar uma desigualdade por -1 **inverte os sinais de desigualdade**, e é por isso que o limite que estava à esquerda vai para a direita. O passo (5) apenas relê a expressão da esquerda para a direita.

Substituindo agora $\bar{X}$ pelo valor observado $\bar{x}$, chegamos ao resultado central:

$$\boxed{\mathrm{IC}_{100(1-\alpha)\%}(\mu) = \left[\bar{x} \mp z_{\alpha/2}\frac{\sigma}{\sqrt{n}}\right] = \left[\bar{x}-z_{\alpha/2}\frac{\sigma}{\sqrt{n}}\;;\;\bar{x}+z_{\alpha/2}\frac{\sigma}{\sqrt{n}}\right]}$$

A estrutura é a que se repetirá o capítulo inteiro:

$$\text{estimativa pontual} \;\mp\; \underbrace{(\text{quantil}) \times (\text{erro-padrão})}_{\text{margem de erro } \varepsilon}.$$

A **margem de erro** $\varepsilon = z_{\alpha/2}\,\sigma/\sqrt{n}$ é a semi-amplitude do intervalo. Vale reter três coisas sobre ela, todas legíveis na fórmula: cresce com a confiança desejada (via $z_{\alpha/2}$), cresce com a dispersão da população (via σ) e decresce com o tamanho da amostra — mas só na razão $1/\sqrt{n}$, de modo que **quadruplicar a amostra reduz a margem à metade**, exatamente o decaimento do erro-padrão de Seção 7.

9.2.3 Os quantis que se usa

Três valores cobrem praticamente todas as aplicações do curso e devem ser memorizados:

Confiança $1-\alpha$	α	$\alpha/2$	$z_{\alpha/2}$
90%	0,10	0,05	1,645
95%	0,05	0,025	1,96
99%	0,01	0,005	2,575

> **i** Ferramenta matemática — lendo a tabela Normal
>
> A tabela da Normal padrão usada no curso (Seção 6) fornece $P(0 \le Z \le z)$, a área entre a média e z. Para achar $z_{\alpha/2}$ procura-se, **no corpo da tabela**, o valor $0{,}5 - \alpha/2$ e lê-se a abscissa correspondente nas margens. Para 95%: $\alpha/2 = 0{,}025$, procura-se $0{,}5 - 0{,}025 = 0{,}475$ no corpo, e encontra-se $z = 1{,}96$.
> No R, o caminho é direto — `qnorm(1 - alpha/2)`, porque o R trabalha sempre com a acumulada à esquerda. Para 95%: `qnorm(0.975) = 1.959964`. A tabela impressa arredonda para 1,96; a diferença é irrelevante na quarta casa da resposta, mas é uma boa razão para conferir os exercícios no computador.
> *(O valor* 2,575 *da linha de 99% é o arredondamento tradicional das tabelas de duas casas; o valor exato é* 2,5758*. Usaremos o valor da tabela nos exemplos à mão, como faz o deck, e o exato no laboratório.)*

Exemplo (alturas, com σ conhecido). Retomamos a amostra de cinco alturas de Seção 8, cuja média foi $\bar{x} = 180$ cm, e supomos saber que o desvio padrão populacional é $\sigma = 6$ cm. O intervalo de 95% para μ é

$$\mathrm{IC}_{95\%}(\mu) = \left[180 \mp 1{,}96 \cdot \frac{6}{\sqrt{5}}\right] = [180 \mp 1{,}96 \cdot 2{,}6833] = [180 \mp 5{,}2592] = [174{,}74 \ ; \ 185{,}26]\,.$$

A margem de erro é de aproximadamente 5,26 cm. Note que a estimativa pontual, 180, está exatamente no **centro** do intervalo — sempre está, e esse fato aparentemente trivial vai derrubar um dos erros de interpretação da próxima seção.

9.2.4 O trade-off entre confiança e precisão

O que aconteceria se quiséssemos mais confiança? Basta trocar o quantil. Com os mesmos dados:

Confiança	$z_{\alpha/2}$	Margem ε	Intervalo	Amplitude
90%	1,645	4,41	[175,59; 184,41]	8,83
95%	1,96	5,26	[174,74; 185,26]	10,52

Confiança	$z_{\alpha/2}$	Margem ε	Intervalo	Amplitude
99%	2,575	6,91	[173,09; 186,91]	13,82

O padrão é inescapável e vale como princípio geral:

> **!** O trade-off fundamental da estimação intervalar
>
> **Quanto maior o grau de confiança, mais amplo o intervalo** — e, portanto, **menos preciso**.
> Confiança e precisão são objetivos que competem entre si. Um intervalo de confiança 100% é sempre disponível — $(-\infty, +\infty)$ — e é perfeitamente inútil, porque não exclui nada. Um intervalo muito estreito é muito informativo, mas raramente conterá o parâmetro. A escolha de α é a escolha de um ponto nesse trade-off, e não há resposta técnica para qual ponto é o certo: é decisão do pesquisador, guiada pelo custo relativo dos dois tipos de erro.
> O único modo de **melhorar os dois ao mesmo tempo** é aumentar n: com mais dados, obtém-se maior confiança *e* menor amplitude. Precisão custa amostra.

9.3 A interpretação: o ponto mais difícil do capítulo

Chegamos ao ponto em que a maior parte dos alunos erra, e erra de maneira previsível. O problema é que a expressão "95% de confiança" **soa** como uma probabilidade, e a tentação de tratá-la como tal é enorme. É preciso resistir.

9.3.1 Onde há variáveis aleatórias e onde não há

Compare com cuidado as duas expressões. A primeira, obtida ao final da derivação **antes** de observar a amostra:

$$P\big(\bar{X} - \varepsilon \leq \mu \leq \bar{X} + \varepsilon\big) = 1 - \alpha.$$

Aqui $\bar{X}$ é uma **variável aleatória** — a média de uma amostra que ainda não foi coletada. Os limites do intervalo são aleatórios; o intervalo inteiro é um objeto aleatório, que se desloca conforme a amostra sorteada. A afirmação probabilística é legítima: ela diz que o intervalo aleatório $[\bar{X} - \varepsilon,\ \bar{X} + \varepsilon]$ tem probabilidade $1 - \alpha$ de **capturar** o ponto fixo μ.

A segunda, obtida **depois** de observar a amostra, substituindo $\bar{X}$ por $\bar{x} = 180$:

$$P(174{,}74 \leq \mu \leq 185{,}26) = \ ?$$

Olhe para essa expressão. O que há dentro dos parênteses? Três **números**: 174,74; μ; e 185,26. O primeiro e o terceiro são números que calculamos. E μ — a altura média da população — também é um número: fixo, determinado, desconhecido por nós, mas fixo. Ele não é uma variável aleatória; não tem distribuição; não flutua.

! Grau de confiança NÃO é probabilidade

Em $P(\bar{X} - \varepsilon \leq \mu \leq \bar{X} + \varepsilon) = 1 - \alpha$ **há variável aleatória**: o $\bar{X}$ maiúsculo. A afirmação probabilística faz sentido.
Em $P(\bar{x} - \varepsilon \leq \mu \leq \bar{x} + \varepsilon)$ **só há números**. Não há variável aleatória. **Logo não se pode falar em probabilidade.**
O parâmetro μ **não é variável aleatória, é constante**. Uma vez calculado o intervalo numérico, uma de duas coisas é verdade — **ou μ está no intervalo, ou não está**. A "probabilidade" desse evento é 1 ou 0, e nós simplesmente não sabemos qual.
É exatamente por isso que se usa a palavra **confiança**, e não **probabilidade**. A troca de vocabulário não é elegância: é a sinalização de que o objeto mudou de natureza no passo 4 da receita, quando trocamos maiúsculas por minúsculas.

A distinção é a mesma de Seção 8 entre **estimador** e **estimativa**, aplicada agora ao intervalo. O *intervalo aleatório* $[\bar{X} \mp \varepsilon]$ é a "estimador"; o *intervalo numérico* [174,74; 185,26] é a "estimativa". Faz sentido perguntar pela probabilidade de um estimador cobrir o parâmetro; não faz sentido perguntar pela probabilidade de um número estar entre outros dois números.

9.3.2 A interpretação correta: amostras repetidas

Se a confiança não é uma probabilidade sobre *este* intervalo, o que ela é? Uma propriedade do **procedimento** que o gerou.

! A interpretação frequentista, em amostras repetidas

Se sorteássemos **repetidas amostras** de tamanho n da mesma população e construíssemos, em cada uma, um intervalo pela fórmula $[\bar{x} \mp z_{\alpha/2}\sigma/\sqrt{n}]$, então $100(1-\alpha)\%$ **desses intervalos conteriam o verdadeiro** μ e $100\alpha\%$ não o conteriam.
Para o exemplo das alturas: de cada 100 amostras de cinco estudantes, cerca de 95 produziriam intervalos que contêm a verdadeira altura média — e cerca de 5 produziriam intervalos que a deixam de fora, sem que tenhamos como saber, olhando para um intervalo isolado, em qual dos dois grupos ele caiu.

A confiança, portanto, qualifica o **método**, não o resultado particular. É como a taxa de acerto de um procedimento cirúrgico: dizer que a cirurgia tem 95% de sucesso não diz nada sobre *este* paciente em particular depois que a operação acabou — ela ou deu certo, ou não deu. A taxa descreve o histórico do procedimento, e é ela que justifica confiar nele.

Note ainda um detalhe importante: os 5% que falham não falham por erro do estatístico. Não há nada a corrigir. Amostras atípicas ocorrem com a frequência exata que a teoria prevê, e o método está funcionando **como projetado** quando erra em 5% dos casos. Um método que nunca errasse teria de produzir intervalos infinitos.

A Figura 15 do laboratório faz exatamente esse experimento — 100 amostras, 100 intervalos, contagem dos que falham — e é a figura mais importante do capítulo.

9.3.3 Os dois erros de interpretação

O professor lista dois enunciados errados que aparecem com regularidade em provas. Vale enunciá-los explicitamente, porque reconhecer o erro é mais fácil do que reconstruir a interpretação correta do zero.

! Erro nº 1 — confundir o parâmetro com a estimativa

"Temos 95% de confiança de que ***a estimativa*** *esteja no intervalo [174,74; 185,26]."*

Errado, e de um modo particularmente instrutivo: a afirmação não é apenas imprecisa, é **vazia**. A estimativa $\bar{x} = 180$ está no intervalo com certeza absoluta, sempre, **por construção** — ela é o *centro* do intervalo, que foi literalmente construído somando e subtraindo ε dela. Não há 95% de confiança nenhuma: há 100%, trivialmente, e a afirmação não contém informação. O objeto sobre o qual há incerteza é o **parâmetro** μ, nunca a estimativa. Quem confunde os dois perdeu de vista a distinção fundamental de Seção 8.

! Erro nº 2 — tratar a confiança como probabilidade

"A probabilidade de que o parâmetro pertença ao intervalo [174,74; 185,26] é 0,95."

Errado, pela razão desenvolvida acima: para que essa frase fizesse sentido, μ teria de ser uma variável aleatória com distribuição sobre a reta — e ele não é. Ele é uma constante desconhecida. A afirmação atribui probabilidade a um evento que já está determinado.
Este é o erro sutil, e é o mais comum. A formulação correta desloca a probabilidade do parâmetro para o **procedimento**: *"o procedimento que gerou este intervalo acerta em 95% das amostras"*.

i Ferramenta matemática — e a interpretação bayesiana?

Um leitor atento notará que o Erro nº 2 é *exatamente* o que um estatístico bayesiano diz — e diz legitimamente. A diferença é de paradigma: na estatística bayesiana, θ **é** tratado como variável aleatória, com uma distribuição *a priori* que os dados atualizam via teorema de Bayes (Seção 3) numa distribuição *a posteriori*. O intervalo construído a partir dela chama-se **intervalo de credibilidade**, e para ele a frase "a probabilidade de que θ esteja no intervalo é 0,95" é a leitura correta.
Este curso é inteiramente **frequentista**: θ é constante e a probabilidade se refere à variabilidade amostral. Nesse arcabouço, o Erro nº 2 é erro. O que se deve reter é que a proibição não é uma esquisitice arbitrária — é uma consequência lógica de tratar o parâmetro como fixo. *(Escolas de inferência e o paradigma bayesiano estão fora do programa; a nota existe apenas para que o leitor não se surpreenda ao encontrar a outra linguagem na literatura.)*

9.4 σ desconhecido: a distribuição t de Student

Volte ao caso realista. Não conhecemos σ. A saída óbvia é estimá-lo pelo desvio padrão amostral S de Seção 8, com divisor $n-1$, e substituir na pivotal:

$$T = \frac{\bar{X} - \mu}{S/\sqrt{n}}.$$

A questão é: essa quantidade ainda é Normal padrão? **Não.** E a razão é clara: no denominador de Z havia uma constante, σ; no de T há uma **variável aleatória**, S, que flutua de amostra para amostra. Introduzimos uma segunda fonte de aleatoriedade, e a distribuição resultante é mais **dispersa** que a Normal — pode acontecer de $\bar{x}$ estar longe de μ *e* de s ser pequeno por acaso, produzindo valores de T muito grandes com frequência maior do que a Normal admitiria.

9.4.1 Gosset, a Guinness e o pseudônimo

A distribuição correta foi descoberta em 1908 por **William Sealy Gosset**, químico e matemático que trabalhava na cervejaria **Guinness**, em Dublin. Gosset lidava com um problema industrial concreto: controlar a qualidade do lúpulo e da cevada a partir de **amostras pequenas** — testar cada lote inteiro seria destrutivo e proibitivamente caro. As aproximações normais disponíveis funcionavam mal com $n = 4$ ou $n = 5$, e ele derivou a distribuição exata da razão acima.

A Guinness proibia seus funcionários de publicar, temendo que revelassem segredos industriais — política adotada depois que um pesquisador anterior divulgara informação sensível. Gosset negociou publicar sob **pseudônimo**, e escolheu "**Student**". Daí o nome pelo qual a distribuição é conhecida até hoje: *t de Student*. É provavelmente o único caso na história da ciência em que uma distribuição de probabilidade leva o nome falso de seu descobridor por razões de política corporativa cervejeira.

O episódio tem uma moral estatística, além da anedótica: a t nasceu de um problema de **amostras pequenas**, e é exatamente aí que ela importa. Com n grande, como veremos, ela se confunde com a Normal.

9.4.2 A distribuição t

! Definição — distribuição t de Student

Se $X_1, \ldots, X_n$ é uma AAS de uma população $N(\mu, \sigma^2)$, então

$$\boxed{T = \frac{\bar{X} - \mu}{S/\sqrt{n}} \sim t_{n-1}}$$

isto é, T tem distribuição t de Student com $\nu = n - 1$ **graus de liberdade**.
A t tem um único parâmetro, ν, que controla a forma. Suas propriedades:

- é **simétrica em torno de zero**, como a Normal padrão;
- tem forma de sino, como a Normal padrão;
- tem **caudas mais pesadas** que a Normal — mais massa nos extremos, menos no centro;
- $\mathbb{E}(T) = 0$ para $\nu > 1$ e $\mathrm{V}(T) = \dfrac{\nu}{\nu - 2} > 1$ para $\nu > 2$;
- **converge para a Normal padrão** quando $\nu \to \infty$.

O parâmetro $\nu = n - 1$ é o mesmo "grau de liberdade" que apareceu no divisor de S^2 em Seção 8, e pela mesma razão: dos n desvios $(X_i - \bar{X})$, apenas $n - 1$ são livres, porque sua soma é forçosamente

nula. A informação disponível para estimar a dispersão é $n-1$, não n, e a distribuição registra isso.

A consequência prática das caudas pesadas é que **os quantis da t são maiores que os da Normal**: $t_{4;0,025} = 2{,}776$ contra $z_{0,025} = 1{,}96$. Ao não conhecer σ, pagamos um preço em amplitude, e o preço é tanto maior quanto menor a amostra. Faz sentido: sabemos menos, devemos afirmar menos.

9.4.3 A relação entre t e Normal

$$\boxed{t_\nu \xrightarrow[\nu\to\infty]{} N(0,1)}$$

A intuição é imediata a partir da origem do problema. Quando n cresce, S estima σ cada vez melhor — $S \to \sigma$, pela lei dos grandes números —, e a segunda fonte de aleatoriedade que criou a diferença simplesmente desaparece. No limite, S é praticamente uma constante, e T volta a ser Z.

Um detalhe operacional confirma isso, e vale conhecer porque aparece em prova: a **última linha da tabela** t, rotulada $\nu = \infty$, contém exatamente os valores da Normal padrão — 1,645; 1,96; 2,576. As tabelas são construídas assim de propósito, e a coincidência é um bom teste de que se está lendo a tabela certa.

A convergência é rápida na prática. A tabela abaixo, gerada no laboratório, mostra $t_{\nu;0,025}$:

ν	1	2	5	10	20	30	60	120	∞
$t_{\nu;0,025}$	12,706	4,303	2,571	2,228	2,086	2,042	2,000	1,980	1,960

Com $\nu = 1$ o quantil é seis vezes o da Normal — a t_1 (a Cauchy) é tão pesada nas caudas que **nem sequer tem média**. Com $\nu = 30$ o erro de usar 1,96 no lugar de 2,042 é de 4%; com $\nu = 120$, de 1%. É essa observação que sustenta a regra prática do professor.

9.4.4 O intervalo para μ com σ desconhecido

A derivação é *idêntica* à do caso σ conhecido — os mesmos cinco passos algébricos —, apenas com Z trocado por T, $z_{\alpha/2}$ por $t_{n-1;\alpha/2}$ e σ por s:

$$\boxed{\mathrm{IC}_{100(1-\alpha)\%}(\mu) = \left[\bar{x} \mp t_{n-1;\alpha/2}\frac{s}{\sqrt{n}}\right]}$$

> **!** A regra do professor — quando Normal, quando t
>
> **Se σ é CONHECIDO** $\Longrightarrow$ usa-se a **tabela Normal**, para **qualquer** n.
> **Se σ é DESCONHECIDO** $\Longrightarrow$ usa-se a **tabela** t com $\nu = n-1$ graus de liberdade para $n \leq 30$; **pode-se usar a Normal para** $n > 30$, pela convergência $t_\nu \to N(0,1)$.

Duas observações sobre a regra. A primeira: o corte em $n = 30$ é uma **convenção prática**, não um teorema — não acontece nada de especial em 30. Ele reflete o julgamento de que, a partir daí, a

diferença entre os quantis (2,042 contra 1,96) é pequena o bastante para ser ignorada em cálculos à mão. A segunda: no computador, **nunca há razão para usar a aproximação**. A função `qt()` custa o mesmo que `qnorm()`, e a t é exata. A regra existe porque tabelas impressas são finitas; softwares não são.

Registre-se também que, quando σ é desconhecido **e a população não é Normal**, a justificativa passa a ser o Teorema Central do Limite de Seção 7: com n grande, $\bar{X}$ é aproximadamente Normal qualquer que seja a população, e o intervalo com z vale aproximadamente. Com n pequeno e população não-Normal, nenhuma das duas fórmulas se aplica — é o caso em que se recorre a métodos não paramétricos ou a *bootstrap*, fora do programa.

9.4.5 Exemplos

Exemplo 1 (alturas, agora sem conhecer σ). Voltemos à amostra concreta de Seção 8:

$$174 \qquad 186 \qquad 186 \qquad 180 \qquad 174$$

Já sabemos que $\bar{x} = 180$. A variância amostral, com divisor $n - 1 = 4$:

$$s^2 = \frac{(174-180)^2 + (186-180)^2 + (186-180)^2 + (180-180)^2 + (174-180)^2}{4} = \frac{36+36+36+0+36}{4} = \frac{144}{4} =$$

de modo que $s = 6$. Curiosamente, é o **mesmo valor** que supusemos conhecido no primeiro exemplo — o que torna a comparação perfeitamente limpa, porque a única coisa que muda é o quantil. Com $\nu = n - 1 = 4$ e $\alpha/2 = 0{,}025$, a tabela t dá $t_{4;0{,}025} = 2{,}776$:

$$\mathrm{IC}_{95\%}(\mu) = \left[180 \mp 2{,}776 \cdot \frac{6}{\sqrt{5}}\right] = [180 \mp 2{,}776 \cdot 2{,}6833] = [180 \mp 7{,}4489] = [172{,}55 \ ; \ 187{,}45]\,.$$

Compare com o intervalo do caso σ conhecido, [174,74; 185,26]. O novo intervalo é **substancialmente mais largo**: amplitude 14,90 contra 10,52, um aumento de 42%. E o aumento é *exatamente* a razão dos quantis:

$$\frac{t_{4;0{,}025}}{z_{0{,}025}} = \frac{2{,}776}{1{,}96} = 1{,}4163,$$

porque tudo o mais na fórmula é idêntico ($\bar{x} = 180$, $s = \sigma = 6$, $n = 5$). O preço de não conhecer σ, com uma amostra de cinco observações, é 42% de amplitude a mais. Com $n = 31$ o mesmo preço seria de 4%; com $n = 121$, de 1%. É a penalidade por ter de estimar a dispersão com poucos dados, e ela é dramática justamente onde Gosset a encontrou: em amostras pequenas.

Exemplo 2 (renda de trabalhadores). Uma amostra de $n = 25$ trabalhadores de certo setor tem renda média $\bar{x} = 400$ e desvio padrão amostral $s = 450$. Um intervalo de **90%** para a renda média do setor. Com $\nu = 24$ e $\alpha/2 = 0{,}05$, a tabela dá $t_{24;0{,}05} = 1{,}711$:

$$\mathrm{IC}_{90\%}(\mu) = \left[400 \mp 1{,}711 \cdot \frac{450}{\sqrt{25}}\right] = [400 \mp 1{,}711 \cdot 90] = [400 \mp 153{,}99] = [246{,}01 \ ; \ 553{,}99]\,.$$

Um intervalo enorme, e por um motivo visível nos dados: o desvio padrão (450) é **maior que a média** (400), um coeficiente de variação de 112% no vocabulário de Seção 2. Rendas são notoriamente dispersas e assimétricas, e nenhuma quantidade de sofisticação estatística compensa uma amostra de 25 observações de uma população tão heterogênea. Vale como advertência prática: a amplitude do intervalo é um diagnóstico honesto: quando ela é grande, os dados simplesmente não sustentam conclusões precisas, e o intervalo tem a virtude de dizer isso em voz alta — coisa que a estimativa pontual "$\bar{x} = 400$" jamais diria.

Exemplo 3 (vida útil de televisores). Testa-se a durabilidade de $n = 16$ aparelhos, obtendo-se $\bar{x} = 8.900$ horas e $s = 500$ horas. Um intervalo de **99%**. Com $\nu = 15$ e $\alpha/2 = 0{,}005$, temos $t_{15;0{,}005} = 2{,}947$:

$$\mathrm{IC}_{99\%}(\mu) = \left[8900 \mp 2{,}947 \cdot \frac{500}{\sqrt{16}}\right] = [8900 \mp 2{,}947 \cdot 125] = [8900 \mp 368{,}4] = [8.531{,}6 \ ; \ 9.268{,}4]\,.$$

Com 99% de confiança, a vida útil média dos televisores dessa linha está entre 8.532 e 9.268 horas. Este é, aliás, o exemplo canônico de por que se amostra: o teste é **destrutivo** — é preciso queimar o aparelho para medir sua vida útil —, e o censo destruiria o estoque (Seção 8).

Exemplo 4 (endividamento, e o efeito da confiança). Uma amostra de $n = 9$ famílias registra dívida média $\bar{x} = 45$ (em unidades monetárias) com $s = 30$. Vejamos dois níveis de confiança, para isolar o efeito do quantil. Aqui $\nu = 8$ e $s/\sqrt{n} = 30/3 = 10$.

Para 95%, $t_{8;0{,}025} = 2{,}306$:

$$\mathrm{IC}_{95\%}(\mu) = [45 \mp 2{,}306 \cdot 10] = [45 \mp 23{,}06] = [21{,}9 \ ; \ 68{,}1]\,.$$

Para 90%, $t_{8;0{,}05} = 1{,}86$:

$$\mathrm{IC}_{90\%}(\mu) = [45 \mp 1{,}86 \cdot 10] = [45 \mp 18{,}6] = [26{,}4 \ ; \ 63{,}6]\,.$$

O intervalo de 90% é mais estreito — amplitude 37,2 contra 46,1 —, exatamente como o trade-off prevê. Se quisermos afirmar menos, podemos afirmar com mais precisão.

9.5 Intervalo de confiança para a proporção

Muitas perguntas econômicas são sobre **proporções**: a taxa de inadimplência de uma carteira, o percentual de desempregados, a fração de peças defeituosas em um lote, a intenção de voto. O parâmetro é p, e o estimador de Seção 8 é a proporção amostral

$$\hat{p} = \frac{\text{número de sucessos}}{n} = \bar{X},$$

que é a média amostral de variáveis Bernoulli(p) e, por isso, não viesado.

A construção do intervalo é a mesma, com um ingrediente do Teorema Central do Limite. Como $\hat{p}$ é uma média de i.i.d., para n grande vale a aproximação Normal

$$\hat{p} \;\dot{\sim}\; N\left(p,\ \frac{p(1-p)}{n}\right),$$

pois $\mathbb{E}(\hat{p}) = p$ e $\mathrm{V}(\hat{p}) = p(1-p)/n$ — a variância da Bernoulli dividida por n (Seção 5, Seção 7). A pivotal é

$$Z = \frac{\hat{p} - p}{\sqrt{p(1-p)/n}} \;\dot{\sim}\; N(0,1),$$

e o isolamento de p seguiria os mesmos cinco passos — exceto por uma dificuldade: p aparece também no **denominador**, de modo que a álgebra não o isola limpa-mente. A saída padrão é substituir p por $\hat{p}$ no erro-padrão, o que é legítimo assintoticamente porque $\hat{p} \to p$. Obtém-se:

$$\boxed{\mathrm{IC}_{100(1-\alpha)\%}(p) = \left[\hat{p} \mp z_{\alpha/2}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}\right]}$$

Três observações. A primeira: usa-se **sempre a Normal**, nunca a t — o resultado é assintótico desde o início, e não há analogia exata com a t para proporções. A segunda: a validade exige **amostra grande**, e a regra prática usual é $n\hat{p} \geq 5$ e $n(1-\hat{p}) \geq 5$ (algumas referências pedem 10). A terceira: o erro-padrão $\sqrt{\hat{p}(1-\hat{p})/n}$ é máximo em $\hat{p} = 0{,}5$ — proporções próximas de meio a meio são as mais difíceis de estimar, e é por isso que pesquisas eleitorais em disputas apertadas precisam de amostras maiores.

Exemplo (peças defeituosas). Em um lote, inspecionam-se $n = 70$ peças e encontram-se 49 defeituosas. Então

$$\hat{p} = \frac{49}{70} = 0{,}7, \qquad \sqrt{\frac{\hat{p}(1-\hat{p})}{n}} = \sqrt{\frac{0{,}7 \times 0{,}3}{70}} = \sqrt{0{,}003} = 0{,}0548.$$

Verificando as condições: $n\hat{p} = 49 \geq 5$ e $n(1-\hat{p}) = 21 \geq 5$. Com 95% de confiança:

$$\mathrm{IC}_{95\%}(p) = [0{,}7 \mp 1{,}96 \times 0{,}0548] = [0{,}7 \mp 0{,}1074] = [0{,}5927\ ;\ 0{,}8073]\,.$$

Entre 59,3% e 80,7% das peças do lote são defeituosas — uma faixa larga, com apenas 70 inspeções, mas larga o bastante em torno de valores todos altos para que a conclusão prática seja inequívoca: o lote deve ser rejeitado.

9.6 Intervalo de confiança para a variância

Até aqui, a dispersão era um estorvo — algo que entrava no erro-padrão e atrapalhava a estimação da média. Mas em economia e finanças a dispersão é frequentemente **o objeto de interesse**: volatilidade de um ativo, variabilidade de um processo produtivo, risco de uma carteira (Seção 7). Precisamos de intervalos para σ^2.

A pivotal, aqui, não é Normal nem t.

9.6.1 A distribuição qui-quadrado

! Definição — distribuição qui-quadrado

Se $Z_1, \dots, Z_\nu$ são $N(0,1)$ independentes, então

$$\boxed{X = \sum_{i=1}^{\nu} Z_i^2 \;\sim\; \chi^2_\nu}$$

tem distribuição **qui-quadrado com ν graus de liberdade**. Sua densidade é

$$f(x) = \frac{1}{2^{\nu/2}\,\Gamma(\nu/2)}\, x^{\nu/2-1} e^{-x/2}, \qquad x > 0,$$

e seus momentos são

$$\boxed{\mathbb{E}(X) = \nu \qquad \text{e} \qquad \mathrm{V}(X) = 2\nu}$$

Três propriedades importam para o que segue. A qui-quadrado é definida apenas em $x > 0$ — é soma de quadrados, não pode ser negativa —, o que já a impede de ser simétrica. Ela é **assimétrica à direita**, com cauda longa nos valores altos, e a assimetria diminui à medida que ν cresce (para ν grande ela se aproxima de uma Normal, o que é o TCL aplicado à soma dos Z_i^2). E sua esperança é exatamente ν, o que dá uma referência mental útil: uma χ^2_{29} tem valores tipicamente em torno de 29.

O resultado que conecta a qui-quadrado à variância amostral é o seguinte, e vale registrá-lo:

$$\boxed{\frac{(n-1)S^2}{\sigma^2} \;\sim\; \chi^2_{n-1}}$$

quando a população é $N(\mu, \sigma^2)$. A justificativa é a mesma dos graus de liberdade que já vimos duas vezes: $\sum (X_i - \bar{X})^2/\sigma^2$ seria uma soma de n quadrados de Normais padrão se centrássemos em μ; ao centrar em $\bar{X}$, perdemos um grau de liberdade e sobram $n-1$. Note, incidentalmente, que $\mathbb{E}[(n-1)S^2/\sigma^2] = n-1$ recupera $\mathbb{E}(S^2) = \sigma^2$, o não-viés de Seção 8.

9.6.2 O intervalo

A construção é a mesma receita, com a diferença de que a distribuição é **assimétrica**: as duas caudas exigem quantis diferentes, não simétricos. Partimos de

$$P\left(\chi^2_{n-1;1-\alpha/2} \leq \frac{(n-1)S^2}{\sigma^2} \leq \chi^2_{n-1;\alpha/2}\right) = 1-\alpha,$$

e isolamos σ^2 invertendo a desigualdade — o que **troca a ordem dos extremos**, porque $x \mapsto 1/x$ é decrescente em $x > 0$:

$$\boxed{\mathrm{IC}_{100(1-\alpha)\%}(\sigma^2) = \left[\frac{(n-1)s^2}{\chi^2_{n-1;\alpha/2}} \; ; \; \frac{(n-1)s^2}{\chi^2_{n-1;1-\alpha/2}}\right]}$$

! Atenção à convenção dos quantis — erro clássico

Na notação do curso, $\chi^2_{\nu;\gamma}$ é o valor tal que

$$P(X > \chi^2_{\nu;\gamma}) = \gamma,$$

isto é, o quantil é indexado pela área **à direita**. Consequentemente:

- $\chi^2_{n-1;\alpha/2}$ é o valor **MAIOR** (área $\alpha/2$ à direita, cauda superior) — e vai no denominador do **limite INFERIOR**;
- $\chi^2_{n-1;1-\alpha/2}$ é o valor **MENOR** (área $1-\alpha/2$ à direita, cauda inferior) — e vai no denominador do **limite SUPERIOR**.

Dividir por um número maior dá um resultado menor: por isso o quantil grande produz o limite pequeno. Se o intervalo obtido tiver o limite inferior maior que o superior, o erro foi este. **Cuidado com o R**, que usa a convenção **oposta**: `qchisq(q, df)` devolve o quantil com área q **à esquerda**. A tradução é

$$\chi^2_{\nu;\alpha/2} = \texttt{qchisq(1 - alpha/2, df)} \qquad \text{e} \qquad \chi^2_{\nu;1-\alpha/2} = \texttt{qchisq(alpha/2, df)}.$$

Uma consequência estrutural, e é o que mais estranha quem vê pela primeira vez: **o intervalo para σ^2 não é centrado em s^2**. Como a qui-quadrado é assimétrica, os dois quantis não são "espelhados", e a estimativa pontual fica **descentrada** — mais próxima do limite inferior do que do superior. Não é erro de cálculo; é a assimetria da distribuição se manifestando, e o painel inferior da Figura 17 torna isso visível.

Um intervalo para o **desvio padrão** se obtém tomando raízes quadradas dos dois limites, o que é legítimo porque $x \mapsto \sqrt{x}$ é crescente e preserva a ordem.

Exemplo (variabilidade de um processo). Uma amostra de $n = 30$ observações de uma população Normal produziu $s^2 = 132{,}7$. Com $\nu = 29$ e 95% de confiança, a tabela qui-quadrado dá

$$\chi^2_{29;0,975} = 16{,}047 \qquad \text{(o menor)}, \qquad \chi^2_{29;0,025} = 45{,}72 \qquad \text{(o maior)}.$$

Note que ambos orbitam $\mathbb{E}(\chi^2_{29}) = 29$, como esperado. Com $(n-1)s^2 = 29 \times 132{,}7 = 3.848{,}3$:

$$\text{IC}_{95\%}(\sigma^2) = \left[\frac{3848{,}3}{45{,}72} \;;\; \frac{3848{,}3}{16{,}047}\right] = [84{,}17 \;;\; 239{,}81]\,.$$

Observe a assimetria: a estimativa $s^2 = 132{,}7$ dista 48,5 do limite inferior e **107,1** do superior, mais que o dobro. O ponto médio do intervalo é 162,0, e não 132,7. Observe também a **amplitude**: o limite superior é quase três vezes o inferior. Variâncias são notoriamente difíceis de estimar com precisão — muito mais que médias — e trinta observações não bastam para fixar σ^2 dentro de uma faixa estreita. Tomando raízes, o intervalo para σ é [9,17; 15,49], bem mais palatável: a raiz comprime a escala.

9.7 Intervalo para a diferença de médias

A pergunta que mais aparece na prática não é sobre um parâmetro isolado, e sim sobre uma **comparação**: o salário médio difere entre dois setores? A durabilidade da marca A supera a da marca B? O tratamento teve efeito? O parâmetro de interesse passa a ser $\mu_1 - \mu_2$.

Sejam duas amostras independentes, de tamanhos n_1 e n_2, de populações $N(\mu_1, \sigma^2)$ e $N(\mu_2, \sigma^2)$ — note que supomos a **mesma** variância nas duas. O estimador natural de $\mu_1 - \mu_2$ é $\bar{X}_1 - \bar{X}_2$, e de Seção 7 sabemos que, para amostras independentes,

$$\mathbb{E}(\bar{X}_1 - \bar{X}_2) = \mu_1 - \mu_2, \qquad \text{V}(\bar{X}_1 - \bar{X}_2) = \frac{\sigma^2}{n_1} + \frac{\sigma^2}{n_2} = \sigma^2\left(\frac{1}{n_1} + \frac{1}{n_2}\right).$$

A variância da **diferença** é a **soma** das variâncias — é a propriedade de Seção 7 que costuma surpreender, e vale relembrar por quê: $\text{V}(-X) = (-1)^2\text{V}(X) = \text{V}(X)$, de modo que subtrair uma variável aleatória adiciona sua variabilidade, não a cancela.

Como σ^2 é desconhecido, precisamos estimá-lo. E aqui entra a razão de ser da hipótese de variâncias iguais: se as duas populações compartilham o **mesmo** σ^2, podemos usar as duas amostras conjuntamente para estimá-lo, obtendo mais graus de liberdade e mais precisão. O estimador é a **variância combinada** (ou *pooled*):

$$\boxed{s_p = \sqrt{\frac{(n_1-1)s_1^2 + (n_2-1)s_2^2}{n_1 + n_2 - 2}}}$$

que é uma **média ponderada** das duas variâncias amostrais, com pesos proporcionais aos respectivos graus de liberdade — a amostra maior pesa mais, como deve ser. Perdemos dois graus de liberdade (um por média estimada), e o total disponível é $\nu = n_1 + n_2 - 2$. O intervalo:

$$\boxed{\text{IC}_{100(1-\alpha)\%}(\mu_1 - \mu_2) = \left[(\bar{x}_1 - \bar{x}_2) \mp t_{n_1+n_2-2;\alpha/2}\; s_p\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}\right]}$$

! As três condições de validade

O intervalo acima requer:

1. **as duas populações são Normais**;
2. **as amostras são independentes** — uma não influencia a outra (o que exclui dados pareados: mesmos indivíduos medidos antes e depois exigem outro tratamento);
3. **as variâncias populacionais são IGUAIS**, $\sigma_1^2 = \sigma_2^2$ (homocedasticidade).

Violada (1), com n grande o TCL socorre parcialmente. Violada (2), a fórmula da variância da diferença muda e é preciso incluir a covariância. Violada (3), usa-se a aproximação de Welch — que estima as variâncias separadamente e ajusta os graus de liberdade por uma fórmula não inteira; é o padrão do `t.test()` do R, mas está fora do programa do curso.

i Nota do professor — variâncias amostrais diferentes não invalidam nada

É comum o aluno olhar para $s_1^2 = 62.500$ e $s_2^2 = 40.000$ e concluir que a hipótese de variâncias iguais foi violada. **Não foi.** A hipótese é sobre as variâncias **POPULACIONAIS** σ_1^2 e σ_2^2, que são desconhecidas. As variâncias **AMOSTRAIS** s_1^2 e s_2^2 são estimativas, e estimativas flutuam — seria espantoso que duas amostras finitas de populações com a mesma variância produzissem exatamente o mesmo s^2.
Aliás, é justamente por serem estimativas da *mesma* quantidade que faz sentido combiná-las em s_p: se supuséssemos $\sigma_1^2 \neq \sigma_2^2$, combiná-las seria absurdo.
A pergunta legítima — *as variâncias amostrais diferem o bastante para duvidar da hipótese?* — tem resposta formal, e é o intervalo para a **razão de variâncias** que veremos na próxima seção: se o intervalo para σ_2^2/σ_1^2 contém 1, a hipótese de igualdade é compatível com os dados.

Exemplo 1 (lâmpadas de duas marcas). Testam-se lâmpadas de duas marcas quanto à duração, em horas:

Marca	n_i	$\bar{x}_i$	s_i	s_i^2
A	8	4.600	250	62.500
B	10	4.000	200	40.000

Queremos um intervalo de **99%** para $\mu_A - \mu_B$. A variância combinada:

$$s_p = \sqrt{\frac{(8-1)(62.500) + (10-1)(40.000)}{8+10-2}} = \sqrt{\frac{437.500 + 360.000}{16}} = \sqrt{\frac{797.500}{16}} = \sqrt{49.843{,}75} = 223{,}26.$$

Repare que 223,26 cai **entre** $s_B = 200$ e $s_A = 250$, como toda média ponderada deve cair. Com $\nu = 16$ e $\alpha/2 = 0{,}005$, temos $t_{16;0{,}005} = 2{,}921$, e

$$s_p\sqrt{\frac{1}{8} + \frac{1}{10}} = 223{,}26\sqrt{0{,}225} = 223{,}26 \times 0{,}4743 = 105{,}89.$$

Portanto

$$\text{IC}_{99\%}(\mu_A - \mu_B) = [600 \mp 2{,}921 \times 105{,}89] = [600 \mp 309{,}33] = [290{,}67\ ;\ 909{,}33]\,.$$

A leitura é o que torna este exemplo instrutivo: **o intervalo inteiro é positivo**. Não contém zero. Com 99% de confiança, as lâmpadas da marca A duram entre 291 e 909 horas a mais que as da marca B. A diferença é, na linguagem que Seção 10 formalizará, **estatisticamente significante** — e o intervalo entrega essa conclusão de graça, junto com uma faixa para a **magnitude** do efeito, que o teste de hipóteses sozinho não daria.

Exemplo 2 (duas turmas). Comparam-se as notas de duas turmas:

$$\text{Turma 1: } 5{,}0 \quad 6{,}0 \quad 3{,}5 \quad 9{,}0 \quad 8{,}5 \qquad\qquad \text{Turma 2: } 4{,}0 \quad 10{,}0 \quad 7{,}0 \quad 5{,}0$$

Calculando as estatísticas descritivas: $n_1 = 5$, $\bar{x}_1 = 32/5 = 6{,}4$, $s_1^2 = 5{,}425$; $n_2 = 4$, $\bar{x}_2 = 26/4 = 6{,}5$, $s_2^2 = 7$. A variância combinada, com $\nu = 5 + 4 - 2 = 7$:

$$s_p = \sqrt{\frac{4(5{,}425) + 3(7)}{7}} = \sqrt{\frac{21{,}7 + 21}{7}} = \sqrt{\frac{42{,}7}{7}} = \sqrt{6{,}1} = 2{,}4698.$$

Com $t_{7;0{,}025} = 2{,}365$ e $\sqrt{1/5 + 1/4} = \sqrt{0{,}45} = 0{,}6708$:

$$\text{IC}_{95\%}(\mu_1 - \mu_2) = [(6{,}4 - 6{,}5) \mp 2{,}365 \times 2{,}4698 \times 0{,}6708] = [-0{,}1 \mp 3{,}9182] = [-4{,}0182\ ;\ 3{,}8182]\,.$$

Agora **o intervalo contém zero** — e a conclusão é oposta à do exemplo anterior. Os dados são compatíveis com $\mu_1 = \mu_2$: não há evidência de que as turmas difiram. Note que a estimativa pontual da diferença é $-0{,}1$, ínfima em comparação com a margem de erro de 3,92, que é quarenta vezes maior. Com quatro e cinco observações, praticamente nada poderia ser detectado.

! A regra de leitura de um IC para diferença

Para intervalos de $\mu_1 - \mu_2$ (ou de $p_1 - p_2$), o valor de referência é **zero**:

- se o intervalo **não contém 0**, há evidência de diferença entre as populações, e o **sinal** do intervalo indica a direção (todo positivo: $\mu_1 > \mu_2$);
- se o intervalo **contém 0**, os dados são compatíveis com a ausência de diferença.

Cuidado com a segunda leitura: "compatível com a ausência de diferença" **não** é "prova de que não há diferença". Um intervalo largo que contém zero pode ser apenas o sintoma de uma amostra pequena demais para detectar um efeito que existe. Ausência de evidência não é evidência de ausência — distinção que Seção 10 retomará sob o nome de *erro tipo II*.

9.8 Intervalo para a diferença de proporções

O caso análogo para proporções segue o mesmo molde, com a Normal no lugar da t (porque a justificativa é assintótica) e sem *pooling* (porque nada nos leva a supor $p_1 = p_2$ — é justamente o que se quer investigar). Com

$$\mathrm{V}(\hat{p}_1 - \hat{p}_2) = \frac{p_1(1-p_1)}{n_1} + \frac{p_2(1-p_2)}{n_2}$$

pela independência, e substituindo os p_i pelos $\hat{p}_i$:

$$\boxed{\mathrm{IC}_{100(1-\alpha)\%}(p_1 - p_2) = \left[(\hat{p}_1 - \hat{p}_2) \mp z_{\alpha/2}\sqrt{\frac{\hat{p}_1(1-\hat{p}_1)}{n_1} + \frac{\hat{p}_2(1-\hat{p}_2)}{n_2}}\right]}$$

Exemplo (inadimplência em duas financeiras). A financeira 1 tem 140 inadimplentes entre 180 clientes; a financeira 2, 220 entre 300. Um intervalo de **90%** para $p_1 - p_2$.

$$\hat{p}_1 = \frac{140}{180} = 0{,}7778, \qquad \hat{p}_2 = \frac{220}{300} = 0{,}7333, \qquad \hat{p}_1 - \hat{p}_2 = 0{,}0444.$$

O erro-padrão da diferença:

$$\sqrt{\frac{0{,}7778(0{,}2222)}{180} + \frac{0{,}7333(0{,}2667)}{300}} = \sqrt{0{,}000960 + 0{,}000652} = \sqrt{0{,}001612} = 0{,}0402.$$

Com $z_{0{,}05} = 1{,}645$:

$$\mathrm{IC}_{90\%}(p_1 - p_2) = [0{,}0444 \mp 1{,}645 \times 0{,}0402] = [0{,}0444 \mp 0{,}0661] = [-0{,}0216 \ ; \ 0{,}1105]\,.$$

O intervalo **contém zero**. Embora a financeira 1 apresente inadimplência amostral 4,4 pontos percentuais maior, essa diferença é pequena demais em relação ao ruído amostral para sustentar qualquer conclusão: os dados são compatíveis com taxas iguais. Repare que o intervalo é **assimétrico em relação a zero** — vai de $-0{,}02$ a $+0{,}11$ —, o que reflete o fato de a estimativa pontual ser positiva; ele é, porém, simétrico em torno de 0,0444, como todo intervalo baseado numa distribuição simétrica.

9.9 A distribuição F e o intervalo para a razão de variâncias

Falta uma comparação: as duas populações têm a mesma dispersão? Aqui não faz sentido olhar para a *diferença* $\sigma_1^2 - \sigma_2^2$; variâncias são grandezas positivas e se comparam naturalmente por **razão**. O valor de referência deixa de ser 0 e passa a ser **1**.

9.9.1 A distribuição F de Snedecor

! Definição — distribuição F

Se $U \sim \chi^2_{\nu_1}$ e $V \sim \chi^2_{\nu_2}$ são independentes, então

$$\boxed{F = \frac{U/\nu_1}{V/\nu_2} \;\sim\; F_{\nu_1,\nu_2}}$$

tem distribuição F **com ν_1 graus de liberdade no NUMERADOR e ν_2 no DENOMINADOR**.
A distribuição tem suporte em $x > 0$, é **assimétrica à direita** e — crucialmente — **a ordem dos dois parâmetros importa**: $F_{4,3}$ e $F_{3,4}$ são distribuições diferentes.

A conexão com variâncias é imediata a partir do resultado da seção da qui-quadrado. Se as duas amostras vêm de populações Normais independentes, então $(n_i - 1)S_i^2/\sigma_i^2 \sim \chi^2_{n_i-1}$, e formando a razão as constantes $(n_i - 1)$ se cancelam:

$$\boxed{F = \frac{S_1^2/\sigma_1^2}{S_2^2/\sigma_2^2} \;\sim\; F_{n_1-1,\; n_2-1}}$$

Eis a pivotal: contém as duas variâncias desconhecidas e tem distribuição conhecida.

9.9.2 A inversão: o problema da tabela

Antes de escrever o intervalo, é preciso resolver uma dificuldade **operacional**, que só existe porque trabalhamos com tabelas impressas — e que é a fonte da maior parte dos erros neste tópico.

Como a F é assimétrica, o intervalo exige os **dois** quantis, $f_{\nu_1,\nu_2;\alpha/2}$ (cauda superior) e $f_{\nu_1,\nu_2;1-\alpha/2}$ (cauda inferior). Mas as tabelas de F trazem **apenas a cauda superior** — valores para $\gamma = 0{,}05$, $0{,}025$, $0{,}01$ — porque tabelar as duas caudas para todos os pares (ν_1, ν_2) dobraria um objeto já volumoso. O quantil inferior precisa ser obtido indiretamente, pela identidade:

$$\boxed{f_{\nu_1,\nu_2;\,1-\alpha/2} = \frac{1}{f_{\nu_2,\nu_1;\,\alpha/2}}}$$

! NOTE A TROCA DOS GRAUS DE LIBERDADE

Na identidade acima, os graus de liberdade **trocam de posição**: no lado esquerdo é (ν_1, ν_2); no lado direito, (ν_2, ν_1).
A razão é a definição: se $F \sim F_{\nu_1,\nu_2}$, então $1/F \sim F_{\nu_2,\nu_1}$, porque inverter a razão troca numerador e denominador — e, com eles, os graus de liberdade. Inverter o valor **sem** trocar os graus de liberdade é o erro mais frequente do capítulo.

Exemplo do mecanismo. Queremos $f_{4,3;0,975}$, com $\nu_1 = 4$ e $\nu_2 = 3$. A tabela não traz a cauda de 0,975. Aplicamos a identidade, **trocando** os graus de liberdade: buscamos $f_{3,4;0,025}$, que a tabela dá como 9,98. Então

$$f_{4,3;0,975} = \frac{1}{f_{3,4;0,025}} = \frac{1}{9{,}98} = 0{,}1002.$$

Confira a plausibilidade: 0,1002 é pequeno, como convém a um quantil da cauda **inferior** de uma distribuição positiva. Se o resultado tivesse dado algo maior que 1, teríamos errado.

9.9.3 O intervalo

Isolando σ_2^2/σ_1^2 na afirmação probabilística sobre a pivotal — e lembrando que inverter troca a ordem dos extremos —, chega-se a

$$\boxed{\text{IC}_{100(1-\alpha)\%}\left(\frac{\sigma_2^2}{\sigma_1^2}\right) = \left[\frac{s_2^2}{s_1^2}\, f_{n_1-1,\, n_2-1;\, 1-\alpha/2} \;;\; \frac{s_2^2}{s_1^2}\, f_{n_1-1,\, n_2-1;\, \alpha/2}\right]}$$

Note que os graus de liberdade dos **dois** quantis são $(n_1 - 1,\ n_2 - 1)$ — a ordem da *amostra 1 primeiro* —, mesmo que o parâmetro estimado seja σ_2^2/σ_1^2. É uma inconsistência aparente que confunde, mas decorre diretamente de como a pivotal foi montada; o importante é seguir a fórmula literalmente.

> **!** Os dois erros mais comuns no IC para razão de variâncias
>
> **Erro 1 — trocar numerador e denominador na consulta à tabela.** A tabela F é indexada por duas entradas, e $f_{4,3;0,025} = 15{,}10$ enquanto $f_{3,4;0,025} = 9{,}98$. **Não são o mesmo número.** Sempre confira qual dos dois é o numerador antes de ler a célula.
> **Erro 2 — inverter o valor sem trocar os graus de liberdade.** Calcular $f_{4,3;0,975}$ como $1/f_{4,3;0,025} = 1/15{,}10 = 0{,}0662$ está **errado**; o correto é $1/f_{3,4;0,025} = 1/9{,}98 = 0{,}1002$. A diferença — 0,0662 contra 0,1002, um valor 34% menor que o correto — propaga-se integralmente para o limite inferior do intervalo.
> Um teste de sanidade: o intervalo deve **conter** a razão s_2^2/s_1^2. Se não contiver, um dos dois erros ocorreu.

Exemplo (as duas turmas, de novo). Retomamos $s_1^2 = 5{,}425$ com $n_1 = 5$ e $s_2^2 = 7$ com $n_2 = 4$, e construímos um intervalo de 95% para σ_2^2/σ_1^2. Os graus de liberdade são $\nu_1 = n_1 - 1 = 4$ e $\nu_2 = n_2 - 1 = 3$. A razão amostral:

$$\frac{s_2^2}{s_1^2} = \frac{7}{5{,}425} = 1{,}2903.$$

Os quantis. Da tabela, cauda superior: $f_{4,3;0,025} = 15{,}10$. Para a cauda inferior, trocando os graus de liberdade: $f_{3,4;0,025} = 9{,}98$, logo $f_{4,3;0,975} = 1/9{,}98 = 0{,}1002$. Então

$$\text{IC}_{95\%}\left(\frac{\sigma_2^2}{\sigma_1^2}\right) = [1{,}2903 \times 0{,}1002 \;;\; 1{,}2903 \times 15{,}10] = [0{,}1293 \;;\; 19{,}4839]\,.$$

Duas conclusões. A primeira, substantiva: **o intervalo contém 1**, de modo que a hipótese $\sigma_1^2 = \sigma_2^2$ é perfeitamente compatível com os dados — o que **valida retroativamente** o uso da variância combinada s_p no intervalo para $\mu_1 - \mu_2$ da seção anterior. É assim que se justifica formalmente a terceira condição de validade.

A segunda, sobre precisão: o intervalo vai de 0,13 a 19,48, uma faixa que abrange **duas ordens de grandeza**. Com quatro e cinco observações, não sabemos praticamente nada sobre a razão das variâncias. O intervalo "contém 1", sim, mas contém quase todo o resto também — é um exemplo límpido de por que "não rejeitar" não é "aceitar". A amostra é pequena demais para discriminar qualquer coisa, e o intervalo tem a honestidade de dizê-lo.

9.10 Tabela-resumo de todos os intervalos

A tabela abaixo reúne os seis intervalos do capítulo. É a tabela que se leva para a prova: a coluna das **condições** determina qual linha usar, e a coluna da **distribuição** determina qual tabela consultar e com quantos graus de liberdade.

Parâmetro	Condições	Fórmula do intervalo	Distribuição e g.l.
μ	População Normal (ou n grande, via TCL); σ^2 **conhecido**	$\bar{x} \mp z_{\alpha/2}\dfrac{\sigma}{\sqrt{n}}$	Normal padrão (qualquer n)
μ	População Normal; σ^2 **desconhecido**	$\bar{x} \mp t_{n-1;\alpha/2}\dfrac{s}{\sqrt{n}}$	t de Student $\nu = n-1$ (z se $n > 30$)
p	Amostra grande: $n\hat{p} \geq 5$ e $n(1-\hat{p}) \geq 5$	$\hat{p} \mp z_{\alpha/2}\sqrt{\dfrac{\hat{p}(1-\hat{p})}{n}}$	Normal padrão (sempre)
σ^2	População Normal	$\left[\dfrac{(n-1)s^2}{\chi^2_{n-1;\alpha/2}} \;;\; \dfrac{(n-1)s^2}{\chi^2_{n-1;1-\alpha/2}}\right]$	Qui-quadrado $\nu = n-1$ (assimétrica)
$\mu_1 - \mu_2$	Populações Normais; amostras independentes; $\sigma_1^2 = \sigma_2^2$	$(\bar{x}_1 - \bar{x}_2) \mp t_{n_1+n_2-2;\alpha/2}\, s_p\sqrt{\dfrac{1}{n_1}+\dfrac{1}{n_2}}$ com $s_p = \sqrt{\dfrac{(n_1-1)s_1^2 + (n_2-1)s_2^2}{n_1+n_2-2}}$	t de Student $\nu = n_1 + n_2 - 2$
$p_1 - p_2$	Amostras grandes e independentes	$(\hat{p}_1 - \hat{p}_2) \mp z_{\alpha/2}\,\hat{\sigma}_d$ com $\hat{\sigma}_d = \sqrt{\dfrac{\hat{p}_1(1-\hat{p}_1)}{n_1} + \dfrac{\hat{p}_2(1-\hat{p}_2)}{n_2}}$	Normal padrão (sempre)
σ_2^2/σ_1^2	Populações Normais; amostras independentes	$\left[\dfrac{s_2^2}{s_1^2} f_{\nu_1,\nu_2;1-\alpha/2} \;;\; \dfrac{s_2^2}{s_1^2} f_{\nu_1,\nu_2;\alpha/2}\right]$ com $f_{\nu_1,\nu_2;1-\alpha/2} = 1/f_{\nu_2,\nu_1;\alpha/2}$	F de Snedecor $\nu_1 = n_1 - 1$ (num.) $\nu_2 = n_2 - 1$ (den.)

Três padrões atravessam a tabela e vale destacá-los, porque reduzem sete fórmulas a uma ideia:

1. **Todos** os intervalos têm a forma *estimativa pontual* $\mp$ *quantil* $\times$ *erro-padrão* — exceto os dois baseados em distribuições assimétricas (variância e razão de variâncias), em que a assimetria impede a forma "$\mp$" e obriga a escrever os limites separadamente.
2. A escolha da **distribuição** obedece a uma lógica única: Normal quando a variabilidade do denominador é conhecida ou desprezível; t quando estimamos uma variância; qui-quadrado quando o parâmetro *é* uma variância; F quando é uma razão de variâncias.
3. Os **graus de liberdade** contam sempre a mesma coisa: observações menos parâmetros de posição estimados. Um μ estimado custa 1 ($\nu = n - 1$); dois μ estimados custam 2 ($\nu = n_1 + n_2 - 2$).

9.11 Laboratório computacional em R

9.11.1 A figura definitiva: o que "95% de confiança" significa

Esta é a figura mais importante do capítulo, e ela resolve visualmente o mal-entendido conceitual que nenhuma quantidade de prosa resolve por completo. O experimento é literalmente a definição frequentista: fixamos uma população com μ **conhecido**, sorteamos 100 amostras, construímos os 100 intervalos de 95% e contamos quantos contêm μ.

O ponto crucial é que só podemos fazer esse experimento **por simulação**. Na prática nunca conhecemos μ, e por isso nunca sabemos se o nosso intervalo é dos que acertam ou dos que erram. Aqui conhecemos, e podemos olhar por cima do ombro da natureza.

```
set.seed(20264)
mu_v <- 100; sigma_v <- 15; n_v <- 25; K <- 100

amostras <- matrix(rnorm(K * n_v, mean = mu_v, sd = sigma_v), nrow = K, ncol = n_v)
xbar_k <- rowMeans(amostras)
s_k    <- apply(amostras, 1, sd)
erro_k <- qt(0.975, df = n_v - 1) * s_k / sqrt(n_v)

d_ic <- data.frame(rep = 1:K, xbar = xbar_k,
                   li = xbar_k - erro_k, ls = xbar_k + erro_k)
d_ic$contem <- d_ic$li <= mu_v & mu_v <= d_ic$ls

c(replicas = K, contem_mu = sum(d_ic$contem), falham = sum(!d_ic$contem),
  proporcao = mean(d_ic$contem))
```

```
 replicas contem_mu    falham proporcao
   100.00     95.00      5.00      0.95
```

Exatamente **cinco** dos cem intervalos não contêm $\mu = 100$ — precisamente os 5% que a teoria prevê. Vejamos quais são.

```r
ggplot(d_ic, aes(x = rep, y = xbar, colour = contem)) +
  geom_hline(yintercept = mu_v, colour = cz, linewidth = 0.7) +
  geom_linerange(aes(ymin = li, ymax = ls), linewidth = 0.5) +
  geom_point(size = 0.8) +
  scale_colour_manual(values = c(`TRUE` = az, `FALSE` = vm),
                      labels = c(`TRUE` = "contem mu", `FALSE` = "nao contem mu")) +
  coord_flip() +
  labs(x = "replica (amostra)", y = "IC de 95% para mu   (verdadeiro: mu = 100)")
```

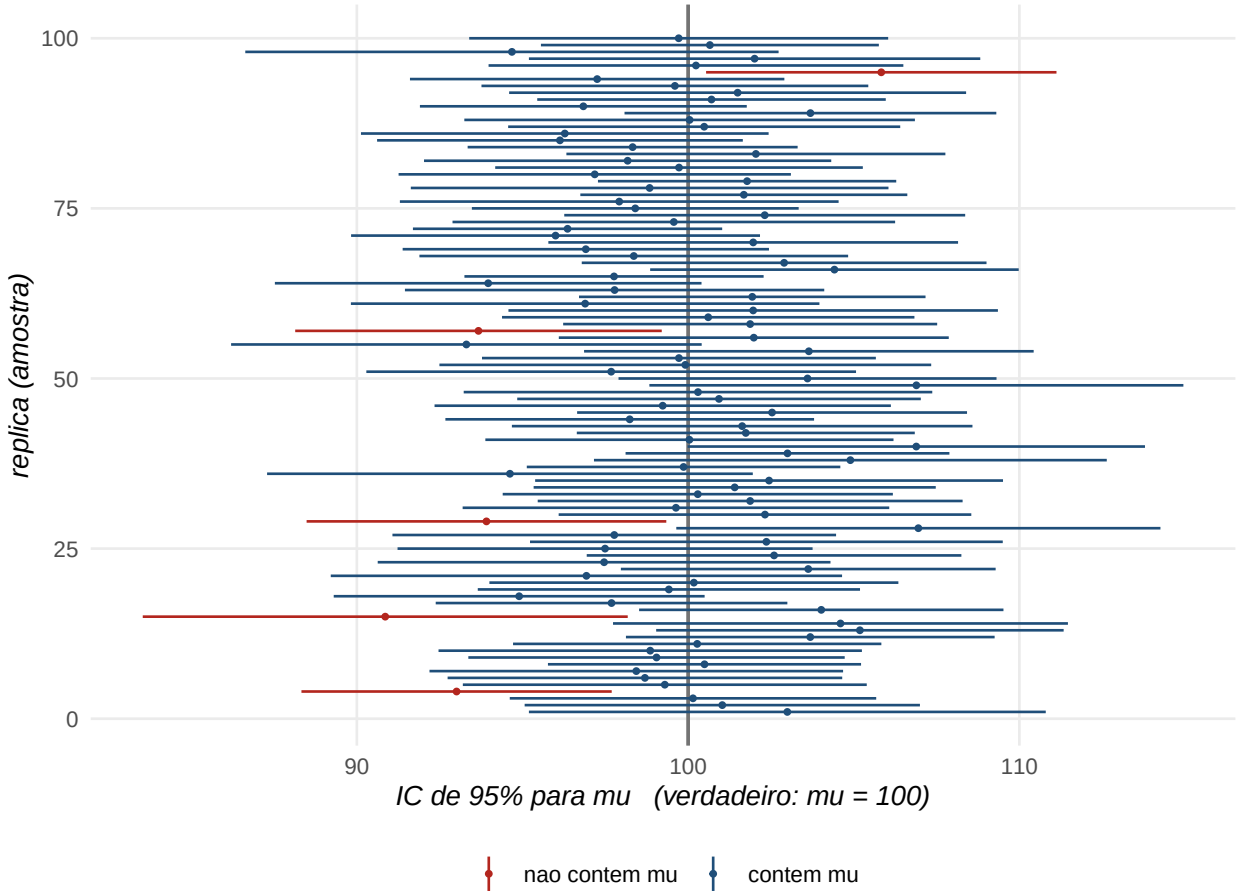


Figura 15: Cem intervalos de confiança de 95% para μ, construídos a partir de cem amostras independentes de tamanho $n = 25$ de uma população $N(100, 15^2)$. A linha vertical marca o verdadeiro $\mu = 100$. Cinco intervalos (vermelhos) não o contêm — exatamente os 5% esperados. O parâmetro está fixo; o que varia de réplica para réplica é o **intervalo**. É essa a interpretação de 'confiança de 95%': não uma probabilidade sobre μ, mas a taxa de acerto do procedimento em amostras repetidas.

Olhe a figura com atenção, porque cada elemento dela corresponde a um item da discussão conceitual.

A **linha vertical não se move**: $\mu = 100$ é uma constante, não uma variável aleatória. O que se desloca de réplica para réplica são os **intervalos**, que sobem e descem conforme a amostra sorteada e

também mudam de comprimento (porque s varia). O objeto aleatório é o intervalo, não o parâmetro — e é exatamente por isso que a afirmação probabilística tem de ser sobre ele.

Tome agora um intervalo qualquer da figura, digamos o de cima. Ele contém μ? Sim — e a "probabilidade" disso é 1, não 0,95. Tome um dos vermelhos: contém μ? Não — probabilidade 0. **Nenhum intervalo individual tem probabilidade 0,95 de conter μ.** Os 95% descrevem a **coleção**: a proporção de barras azuis. Este é o Erro nº 2 desfeito em imagem.

E note que os intervalos vermelhos não têm nada de errado. Foram construídos com a mesma fórmula, sobre amostras igualmente legítimas. Simplesmente calharam de vir de amostras cuja média se afastou mais do que o usual. Errar 5% das vezes é o comportamento **projetado** do procedimento, não uma falha dele.

A cobertura de longo prazo se confirma com um número maior de réplicas:

```
set.seed(20265)
M <- 20000
A  <- matrix(rnorm(M * n_v, mean = mu_v, sd = sigma_v), nrow = M)
xb <- rowMeans(A); sm <- apply(A, 1, sd)

et <- qt(0.975, df = n_v - 1) * sm / sqrt(n_v)   # correto: quantil t
ez <- qnorm(0.975)            * sm / sqrt(n_v)   # errado:  quantil z com s amostral

round(c(nominal          = 0.95,
        cobertura_com_t  = mean(xb - et <= mu_v & mu_v <= xb + et),
        cobertura_com_z  = mean(xb - ez <= mu_v & mu_v <= xb + ez)), 4)
```

```
        nominal cobertura_com_t cobertura_com_z
         0.9500          0.9510          0.9394
```

A cobertura com o quantil t é 0,9510 — o valor nominal a menos de ruído de simulação. A terceira linha é o diagnóstico do erro de usar z quando σ é desconhecido: a cobertura cai para 0,9394. O intervalo fica **estreito demais** e promete uma confiança que não entrega. Com $n = 25$ o prejuízo é de cerca de 1 ponto percentual; com $n = 5$ seria muito maior. É a regra do professor verificada por simulação.

9.11.2 A t de Student e sua convergência para a Normal

```
nus <- c(1, 2, 3, 4, 5, 8, 10, 15, 20, 30, 60, 120, 1000)
tab_t <- data.frame(nu = nus,
                    t_0.025  = round(qt(0.975, nus), 4),
                    razao_ao_z = round(qt(0.975, nus) / qnorm(0.975), 4))
tab_t
```

```
     nu t_0.025 razao_ao_z
1     1 12.7062     6.4829
```

```
2      2  4.3027     2.1953
3      3  3.1824     1.6237
4      4  2.7764     1.4166
5      5  2.5706     1.3115
6      8  2.3060     1.1766
7     10  2.2281     1.1368
8     15  2.1314     1.0875
9     20  2.0860     1.0643
10    30  2.0423     1.0420
11    60  2.0003     1.0206
12   120  1.9799     1.0102
13  1000  1.9623     1.0012
```

```r
c(normal_z_0.025 = round(qnorm(0.975), 4))
```

```
normal_z_0.025
          1.96
```

A coluna `razao_ao_z` quantifica o "preço" de não conhecer σ. Com $\nu = 4$ — o caso do exemplo das alturas — o preço é 42% de amplitude a mais, precisamente o que calculamos à mão. Com $\nu = 30$ o preço já caiu para 4%, e é essa a justificativa numérica do corte em $n = 30$ da regra do professor. A última linha, $\nu = 1000$, dá 1,9623 contra 1,9600 da Normal: é a "última linha da tabela t" convergindo para os valores normais.

```r
grid_t <- seq(-4.5, 4.5, length.out = 600)
nus_g  <- c(1, 2, 5, 30)
d_dens <- do.call(rbind, lapply(nus_g, function(v)
  data.frame(x = grid_t, dens = dt(grid_t, df = v),
             curva = paste0("t com nu = ", v))))
d_dens <- rbind(d_dens,
                data.frame(x = grid_t, dens = dnorm(grid_t), curva = "Normal(0,1)"))
d_dens$curva <- factor(d_dens$curva,
                       levels = c(paste0("t com nu = ", nus_g), "Normal(0,1)"))

library(patchwork)

gA <- ggplot(d_dens, aes(x, dens, colour = curva, linewidth = curva)) +
  geom_line() +
  scale_colour_manual(values = c(vm, lr, vd, az, "black")) +
  scale_linewidth_manual(values = c(0.6, 0.6, 0.6, 0.6, 1.1)) +
  guides(linewidth = "none") +
  labs(x = "t", y = "densidade")

nu_seq <- c(1:10, 12, 15, 20, 30, 50, 100, 200, 500)
gB <- ggplot(data.frame(nu = nu_seq, q = qt(0.975, nu_seq)), aes(nu, q)) +
  geom_hline(yintercept = qnorm(0.975), colour = vm, linetype = "dashed") +
```

```
  geom_line(colour = az, linewidth = 0.7) +
  geom_point(colour = az, size = 1) +
  scale_x_log10() +
  labs(x = "nu  (escala log)", y = "quantil 0,975")

gA + gB
```

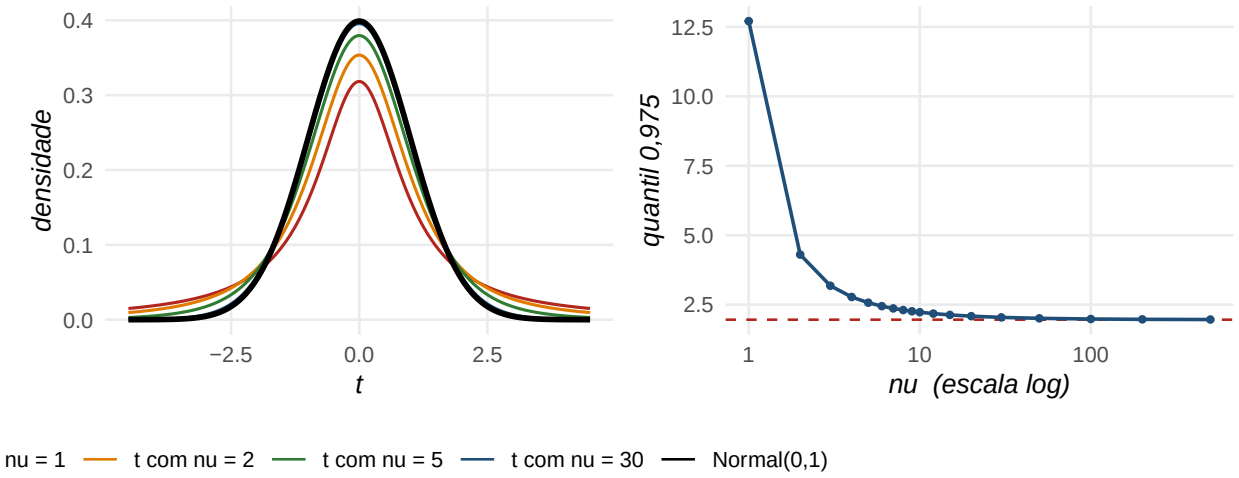


Figura 16: A t de Student e a Normal. Esquerda: densidades da t_ν para $\nu = 1, 2, 5, 30$ contra a $N(0,1)$ (preta, mais grossa). Graus de liberdade baixos produzem picos mais baixos e caudas mais pesadas; em $\nu = 30$ a curva já é quase indistinguível da Normal. Direita: o quantil $t_{\nu;0,025}$ em função de ν (escala log), convergindo para a linha tracejada em 1,96.

O painel esquerdo mostra o que "caudas mais pesadas" quer dizer concretamente. A t_1 (a Cauchy) é tão dispersa que sequer possui média; a t_5 já tem forma reconhecível de sino mas com ombros visivelmente mais largos que a Normal; a t_{30} praticamente se sobrepõe à curva preta. Como toda densidade integra a 1, mais massa nas caudas implica **menos** massa no centro: as curvas de ν baixo têm picos mais baixos.

A consequência para a inferência é direta:

```
data.frame(nu = c(1, 4, 10, 30, 120),
           P_abs_T_maior_1.96 = round(2 * pt(-1.96, df = c(1, 4, 10, 30, 120)), 4),
           referencia_normal  = 0.05)
```

```
   nu P_abs_T_maior_1.96 referencia_normal
1   1             0.3003              0.05
2   4             0.1216              0.05
3  10             0.0784              0.05
4  30             0.0593              0.05
5 120             0.0523              0.05
```

Sob $\nu = 4$, o intervalo $[-1{,}96;\ 1{,}96]$ deixa de fora 12,2% da massa, não 5%. Usar o quantil normal com quatro graus de liberdade produziria um intervalo de 87,8% de confiança rotulado como 95% — a discrepância que a simulação de cobertura já havia detectado.

9.11.3 Recalculando todos os exemplos do capítulo

Verificamos agora, com os quantis exatos do R, todos os intervalos calculados à mão. As pequenas diferenças na última casa decimal vêm dos quantis arredondados das tabelas impressas (1,96 contra 1,959964; 2,776 contra 2,776445), não de erro nas contas.

```
# funcao geral: IC para mu, com sigma conhecido (z) ou desconhecido (t)
ic_media <- function(xbar, desvio, n, conf = 0.95, sigma_conhecido = FALSE) {
  a <- 1 - conf
  q <- if (sigma_conhecido) qnorm(1 - a/2) else qt(1 - a/2, df = n - 1)
  c(quantil = q, li = xbar - q*desvio/sqrt(n), ls = xbar + q*desvio/sqrt(n))
}

round(rbind(
  `alturas, sigma=6 conhecido, 95%`   = ic_media(180, 6, 5, 0.95, TRUE),
  `alturas, s=6 desconhecido, 95%`    = ic_media(180, 6, 5, 0.95, FALSE),
  `trabalhadores, n=25, 90%`          = ic_media(400, 450, 25, 0.90),
  `televisores, n=16, 99%`            = ic_media(8900, 500, 16, 0.99),
  `endividamento, n=9, 95%`           = ic_media(45, 30, 9, 0.95),
  `endividamento, n=9, 90%`           = ic_media(45, 30, 9, 0.90)), 4)
```

```
                              quantil        li        ls
alturas, sigma=6 conhecido, 95%  1.9600  174.7409  185.2591
alturas, s=6 desconhecido, 95%   2.7764  172.5500  187.4500
trabalhadores, n=25, 90%         1.7109  246.0206  553.9794
televisores, n=16, 99%           2.9467 8531.6609 9268.3391
endividamento, n=9, 95%          2.3060   21.9400   68.0600
endividamento, n=9, 90%          1.8595   26.4045   63.5955
```

Todos conferem com o deck: $[174{,}74;\ 185{,}26]$, $[172{,}55;\ 187{,}45]$, $[246{,}02;\ 553{,}98]$ (deck: 246,01 e 553,99, com $t = 1{,}711$ arredondado), $[8.531{,}66;\ 9.268{,}34]$, $[21{,}94;\ 68{,}06]$ e $[26{,}40;\ 63{,}60]$.

Vale conferir os dados brutos das alturas, para verificar que $s^2 = 36$ de fato:

```
alturas <- c(174, 186, 186, 180, 174)
c(n = length(alturas), xbar = mean(alturas),
  s2 = var(alturas), s = sd(alturas),
  t_4_0.025 = qt(0.975, 4), z_0.025 = qnorm(0.975),
  razao_amplitudes = qt(0.975, 4) / qnorm(0.975))
```

```
                n             xbar               s2                s
        5.000000       180.000000        36.000000         6.000000
```

```
      t_4_0.025          z_0.025 razao_amplitudes
       2.776445         1.959964         1.416580
```

A razão das amplitudes é 1,4166: os 42% a mais de amplitude, confirmados.

Proporção e diferença de proporções:

```
ic_prop <- function(x, n, conf = 0.95) {
  ph <- x/n; ep <- sqrt(ph*(1-ph)/n); q <- qnorm(1 - (1-conf)/2)
  c(phat = ph, ep = ep, li = ph - q*ep, ls = ph + q*ep,
    cond_nphat = n*ph, cond_n1mphat = n*(1-ph))
}
ic_dif_prop <- function(x1, n1, x2, n2, conf = 0.90) {
  p1 <- x1/n1; p2 <- x2/n2
  ep <- sqrt(p1*(1-p1)/n1 + p2*(1-p2)/n2); q <- qnorm(1 - (1-conf)/2)
  c(p1 = p1, p2 = p2, dif = p1 - p2, ep = ep,
    li = (p1-p2) - q*ep, ls = (p1-p2) + q*ep)
}
round(ic_prop(49, 70, 0.95), 4)
```

```
      phat         ep         li         ls cond_nphat cond_n1mphat
    0.7000     0.0548     0.5926     0.8074    49.0000    21.0000
```

```
round(ic_dif_prop(140, 180, 220, 300, 0.90), 4)
```

```
    p1     p2    dif     ep     li     ls
0.7778 0.7333 0.0444 0.0402 -0.0216 0.1105
```

[0,5926; 0,8074] para a proporção (deck: 0,5927 e 0,8073, com $z = 1{,}96$) e [−0,0216; 0,1105] para a diferença — este último idêntico ao deck. O intervalo da diferença contém zero, como discutido.

Variância, com atenção à convenção invertida do `qchisq`:

```
ic_var <- function(s2, n, conf = 0.95) {
  a <- 1 - conf; gl <- n - 1
  chi_sup <- qchisq(1 - a/2, gl)   # o MAIOR: area a/2 a DIREITA
  chi_inf <- qchisq(a/2,     gl)   # o MENOR: area a/2 a ESQUERDA
  c(chi2_maior = chi_sup, chi2_menor = chi_inf,
    li = gl*s2/chi_sup, ls = gl*s2/chi_inf)
}
round(ic_var(132.7, 30, 0.95), 4)
```

```
chi2_maior chi2_menor         li         ls
   45.7223    16.0471    84.1668   239.8132
```

Quantis 45,7223 e 16,0471 (deck: 45,72 e 16,047) e intervalo [84,1668; 239,8132] (deck: 84,17 e 239,81). Confere.

Diferença de médias, com a variância combinada:

```
ic_dif_media <- function(m1, s1, n1, m2, s2, n2, conf = 0.95) {
  gl <- n1 + n2 - 2
  sp <- sqrt(((n1-1)*s1^2 + (n2-1)*s2^2) / gl)
  ep <- sp * sqrt(1/n1 + 1/n2)
  q  <- qt(1 - (1-conf)/2, df = gl)
  c(sp = sp, gl = gl, t = q, ep = ep,
    li = (m1-m2) - q*ep, ls = (m1-m2) + q*ep)
}
T1 <- c(5.0, 6.0, 3.5, 9.0, 8.5)
T2 <- c(4.0, 10.0, 7.0, 5.0)

round(rbind(
  `lampadas A x B, 99%` = ic_dif_media(4600, 250, 8, 4000, 200, 10, 0.99),
  `turmas 1 x 2, 95%`   = ic_dif_media(mean(T1), sd(T1), length(T1),
                                       mean(T2), sd(T2), length(T2), 0.95)), 4)
```

```
                          sp gl      t       ep       li       ls
lampadas A x B, 99% 223.2571 16 2.9208 105.9002 290.6888 909.3112
turmas 1 x 2, 95%     2.4698  7 2.3646   1.6568  -4.0177   3.8177
```

Para as lâmpadas: $s_p = 223{,}2571$ (deck: 223,26), $t_{16;0{,}005} = 2{,}9208$ (deck: 2,921) e [290,69; 909,31] (deck: 290,67 e 909,33). Para as turmas: $s_p = 2{,}4698$, $t_{7;0{,}025} = 2{,}3646$ (deck: 2,365) e [−4,0177; 3,8177] (deck: −4,0182 e 3,8182). Todas as diferenças são de arredondamento do quantil tabelado.

Confirmando as estatísticas das turmas, que o deck fornece prontas:

```
rbind(turma_1 = c(n = length(T1), xbar = mean(T1), s2 = var(T1), s = sd(T1)),
      turma_2 = c(length(T2), mean(T2), var(T2), sd(T2)))
```

```
        n xbar    s2        s
turma_1 5  6.4 5.425 2.329163
turma_2 4  6.5 7.000 2.645751
```

$\bar{x}_1 = 6{,}4$ com $s_1^2 = 5{,}425$, e $\bar{x}_2 = 6{,}5$ com $s_2^2 = 7$: exatamente o deck.

Razão de variâncias, o caso mais delicado. Note o cuidado com a ordem dos graus de liberdade:

```
gl1 <- length(T1) - 1   # 4, numerador
gl2 <- length(T2) - 1   # 3, denominador

# a tabela so traz a cauda superior; a inferior sai pela inversao COM TROCA dos g.l.
```

```r
f_sup <- qf(0.975, gl1, gl2)      # f_{4,3;0,025}
f_aux <- qf(0.975, gl2, gl1)      # f_{3,4;0,025}  <- graus de liberdade TROCADOS
f_inf <- 1 / f_aux                # f_{4,3;0,975}

round(c(f_4_3_0.025 = f_sup, f_3_4_0.025 = f_aux,
        f_4_3_0.975_por_inversao = f_inf,
        f_4_3_0.975_direto       = qf(0.025, gl1, gl2),
        ERRADO_inverter_sem_trocar = 1 / f_sup), 4)
```

```
              f_4_3_0.025                f_3_4_0.025
                  15.1010                     9.9792
  f_4_3_0.975_por_inversao          f_4_3_0.975_direto
                   0.1002                     0.1002
ERRADO_inverter_sem_trocar
                   0.0662
```

Os quantis batem com o deck (15,101 e 9,9792 contra 15,10 e 9,98), e a inversão com troca dá 0,1002 — idêntico ao `qf(0.025, 4, 3)` calculado diretamente, o que **valida a identidade**. A última entrada é o Erro 2 da caixa: inverter sem trocar dá 0,0662, um valor 34% abaixo do correto.

```r
razao <- var(T2) / var(T1)
c(razao_amostral = razao,
  li = razao * f_inf, ls = razao * f_sup,
  contem_1 = (razao * f_inf < 1) && (1 < razao * f_sup))
```

```
razao_amostral             li             ls       contem_1
     1.2903226      0.1293012     19.4851341      1.0000000
```

[0,1293; 19,4851] (deck: 0,1293 e 19,4839, com $f = 15{,}10$ tabelado). O intervalo contém 1: a hipótese de variâncias iguais que sustentou o s_p da seção anterior está validada. E contém também a razão amostral 1,2903, o teste de sanidade da caixa.

9.11.4 A assimetria do intervalo para a variância

A qui-quadrado é assimétrica, e a consequência — o intervalo para σ^2 **não** ser centrado em s^2 — merece uma figura.

```r
gl <- 29; s2_obs <- 132.7
lim <- ic_var(s2_obs, 30, 0.95)
round(c(s2 = s2_obs, li = lim[["li"]], ls = lim[["ls"]],
        dist_ao_li = s2_obs - lim[["li"]],
        dist_ao_ls = lim[["ls"]] - s2_obs,
        ponto_medio = (lim[["li"]] + lim[["ls"]])/2,
        razao_ls_li = lim[["ls"]]/lim[["li"]]), 4)
```

```
        s2          li          ls  dist_ao_li  dist_ao_ls ponto_medio
  132.7000     84.1668    239.8132     48.5332    107.1132    161.9900
razao_ls_li
     2.8493
```

A estimativa dista 48,5 do limite inferior e 107,1 do superior: a "cauda direita" do intervalo é **2,2 vezes** mais longa. O ponto médio, 162,0, está bem acima de $s^2 = 132{,}7$. E o limite superior é 2,85 vezes o inferior — variâncias são difíceis de estimar.

```
grid_chi <- seq(0.01, 75, length.out = 800)
d_chi <- do.call(rbind, lapply(c(3, 10, 29), function(v)
  data.frame(x = grid_chi, dens = dchisq(grid_chi, df = v),
             curva = paste0("nu = ", v))))
medias_chi <- data.frame(curva = paste0("nu = ", c(3, 10, 29)), m = c(3, 10, 29))

gC <- ggplot(d_chi, aes(x, dens, colour = curva)) +
  geom_line(linewidth = 0.75) +
  geom_vline(data = medias_chi, aes(xintercept = m, colour = curva),
             linetype = "dashed", linewidth = 0.4) +
  scale_colour_manual(values = c(vm, lr, az)) +
  coord_cartesian(xlim = c(0, 75), ylim = c(0, 0.25)) +
  labs(x = "x", y = "densidade")

d_seg <- data.frame(li = lim[["li"]], ls = lim[["ls"]], y = 1)
d_pts <- data.frame(x = c(s2_obs, (lim[["li"]] + lim[["ls"]])/2), y = c(1, 1),
                    o = c("s2 observado", "ponto medio do IC"))

gD <- ggplot() +
  geom_segment(data = d_seg, aes(x = li, xend = ls, y = y, yend = y),
               colour = az, linewidth = 1.4) +
  geom_point(data = d_seg, aes(x = li, y = y), colour = az, size = 3) +
  geom_point(data = d_seg, aes(x = ls, y = y), colour = az, size = 3) +
  geom_point(data = d_pts, aes(x = x, y = y, shape = o, colour = o), size = 3) +
  scale_shape_manual(values = c(`s2 observado` = 17, `ponto medio do IC` = 18)) +
  scale_colour_manual(values = c(`s2 observado` = vm, `ponto medio do IC` = cz)) +
  scale_y_continuous(breaks = NULL, limits = c(0.9, 1.1)) +
  labs(x = "sigma^2   |   IC de 95% = [84,17 ; 239,81]", y = NULL)

gC / gD + plot_layout(heights = c(3, 1.4))
```

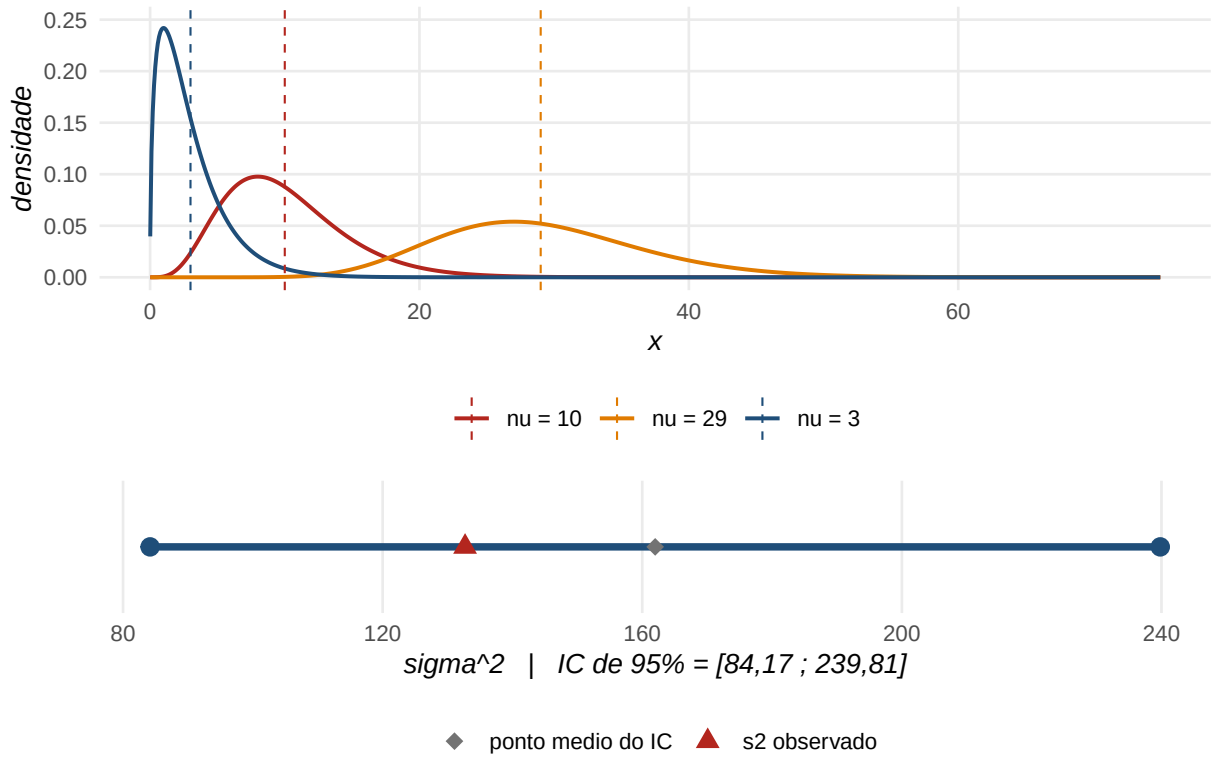


Figura 17: Acima: densidades qui-quadrado para $\nu = 3, 10, 29$. A assimetria à direita é acentuada com poucos graus de liberdade e vai diminuindo à medida que ν cresce (a linha tracejada marca $\mathbb{E}(X) = \nu$ em cada caso). Abaixo: o intervalo de 95% para σ^2 do exemplo, [84,17; 239,81]. A estimativa pontual $s^2 = 132{,}7$ (vermelho) NÃO está no centro — a assimetria da qui-quadrado empurra o limite superior para muito mais longe do que o inferior. O losango cinza marca o ponto médio do intervalo, 162,0.

Compare mentalmente com o intervalo para μ: lá, a estimativa está sempre no centro exato, porque a Normal e a t são simétricas e os dois quantis são $\mp$ o mesmo número. Aqui não há simetria a explorar, e a fórmula precisa listar os dois limites separadamente. É a razão estrutural pela qual as duas últimas linhas da tabela-resumo não usam a notação "$\mp$".

9.11.5 O trade-off confiança-precisão, quantificado

Fechamos com a verificação numérica do trade-off, usando os dados das alturas:

```
confs <- c(0.80, 0.90, 0.95, 0.99, 0.999)
d_tr <- data.frame(
  confianca = confs,
  z         = round(qnorm(1 - (1 - confs)/2), 4),
  margem    = round(qnorm(1 - (1 - confs)/2) * 6/sqrt(5), 3),
  amplitude = round(2 * qnorm(1 - (1 - confs)/2) * 6/sqrt(5), 3))
d_tr
```

```
  confianca      z margem amplitude
1     0.800 1.2816  3.439     6.878
2     0.900 1.6449  4.414     8.827
3     0.950 1.9600  5.259    10.518
4     0.990 2.5758  6.912    13.823
5     0.999 3.2905  8.829    17.659
```

Passar de 80% para 99,9% de confiança **mais que dobra** a amplitude: de 6,88 para 17,66. E o crescimento é acelerado nas confianças altas — os últimos pontos percentuais custam desproporcionalmente caro, porque a cauda da Normal decai rapidamente e é preciso ir muito longe para capturar mais um pouco de massa. A confiança de 100% exigiria amplitude infinita.

O único caminho para escapar do trade-off é a amostra:

```r
ns <- c(5, 10, 25, 50, 100, 400, 1600)
data.frame(n = ns,
           margem_95 = round(qnorm(0.975) * 6/sqrt(ns), 4),
           n_relativo = ns/5,
           reducao_da_margem = round(sqrt(5/ns), 4))
```

```
     n margem_95 n_relativo reducao_da_margem
1    5    5.2591          1            1.0000
2   10    3.7188          2            0.7071
3   25    2.3520          5            0.4472
4   50    1.6631         10            0.3162
5  100    1.1760         20            0.2236
6  400    0.5880         80            0.1118
7 1600    0.2940        320            0.0559
```

A última coluna é $\sqrt{5/n}$: quadruplicar a amostra (de 5 para 20, de 100 para 400) reduz a margem exatamente à metade. Sair de $n = 5$ para $n = 1600$ — 320 vezes mais dados — reduz a margem a 1/17,9. É o decaimento $1/\sqrt{n}$ de Seção 7, e é a razão econômica pela qual pesquisas de opinião nacionais raramente ultrapassam alguns milhares de entrevistas: o retorno marginal em precisão decai com a raiz, enquanto o custo cresce linearmente.

A ponte para o próximo capítulo já está construída. Vimos duas vezes — nas lâmpadas e nas turmas — que a pergunta "o intervalo contém o valor de referência?" produz conclusões substantivas. Formalizar essa pergunta, escolher o valor de referência antes de ver os dados e quantificar os erros que se pode cometer ao respondê-la é o programa de Seção 10.

9.12 Exercícios

1. **A derivação, refeita.** Partindo de $Z = (\bar{X} - \mu)/(\sigma/\sqrt{n}) \sim N(0,1)$, reproduza os cinco passos algébricos que isolam μ, justificando cada um. (a) Em que passo a desigualdade se inverte, e por quê? (b) Refaça a derivação isolando σ em vez de μ — supondo agora μ conhecido — e explique por que o resultado **não** é um intervalo de confiança utilizável na prática. (c) Mostre que o intervalo simétrico $[-z_{\alpha/2}, z_{\alpha/2}]$ é o mais curto entre todos os intervalos $[a, b]$ com $P(a \le Z \le b) = 1 - \alpha$.

2. **Interpretação.** Um pesquisador calcula, a partir de sua amostra, o intervalo $\text{IC}_{95\%}(\mu) = [12{,}3;\ 15{,}7]$. Classifique cada afirmação como correta ou incorreta, justificando: (a) "há 95% de probabilidade de que μ esteja entre 12,3 e 15,7";

   (b) "95% das observações da população estão entre 12,3 e 15,7"; (c) "se repetíssemos o estudo muitas vezes, 95% dos intervalos assim construídos conteriam μ"; (d) "há 95% de confiança de que $\bar{x}$ esteja entre 12,3 e 15,7"; (e) "μ está entre 12,3 e 15,7 ou não está — não sabemos qual". Para cada incorreta, identifique qual dos dois erros do capítulo ela comete (ou se comete um terceiro).

3. **Conhecido versus desconhecido.** Uma amostra de $n = 16$ de uma população Normal deu $\bar{x} = 50$ e $s = 8$. (a) Construa o $\text{IC}_{95\%}(\mu)$ supondo $\sigma = 8$ **conhecido**.

   (b) Construa-o supondo σ **desconhecido**, com $s = 8$. (c) Calcule a razão das amplitudes e explique-a em termos da razão dos quantis. (d) A partir de que n a diferença entre os dois intervalos cairia abaixo de 2%? (Use a tabela de quantis t do laboratório.)
   (c) Segundo a regra do professor, qual dos dois você usaria com $n = 16$? E com $n = 100$?

4. **Determinação do tamanho da amostra.** Deseja-se estimar μ com margem de erro no máximo ε e confiança $1 - \alpha$, sabendo-se σ. (a) Isole n na expressão $\varepsilon = z_{\alpha/2}\sigma/\sqrt{n}$ e obtenha $n \geq \left(z_{\alpha/2}\sigma/\varepsilon\right)^2$. (b) Com $\sigma = 6$, quantas observações são necessárias para uma margem de 1 cm a 95%? E a 99%? (c) Se a margem exigida for reduzida à metade, o que acontece com n? (d) Por que esse cálculo é mais difícil quando σ é desconhecido?

5. **Proporção.** Numa pesquisa com $n = 400$ eleitores, 220 declaram intenção de voto no candidato A. (a) Verifique as condições de validade e construa o $\text{IC}_{95\%}(p)$.

   (b) O intervalo contém 0,5? Que conclusão substantiva se extrai disso? (c) Mostre que o erro-padrão $\sqrt{\hat{p}(1-\hat{p})/n}$ é máximo em $\hat{p} = 0{,}5$, derivando em relação a $\hat{p}$. (d) Usando o resultado de (c), obtenha o n que garante margem de no máximo 2 pontos percentuais a 95%, **qualquer que seja** $\hat{p}$ (a chamada "margem conservadora" das pesquisas eleitorais).

6. **Variância e a assimetria.** Uma amostra de $n = 20$ de uma população Normal deu $s^2 = 45$. (a) Construa o $\text{IC}_{95\%}(\sigma^2)$, indicando claramente qual quantil qui-quadrado vai em cada denominador. (b) Verifique que s^2 **não** é o ponto médio do intervalo, e calcule as duas distâncias. (c) Obtenha o intervalo para σ e verifique que ele contém $s = \sqrt{45}$. (d) Repita (a) com $n = 100$ e $s^2 = 45$: o que acontece com a amplitude relativa LS/LI e com a assimetria? Relacione com $\text{V}(\chi^2_\nu) = 2\nu$.

7. **Diferença de médias e as condições de validade.** Dois processos produtivos foram amostrados: processo I com $n_1 = 12$, $\bar{x}_1 = 105$, $s_1 = 8$; processo II com $n_2 = 15$, $\bar{x}_2 = 98$, $s_2 = 12$. (a) Construa o $\text{IC}_{95\%}(\mu_1 - \mu_2)$, calculando s_p explicitamente. (b) O intervalo contém zero? Qual a conclusão? (c) Um colega observa que s_2 é 50% maior que s_1 e conclui que a hipótese de variâncias iguais foi violada. Responda-lhe. (d) Construa o $\text{IC}_{95\%}(\sigma_2^2/\sigma_1^2)$ e use-o para avaliar formalmente a objeção do colega. (e) Verifique que s_p está entre s_1 e s_2, e explique por que isso sempre acontece.

8. **A tabela F e suas armadilhas.** (a) Usando a identidade $f_{\nu_1,\nu_2;1-\gamma} = 1/f_{\nu_2,\nu_1;\gamma}$, obtenha $f_{10,6;0,95}$ a partir de $f_{6,10;0,05} = 3{,}22$. (b) Um aluno calcula $f_{10,6;0,95}$ como $1/f_{10,6;0,05}$ com $f_{10,6;0,05} = 4{,}06$, obtendo 0,2463. Qual é o valor correto, e de quanto é o erro percentual? (c) Demonstre a identidade, partindo de $F \sim F_{\nu_1,\nu_2} \Longrightarrow 1/F \sim F_{\nu_2,\nu_1}$. (d) Explique por que o

intervalo de confiança para σ_2^2/σ_1^2 deve necessariamente conter a razão amostral s_2^2/s_1^2, e use esse fato como teste de sanidade para os itens anteriores.

10 Testes de Hipóteses

Em Seção 8 aprendemos a produzir um número a partir dos dados; em Seção 9, uma faixa de valores plausíveis para o parâmetro. Falta o terceiro e último movimento da inferência, e é o que a prática mais demanda: **decidir**. Um economista raramente é pago para dizer "a elasticidade estimada é $-0{,}83$ com intervalo $[-1{,}12; -0{,}54]$"; ele é pago para responder se a demanda é elástica, se a política teve efeito, se o novo processo reduziu o custo médio, se a corretora mentiu sobre a fração de clientes avessos ao risco. Essas são perguntas de **sim ou não**, formuladas sobre um parâmetro desconhecido e respondidas com uma amostra finita — e, portanto, respondidas **sob risco de erro**.

O aparato que organiza esse tipo de decisão é o **teste de hipóteses**, e ele encerra estas notas. O capítulo tem quatro partes. Primeiro, o vocabulário: o que é uma hipótese estatística, como se formulam H_0 e H_1, e por que a assimetria entre as duas é o ponto mais importante — e mais mal compreendido — de todo o assunto. Segundo, os dois tipos de erro que se pode cometer, o nível de significância α, e a noção de **poder**. Terceiro, os **três métodos** de decisão — intervalo de confiança, região crítica e p-valor —, que são três linguagens para o mesmo procedimento e que, quando aplicáveis, **têm de** dar a mesma resposta. Quarto, o catálogo dos testes do curso: média, proporção, variância, diferença de médias, diferença de proporções e razão de variâncias.

Uma advertência de leitura. Este capítulo é, ao mesmo tempo, o mais mecânico e o mais conceitualmente escorregadio do curso. Mecânico porque tudo se reduz a montar uma estatística, consultar um quantil e comparar; escorregadio porque quase todas as frases que os alunos escrevem sobre testes — "a probabilidade de H_0 ser verdadeira é 5%", "aceito H_0", "o p-valor é a probabilidade de a hipótese ser falsa" — são erradas. Vamos ser cuidadosos com a linguagem, porque nela mora o entendimento.

10.1 O que é uma hipótese estatística

! Definição — hipótese estatística

Uma **hipótese estatística** é uma afirmação sobre um **parâmetro** de uma população — sobre μ, sobre p, sobre σ^2, sobre a diferença $\mu_1 - \mu_2$. Um **teste de hipóteses** é um procedimento que, com base em uma amostra, decide entre duas hipóteses concorrentes:

$$\begin{array}{|l} H_0 : \text{a } \textbf{hipótese nula}, \\ H_1 : \text{a } \textbf{hipótese alternativa}. \end{array}$$

O teste admite exatamente **duas** decisões possíveis: **rejeitar** H_0 ou **não rejeitar** H_0.

Três observações, cada uma com consequências.

A primeira é que hipóteses são afirmações sobre **parâmetros**, não sobre estatísticas. Escrever "$H_0 : \bar{X} = 175$" é um erro de categoria: $\bar{X}$ é o que observamos, não o que está em questão. A hipótese é sobre μ, que é fixo e desconhecido; a amostra é o instrumento com que a interrogamos. Essa distinção é a mesma de Seção 8 — parâmetro versus estatística — e continua sendo a fonte número um de confusão.

A segunda é que as duas hipóteses são **mutuamente exclusivas e exaustivas** no espaço de parâmetros relevante: ou o parâmetro satisfaz H_0, ou satisfaz H_1, nunca ambas. Não há terceira via, e o teste não a oferece.

A terceira, e é a decisiva, está no enunciado das duas decisões possíveis. Note o que **não** está escrito ali: não se pode "aceitar H_0". As alternativas são *rejeitar* e *não rejeitar*, e a diferença entre "não rejeitar" e "aceitar" é toda a diferença entre entender e não entender um teste de hipóteses.

> **!** Não se aceita H_0
>
> Quando o teste não rejeita H_0, a conclusão correta é: **os dados não forneceram evidência suficiente contra** H_0. Isso é logicamente muito mais fraco do que "os dados demonstraram que H_0 é verdadeira".
> A razão é que a falta de evidência pode ter duas causas distintas e indistinguíveis a partir do resultado: ou H_0 é de fato verdadeira, ou H_0 é falsa e a nossa amostra foi pequena demais (ou ruidosa demais) para detectá-lo. Um teste com $n = 5$ não rejeita quase nada; isso não é prova de nada. Escreva sempre "não rejeitamos H_0" e nunca "aceitamos H_0" — é uma disciplina de vocabulário que protege contra um erro de raciocínio.

10.1.1 A analogia do julgamento

A estrutura lógica do teste de hipóteses é exatamente a de um julgamento criminal sob presunção de inocência, e vale a pena percorrer o paralelo com cuidado, porque ele resolve sozinho quase todas as dúvidas conceituais do capítulo.

Num tribunal, o réu é **presumido inocente**. Essa presunção é a hipótese nula:

$$H_0 : \text{o réu é inocente} \qquad \text{contra} \qquad H_1 : \text{o réu é culpado.}$$

A acusação carrega o **ônus da prova**: cabe a ela apresentar evidência suficiente para derrubar a presunção. O réu não precisa provar sua inocência — ela é o ponto de partida. Se a evidência é convincente "além de dúvida razoável", o júri condena; caso contrário, absolve.

Repare agora nas quatro correspondências, uma a uma:

1. H_0 **é a hipótese protegida.** Só a abandonamos diante de evidência forte. É o *status quo*, a posição conservadora, aquilo que se mantém na ausência de razão para mudar.
2. **A absolvição não é declaração de inocência.** O veredicto é "não culpado", nunca "inocente" — o júri diz que a acusação não provou, não que o réu não fez. É exatamente o "não rejeitar H_0" da caixa anterior, e é literalmente a mesma distinção lógica.
3. **Condenar um inocente é o erro grave.** O sistema jurídico é deliberadamente construído para tornar esse erro raro, aceitando como preço que alguns culpados sejam absolvidos. Veremos na próxima seção que esse é o **erro tipo I**, e que α é precisamente o quanto dele estamos dispostos a tolerar.
4. **"Além de dúvida razoável" é o α.** É o padrão de prova, fixado *antes* de ver as evidências deste caso particular. Não se decide o quão exigente ser depois de olhar os dados — o padrão vem primeiro. Voltaremos a esse ponto quando discutirmos o p-valor.

A analogia também esclarece por que a escolha de quem é H_0 não é simétrica nem arbitrária. Trocar as hipóteses — presumir o réu culpado até prova de inocência — produziria um sistema totalmente diferente, com erros de outro tipo. A mesma coisa vale em estatística: H_0 e H_1 não são intercambiáveis, e escolher errado inverte o sentido da conclusão.

10.1.2 Como formular as hipóteses

Chegamos então à pergunta operacional: dado um problema, qual afirmação vai em H_0 e qual vai em H_1? A diretriz do curso é uma só, e é suficiente para todos os exercícios:

> **!** Diretriz de formulação
>
> **H_1 é aquilo que se quer evidenciar.**
>
> A hipótese alternativa recebe a afirmação que o pesquisador pretende **demonstrar** com os dados; a nula recebe seu complemento, isto é, o *status quo*, a afirmação a ser derrubada. Consequência técnica: a **igualdade fica sempre em** H_0. Escrevemos $H_0 : \mu = k$, $H_0 : \mu \leq k$ ou $H_0 : \mu \geq k$, mas nunca $H_0 : \mu \neq k$ ou $H_0 : \mu > k$. A razão é operacional: para calcular probabilidades sob H_0 precisamos de **um** valor de μ com que trabalhar, e é a fronteira $\mu = k$ que o fornece.

A diretriz combina com a analogia do julgamento de modo transparente. O que se quer evidenciar é o papel da *acusação*, e a acusação tem o ônus da prova; se ela não o cumprir, prevalece o *status quo*. Colocar a afirmação de interesse em H_1 é, portanto, submetê-la ao padrão de prova mais exigente — o que é a postura cética correta.

A forma de H_1 determina se o teste é **bilateral** ou **unilateral**:

Se queremos evidenciar	H_0	H_1	Teste
que μ **difere** de k	$\mu = k$	$\mu \neq k$	bilateral
que μ é **maior** que k	$\mu \leq k$	$\mu > k$	unilateral à direita
que μ é **menor** que k	$\mu \geq k$	$\mu < k$	unilateral à esquerda

Exemplo. Uma indústria afirma que suas embalagens contêm, em média, 500 g. Um órgão de defesa do consumidor suspeita que contenham **menos**. O que o órgão quer evidenciar é $\mu < 500$; logo $H_1 : \mu < 500$ e $H_0 : \mu \geq 500$. Note que a suspeita — a acusação — foi para a alternativa, e a afirmação do fabricante — o *status quo* — para a nula. Se o órgão fosse processado pela indústria e tivesse de provar a fraude em juízo, a estrutura seria exatamente essa.

10.2 Os dois tipos de erro

Decidir com base em uma amostra é decidir sob incerteza, e portanto errar às vezes é inevitável. A contribuição da estatística não é eliminar o erro — é **classificá-lo, quantificá-lo e controlá-lo**.

Como há dois estados possíveis da natureza (H_0 verdadeira ou falsa) e duas decisões possíveis (rejeitar ou não rejeitar), há quatro combinações, das quais duas são acertos e duas são erros:

	H_0 é **verdadeira**	H_0 é **falsa**
Não rejeitar H_0	Decisão correta $(1-\alpha)$	**Erro tipo II** (β)
Rejeitar H_0	**Erro tipo I** (α)	Decisão correta $(1-\beta)$

! Definição — erros tipo I e tipo II

O **erro tipo I** consiste em **rejeitar** H_0 **quando** H_0 **é verdadeira**. Sua probabilidade é o **nível de significância**:

$$\boxed{\alpha = P(\text{rejeitar } H_0 \mid H_0 \text{ verdadeira})}$$

O **erro tipo II** consiste em **não rejeitar** H_0 **quando** H_0 **é falsa**, com

$$\boxed{\beta = P(\text{não rejeitar } H_0 \mid H_0 \text{ falsa})}$$

No tribunal, o erro tipo I é condenar um inocente e o erro tipo II é absolver um culpado. A assimetria de gravidade que o direito penal reconhece — melhor dez culpados soltos que um inocente preso — é a mesma que a estatística adota: **é o erro tipo I que se controla diretamente**.

E controla-se de maneira peculiar: α é **escolhido pelo pesquisador antes de ver os dados**. Não é um resultado do experimento; é um parâmetro do procedimento, o padrão de prova que se decide adotar. Os valores usuais são

$$\alpha = 0{,}01, \qquad \alpha = 0{,}05, \qquad \alpha = 0{,}10,$$

com 0,05 como praxe. Não há nada de sagrado nesses números — são convenção, e uma convenção que já rendeu litros de tinta crítica —, mas a lógica de fixá-los *a priori* é sólida: se pudéssemos escolher α depois de olhar os dados, escolheríamos sempre o valor que produz a conclusão que queríamos, e o procedimento inteiro perderia o sentido.

! α e β não somam 1

Esta é uma advertência explícita do professor, e é um erro frequentíssimo. Não existe relação $\alpha + \beta = 1$, e nem poderia existir: as duas probabilidades são calculadas **sob condicionantes diferentes** — α supõe H_0 verdadeira, β supõe H_0 falsa. São probabilidades de dois experimentos distintos, e não há nenhuma razão para que somem coisa alguma.

O que se pode garantir é uma relação **qualitativa**, e apenas ela: mantido n fixo, **quando um diminui, o outro aumenta**. Tornar o teste mais exigente para rejeitar (reduzir α) o torna mecanicamente menos capaz de detectar falsidades de H_0 (aumenta β). É um *trade-off*, não uma identidade.

A única maneira de reduzir os **dois** simultaneamente é **aumentar** n — mais informação melhora as duas margens ao mesmo tempo. A Figura 18 mostra as três coisas em um só gráfico.

Note ainda que β, ao contrário de α, **não é um número só**: ele depende de *quão falsa* H_0 é. Se $H_0 : \mu = 175$ e a verdade é $\mu = 175{,}01$, praticamente nenhum teste detectará a diferença e $\beta \approx 1 - \alpha$; se a verdade é $\mu = 250$, qualquer teste a detecta e $\beta \approx 0$. Por isso β é uma **função** do valor verdadeiro do parâmetro, e não uma constante que se possa fixar de antemão como se faz com α.

10.2.1 Poder do teste

> **i** Ferramenta matemática — nota sobre o escopo
>
> Os slides da disciplina mencionam β mas **não definem o poder**; o programa de 2026.1 registra "poder" como item de **definição somente**. A subseção a seguir é, portanto, um acréscimo meu: dou a definição pedida e a exploro um pouco além, porque é o conceito que torna o erro tipo II utilizável na prática, e porque a curva de poder é o melhor instrumento pedagógico que conheço para mostrar o *trade-off* da caixa anterior em funcionamento. Nada aqui vai além do que se exige de vocabulário.

> **!** Definição — poder do teste
>
> O **poder** de um teste é a probabilidade de rejeitar H_0 **quando** H_0 **é falsa**, isto é, a probabilidade de detectar corretamente um efeito que existe:
>
> $$\boxed{\pi(\theta) = 1 - \beta(\theta) = P(\text{rejeitar } H_0 \mid \theta)}$$
>
> Como β, o poder é uma **função** do valor verdadeiro θ do parâmetro. O gráfico de $\pi(\theta)$ contra θ é a **curva de poder** do teste.

Três propriedades da curva de poder, todas visíveis na Figura 19:

1. **Vale α na fronteira de H_0.** Em $\theta = k$, rejeitar é cometer erro tipo I, e por construção isso ocorre com probabilidade α. O poder de um teste no ponto em que H_0 é exatamente verdadeira é o próprio nível de significância.
2. **Cresce com a distância a H_0.** Quanto mais falsa a nula, mais fácil detectá-la.
3. **Cresce com n.** Mais dados, mais poder — para todo valor de θ simultaneamente. É a versão formal de "o jeito de reduzir os dois erros ao mesmo tempo é coletar mais informação".

A terceira propriedade tem um corolário desconfortável, que vale registrar: com n suficientemente grande, **qualquer** desvio de H_0, por menor e mais irrelevante que seja, acaba sendo rejeitado. Significância estatística e relevância econômica são coisas diferentes, e confundi-las é um dos vícios mais persistentes da literatura aplicada. Um efeito de 0,0001% pode ser "altamente significante" com um milhão de observações e não significar absolutamente nada para nenhuma decisão.

10.3 Os três métodos

Fixadas as hipóteses e o nível α, resta o procedimento de decisão. O curso apresenta **três** métodos, nesta ordem, e é importante entender desde já a relação entre eles: são **três linguagens para a**

mesma decisão, não três testes diferentes. Quando os três são aplicáveis, os três **têm de** produzir a mesma conclusão — se não produzirem, há erro de conta.

10.3.1 Método 1: o intervalo de confiança

O primeiro método é o mais direto, e reaproveita integralmente a máquina de Seção 9.

> **!** Método do intervalo de confiança
>
> Para testar $H_0 : \theta = k$ contra $H_1 : \theta \neq k$ ao nível de significância α:
>
> construa o IC de $100(1-\alpha)\%$ para θ;
> se $k \notin$ IC $\implies$ **rejeita-se** H_0;
> se $k \in$ IC $\implies$ **não se rejeita** H_0.
>
> **Este método vale apenas para testes bilaterais.**

Eis por que a notação $100(1-\alpha)\%$ foi adotada lá atrás, em Seção 9, quando um simples "95%" teria bastado: ela já antecipava este capítulo. O α do intervalo de confiança e o α do teste são **o mesmo** α. Um teste ao nível de 5% corresponde a um intervalo de 95%; um teste a 1%, a um intervalo de 99%. A coincidência não é notacional — é estrutural, e a razão aparece na próxima subseção.

A intuição é imediata: o intervalo de confiança é o conjunto dos valores do parâmetro **compatíveis com os dados** ao nível escolhido. Se o valor hipotetizado k está entre eles, os dados não o contradizem e não há por que abandonar H_0. Se k ficou de fora, os dados o consideram implausível, e H_0 cai.

A restrição a testes bilaterais é essencial e frequentemente esquecida. O intervalo de Seção 9 é simétrico, com $\alpha/2$ em cada cauda; um teste unilateral põe α inteiro de um lado só. As geometrias não batem, e usar o intervalo bilateral para decidir um teste unilateral dá a resposta errada. Para testes unilaterais, use o método 2 ou o 3.

10.3.2 Método 2: a região crítica

O segundo método é o de uso mais geral, e é o que o curso adota como padrão.

A ideia: em vez de perguntar se k está no intervalo em torno de $\bar{x}$, padronizamos a distância entre $\bar{x}$ e k e perguntamos se ela é grande demais para ser atribuída ao acaso. A quantidade padronizada é a **estatística de teste**, calculada **sob** H_0 — isto é, supondo que $\mu = k$ de fato:

$$z_0 = \frac{\bar{x} - k}{\sigma/\sqrt{n}} \quad (\sigma \text{ conhecido}), \qquad t_0 = \frac{\bar{x} - k}{s/\sqrt{n}} \quad (\sigma \text{ desconhecido}).$$

Sob H_0, sabemos a distribuição dessas quantidades: $z_0 \sim N(0,1)$ e $t_0 \sim t_{n-1}$, pelos resultados de Seção 7 e Seção 9. Valores próximos de zero são o que se espera se H_0 é verdadeira; valores muito afastados de zero são improváveis sob H_0 e constituem evidência contra ela.

! Método da região crítica

A **região crítica** (ou *região de rejeição*) RC é o conjunto de valores da estatística de teste que levam a rejeitar H_0. A regra é:

$$\boxed{\text{estatística} \in \text{RC} \implies \textbf{rejeita-se } H_0.}$$

Para o teste da média com σ conhecido, ao nível α:

$$\begin{aligned}
H_1 : \mu \neq k &\implies \text{RC} = (-\infty,\, -z_{\alpha/2}\,] \cup [\, z_{\alpha/2},\, +\infty) && \text{(bilateral)},\\
H_1 : \mu > k &\implies \text{RC} = [\, z_{\alpha},\, +\infty) && \text{(unilateral à direita)},\\
H_1 : \mu < k &\implies \text{RC} = (-\infty,\, -z_{\alpha}\,] && \text{(unilateral à esquerda)}.
\end{aligned}$$

A construção da região crítica é o que faz α ser α. Como sob H_0 temos $z_0 \sim N(0,1)$, a área da RC bilateral sob a normal padrão é exatamente $\alpha/2 + \alpha/2 = \alpha$; logo $P(z_0 \in \text{RC} \mid H_0) = \alpha$, que é a definição do erro tipo I. A região crítica é, literalmente, "os $100\alpha\%$ resultados mais extremos que H_0 ainda produziria".

! No teste unilateral, α **não** se divide por 2

É o erro mecânico mais comum do capítulo. No teste bilateral há duas caudas a cobrir, e o α se reparte igualmente entre elas: $\alpha/2$ de cada lado, quantil $z_{\alpha/2}$. No teste unilateral há **uma cauda só**, e ela recebe o α **inteiro**: quantil z_{α}.
Concretamente, a $\alpha = 0{,}05$: bilateral usa $z_{0{,}025} = 1{,}96$; unilateral usa $z_{0{,}05} = 1{,}645$. O valor crítico unilateral é **menor**, e portanto o teste unilateral é **mais fácil de rejeitar** naquela direção — o que é justo, porque ele abriu mão de detectar desvios na direção oposta.

A escolha entre z e t segue exatamente a regra de Seção 9: usa-se z quando σ é **conhecido**, e t com $n-1$ graus de liberdade quando σ é **desconhecido** e estimado por s. Como a t tem caudas mais pesadas que a normal, seus valores críticos são **maiores**, e o teste com t é **mais conservador** — rejeita menos. É o preço de ter de estimar mais um parâmetro.

10.3.3 Método 3: o p-valor

Os dois primeiros métodos dão uma resposta binária: rejeita ou não rejeita, dado o α que se escolheu. Perde-se com isso uma informação relevante — a de *quão perto* a decisão esteve de mudar. Uma estatística $z_0 = 1{,}97$ e uma $z_0 = 8{,}4$ ambas rejeitam a 5%, mas são evidências de força incomparável. O p-valor recupera essa informação.

! Definição — p-valor

O **p-valor** (ou **nível descritivo**) é o **menor nível de significância** α que levaria à rejeição de H_0 com os dados observados.
Equivalentemente, e é assim que se calcula: é a probabilidade, **calculada sob** H_0, de a estatística de teste assumir um valor **igual ou mais extremo** do que o efetivamente observado.
A regra de decisão é

$$\boxed{p \le \alpha \implies \textbf{rejeita-se } H_0; \qquad p > \alpha \implies \textbf{não se rejeita } H_0.}$$

"Mais extremo" significa *na direção de* H_1, e por isso o cálculo depende do tipo de teste. Para a estatística observada z_0:

$$\begin{aligned} H_1 : \mu > k &\implies p = P(Z \ge z_0), \\ H_1 : \mu < k &\implies p = P(Z \le z_0), \\ H_1 : \mu \ne k &\implies p = 2\,P(Z \ge |z_0|). \end{aligned}$$

No caso bilateral, **multiplica-se por 2** o p-valor unilateral, porque "mais extremo" inclui as duas caudas: um afastamento de mesma magnitude no sentido oposto seria igualmente contrário a H_0.

As duas definições — "menor α que rejeita" e "probabilidade do que é igual ou mais extremo" — são equivalentes, e vale ver por quê. À medida que α diminui, o valor crítico z_α aumenta e a região crítica encolhe. Existe um α exato em que a fronteira da RC coincide com a estatística observada; abaixo dele, a estatística fica fora e não se rejeita mais. Esse α de fronteira é o p-valor, e a área de cauda além da estatística observada é precisamente ele.

A vantagem prática é clara: reportando o p-valor, o leitor decide com o α que julgar apropriado, sem precisar dos seus dados. É por isso que todo software estatístico reporta p-valores em vez de "rejeita/não rejeita" — e é por isso que o R será a nossa ferramenta natural de conferência no laboratório.

! O que o p-valor **não** é

O p-valor **não é** a probabilidade de H_0 ser verdadeira. Não é a probabilidade de a conclusão estar errada. Não é a probabilidade de o resultado ter ocorrido por acaso.
Todas essas leituras cometem o mesmo erro: inverter o condicionamento. O p-valor é $P(\text{dados tão ou mais extremos} \mid H_0)$; as leituras erradas querem $P(H_0 \mid \text{dados})$. São objetos diferentes — a diferença entre $P(A \mid B)$ e $P(B \mid A)$ de Seção 3 —, e nada na estatística clássica calcula o segundo, que exigiria atribuir uma probabilidade *a priori* a H_0 e é, portanto, terreno bayesiano. Guarde a fórmula: o p-valor condiciona em H_0, sempre.

10.3.4 Monotonicidade da decisão

Uma propriedade simples e muito útil, que decorre imediatamente do p-valor e que o professor destaca:

! Monotonicidade

Se rejeita-se H_0 ao nível α, rejeita-se a todos os níveis **superiores**;

Se **não** se rejeita H_0 ao nível α, não se rejeita a todos os níveis **inferiores**.

A demonstração cabe em uma linha: a regra é $p \le \alpha$, e se $p \le \alpha$ então $p \le \alpha'$ para todo $\alpha' > \alpha$; se $p > \alpha$ então $p > \alpha'$ para todo $\alpha' < \alpha$.

Geometricamente, é o encolhimento da região crítica: aumentar α **amplia** a RC, de modo que uma estatística que já estava dentro continua dentro; diminuir α a **encolhe**, de modo que uma estatística que já estava fora continua fora.

Na prática, isso permite responder a várias perguntas de uma vez. Se um exercício informa que rejeitamos a 5%, sabemos sem contas que também rejeitaríamos a 10%. Se não rejeitamos a 5%, não rejeitaríamos a 1%. O que **não** se pode inferir é a direção contrária: rejeitar a 10% nada diz sobre 5%, e é justamente aí que o p-valor resolve tudo de uma vez, ao localizar exatamente o ponto de virada.

10.4 Teste para a média

Com o aparato montado, passemos aos casos. A estrutura é sempre a mesma — hipóteses, estatística, distribuição sob H_0, região crítica, decisão —, e o que muda de teste para teste é apenas a estatística e sua distribuição.

! Teste para a média

Hipóteses: $H_0 : \mu = k$ contra $H_1 : \mu \neq k$ (ou as versões unilaterais).

$$\boxed{\sigma \text{ conhecido: } z_0 = \frac{\bar{x} - k}{\sigma/\sqrt{n}} \sim N(0,1) \qquad \sigma \text{ desconhecido: } t_0 = \frac{\bar{x} - k}{s/\sqrt{n}} \sim t_{n-1}}$$

com região crítica bilateral $(-\infty, -c_{\alpha/2}] \cup [c_{\alpha/2}, +\infty)$, onde c é o quantil da distribuição correspondente.

10.4.1 Exemplo: alturas (σ conhecido, bilateral)

Retomamos as cinco alturas de Seção 8, que deram $\bar{x} = 180$ cm. Suponha $\sigma = 6$ conhecido, e testemos, ao nível $\alpha = 0{,}05$,

$$H_0 : \mu = 175 \qquad \text{contra} \qquad H_1 : \mu \neq 175.$$

Pelo método do intervalo de confiança. O intervalo de 95% para μ com σ conhecido é

$$\text{IC}(\mu; 95\%) = \bar{x} \pm z_{0{,}025}\,\frac{\sigma}{\sqrt{n}} = 180 \pm 1{,}96 \cdot \frac{6}{\sqrt{5}} = 180 \pm 5{,}2591 = [\,174{,}74\,;\,185{,}26\,].$$

Como $175 \in [174{,}74; 185{,}26]$, **não se rejeita** H_0. Note como foi por pouco: o valor hipotetizado está a menos de 0,3 cm da borda inferior.

Pelo método da região crítica. A estatística de teste é

$$z_0 = \frac{\bar{x} - k}{\sigma/\sqrt{n}} = \frac{180 - 175}{6/\sqrt{5}} = \frac{5}{2{,}6833} = 1{,}8634,$$

e a região crítica bilateral a 5% é RC $= (-\infty; -1{,}96\,] \cup [\,1{,}96; +\infty)$. Como $1{,}8634 \notin$ RC, **não se rejeita** H_0 — a mesma conclusão, como tinha de ser.

Com σ desconhecido. Suponha agora que o 6 não seja o desvio padrão populacional, mas o **amostral** $s = 6$. A estatística tem o mesmo valor numérico, $t_0 = 1{,}8634$, mas a distribuição de referência muda: t com $n - 1 = 4$ graus de liberdade, cujo quantil é $t_{4;0{,}025} = 2{,}7764$. A região crítica passa a RC $= (-\infty; -2{,}7764\,] \cup [\,2{,}7764; +\infty)$, e de novo não se rejeita.

A comparação entre os dois cenários é instrutiva: a mesma estatística, a mesma conclusão, mas por margens muito diferentes. O valor crítico saltou de 1,96 para 2,78 — a região crítica **encolheu** substancialmente. Com $n = 5$, desconhecer σ custa caro, e o teste fica bem mais conservador. É o mesmo alargamento que o intervalo de confiança sofre em Seção 9, visto do outro lado.

10.4.2 Exemplo: salários (σ desconhecido, bilateral)

Uma amostra de $n = 25$ profissionais registrou salário médio $\bar{x} = 400$ com desvio padrão amostral $s = 450$. Um sindicato afirma que o salário médio da categoria é 600. Testemos, a $\alpha = 0{,}10$,

$$H_0 : \mu = 600 \qquad \text{contra} \qquad H_1 : \mu \neq 600.$$

O erro padrão é $s/\sqrt{n} = 450/5 = 90$, e o quantil é $t_{24;0{,}05} = 1{,}7109$.

Pelo intervalo. IC$(\mu; 90\%) = 400 \pm 1{,}7109 \times 90 = [\,246{,}02\,;\,553{,}98\,]$. Como $600 \notin$ IC, **rejeita-se** H_0: aos dados, a afirmação do sindicato é implausível.

Pela região crítica.

$$t_0 = \frac{400 - 600}{90} = -2{,}2222,$$

e RC $= (-\infty; -1{,}7109\,] \cup [\,1{,}7109; +\infty)$. Como $-2{,}2222 \in$ RC, **rejeita-se** H_0. As duas conclusões coincidem, como devem.

O p-valor é $2\,P(T_{24} \leq -2{,}2222) = 0{,}0359$. Pela monotonicidade, sabemos agora tudo de uma vez: rejeitaríamos também a $\alpha = 0{,}05$ (pois $0{,}0359 \leq 0{,}05$), mas **não** a $\alpha = 0{,}01$.

10.4.3 Exemplo: TVs

Uma amostra de $n = 16$ televisores tem preço médio $\bar{x} = 8900$ com $s = 500$. Testemos $H_0 : \mu = 9000$ contra $H_1 : \mu \neq 9000$ a $\alpha = 0{,}01$. Com $t_{15;0{,}005} = 2{,}9467$ e erro padrão $500/4 = 125$:

$$\text{IC}(\mu; 99\%) = 8900 \pm 2{,}9467 \times 125 = [\,8531{,}66\,;\,9268{,}34\,].$$

Como $9000 \in$ IC, **não se rejeita** H_0. Pela região crítica, $t_0 = (8900 - 9000)/125 = -0{,}80$, muito longe de RC $= (-\infty; -2{,}9467] \cup [2{,}9467; \infty)$ — mesma conclusão, e desta vez com folga larguíssima ($p = 0{,}4362$).

10.4.4 Exemplo: nicotina (unilateral, com p-valor)

Um fabricante garante que seus cigarros contêm no máximo 30 mg de nicotina. Um laboratório suspeita que contenham **mais**, e é isso que ele quer evidenciar. Logo

$$H_0 : \mu \leq 30 \qquad \text{contra} \qquad H_1 : \mu > 30.$$

Com $n = 25$, $\bar{x} = 31{,}5$, $\sigma = 3$ conhecido e $\alpha = 0{,}05$, a estatística é

$$z_0 = \frac{31{,}5 - 30}{3/\sqrt{25}} = \frac{1{,}5}{0{,}6} = 2{,}5.$$

A região crítica é **unilateral à direita**, com o α inteiro numa cauda só: $\text{RC} = [\,1{,}645; +\infty)$. Como $2{,}5 \in \text{RC}$, **rejeita-se** H_0: há evidência de que o teor médio excede 30 mg.

O p-valor é a área de cauda além da estatística observada,

$$p = P(Z \geq 2{,}5) = 1 - \Phi(2{,}5) = 0{,}0062,$$

confortavelmente menor que 0,05 — e menor até que 0,01, de modo que rejeitaríamos também ao nível mais exigente.

Vale registrar quanto custaria a formulação bilateral. Se as hipóteses fossem $H_0 : \mu = 30$ contra $H_1 : \mu \neq 30$, o p-valor seria $2 \times 0{,}00621 = 0{,}01242$ — o dobro. Continuaria abaixo de 0,05, mas passaria a ficar **acima** de 0,01, e a conclusão ao nível de 1% se inverteria. É a ilustração exata de que o teste unilateral é mais poderoso na direção que ele escolheu observar.

10.4.5 Exemplo: endividamento (a monotonicidade em ação)

Sobre o endividamento médio de uma carteira de clientes, testa-se $H_0 : \mu \leq 30$ contra $H_1 : \mu > 30$, com $n = 9$, $\sigma = 30$ e $\bar{x} = 45$:

$$z_0 = \frac{45 - 30}{30/\sqrt{9}} = \frac{15}{10} = 1{,}5.$$

Comparemos com os três níveis usuais:

α	z_α	RC	$z_0 = 1{,}5 \in$ RC?	Decisão
0,01	2,3263	$[2{,}3263; \infty)$	não	não rejeita
0,05	1,6449	$[1{,}6449; \infty)$	não	não rejeita
0,10	1,2816	$[1{,}2816; \infty)$	**sim**	**rejeita**

Rejeita-se apenas ao nível de 10%. O p-valor confirma e explica tudo: $p = P(Z \geq 1{,}5) = 0{,}0668$, que está entre 0,05 e 0,10. A monotonicidade é visível na tabela: à medida que α cresce, a RC se amplia (a fronteira desce), e há um ponto — exatamente $\alpha = 0{,}0668$ — em que ela alcança a estatística observada.

10.4.6 Exemplo: fundos de investimento

Um fundo teve, em $n = 16$ meses, retorno médio $\bar{x} = -1\%$ com desvio padrão amostral $s = 2\%$. Queremos evidenciar que o retorno médio é **negativo**:

$$H_0 : \mu \geq 0 \qquad \text{contra} \qquad H_1 : \mu < 0, \qquad t_0 = \frac{-1 - 0}{2/\sqrt{16}} = \frac{-1}{0{,}5} = -2.$$

Com $n - 1 = 15$ graus de liberdade, e regiões críticas unilaterais à esquerda:

α	$t_{15;\alpha}$	RC	$t_0 = -2 \in$ RC?	Decisão
0,01	2,6025	$(-\infty; -2{,}6025]$	não	não rejeita
0,05	1,7531	$(-\infty; -1{,}7531]$	**sim**	**rejeita**
0,10	1,3406	$(-\infty; -1{,}3406]$	**sim**	**rejeita**

Rejeita-se a 5% e a 10%, mas não a 1%. O p-valor é $P(T_{15} \leq -2) = 0{,}0320$, entre 0,01 e 0,05 — precisamente o que a tabela mostra. Note mais uma vez a monotonicidade: rejeitou a 5%, então rejeita a 10%; não rejeitou a 1%, então não rejeitaria a 0,5%.

10.5 Teste para a proporção

Para uma proporção populacional p, o estimador é $\hat{p} = X/n$ e, pelo Teorema Central do Limite (Seção 7), sua distribuição é aproximadamente normal para n razoavelmente grande.

! Teste para a proporção

Hipóteses: $H_0 : p = k$ contra $H_1 : p \neq k$ (ou unilaterais).

$$\boxed{z_0 = \frac{\hat{p} - k}{\sqrt{\dfrac{k(1-k)}{n}}} \;\sim\; N(0,1) \;\text{ sob } H_0}$$

com as mesmas regiões críticas do teste da média com σ conhecido.

! Atenção: o denominador usa k, não $\hat{p}$

Aqui há uma **assimetria** em relação a Seção 9 que é fonte constante de erro.
No **intervalo de confiança** para uma proporção, o erro padrão é estimado com $\hat{p}$:

$$\text{IC}(p) = \hat{p} \pm z_{\alpha/2}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}.$$

No **teste de hipóteses**, o erro padrão usa o valor **hipotetizado** k:

$$z_0 = \frac{\hat{p} - k}{\sqrt{k(1-k)/n}}.$$

A razão é conceitual, e é a mesma que rege todo o método da região crítica: **a estatística de teste é calculada supondo H_0 verdadeira**. Se H_0 diz que $p = k$, então a variância de $\hat{p}$ sob H_0 é $k(1-k)/n$, e é essa que devemos usar — não há por que estimar o que a hipótese já especifica. No intervalo de confiança não há hipótese nenhuma a supor, e só resta estimar. Consequência prática, e é ela que confunde: para proporções, o método do intervalo e o método da região crítica podem, em casos de fronteira, **discordar** ligeiramente, porque usam denominadores diferentes. Não é contradição lógica — são duas aproximações distintas do mesmo procedimento assintótico. Nos exercícios do curso, siga sempre a fórmula com k para o teste.

10.5.1 Exemplo: clientes avessos ao risco

Uma corretora afirma que 30% de seus clientes são avessos ao risco. Numa amostra de $n = 64$ clientes, 20 se revelaram avessos, de modo que $\hat{p} = 20/64 = 0{,}3125$. Testemos, a $\alpha = 0{,}10$,

$$H_0 : p = 0{,}30 \qquad \text{contra} \qquad H_1 : p \neq 0{,}30.$$

Com $k = 0{,}30$, o erro padrão **sob** H_0 é

$$\sqrt{\frac{k(1-k)}{n}} = \sqrt{\frac{0{,}3 \times 0{,}7}{64}} = \sqrt{0{,}00328125} = 0{,}05728,$$

e portanto

$$z_0 = \frac{0{,}3125 - 0{,}30}{0{,}05728} = \frac{0{,}0125}{0{,}05728} = 0{,}2182.$$

A região crítica bilateral a 10% é $\text{RC} = (-\infty; -1{,}6449] \cup [1{,}6449; \infty)$. Como $0{,}2182 \notin \text{RC}$, **não se rejeita** H_0: os dados são inteiramente compatíveis com a afirmação da corretora. O p-valor é $2P(Z \geq 0{,}2182) = 0{,}8273$ — praticamente nenhuma evidência contra H_0.

(A título de comparação, se tivéssemos usado $\hat{p}$ no denominador obteríamos $z_0 = 0{,}2157$ em vez de 0,2182. A diferença é irrelevante aqui, mas registre que ela existe.)

10.5.2 Exemplo: audiência (unilateral, passo a passo)

Este exemplo merece a resolução completa, porque combina tudo: formulação de hipóteses unilaterais, região crítica de uma cauda e p-valor.

Uma emissora afirma que **pelo menos 60%** dos domicílios assistem a seu programa. Uma agência de publicidade duvida, e quer evidenciar que a audiência é **menor**. Numa amostra de $n = 400$ domicílios, 220 assistem.

Passo 1 — hipóteses. O que se quer evidenciar é $p < 0{,}6$; logo essa é a alternativa, e a afirmação da emissora — o *status quo* — é a nula:

$$H_0 : p \geq 0{,}60 \qquad \text{contra} \qquad H_1 : p < 0{,}60.$$

Passo 2 — nível e estatística. Com $\alpha = 0{,}05$ e $\hat{p} = 220/400 = 0{,}55$:

$$z_0 = \frac{\hat{p} - k}{\sqrt{k(1-k)/n}} = \frac{0{,}55 - 0{,}60}{\sqrt{\dfrac{0{,}6 \times 0{,}4}{400}}} = \frac{-0{,}05}{\sqrt{0{,}0006}} = \frac{-0{,}05}{0{,}024495} = -2{,}0412.$$

Passo 3 — região crítica. Teste unilateral **à esquerda**, com α inteiro na cauda esquerda:

$$\text{RC} = (-\infty;\, -z_{0{,}05}\,] = (-\infty;\, -1{,}6449\,].$$

Passo 4 — decisão. Como $-2{,}0412 \in \text{RC}$, **rejeita-se** H_0. Há evidência, ao nível de 5%, de que a audiência é inferior a 60% — a agência tem razão em duvidar.

Passo 5 — p-valor. Na direção de H_1 (cauda esquerda):

$$p = P(Z \leq -2{,}0412) = 0{,}0206,$$

menor que 0,05, o que confirma a rejeição. Note que $p > 0{,}01$: ao nível de 1% **não** rejeitaríamos. A evidência é boa, não esmagadora.

10.6 Teste para a variância

Quando o parâmetro de interesse é a dispersão — e em economia ele frequentemente é, porque dispersão é risco —, a distribuição de referência é a **qui-quadrado**, exatamente como no intervalo de confiança para σ^2 de Seção 9.

! Teste para a variância

Hipóteses: $H_0 : \sigma^2 = k$ contra $H_1 : \sigma^2 \neq k$ (ou unilaterais).

$$\boxed{q_0 = \frac{(n-1)\,s^2}{k} \;\sim\; \chi^2_{n-1} \;\text{ sob } H_0}$$

Região crítica bilateral, com as duas caudas:

$$\boxed{\text{RC} = (0;\; \chi^2_{n-1;\,1-\alpha/2}] \;\cup\; [\,\chi^2_{n-1;\,\alpha/2};\; +\infty)}$$

Duas observações sobre a mecânica. A primeira: a qui-quadrado é **assimétrica** e definida apenas em $(0, \infty)$, de modo que a região crítica **não** é simétrica — as duas caudas têm a mesma área $\alpha/2$, mas os valores críticos não são simétricos em torno de nada. A segunda: mantemos a convenção de Seção 9, em que $\chi^2_{\nu;\gamma}$ denota o valor que deixa área γ **à direita**. Assim $\chi^2_{n-1;\,\alpha/2}$ é o valor crítico **superior** e $\chi^2_{n-1;\,1-\alpha/2}$ o **inferior** — atenção à inversão dos índices, que é onde se erra.

10.6.1 Exemplo: variância de um processo

Uma amostra de $n = 30$ observações produziu $s = 3{,}63$. Testemos, a $\alpha = 0{,}05$,

$$H_0 : \sigma^2 = 16 \qquad \text{contra} \qquad H_1 : \sigma^2 \neq 16.$$

Com $n - 1 = 29$ graus de liberdade, os dois valores críticos são

$$\chi^2_{29;\,0{,}975} = 16{,}047 \qquad \text{e} \qquad \chi^2_{29;\,0{,}025} = 45{,}722,$$

de modo que $\text{RC} = (0;\, 16{,}047\,] \cup [\,45{,}722;\, +\infty)$. A estatística é

$$q_0 = \frac{(n-1)s^2}{k} = \frac{29 \times (3{,}63)^2}{16} = \frac{29 \times 13{,}1769}{16} = \frac{382{,}13}{16} = 23{,}883.$$

Como $23{,}883 \notin \text{RC}$ — está confortavelmente entre os dois valores críticos —, **não se rejeita** H_0. O p-valor bilateral é 0,5307: nenhuma evidência contra a hipótese de que $\sigma^2 = 16$.

Repare que $s^2 = 13{,}18$ é sensivelmente menor que o hipotetizado 16 — uma diferença de quase 18% —, e ainda assim o teste não rejeita. Estimar variâncias é difícil: a distribuição amostral de s^2 é larga, e são necessárias amostras grandes para distinguir dispersões que parecem bem diferentes. É uma lição prática relevante para quem trabalha com risco.

10.7 Teste para a diferença de médias

Passamos aos testes de **duas amostras**, que são os que respondem às perguntas comparativas: o tratamento funciona melhor que o controle? A produtividade das duas plantas difere? O curso trata o caso de **variâncias desconhecidas mas supostas iguais**, com variância combinada, exatamente como em Seção 9.

! Teste para a diferença de médias

Hipóteses: $H_0 : \mu_1 - \mu_2 = 0$ contra $H_1 : \mu_1 - \mu_2 \neq 0$ (ou unilaterais).

$$\boxed{\,t_0 = \frac{\bar{x}_1 - \bar{x}_2}{s_p\sqrt{\dfrac{1}{n_1} + \dfrac{1}{n_2}}} \;\sim\; t_{\,n_1+n_2-2} \;\text{ sob } H_0\,}$$

com a **variância combinada** (*pooled*)

$$\boxed{\,s_p^2 = \frac{(n_1-1)s_1^2 + (n_2-1)s_2^2}{n_1 + n_2 - 2}\,}$$

Graus de liberdade: $\nu = n_1 + n_2 - 2$.

As **três condições de validade** são as mesmas de Seção 9, e convém repeti-las porque a última é a que motiva o teste F da próxima seção:

1. as duas amostras são **independentes** entre si;
2. as populações são **normais** (ou os n são grandes o bastante para o TCL operar);
3. as **variâncias populacionais são iguais**, $\sigma_1^2 = \sigma_2^2$.

A terceira condição é o que autoriza combinar s_1^2 e s_2^2 em um único s_p^2: se as duas amostras estimam a *mesma* variância, faz sentido juntá-las numa estimativa mais precisa, ponderada pelos respectivos graus de liberdade. Se as variâncias fossem diferentes, a combinação não faria sentido algum — e é por isso que precisaremos de um teste para verificar esse pressuposto.

10.7.1 Exemplo: durabilidade de lâmpadas

Dois fornecedores de lâmpadas. Do fornecedor 1 testaram-se $n_1 = 8$ unidades, com duração média $\bar{x}_1 = 4600$ horas e $s_1 = 250$; do fornecedor 2, $n_2 = 10$ unidades, com $\bar{x}_2 = 4000$ e $s_2 = 200$. Testemos, a $\alpha = 0{,}01$, se as durações médias diferem:

$$H_0 : \mu_1 = \mu_2 \qquad \text{contra} \qquad H_1 : \mu_1 \neq \mu_2.$$

Variância combinada.

$$s_p^2 = \frac{(8-1)(250)^2 + (10-1)(200)^2}{8+10-2} = \frac{7 \times 62500 + 9 \times 40000}{16} = \frac{437500 + 360000}{16} = \frac{797500}{16} = 49843{,}75,$$

de onde $s_p = \sqrt{49843{,}75} = 223{,}26$ horas.

Estatística. Com $\sqrt{1/8 + 1/10} = \sqrt{0{,}225} = 0{,}47434$:

$$t_0 = \frac{4600 - 4000}{223{,}26 \times 0{,}47434} = \frac{600}{105{,}90} = 5{,}666.$$

Decisão. Com $\nu = 16$ graus de liberdade, $t_{16;0{,}005} = 2{,}9208$ e RC $= (-\infty; -2{,}9208] \cup [2{,}9208; \infty)$. Como $5{,}666 \in$ RC, **rejeita-se** H_0 com folga: há forte evidência de que as durações médias diferem. O p-valor é 0,0000351 — evidência esmagadora, e note que rejeitaríamos a qualquer nível usual.

10.7.2 Exemplo: duas turmas (a equivalência IC $\leftrightarrow$ teste)

Comparam-se as notas de duas turmas. Da turma 1, $n_1 = 5$ alunos com variância amostral $s_1^2 = 5{,}425$; da turma 2, $n_2 = 4$ alunos com $s_2^2 = 7$. A diferença das médias amostrais é $\bar{x}_1 - \bar{x}_2 = -0{,}1$. Testemos a $\alpha = 0{,}05$ se as médias diferem.

A variância combinada é

$$s_p^2 = \frac{4 \times 5{,}425 + 3 \times 7}{5+4-2} = \frac{21{,}7 + 21}{7} = \frac{42{,}7}{7} = 6{,}10, \qquad s_p = 2{,}4698,$$

com $\nu = n_1 + n_2 - 2 = 7$ graus de liberdade e $t_{7;0{,}025} = 2{,}3646$. O erro padrão da diferença é $s_p\sqrt{1/5 + 1/4} = 2{,}4698 \times 0{,}67082 = 1{,}6567$, e portanto

$$t_0 = \frac{-0{,}1}{1{,}6567} = -0{,}0604.$$

Pela região crítica. $\text{RC} = (-\infty; -2{,}3646] \cup [2{,}3646; \infty)$, e $-0{,}0604 \notin \text{RC}$: **não se rejeita** H_0.

Pelo intervalo de confiança.

$$\text{IC}(\mu_1 - \mu_2; 95\%) = (\bar{x}_1 - \bar{x}_2) \pm t_{7;0{,}025}\, s_p \sqrt{\tfrac{1}{5} + \tfrac{1}{4}} = -0{,}1 \pm 2{,}3646 \times 1{,}6567 = [\,-4{,}0182\,;\, 3{,}8182\,].$$

Como $0 \in [-4{,}0182; 3{,}8182]$, **não se rejeita** H_0.

> **!** A conclusão **tem** que ser igual
>
> Os dois métodos deram a mesma resposta, e isso não é coincidência nem sorte: é uma identidade algébrica. Para o teste **bilateral**, "k está no intervalo" e "a estatística está fora da região crítica" são a **mesma desigualdade**, escrita de duas maneiras.
> Veja, no caso da média com σ desconhecido:
>
> $$k \in \text{IC} \iff |\bar{x} - k| \le t_{n-1;\alpha/2} \frac{s}{\sqrt{n}} \iff \left|\frac{\bar{x} - k}{s/\sqrt{n}}\right| \le t_{n-1;\alpha/2} \iff |t_0| \le t_{n-1;\alpha/2} \iff t_0 \notin \text{RC}.$$
>
> Dividir os dois lados por $s/\sqrt{n}$ leva de uma formulação à outra. Portanto, **se os dois métodos discordarem, há erro de conta em algum dos dois** — não há caso em que a divergência seja legítima (à exceção do teste de proporção, pela assimetria de denominadores discutida acima, e dos testes unilaterais, aos quais o método do IC não se aplica). O laboratório verifica essa identidade numericamente.

O p-valor deste teste é $2P(T_7 \le -0{,}0604) = 0{,}9536$ — praticamente 1. As duas turmas são, para todos os efeitos, indistinguíveis nas médias.

10.8 Teste para a diferença de proporções

> **!** Teste para a diferença de proporções
>
> Hipóteses: $H_0 : p_1 - p_2 = 0$ contra $H_1 : p_1 - p_2 \neq 0$ (ou unilaterais).
>
> $$z_0 = \frac{\hat{p}_1 - \hat{p}_2}{\sqrt{\dfrac{\hat{p}_1(1-\hat{p}_1)}{n_1} + \dfrac{\hat{p}_2(1-\hat{p}_2)}{n_2}}} \sim N(0,1)$$
>
> Aqui o denominador usa $\hat{p}_1$ e $\hat{p}_2$ **separadamente** — não há combinação (*pooling*) das duas proporções.

Vale explicitar o contraste com o teste de uma proporção, porque ele parece inconsistente e não é. Lá, H_0 especificava um valor numérico k para p, e usamos esse valor no denominador. Aqui, H_0

especifica apenas que $p_1 = p_2$, **sem dizer qual é o valor comum** — e como não há número a usar, resta estimar as duas variâncias separadamente a partir dos dados. A regra por trás dos dois casos é a mesma: use sob H_0 tudo que H_0 fornece, e estime o que ela não fornece. (Há uma variante *pooled* que estima o valor comum por $\hat{p} = (x_1 + x_2)/(n_1 + n_2)$; ela existe na literatura, mas **não é a do curso**.)

10.8.1 Exemplo: inadimplência em duas financeiras

Na financeira 1, 140 de $n_1 = 180$ contratos foram quitados em dia; na financeira 2, 220 de $n_2 = 300$. Testemos, a $\alpha = 0{,}10$, se as taxas diferem.

$$\hat{p}_1 = \frac{140}{180} = 0{,}7778, \qquad \hat{p}_2 = \frac{220}{300} = 0{,}7333, \qquad \hat{p}_1 - \hat{p}_2 = 0{,}0444.$$

O denominador é

$$\sqrt{\frac{0{,}7778 \times 0{,}2222}{180} + \frac{0{,}7333 \times 0{,}2667}{300}} = \sqrt{0{,}00096022 + 0{,}00065185} = \sqrt{0{,}00161207} = 0{,}040151,$$

de onde

$$z_0 = \frac{0{,}0444}{0{,}040151} = 1{,}1064.$$

Com $\mathrm{RC} = (-\infty; -1{,}6449] \cup [1{,}6449; \infty)$ e $1{,}1064 \notin \mathrm{RC}$, **não se rejeita** H_0: as taxas de quitação não diferem significativamente.

Versão unilateral. Se quiséssemos evidenciar especificamente que $p_1 > p_2$, teríamos $H_1 : p_1 > p_2$ e $\mathrm{RC} = [\, z_{0,10}; \infty) = [\, 1{,}2816; \infty)$ — note que a fronteira, com α inteiro numa cauda, é **menor** que a bilateral. Ainda assim $1{,}1064 \notin \mathrm{RC}$, e a conclusão não muda. O p-valor unilateral é $P(Z \geq 1{,}1064) = 0{,}1342$, maior que 0,10: o teste unilateral era mais fácil de rejeitar e mesmo assim não rejeitou.

10.9 Teste para a razão de variâncias (teste F)

Chegamos ao último teste do curso, e ele existe por uma razão muito concreta: **verificar o pressuposto do teste de diferença de médias**. A terceira condição de validade daquele teste era $\sigma_1^2 = \sigma_2^2$, e até aqui a supusemos verdadeira sem checar. O teste F é a checagem.

A ideia é natural. Se as duas variâncias populacionais são iguais, sua **razão** é 1, e as variâncias amostrais devem ser parecidas — sua razão deve ficar perto de 1. A distribuição dessa razão, sob normalidade, é a F de Snedecor.

! Teste para a razão de variâncias

Hipóteses: $H_0 : \sigma_1^2 = \sigma_2^2$ contra $H_1 : \sigma_1^2 \neq \sigma_2^2$ — equivalentemente, $H_0 : \sigma_1^2/\sigma_2^2 = 1$.

$$\boxed{f_0 = \frac{s_1^2}{s_2^2} \;\sim\; F_{n_1-1,\; n_2-1} \;\text{ sob } H_0}$$

Região crítica bilateral:

$$\boxed{\text{RC} = (0;\; f_{n_1-1,\,n_2-1;\,1-\alpha/2}] \;\cup\; [\, f_{n_1-1,\,n_2-1;\,\alpha/2};\; +\infty)}$$

com $f_{\nu_1,\nu_2;\gamma}$ denotando o valor que deixa área γ **à direita** numa F com ν_1 graus de liberdade no numerador e ν_2 no denominador.

i Ferramenta matemática — a inversão da F e a troca dos graus de liberdade

As tabelas impressas de F trazem apenas as caudas superiores (γ pequeno), porque imprimir as duas caudas dobraria seu tamanho. O quantil inferior obtém-se pela identidade

$$\boxed{f_{\nu_1,\nu_2;\,1-\alpha/2} = \frac{1}{f_{\nu_2,\nu_1;\,\alpha/2}}}$$

Note que os graus de liberdade **trocam de posição** na inversão. A razão é imediata: se $F = \frac{S_1^2/\sigma_1^2}{S_2^2/\sigma_2^2} \sim F_{\nu_1,\nu_2}$, então $1/F \sim F_{\nu_2,\nu_1}$ — inverter a razão inverte os papéis de numerador e denominador. Passando essa relação para os quantis, $P(F \le c) = P(1/F \ge 1/c)$, e o quantil inferior de F_{ν_1,ν_2} é o inverso do quantil superior de F_{ν_2,ν_1}.
Inverter sem trocar os graus de liberdade é erro, e é o mais comum deste teste. Salvo o caso $\nu_1 = \nu_2$, em que a troca é inócua, o valor resultante estará errado.

10.9.1 Exemplo: as variâncias das duas turmas

Voltemos ao exemplo das turmas, onde supusemos $\sigma_1^2 = \sigma_2^2$ para aplicar o teste de diferença de médias. Vamos agora testar esse pressuposto, a $\alpha = 0{,}05$. Os dados: $n_1 = 5$ com $s_1^2 = 5{,}425$; $n_2 = 4$ com $s_2^2 = 7$.

Estatística.

$$f_0 = \frac{s_1^2}{s_2^2} = \frac{5{,}425}{7} = 0{,}775.$$

Valores críticos. Com $\nu_1 = n_1 - 1 = 4$ e $\nu_2 = n_2 - 1 = 3$:

$$f_{4,3;\,0{,}025} = 15{,}101 \qquad \text{(cauda superior, direto da tabela)},$$

$$f_{4,3;\,0{,}975} = \frac{1}{f_{3,4;\,0{,}025}} = \frac{1}{9{,}9792} = 0{,}1002 \qquad \text{(cauda inferior, pela inversão } \textbf{com troca} \text{ de g.l.)}.$$

Logo RC $= (0;\ 0{,}1002\,] \cup [\,15{,}101;\ +\infty)$.

Decisão. Como $f_0 = 0{,}775 \notin$ RC, **não se rejeita** H_0: não há evidência de que as variâncias difiram. O pressuposto do teste de diferença de médias da seção anterior está **validado**, e a análise que fizemos lá é legítima. O p-valor bilateral é 0,7844.

Observe a largura da região de não-rejeição: qualquer razão entre 0,10 e 15,10 passa no teste. Com amostras de tamanho 5 e 4, o teste F praticamente não tem poder — só detectaria diferenças brutais de variância. É a mesma lição do teste de variância: dispersão é um parâmetro difícil de estimar, e mais ainda de comparar. Na prática, "não rejeitar variâncias iguais" com n pequeno é evidência fraquíssima a favor do pressuposto, e o resultado deve ser lido com essa reserva.

! Erros conceituais frequentes

O professor lista explicitamente os enganos abaixo. Todos aparecem em provas, todos são evitáveis.
Sobre o método do intervalo de confiança:

1. **Não rejeitar porque a *estimativa* pertence ao intervalo.** A estimativa $\bar{x}$ pertence ao intervalo **sempre**, por construção — o intervalo é $\bar{x} \pm$ (margem), e está centrado nela. Verificar isso não é verificar coisa alguma. O que se testa é se o **valor hipotetizado** k pertence ao intervalo.
2. **Não rejeitar porque t_0 (ou z_0) pertence ao intervalo.** Isto mistura os dois métodos e não faz sentido dimensional: a estatística de teste é um número padronizado, adimensional, na escala da t ou da normal, ao passo que o intervalo está na escala do parâmetro (reais, centímetros, horas). Comparar $t_0 = -2{,}22$ com o intervalo $[246; 554]$ é comparar coisas incomparáveis. A estatística se compara com a **região crítica**; o valor hipotetizado, com o **intervalo**. Nunca cruze.

Sobre o teste F:

3. **Confundir "$1 \in$ RC" com "$1 \in$ IC".** São critérios de métodos diferentes e apontam em sentidos **opostos**: pelo intervalo, rejeita-se quando $1 \notin$ IC; pela região crítica, rejeita-se quando $f_0 \in$ RC. Trocar um pelo outro inverte a conclusão.
4. **Inverter as variâncias amostrais no cálculo de f_0.** Fixe a ordem no início e não a mude: $f_0 = s_1^2/s_2^2$, com o índice 1 sendo o que você chamou de amostra 1.
5. **Trocar numerador e denominador ao consultar a tabela.** $f_{4,3}$ e $f_{3,4}$ são valores diferentes (15,101 e 9,979 no exemplo); a primeira posição é sempre o g.l. do **numerador**.
6. **Inverter sem trocar os graus de liberdade.** Como visto na caixa anterior, $f_{\nu_1,\nu_2;1-\alpha/2} = 1/f_{\nu_2,\nu_1;\alpha/2}$, com a troca. Usar $1/f_{\nu_1,\nu_2;\alpha/2}$ daria $1/15{,}101 = 0{,}0662$ em vez do correto 0,1002 — uma região crítica errada.

E a assimetria que fecha o capítulo: em Seção 9, o intervalo de confiança é construído para a razão σ_2^2/σ_1^2 (na ordem invertida, por conveniência algébrica da derivação), enquanto aqui a estatística do teste é $f_0 = s_1^2/s_2^2$ (na ordem direta). As duas convenções coexistem, e é preciso ter cuidado ao passar de uma à outra — sobretudo em testes unilaterais, onde a inversão troca o lado da alternativa.

10.10 Tabela-resumo dos testes

Reunimos tudo. Esta é a tabela de prova: dado o problema, identifique o parâmetro, leia a linha, monte o teste. Em todas as linhas, k é o valor hipotetizado e o quantil indicado é o que deixa a área correspondente **à direita** na distribuição de referência; as regiões críticas listadas são as **bilaterais**, e para as unilaterais basta usar α inteiro em uma cauda só — a do lado de H_1.

Parâmetro	**Hipóteses**	**Estatística de teste**	**Distribuição**	**Região crítica (bilateral)**
Média, σ conhecido	$H_0 : \mu = k$ $H_1 : \mu \neq k$	$z_0 = \dfrac{\bar{x} - k}{\sigma/\sqrt{n}}$	$N(0,1)$	$(-\infty, -z_{\alpha/2}] \cup [z_{\alpha/2}, \infty)$
Média, σ desconhecido	$H_0 : \mu = k$ $H_1 : \mu \neq k$	$t_0 = \dfrac{\bar{x} - k}{s/\sqrt{n}}$	t_{n-1}	$(-\infty, -t_{n-1;\alpha/2}] \cup [t_{n-1;\alpha/2}, \infty)$
Proporção	$H_0 : p = k$ $H_1 : p \neq k$	$z_0 = \dfrac{\hat{p} - k}{\sqrt{k(1-k)/n}}$ (denominador com k!)	$N(0,1)$	$(-\infty, -z_{\alpha/2}] \cup [z_{\alpha/2}, \infty)$
Variância	$H_0 : \sigma^2 = k$ $H_1 : \sigma^2 \neq k$	$q_0 = \dfrac{(n-1)s^2}{k}$	χ^2_{n-1}	$(0, \chi^2_{n-1;1-\alpha/2}] \cup [\chi^2_{n-1;\alpha/2}, \infty)$
Diferença de médias	$H_0 : \mu_1 = \mu_2$ $H_1 : \mu_1 \neq \mu_2$	$t_0 = \dfrac{\bar{x}_1 - \bar{x}_2}{s_p\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}$ $s_p^2 = \frac{(n_1-1)s_1^2 + (n_2-1)s_2^2}{n_1+n_2-2}$	$t_{n_1+n_2-2}$	$(-\infty, -t_{\nu;\alpha/2}] \cup [t_{\nu;\alpha/2}, \infty)$, $\nu = n_1 + n_2 - 2$
Diferença de proporções	$H_0 : p_1 = p_2$ $H_1 : p_1 \neq p_2$	$z_0 = \dfrac{\hat{p}_1 - \hat{p}_2}{\sqrt{v}}$, com $v = \frac{\hat{p}_1(1-\hat{p}_1)}{n_1} + \frac{\hat{p}_2(1-\hat{p}_2)}{n_2}$ (sem *pooling*)	$N(0,1)$	$(-\infty, -z_{\alpha/2}] \cup [z_{\alpha/2}, \infty)$
Razão de variâncias	$H_0 : \sigma_1^2 = \sigma_2^2$ $H_1 : \sigma_1^2 \neq \sigma_2^2$	$f_0 = \dfrac{s_1^2}{s_2^2}$	$F_{n_1-1,\, n_2-1}$	$(0, f_{\nu_1,\nu_2;1-\alpha/2}] \cup [f_{\nu_1,\nu_2;\alpha/2}, \infty)$ com $f_{\nu_1,\nu_2;1-\alpha/2} = 1/f_{\nu_2,\nu_1;\alpha/2}$ (troca de g.l.)

Três lembretes que a tabela não comporta:

- **Unilateral não divide α por 2.** Substitua $\alpha/2$ por α e use uma cauda só — a do lado de H_1.
- **O método do IC só serve para testes bilaterais**, e para a proporção pode divergir ligeiramente pela assimetria dos denominadores.
- **p-valor bilateral $= 2\times$ unilateral**, sempre calculado sob H_0 na direção de H_1.

10.11 Laboratório computacional em R

O R torna este capítulo quase trivial de conferir: as quatro funções `pnorm`, `pt`, `pchisq` e `pf` (com suas versões `q*` para quantis) cobrem todos os testes da tabela. Mas o laboratório serve a mais que conferência: as três primeiras subseções tornam visíveis os conceitos — erros, poder, taxa de rejeição — que a álgebra apenas descreve.

10.11.1 Os dois erros, visualizados

Esta é a figura central do capítulo. Considere o teste $H_0 : \mu = 100$ contra $H_1 : \mu > 100$, com $\sigma = 15$ conhecido e $n = 25$, de modo que o erro padrão é $\sigma/\sqrt{n} = 3$. A regra de decisão rejeita quando $\bar{X}$ excede o valor crítico $\bar{x}_c = 100 + z_\alpha \cdot 3$.

O gráfico sobrepõe **duas** densidades: a de $\bar{X}$ sob H_0 (centrada em 100) e a de $\bar{X}$ sob uma alternativa específica $\mu_1 = 106$. A área da cauda direita da primeira, além do valor crítico, é α; a área da cauda esquerda da segunda, aquém dele, é β.

```
mu0 <- 100; mu1 <- 106; sigma <- 15; n <- 25
ep  <- sigma / sqrt(n)                       # erro padrao de Xbar
crit <- function(a) mu0 + qnorm(1 - a) * ep  # valor critico de Xbar

tab_erros <- do.call(rbind, lapply(c(0.01, 0.05, 0.10), function(a) {
  xc <- crit(a)
  data.frame(alpha = a,
             xbar_critico = xc,
             alpha_real   = 1 - pnorm(xc, mu0, ep),   # confere: deve dar alpha
             beta         = pnorm(xc, mu1, ep),
             poder        = 1 - pnorm(xc, mu1, ep))
}))
round(tab_erros, 4)
```

```
  alpha xbar_critico alpha_real   beta  poder
1  0.01     106.9790       0.01 0.6279 0.3721
2  0.05     104.9346       0.05 0.3612 0.6388
3  0.10     103.8447       0.10 0.2362 0.7638
```

A coluna `alpha_real` reproduz exatamente o α pedido — é a verificação de que a região crítica foi construída corretamente. E a tabela exibe o *trade-off* da caixa conceitual em números: ao reduzir α de 0,10 para 0,01, o valor crítico sobe de 103,84 para 106,98, e β salta de 0,236 para 0,628 — o poder despenca de 76% para 37%. Note também que $\alpha + \beta$ dá 0,638, 0,411 e 0,336 nas três linhas: **não somam 1**, nem somam nada em particular, exatamente como advertido.

```
grade <- seq(88, 118, length.out = 800)

painel <- function(a, nn, rotulo) {
  epn <- sigma / sqrt(nn)
  xc  <- mu0 + qnorm(1 - a) * epn
  d0  <- data.frame(x = grade, y = dnorm(grade, mu0, epn))
  d1  <- data.frame(x = grade, y = dnorm(grade, mu1, epn))
  ggplot() +
    geom_area(data = subset(d0, x >= xc), aes(x, y), fill = vm, alpha = 0.55) +
    geom_area(data = subset(d1, x <= xc), aes(x, y), fill = lr, alpha = 0.45) +
    geom_line(data = d0, aes(x, y), colour = az, linewidth = 0.7) +
    geom_line(data = d1, aes(x, y), colour = vd, linewidth = 0.7) +
```

```
    geom_vline(xintercept = xc, colour = cz, linetype = "dashed") +
    labs(x = rotulo, y = "densidade")
}

p1 <- painel(0.10, 25, "alpha = 0,10   (n = 25)")
p2 <- painel(0.05, 25, "alpha = 0,05   (n = 25)")
p3 <- painel(0.01, 25, "alpha = 0,01   (n = 25)")
p4 <- painel(0.05, 100, "alpha = 0,05   (n = 100)")
(p1 + p2) / (p3 + p4)
```

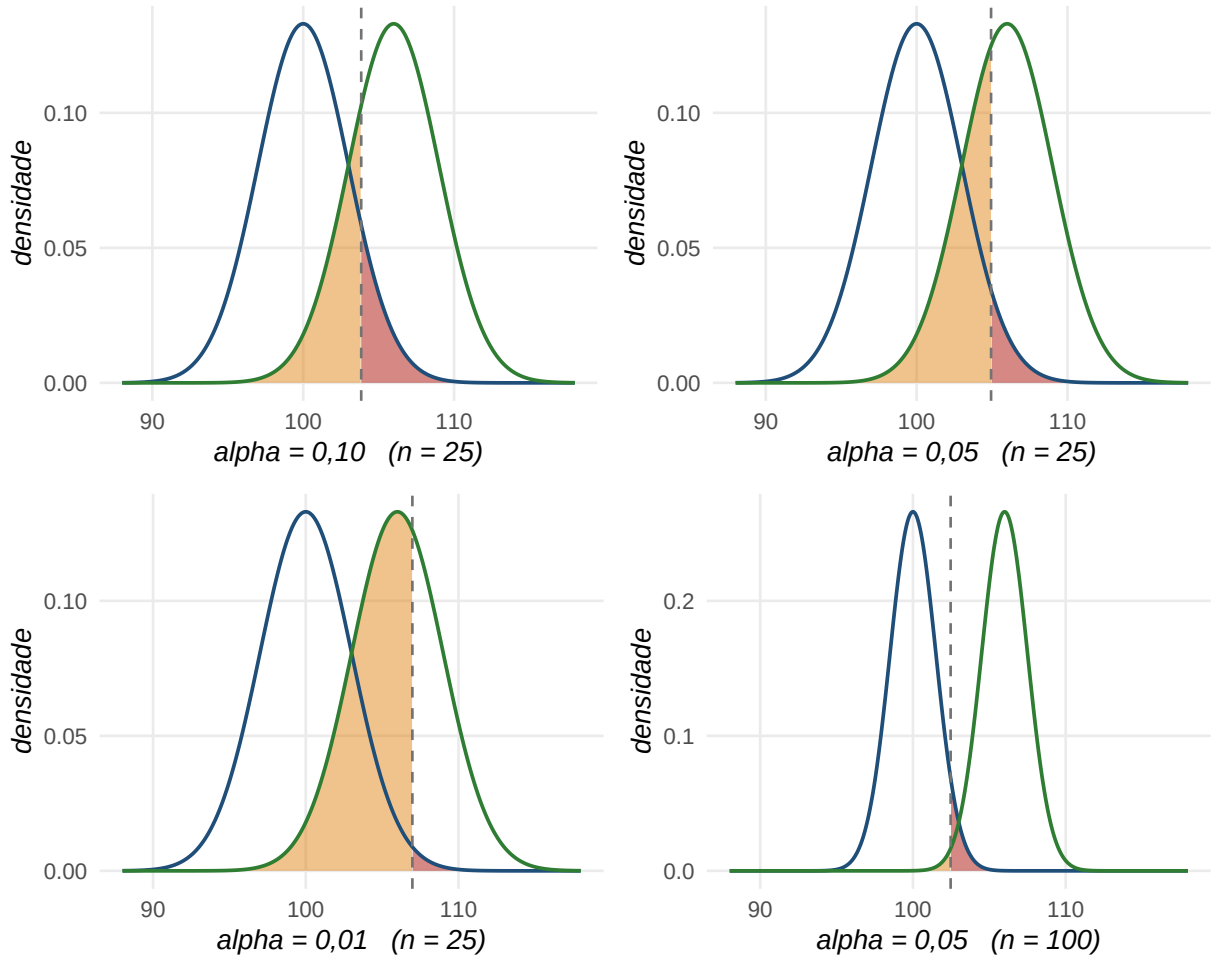


Figura 18: Os dois tipos de erro. Em cada painel, a curva azul é a distribuição de $\bar{X}$ sob H_0 ($\mu = 100$) e a verde é sob a alternativa $\mu_1 = 106$; a linha vertical é o valor crítico. A área vermelha à direita é α (erro tipo I); a área laranja à esquerda é β (erro tipo II). Ao endurecer o teste — de $\alpha = 0{,}10$ para $\alpha = 0{,}01$ —, a área vermelha encolhe e a laranja cresce: é o *trade-off*. O quarto painel mostra a única saída, aumentar n: com $n = 100$ as duas curvas se estreitam e as duas áreas diminuem simultaneamente.

A leitura da figura é a lição inteira da seção sobre erros. Descendo os três primeiros painéis, a linha

crítica desliza para a direita: a área vermelha (α) diminui, e a laranja (β) necessariamente aumenta, porque a linha que separa as duas é a mesma. Não há como empurrá-la e melhorar os dois lados.

O quarto painel mostra a saída. Com $n = 100$ em vez de 25, o erro padrão cai pela metade, as duas densidades se estreitam e se separam — e agora as duas áreas são pequenas ao mesmo tempo. Mais informação, menos erro de ambos os tipos. É a razão pela qual a resposta certa para "meu teste não rejeitou, e agora?" é quase sempre "colete mais dados", e não "afrouxe o α".

10.11.2 A curva de poder

O gráfico anterior fixou uma alternativa, $\mu_1 = 106$. Mas β depende de *qual* é a verdade, e o objeto que descreve isso por completo é a curva de poder.

```
mus <- seq(94, 114, length.out = 400)
poder_ptual <- function(mu, nn, a = 0.05) {
  epn <- sigma / sqrt(nn)
  1 - pnorm(mu0 + qnorm(1 - a) * epn, mean = mu, sd = epn)
}
d_pod <- do.call(rbind, lapply(c(10, 25, 50, 100), function(nn)
  data.frame(mu = mus, poder = poder_ptual(mus, nn), n = factor(nn))))

ggplot(d_pod, aes(mu, poder, colour = n)) +
  geom_hline(yintercept = 0.05, colour = cz, linetype = "dotted") +
  geom_vline(xintercept = mu0,  colour = cz, linetype = "dashed") +
  geom_line(linewidth = 0.8) +
  scale_colour_manual(values = c(cz, lr, vd, az)) +
  scale_y_continuous(limits = c(0, 1)) +
  labs(x = "verdadeiro valor de mu", y = "poder = 1 - beta")
```

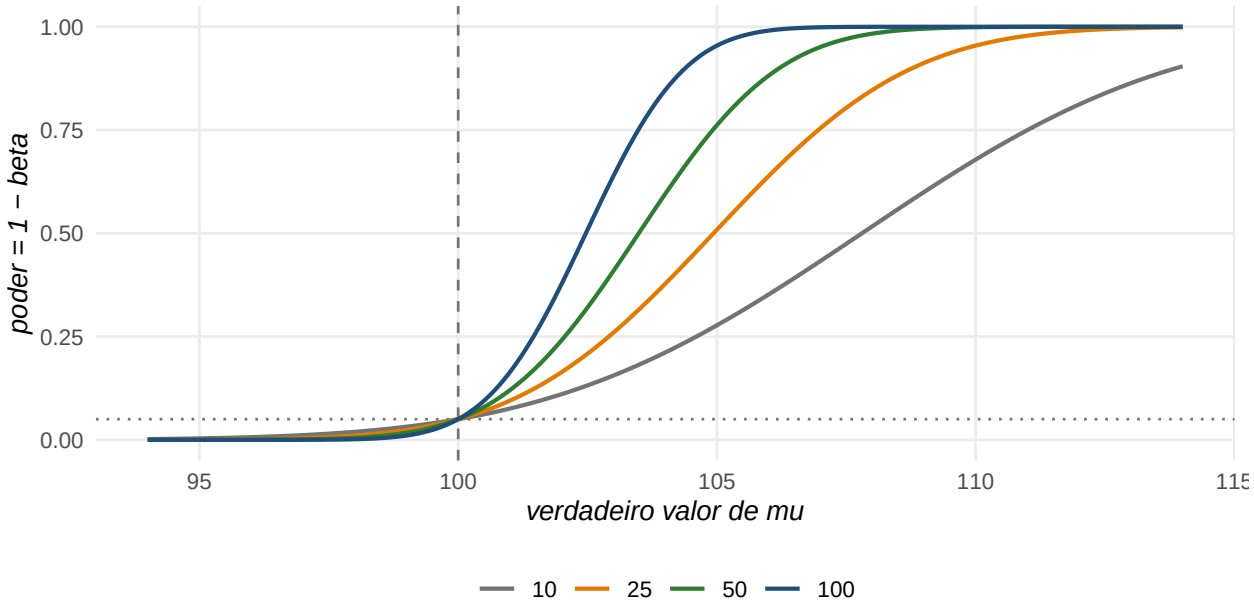


Figura 19: Curva de poder do teste unilateral $H_0 : \mu \leq 100$ contra $H_1 : \mu > 100$, com $\sigma = 15$ e $\alpha = 0{,}05$, para quatro tamanhos de amostra. Em $\mu = 100$ (linha cinza vertical) todas as curvas valem exatamente $\alpha = 0{,}05$ (linha cinza pontilhada): rejeitar ali é o erro tipo I. O poder cresce com a distância a H_0 e com n — em $\mu = 106$ o teste com $n = 100$ detecta em 99% das vezes, ao passo que o com $n = 10$ detecta em apenas 35%.

As três propriedades enunciadas na seção teórica estão todas na figura. Em $\mu = 100$ as quatro curvas se cruzam sobre a linha pontilhada em 0,05 — o poder na fronteira de H_0 é o próprio α. À direita, todas sobem monotonicamente rumo a 1: quanto mais falsa a nula, mais fácil detectá-la. E as curvas se ordenam por n, com a de $n = 100$ acima de todas em qualquer ponto.

À **esquerda** de 100 o poder é menor que α, e isso não é anomalia: o teste é unilateral à direita, e valores de μ abaixo de 100 estão na direção que ele deliberadamente não observa. Ele quase nunca rejeitará ali — corretamente, aliás, já que $H_0 : \mu \leq 100$ é verdadeira nessa região.

Vale quantificar o que a figura sugere sobre planejamento amostral:

```
alvos <- c(102, 104, 106)
tab_pod <- sapply(c(10, 25, 50, 100, 400),
                  function(nn) poder_ptual(alvos, nn))
dimnames(tab_pod) <- list(paste0("mu = ", alvos),
                          paste0("n = ", c(10, 25, 50, 100, 400)))
round(tab_pod, 3)
```

```
         n = 10 n = 25 n = 50 n = 100 n = 400
mu = 102  0.111  0.164  0.241   0.378   0.847
mu = 104  0.211  0.378  0.595   0.847   1.000
mu = 106  0.352  0.639  0.882   0.991   1.000
```

Para detectar um desvio pequeno ($\mu = 102$, pouco mais de um sétimo de desvio padrão), $n = 25$ dá poder de apenas 16,4% — o teste falharia em cinco de cada seis tentativas. É preciso $n = 400$ para chegar a 84,7%. Já um desvio grande ($\mu = 106$) é detectado em 63,9% das vezes com $n = 25$, e em 99,1% com $n = 100$. É o cálculo que se faz **antes** de coletar dados, e cuja ausência explica boa parte dos resultados inconclusivos na literatura aplicada.

10.11.3 Simulando as taxas de erro

Nada convence tanto quanto ver o α acontecer. Geramos 10.000 amostras de uma população em que H_0 é **verdadeira**, rodamos o teste em cada uma e contamos as rejeições. A frequência deve ficar em torno de 5%.

```
set.seed(20261)
M <- 10000; n_sim <- 25; mu_H0 <- 100; sd_pop <- 15
crit_t <- qt(0.975, n_sim - 1)      # teste bilateral com sigma desconhecido

roda_teste <- function(mu_verdadeiro) {
  A  <- matrix(rnorm(M * n_sim, mu_verdadeiro, sd_pop), nrow = M)
  t0 <- (rowMeans(A) - mu_H0) / (apply(A, 1, sd) / sqrt(n_sim))
  mean(abs(t0) >= crit_t)           # proporcao de rejeicoes
}

c(taxa_sob_H0     = roda_teste(100),   # deve ficar perto de 0,05 = alpha
  poder_mu_103    = roda_teste(103),
  poder_mu_106    = roda_teste(106),
  poder_mu_110    = roda_teste(110))
```

```
 taxa_sob_H0 poder_mu_103 poder_mu_106 poder_mu_110
      0.0514       0.1653       0.4878       0.8981
```

A primeira linha é a verificação essencial: sob H_0 verdadeira, o teste rejeitou em 5,14% das 10.000 simulações — contra os 5% nominais. A diferença é ruído de simulação (com $M = 10.000$, o erro padrão dessa proporção é $\sqrt{0{,}05 \times 0{,}95/10000} \approx 0{,}2$ ponto percentual, de modo que 5,14% está bem dentro do esperado). Não é aproximação grosseira: é o α nominal reproduzido pela frequência empírica, exatamente como a definição $\alpha = P(\text{rejeitar} \mid H_0)$ promete. E é o sentido em que "nível de significância" é uma **taxa de erro de longo prazo do procedimento**, não uma afirmação sobre a amostra que temos em mãos.

As demais linhas medem o poder empírico. Com a verdade em $\mu = 103$, o teste bilateral detecta em cerca de 16,5% dos casos; em $\mu = 106$, em 48,8%; em $\mu = 110$, em 89,8%. Note que esses valores são menores que os da curva teórica da subseção anterior (63,9% em $\mu = 106$ com o mesmo $n = 25$), e há duas razões: este teste é **bilateral** (gasta metade de α na cauda errada) e usa t em vez de z (porque σ é estimado). Ambos os efeitos custam poder, e a simulação os captura sem que precisemos de fórmula alguma.

10.11.4 Conferindo todos os exemplos do capítulo

Passemos à conferência sistemática. Começamos pelos testes de média, com uma função que devolve estatística, valor crítico e p-valor de uma vez.

```
teste_media <- function(xbar, k, dp, n, alpha, tipo = c("bi", "dir", "esq"),
                        dist = c("t", "z")) {
  tipo <- match.arg(tipo); dist <- match.arg(dist)
  gl   <- n - 1
  est  <- (xbar - k) / (dp / sqrt(n))
  qf_  <- function(q) if (dist == "t") qt(q, gl) else qnorm(q)
  pf_  <- function(x) if (dist == "t") pt(x, gl) else pnorm(x)
  crit <- switch(tipo, bi = qf_(1 - alpha/2), dir = qf_(1 - alpha), esq = qf_(1 - alpha))
  p    <- switch(tipo, bi  = 2 * pf_(-abs(est)),
                       dir = 1 - pf_(est),
                       esq = pf_(est))
  rejeita <- switch(tipo, bi = abs(est) >= crit, dir = est >= crit, esq = est <= -crit)
  c(estatistica = est, critico = crit, p_valor = p, rejeita = as.numeric(rejeita))
}

round(rbind(
  `alturas (z, bi, 5%)`   = teste_media(180, 175, 6, 5, 0.05, "bi", "z"),
  `alturas (t, bi, 5%)`   = teste_media(180, 175, 6, 5, 0.05, "bi", "t"),
  `salarios (t, bi, 10%)` = teste_media(400, 600, 450, 25, 0.10, "bi", "t"),
  `TVs (t, bi, 1%)`       = teste_media(8900, 9000, 500, 16, 0.01, "bi", "t"),
  `nicotina (z, dir, 5%)` = teste_media(31.5, 30, 3, 25, 0.05, "dir", "z"),
  `endivid. (z, dir, 5%)` = teste_media(45, 30, 30, 9, 0.05, "dir", "z"),
  `fundos (t, esq, 5%)`   = teste_media(-1, 0, 2, 16, 0.05, "esq", "t")), 5)
```

```
                     estatistica critico p_valor rejeita
alturas (z, bi, 5%)      1.86339 1.95996 0.06241       0
alturas (t, bi, 5%)      1.86339 2.77645 0.13587       0
salarios (t, bi, 10%)   -2.22222 1.71088 0.03595       1
TVs (t, bi, 1%)         -0.80000 2.94671 0.43620       0
nicotina (z, dir, 5%)    2.50000 1.64485 0.00621       1
endivid. (z, dir, 5%)    1.50000 1.64485 0.06681       0
fundos (t, esq, 5%)     -2.00000 1.75305 0.03197       1
```

Todos os valores do capítulo se confirmam: $z_0 = 1{,}8634$ nas alturas, com crítico 1,96 (e 2,7764 na versão t); $t_0 = -2{,}2222$ nos salários contra 1,7109; $t_0 = -0{,}80$ nas TVs; $z_0 = 2{,}5$ na nicotina com crítico 1,6449 e $p = 0{,}00621$; $z_0 = 1{,}5$ no endividamento; $t_0 = -2$ nos fundos com $p = 0{,}0320$. A coluna `rejeita` resume as decisões — 1 significa rejeitar — e reproduz cada conclusão do texto.

O caso do endividamento merece a varredura em α, que é a monotonicidade em código:

```r
z0_end <- (45 - 30) / (30 / sqrt(9))
p_end  <- 1 - pnorm(z0_end)
alphas <- c(0.01, 0.025, 0.05, p_end - 1e-6, p_end + 1e-6, 0.10, 0.20)
data.frame(alpha   = round(alphas, 6),
           critico = round(qnorm(1 - alphas), 4),
           rejeita = z0_end >= qnorm(1 - alphas),
           p_valor = round(p_end, 6))
```

```
     alpha critico rejeita  p_valor
1 0.010000  2.3263   FALSE 0.066807
2 0.025000  1.9600   FALSE 0.066807
3 0.050000  1.6449   FALSE 0.066807
4 0.066806  1.5000   FALSE 0.066807
5 0.066808  1.5000    TRUE 0.066807
6 0.100000  1.2816    TRUE 0.066807
7 0.200000  0.8416    TRUE 0.066807
```

A coluna `rejeita` muda de `FALSE` para `TRUE` **uma única vez**, e a virada ocorre exatamente sobre o p-valor 0,066807: as duas linhas do meio, que o cercam por um milionésimo para baixo e para cima, já estão de lados opostos da decisão. Abaixo do p-valor nunca se rejeita; acima, sempre. É a monotonicidade tornada tabela, e é também a definição do p-valor como "o menor α que leva à rejeição", verificada numericamente.

Passemos às proporções, atentos ao denominador com k:

```r
teste_prop <- function(x, n, k, alpha, tipo = c("bi", "dir", "esq")) {
  tipo <- match.arg(tipo)
  ph   <- x / n
  est  <- (ph - k) / sqrt(k * (1 - k) / n)          # denominador SOB H0: usa k
  crit <- if (tipo == "bi") qnorm(1 - alpha/2) else qnorm(1 - alpha)
  p    <- switch(tipo, bi = 2 * pnorm(-abs(est)),
                       dir = 1 - pnorm(est), esq = pnorm(est))
  c(phat = ph, estatistica = est, critico = crit, p_valor = p)
}

round(rbind(
  `corretora (bi, 10%)` = teste_prop(20, 64, 0.30, 0.10, "bi"),
  `audiencia (esq, 5%)` = teste_prop(220, 400, 0.60, 0.05, "esq")), 5)
```

```
                     phat estatistica critico p_valor
corretora (bi, 10%) 0.3125     0.21822 1.64485 0.82726
audiencia (esq, 5%) 0.5500    -2.04124 1.64485 0.02061
```

```r
# a assimetria denominador-com-k versus denominador-com-phat
ph_c <- 20/64
c(com_k     = (ph_c - 0.3) / sqrt(0.3 * 0.7 / 64),
  com_phat  = (ph_c - 0.3) / sqrt(ph_c * (1 - ph_c) / 64))
```

```
    com_k  com_phat
0.2182179 0.2157440
```

Confirmam-se $\hat{p} = 0{,}3125$ e $z_0 = 0{,}2182$ na corretora (contra crítico 1,6449, não rejeita, $p = 0{,}827$), e $\hat{p} = 0{,}55$ com $z_0 = -2{,}0412$ na audiência. Sobre este último, um registro de precisão: **o p-valor exato é** 0,02061, ao passo que o valor do slide, 0,0207, decorre de arredondar a estatística para $-2{,}04$ antes de consultar a tabela. A diferença é imaterial para a decisão, mas vale saber de onde vem.

O bloco final mostra a assimetria em números: 0,2182 com k contra 0,2157 com $\hat{p}$. Perto de H_0 os dois quase coincidem; longe dela, podem divergir bem mais.

Variância, diferença de médias, diferença de proporções e razão de variâncias:

```r
# --- variancia: n = 30, s = 3,63, H0: sigma^2 = 16, alpha = 5%
n_v <- 30; s_v <- 3.63; k_v <- 16
q0  <- (n_v - 1) * s_v^2 / k_v
c(q0        = q0,
  inferior  = qchisq(0.025, n_v - 1),   # chi2_{29; 0,975}
  superior  = qchisq(0.975, n_v - 1),   # chi2_{29; 0,025}
  p_valor   = 2 * min(pchisq(q0, n_v - 1), 1 - pchisq(q0, n_v - 1)))
```

```
        q0   inferior   superior    p_valor
23.8831312 16.0470717 45.7222858  0.5306542
```

```r
teste_dif_medias <- function(m1, s1, n1, m2, s2, n2, alpha) {
  gl  <- n1 + n2 - 2
  sp2 <- ((n1 - 1) * s1^2 + (n2 - 1) * s2^2) / gl
  ep  <- sqrt(sp2) * sqrt(1/n1 + 1/n2)
  t0  <- (m1 - m2) / ep
  ic  <- (m1 - m2) + c(-1, 1) * qt(1 - alpha/2, gl) * ep
  c(sp = sqrt(sp2), gl = gl, t0 = t0, critico = qt(1 - alpha/2, gl),
    p_valor = 2 * pt(-abs(t0), gl), IC_inf = ic[1], IC_sup = ic[2])
}

round(rbind(
  `lampadas (1%)` = teste_dif_medias(4600, 250, 8, 4000, 200, 10, 0.01),
  `turmas (5%)`   = teste_dif_medias(-0.1, sqrt(5.425), 5, 0, sqrt(7), 4, 0.05)), 4)
```

```
                   sp gl      t0 critico p_valor   IC_inf   IC_sup
lampadas (1%) 223.2571 16  5.6657  2.9208  0.0000 290.6888 909.3112
turmas (5%)     2.4698  7 -0.0604  2.3646  0.9536  -4.0177   3.8177
```

Nas lâmpadas: $s_p = 223{,}2571$, $\nu = 16$, $t_0 = 5{,}6657$ contra crítico 2,9208 — rejeita-se, com $p = 0{,}000035$. Nas turmas (onde passamos a diferença de médias $-0{,}1$ como `m1` e zero como `m2`, o que preserva a estatística): $s_p = 2{,}4698$, $\nu = 7$, $t_0 = -0{,}0604$ contra 2,3646, $p = 0{,}9536$ — não se rejeita.

E o intervalo confirma a equivalência anunciada: $[-4{,}0177; 3{,}8177]$, que contém o zero. (O texto reporta $[-4{,}0182; 3{,}8182]$; a diferença na quarta casa vem de o slide usar $t_{7;0{,}025} = 2{,}365$ arredondado, contra 2,36462 exato. Rigorosamente idêntico em tudo que importa.)

```
# --- diferenca de proporcoes: financeiras, alpha = 10%
x1 <- 140; n1 <- 180; x2 <- 220; n2 <- 300
p1 <- x1/n1; p2 <- x2/n2
z0 <- (p1 - p2) / sqrt(p1*(1-p1)/n1 + p2*(1-p2)/n2)
c(p1 = p1, p2 = p2, z0 = z0,
  crit_bilateral  = qnorm(0.95),  p_bilateral  = 2 * (1 - pnorm(z0)),
  crit_unilateral = qnorm(0.90),  p_unilateral = 1 - pnorm(z0))
```

```
            p1             p2             z0  crit_bilateral     p_bilateral
     0.7777778      0.7333333      1.1069432       1.6448536       0.2683185
crit_unilateral    p_unilateral
     1.2815516      0.1341592
```

```
teste_F <- function(s1_2, n1, s2_2, n2, alpha) {
  v1 <- n1 - 1; v2 <- n2 - 1
  f0 <- s1_2 / s2_2
  sup <- qf(1 - alpha/2, v1, v2)              # f_{v1,v2; alpha/2}
  inf <- qf(alpha/2, v1, v2)                  # f_{v1,v2; 1-alpha/2}
  c(f0 = f0, gl1 = v1, gl2 = v2, inferior = inf, superior = sup,
    inf_por_inversao = 1 / qf(1 - alpha/2, v2, v1),   # a identidade, COM troca
    p_valor = 2 * min(pf(f0, v1, v2), 1 - pf(f0, v1, v2)))
}

round(rbind(
  `turmas (5%)`    = teste_F(5.425, 5, 7, 4, 0.05),
  `exercicio (10%)`= teste_F(40, 6, 37, 6, 0.10)), 5)
```

```
                    f0 gl1 gl2 inferior superior inf_por_inversao p_valor
turmas (5%)     0.77500   4   3  0.10021 15.10098          0.10021 0.78439
exercicio (10%) 1.08108   5   5  0.19801  5.05033          0.19801 0.93391
```

O teste F das turmas confirma tudo: $f_0 = 0{,}775$, RC $= (0; 0{,}1002] \cup [15{,}101; \infty)$ e $p = 0{,}7844$ — não se rejeita, e o pressuposto de variâncias iguais do teste de médias está validado. As colunas `inferior` e `inf_por_inversao` são **idênticas** (0,10021), o que verifica numericamente a identidade $f_{\nu_1,\nu_2;1-\alpha/2} = 1/f_{\nu_2,\nu_1;\alpha/2}$ **com a troca dos graus de liberdade**. Vale ver o que daria o erro comum de inverter sem trocar:

```
c(correto_com_troca  = 1 / qf(0.975, 3, 4),   # f_{4,3;0,975} = 1 / f_{3,4;0,025}
  errado_sem_troca   = 1 / qf(0.975, 4, 3),   # 1 / f_{4,3;0,025}  <- ERRADO
  f_4_3_sup          = qf(0.975, 4, 3),
  f_3_4_sup          = qf(0.975, 3, 4))
```

```
correto_com_troca  errado_sem_troca         f_4_3_sup         f_3_4_sup
       0.10020845        0.06622087       15.10097893        9.97919853
```

O correto é 0,10021; o errado, 0,06622 — uma região crítica inferior 34% menor, que mudaria a decisão para qualquer f_0 entre os dois valores. A quarta e a terceira linhas mostram por que: $f_{4,3} = 15{,}101$ e $f_{3,4} = 9{,}979$ são bem diferentes, e é justamente essa assimetria que a troca de índices corrige.

O exercício final da lista também confere: $f_0 = 1{,}0811$, com RC $= (0; 0{,}198] \cup [5{,}0503; \infty)$ e $p = 0{,}9339$ — não se rejeita, como esperado para variâncias tão próximas quanto 40 e 37.

10.11.5 A equivalência IC ↔ teste bilateral

Fechamos verificando, de forma sistemática, a identidade demonstrada na caixa das turmas: no teste bilateral, os dois métodos são a mesma desigualdade. Em vez de um exemplo, testamos **mil** valores hipotetizados diferentes e conferimos se as duas decisões jamais divergem.

```r
xbar_s <- 400; s_s <- 450; n_s <- 25; alpha_s <- 0.10
ep_s   <- s_s / sqrt(n_s)
tc     <- qt(1 - alpha_s/2, n_s - 1)
IC     <- xbar_s + c(-1, 1) * tc * ep_s

ks <- seq(150, 650, length.out = 1000)
dentro_do_IC  <- ks >= IC[1] & ks <= IC[2]              # metodo 1
fora_da_RC    <- abs((xbar_s - ks) / ep_s) <  tc        # metodo 2
p_maior_alpha <- 2 * pt(-abs((xbar_s - ks) / ep_s), n_s - 1) >= alpha_s  # metodo 3

c(IC_inf = IC[1], IC_sup = IC[2],
  discordancias_IC_vs_RC = sum(dentro_do_IC != fora_da_RC),
  discordancias_IC_vs_p  = sum(dentro_do_IC != p_maior_alpha))
```

```
              IC_inf               IC_sup discordancias_IC_vs_RC
            246.0206             553.9794                  0.0000
discordancias_IC_vs_p
                0.0000
```

Zero discordâncias, em mil valores testados: os três métodos não apenas concordam no exemplo dos salários — eles concordam para *todo* valor hipotetizado possível. O intervalo de confiança é, literalmente, **o conjunto dos k que não seriam rejeitados** por um teste bilateral ao nível α. Essa é a formulação mais precisa da relação entre os dois capítulos, e ela justifica retroativamente a notação $100(1-\alpha)\%$ de Seção 9.

```r
d_eq <- data.frame(k = ks, p = 2 * pt(-abs((xbar_s - ks) / ep_s), n_s - 1))
ggplot(d_eq, aes(k, p)) +
  annotate("rect", xmin = IC[1], xmax = IC[2], ymin = -Inf, ymax = Inf,
           fill = az, alpha = 0.12) +
```

```
  geom_line(colour = az, linewidth = 0.8) +
  geom_hline(yintercept = alpha_s, colour = vm, linetype = "dashed") +
  geom_vline(xintercept = IC, colour = cz, linetype = "dotted") +
  annotate("point", x = 600,
           y = 2 * pt(-abs((xbar_s - 600) / ep_s), n_s - 1),
           colour = vm, size = 2.2) +
  labs(x = "valor hipotetizado k   (faixa azul = IC de 90%)", y = "p-valor")
```

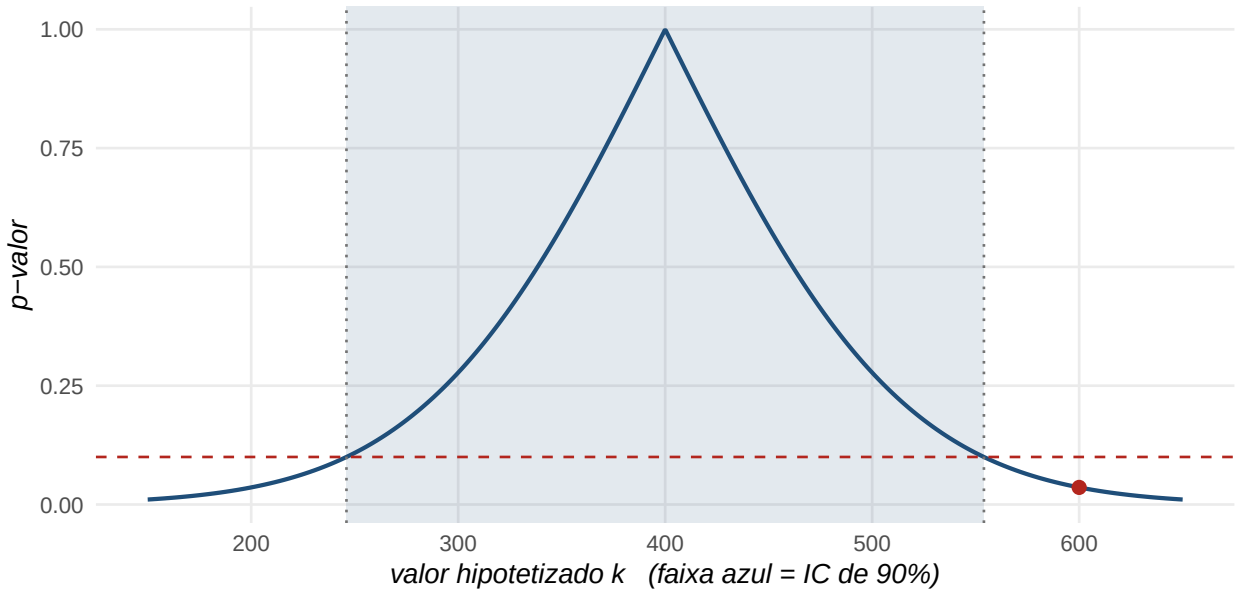


Figura 20: A equivalência entre os três métodos, no exemplo dos salários ($\bar{x} = 400$, $s = 450$, $n = 25$, $\alpha = 0{,}10$). O eixo horizontal percorre valores hipotetizados k; a curva é o p-valor do teste bilateral de $H_0 : \mu = k$. A faixa azul é o intervalo de confiança de 90%, e a linha tracejada é $\alpha = 0{,}10$. A curva cruza α **exatamente** nas bordas do intervalo: dentro dele, $p > \alpha$ e não se rejeita; fora, $p \leq \alpha$ e se rejeita. O ponto vermelho é o valor hipotetizado do exemplo, $k = 600$, que cai fora — $p = 0{,}036$.

A figura é, talvez, o melhor resumo visual do capítulo inteiro. O p-valor é máximo (igual a 1) quando $k = \bar{x} = 400$ — hipotetizar exatamente o que se observou é a hipótese mais compatível possível com os dados — e decresce simetricamente em ambas as direções. Ele cruza a linha de α precisamente nas bordas do intervalo de confiança, que é o que a identidade algébrica previa. E o valor hipotetizado do exemplo, $k = 600$, está bem à direita da borda, com $p = 0{,}036$: rejeita-se.

10.12 Exercícios

1. **Formulação de hipóteses.** Para cada situação, escreva H_0 e H_1, indique se o teste é uni ou bilateral e justifique quem foi para a nula: (a) uma montadora quer evidenciar que seu novo motor consome menos de 8 litros por 100 km; (b) um órgão regulador suspeita que uma seguradora demora, em média, mais de 30 dias para pagar sinistros; (c) um pesquisador quer saber se a taxa de desemprego mudou em relação aos 7,2% do trimestre anterior; (d) uma

fábrica quer verificar se a variabilidade do processo se manteve em $\sigma^2 = 4$. (e) Em qual dos itens a analogia do julgamento é mais esclarecedora, e por quê?

2. **Os dois erros.** Uma máquina deve encher pacotes com $\mu = 500$ g. Testa-se $H_0 : \mu = 500$ contra $H_1 : \mu \neq 500$ com $\sigma = 10$ conhecido e $n = 25$, a $\alpha = 0{,}05$. (a) Obtenha a região crítica em termos de $\bar{x}$ (não da estatística padronizada). (b) Calcule β supondo que a verdade seja $\mu = 505$. (c) Repita para $\mu = 510$ e para $\mu = 500{,}5$; descreva como β varia com a distância a H_0.

   (d) Verifique que $\alpha + \beta \neq 1$ nos três casos e explique por que isso não é um erro. (e) Que n seria necessário para que $\beta \leq 0{,}10$ quando $\mu = 505$?

3. **Os três métodos.** Uma amostra de $n = 36$ contratos tem prazo médio $\bar{x} = 42$ dias com $s = 12$. Testa-se $H_0 : \mu = 45$ contra $H_1 : \mu \neq 45$ a $\alpha = 0{,}05$. (a) Decida pelo método do intervalo de confiança. (b) Decida pela região crítica. (c) Calcule o p-valor e decida por ele. (d) Mostre algebricamente, para este caso, que (a) e (b) são a mesma desigualdade. (e) A que níveis α a conclusão se inverteria?

4. **Unilateral e o α inteiro.** Retome o exemplo da nicotina ($n = 25$, $\bar{x} = 31{,}5$, $\sigma = 3$, $H_0 : \mu \leq 30$). (a) Refaça o teste a $\alpha = 0{,}01$.

   (b) Refaça-o na versão bilateral, também a $\alpha = 0{,}01$, e compare as duas conclusões.
   (c) Explique, em termos de área de cauda, por que a versão unilateral rejeita e a bilateral não. (d) Um pesquisador que olhou os dados, viu $\bar{x} > 30$ e só *então* decidiu fazer o teste unilateral está agindo corretamente? Relacione com a exigência de fixar α (e as hipóteses) *a priori*.

5. **Proporção e o denominador.** Um banco afirma que 80% de seus clientes usam o aplicativo. Numa amostra de $n = 150$, 108 usam. (a) Teste $H_0 : p = 0{,}80$ contra $H_1 : p \neq 0{,}80$ a $\alpha = 0{,}05$, com o denominador correto. (b) Refaça usando $\hat{p}$ no denominador e compare as estatísticas. (c) Construa o intervalo de 95% para p e verifique se ele contém 0,80. (d) Se as duas abordagens discordassem, qual você reportaria e por quê?

6. **Variância e poder.** Uma corretora exige que a volatilidade de um fundo não exceda $\sigma^2 = 25$. Numa amostra de $n = 20$ meses, $s^2 = 38$. (a) Teste $H_0 : \sigma^2 \leq 25$ contra $H_1 : \sigma^2 > 25$ a $\alpha = 0{,}05$ (atenção: unilateral, α inteiro numa cauda). (b) Calcule o p-valor. (c) Refaça com $n = 60$ e o mesmo $s^2 = 38$; a conclusão muda? (d) O que o par de resultados ilustra sobre a relação entre n e poder?

7. **Diferença de médias e o pressuposto.** Duas filiais registraram vendas mensais: filial A com $n_1 = 12$, $\bar{x}_1 = 84$ e $s_1 = 15$; filial B com $n_2 = 15$, $\bar{x}_2 = 76$ e $s_2 = 9$. (a) Teste $H_0 : \sigma_1^2 = \sigma_2^2$ a $\alpha = 0{,}10$ pelo teste F, obtendo os dois valores críticos (use a identidade da inversão, com troca de graus de liberdade). (b) À luz de (a), é legítimo aplicar o teste de diferença de médias com s_p? (c) Faça-o, a $\alpha = 0{,}05$, bilateral. (d) Confirme a conclusão pelo intervalo de confiança. (e) Por que a ordem — primeiro o F, depois o t — é a correta?

8. **Simulação.** Adapte o código do laboratório para investigar o que acontece quando as hipóteses do teste são violadas. (a) Gere 10.000 amostras de tamanho $n = 10$ de uma **exponencial** de média 100 (fortemente assimétrica), aplique o teste t bilateral de $H_0 : \mu = 100$ a $\alpha = 0{,}05$ e compare a taxa de rejeição empírica com 0,05: o teste mantém seu nível nominal? (b) Repita com $n = 100$ e depois $n = 1000$; o que o TCL de Seção 7 prevê e o que a simulação mostra? (c)

Repita (a) com uma população Normal, para confirmar que ali o nível é exato para qualquer n. (d) Que lição prática você extrai sobre o uso do teste t com dados assimétricos e amostras pequenas?

Assim se fecha o arco destas notas. Começamos em Seção 2 com trinta números numa página e a pergunta de como resumi-los sem destruí-los — **descrever**. Passamos por Seção 3, Seção 4 e pelas distribuições, construindo o vocabulário que permite atribuir estrutura ao que parecia arbitrário — **modelar o acaso**. E na terça parte final, de Seção 8 até aqui, invertemos a seta: dos dados de volta ao modelo, do particular ao geral, com estimativas pontuais, intervalos e, agora, decisões — **inferir**. O que fica não é uma coleção de fórmulas, mas um hábito: exibir a incerteza em vez de escondê-la, e nunca afirmar mais do que os dados sustentam. É esse hábito que a econometria dos próximos cursos vai levar adiante.

Referências

Bussab, Wilton O., e Pedro A. Morettin. 2010. *Estatística Básica*. 5º ed. Saraiva.

Casella, George, e Roger L. Berger. 2002. *Statistical Inference*. 2º ed. Duxbury.

DeGroot, Morris H., e Mark J. Schervish. 2012. *Probability and Statistics*. 4º ed. Addison-Wesley.

Hoffmann, Rodolfo. 2006. *Estatística para Economistas*. 4º ed. Cengage Learning.

Keller, Gerald. 2011. *Statistics for Management and Economics*. 9º ed. Thomson South-Western.

Larson, Harold J. 1982. *Introduction to Probability Theory and Statistical Inference*. 3º ed. John Wiley & Sons.

Meyer, Paul L. 1983. *Probabilidade: Aplicações à Estatística*. 2º ed. LTC.