Capítulo 10

Introdução à Inferência Estatística

10.1 Introdução

Vimos, na Parte 1, como resumir descritivamente variáveis associadas a um ou mais conjuntos de dados. Na Parte 2, construímos modelos teóricos (probabilísticos), identificados por parâmetros, capazes de representar adequadamente o comportamento de algumas variáveis. Nesta terceira parte apresentaremos os argumentos estatísticos para fazer afirmações sobre as características de uma população, com base em informações dadas por amostras.

O uso de informações de uma amostra para concluir sobre o todo faz parte da atividade diária da maioria das pessoas. Basta observar como uma cozinheira verifica se o prato que ela está preparando tem ou não a quantidade adequada de sal. Ou, ainda, quando um comprador, após experimentar um pedaço de laranja numa banca de feira, decide se vai comprar ou não as laranjas. Essas são decisões baseadas em procedimentos amostrais.

Nosso objetivo nos capítulos seguintes é procurar dar a conceituação formal a esses princípios intuitivos do dia-a-dia para que possam ser utilizados cientificamente em situações mais complexas.

10.2 População e Amostra

Nos capítulos anteriores, tomamos conhecimento de alguns modelos probabilísticos que procuram medir a variabilidade de fenômenos casuais de acordo com suas ocorrências: as distribuições de probabilidades de variáveis aleatórias (qualitativas ou quantitativas). Na prática, freqüentemente o pesquisador tem alguma idéia sobre a forma da distribuição, mas não dos valores exatos dos parâmetros que a especificam.

Por exemplo, parece razoável supor que a distribuição das alturas dos brasileiros adultos possa ser representada por um modelo normal (embora as alturas não possam assumir valores negativos). Mas essa afirmação não é suficiente para determinar qual a distribuição normal correspondente; precisaríamos conhecer os parâmetros (média e variância) dessa normal para que ela ficasse completamente especificada. O propósito do pesquisador seria, então, descobrir (estimar) os parâmetros da distribuição para sua posterior utilização.

Daqui para frente, a menos que esteja especificada de outra maneira, sempre que mencionarmos a palavra amostra, estaremos entendendo a amostra obtida pelo processo probabilístico AAS, ou seja, o vetor aleatório $(X_1, X_2, ..., X_n)$ definido acima.

Problemas

3. A distribuição do número de filhos, por família, de uma zona rural está no quadro abaixo.

Nº de filhos	Porcentagem
0	10
1	20
2	30
3	25
4	15
Total	100

(a) Sugira um procedimento para sortear uma observação ao acaso dessa população.
(b) Dê, na forma de uma tabela de dupla entrada, as possíveis amostras do número de filhos de duas famílias que podem ser sorteadas e as respectivas probabilidades de ocorrência.
(c) Se fosse escolhida uma amostra de tamanho 4, qual seria a probabilidade de se observar a quádrupla ordenada (2, 3, 3, 1)?

10.6 Estatísticas e Parâmetros

Obtida uma amostra, muitas vezes desejamos usá-la para produzir alguma característica específica. Por exemplo, se quisermos calcular a média da amostra $(X_1, X_2, ..., X_n)$, esta será dada por

$$\overline{X} = \frac{1}{n}\{X_1 + X_2 + ... + X_n\}.$$

É fácil verificar que $\overline{X}$ é também uma variável aleatória. Podemos também estar interessados em qualquer outra característica da amostra, que será sempre uma função do vetor aleatório $(X_1, ..., X_n)$.

Definição. Uma *estatística* é uma característica da amostra, ou seja, uma estatística T é uma função de $X_1, X_2, ..., X_n$.

As estatísticas mais comuns são:

$$\overline{X} = 1/n \sum_{i=1}^{n} X_i \ : \text{ média da amostra,}$$

$$S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \overline{X})^2 \ : \text{variância da amostra,}$$

$$X_{(1)} = \min(X_1, X_2, ..., X_n) \ : \text{o menor valor da amostra,}$$

$$X_{(n)} = \max\,(X_1, X_2, \ldots, X_n) \text{ : o maior valor da amostra,}$$

$$W = X_{(n)} - X_{(1)} \text{ : amplitude amostral,}$$

$$X_{(i)} = \text{a } i\text{-ésima maior observação da amostra.}$$

Em geral, como já vimos no Capítulo 3, podemos considerar as *estatísticas de ordem*,

$$X_{(1)} \leq X_{(2)} \leq \ldots \leq X_{(n)},$$

ou seja, os elementos da amostra ordenados.

Outras estatísticas importantes são os quantis (empíricos), $q(p)$, $0 < p < 1$, definidos no Capítulo 3, especialmente os três quartis q_1, q_2 e q_3.

Para facilitar a linguagem usada em Inferência Estatística, iremos diferenciar as características da amostra e da população.

Definição. Um *parâmetro* é uma medida usada para descrever uma característica da população.

Assim, se estivermos colhendo amostras de uma população, identificada pela v.a. X, seriam parâmetros a média $E(X)$ e sua variância $\text{Var}(X)$.

Os símbolos mais comuns são dados na tabela a seguir.

Denominação	População	Amostra
Média	$\mu = E(X)$	$\bar{X} = \sum X_i/n$
Mediana	$\text{Md} = Q_2$	$\text{md} = q_2$
Variância	$\sigma^2 = \text{Var}(X)$	$S^2 = \sum (X_i - \bar{X})^2/(n-1)$
Nº de elementos	N	n
Proporção	p	$\hat{p}$
Quantil	$Q(p)$	$q(p)$
Quartis	Q_1, Q_2, Q_3	q_1, q_2, q_3
Intervalo inter-quartil	$d_Q = Q_3 - Q_1$	$d_q = q_3 - q_1$
Função densidade	$f(x)$	histograma
Função de distribuição	$F(x)$	$F_e(x)$

10.7 Distribuições Amostrais

Vimos na seção 10.3 que o problema da inferência estatística é fazer uma afirmação sobre os parâmetros da população através da amostra. Digamos que nossa afirmação deva ser feita sobre um parâmetro θ da população (por exemplo, a média, a variância

ou qualquer outra medida). Decidimos que usaremos uma AAS de n elementos sorteados dessa população. Nossa decisão será baseada na estatística T, que será uma função da amostra $(X_1, X_2, ..., X_n)$, ou seja, $T = f(X_1, ..., X_n)$. Colhida essa amostra, teremos observado um particular valor de T, digamos t_0, e baseados nesse valor é que faremos a afirmação sobre θ, o parâmetro populacional. Veja a Figura 10.1 (a).

A validade da nossa resposta seria melhor compreendida se soubéssemos o que acontece com a estatística T, quando retiramos todas as amostras de uma população conhecida segundo o plano amostral adotado. Isto é, qual a distribuição de T quando $(X_1, ..., X_n)$ assume todos os valores possíveis. Essa distribuição é chamada *distribuição amostral da estatística* T e desempenha papel fundamental na teoria da inferência estatística. Esquematicamente, teríamos o procedimento representado na Figura 10.1, onde temos:

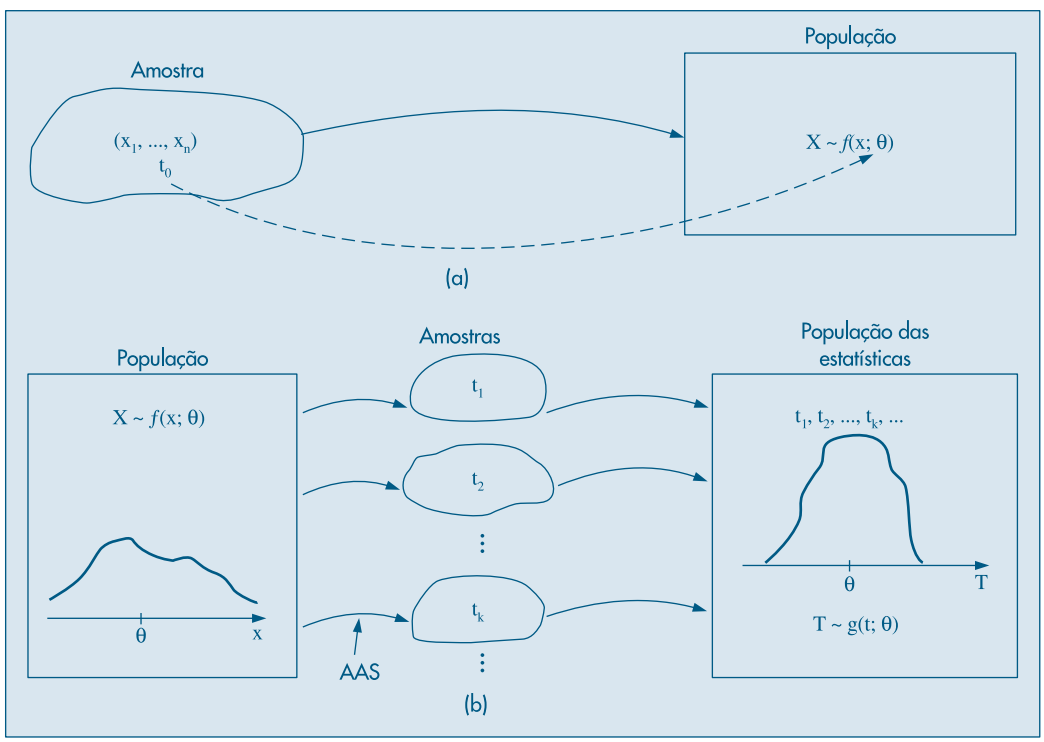
(a) uma população X, com determinado parâmetro de interesse θ;

(b) todas as amostras retiradas da população, de acordo com certo procedimento;

(c) para cada amostra, calculamos o valor t da estatística T; e

(d) os valores t formam uma nova população, cuja distribuição recebe o nome de distribuição amostral de T.

Figura 10.1: (a) Esquema de inferência sobre θ.
(b) Distribuição amostral da estatística T.



Vejamos alguns exemplos simples para aclarar um pouco mais o conceito de distribuição amostral de uma estatística. Nosso principal objetivo é identificar um modelo que explique bem a distribuição amostral de T. É evidente que a distribuição de T irá depender da distribuição de X e do plano amostral, em nosso caso reduzido a AAS.

Exemplo 10.9. Voltemos ao Exemplo 10.7, no qual selecionamos todas as amostras de tamanho 2, com reposição, da população {1, 3, 5, 5, 7}. A distribuição conjunta da variável bidimensional (X_1, X_2) é dada na Tabela 10.2.

Vejamos qual é a distribuição da estatística

$$\overline{X} = \frac{X_1 + X_2}{2}. \tag{10.1}$$

Essa distribuição é obtida por meio da Tabela 10.2. Por exemplo, quando a amostra selecionada é o par (1, 1), a média será 1; então, temos que $P(\overline{X} = 1) = 1/25$. Obteremos a média igual a 3 quando ocorrer o evento $A = \{(1, 5),(3, 3),(5, 1)\}$, logo

$$P(\overline{X} = 3) = P(A) = \frac{2}{25} + \frac{1}{25} + \frac{2}{25} + \frac{5}{25} = \frac{1}{5}.$$

Tabela 10.2: Distribuição das probabilidades das possíveis amostras de tamanho 2 que podem ser selecionadas com reposição da população {1, 3, 5, 5, 7}.

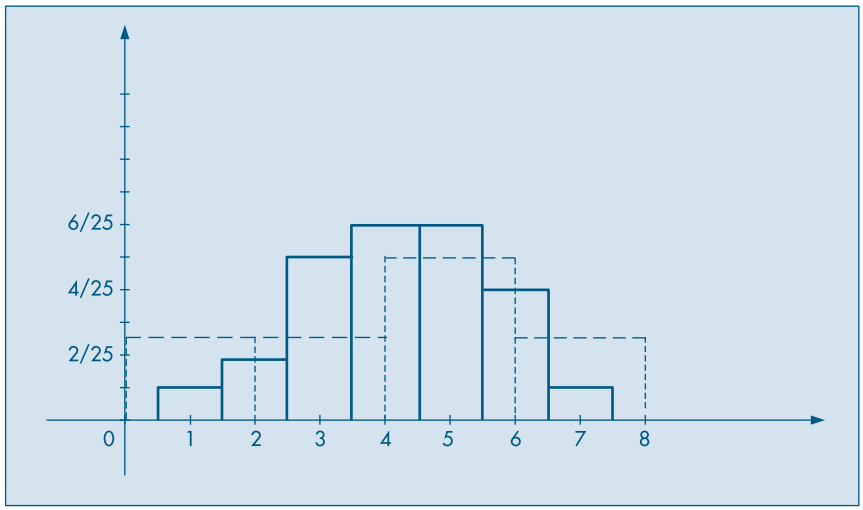
X_2 \ X_1	1	3	5	7	Total
1	1/25	1/25	2/25	1/25	1/5
3	1/25	1/25	2/25	1/25	1/5
5	2/25	2/25	4/25	2/25	2/5
7	1/25	1/25	2/25	1/25	1/5
Total	1/5	1/5	2/5	1/5	1

Procedendo de maneira análoga para os demais valores que $\overline{X}$ pode assumir, obtemos a Tabela 10.3, que dá a distribuição da v.a. $\overline{X}$. Na Figura 10.2 temos as distribuições de X e de $\overline{X}$.

Tabela 10.3: Distribuição amostral da estatística $\overline{X}$.

$\overline{x}$	1	2	3	4	5	6	7	Total
$P(\overline{X} = \overline{x})$	1/25	2/25	5/25	6/25	6/25	4/25	1/25	1,00

Figura 10.2: Distribuição de X(– – – –) e $\bar{X}$(——), obtida de 25 amostras de tamanho 2 de {1, 3, 5, 5, 7}.



Com um procedimento análogo podemos obter as distribuições amostrais de outras estatísticas de interesse. As Tabelas 10.4 e 10.5 trazem as distribuições amostrais das estatísticas W = amplitude total e $S^2 = \sum(X_i - \bar{X})^2/(n-1)$, respectivamente.

Tabela 10.4: Distribuição amostral de W.

w	0	2	4	6	Total
$P(W = w)$	7/25	10/25	6/25	2/25	1,00

Tabela 10.5: Distribuição amostral de S^2.

s^2	0	2	8	18	Total
$P(S^2 = s^2)$	7/25	10/25	6/25	2/25	1,00

Exemplo 10.5. (continuação) No caso do lançamento de uma moeda 50 vezes, usando como estatística X = número de caras obtidas, a obtenção da distribuição amostral, que já foi vista, é feita por meio do modelo binomial $b(50, p)$, qualquer que seja p = probabilidade de ocorrência de cara num lançamento, $0 < p < 1$. Se estivermos interessados em julgar a "honestidade" da moeda, estaremos verificando se $p = 0{,}5$. Nessas condições, a $P(X \geqslant 36 | n = 50,\ p = 0{,}5) = 0{,}0013 = 0{,}13\%$.

Portanto, caso a moeda seja honesta, em 50 lançamentos, a probabilidade de se obterem 36 ou mais caras é da ordem de 1 por 1.000. Ou seja, se a moeda fosse honesta, o resultado observado (36 caras) seria muito pouco provável, evidenciando que $p > 0{,}5$.

Comparando os dois últimos exemplos, vemos que nos interessa determinar propriedades das distribuições amostrais que possam ser aplicadas em situações mais gerais (como no caso binomial) e não em situações muito particulares (como no Exemplo 10.7). Iremos, agora, estudar as distribuições amostrais de algumas estatísticas importantes. Nos capítulos seguintes essas distribuições serão usadas para fazer inferências sobre populações.

Quando estivermos trabalhando com populações identificadas pela distribuição de probabilidades, não poderemos gerar *todas as amostras possíveis*. Devemos contentar-nos em simular um número "grande" de amostras e ter uma idéia do que acontece com a estatística de interesse.

Exemplo 10.8. (continuação) Qual seria a distribuição amostral da mediana das alturas de amostras de 5 mulheres retiradas da população $X \sim N(167; 25)$? Como não podemos gerar todas as possíveis amostras de tamanho 5 dessa população, simulamos, via Excel, 200 amostras de tamanho 5 e obtivemos os seguintes resultados:

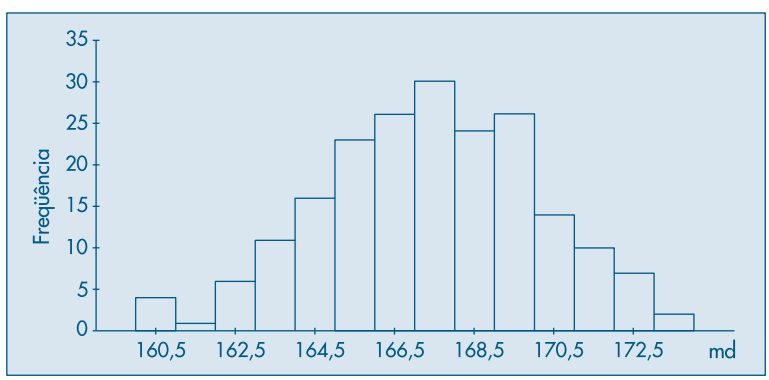
$$E(\text{md}) = 166{,}88, \quad \text{Var(md)} = 7{,}4289, \quad \text{dp(md)} = 2{,}72,$$

$$x_{(1)} = \min(X_1, \ldots, X_{200}) = 160, \quad x_{(200)} = \max(X_1, \ldots, X_{200}) = 173.$$

Observando os resultados somos levados a pensar que a distribuição amostral de md deve ser próxima de uma normal, com média próxima de $\mu = 167$ e desvio padrão menor do que $\sigma = 5$. Veja a Figura 10.3.

Voltaremos a falar na distribuição da mediana amostral em seções futuras.

Figura 10.3: Distribuição amostral da mediana, obtida de 200 amostras de tamanho 5 de X ~ N (167; 25).



Problemas

4. Usando os dados da Tabela 10.2, construa a distribuição amostral da estatística $\hat{\sigma}^2 = \dfrac{\sum (X_i - \bar{X})^2}{n}$.

5. No Problema 3, se X indicar o número de filhos na população, X_1 o número de filhos observados na primeira extração e X_2 na segunda:
   (a) calcule a média e a variância de X;
   (b) calcule $E(X_i)$ e $\text{Var}(X_i)$, $i = 1, 2$;
   (c) construa a distribuição amostral de $\overline{X} = \dfrac{(X_1 + X_2)}{2}$;
   (d) calcule $E(\overline{X})$ e $\text{Var}(\overline{X})$;
   (e) faça num mesmo gráfico os histogramas de X e de $\overline{X}$;
   (f) construa as distribuições amostrais de $S^2 = \sum_{i=1}^{2}(X_i - \overline{X})^2$ e $\hat{\sigma}^2 = \sum_{i=1}^{2}(X_i - \overline{X})^2/2$;
   (g) baseado no resultado de (f), qual dos dois estimadores você usaria para estimar a variância de X? Por quê?
   (h) calcule $P(|\overline{X} - \mu| > 1)$.
6. Ainda com os dados do Problema 3, e para amostras de tamanho 3:
   (a) determine a distribuição amostral de $\overline{X}$ e faça o histograma;
   (b) calcule a média e variância de $\overline{X}$;
   (c) calcule $P(|\overline{X} - \mu| > 1)$.
   (d) se as amostras fossem de tamanho 4, a $P(|\overline{X} - \mu| > 1)$ seria maior ou menor do que a probabilidade encontrada em (c)? Por quê?

10.8 Distribuição Amostral da Média

Vamos estudar agora a distribuição amostral da estatística $\overline{X}$, a média da amostra. Consideremos uma população identificada pela variável X, cujos parâmetros média populacional $\mu = E(X)$ e variância populacional $\sigma^2 = \text{Var}(X)$ são supostos conhecidos. Vamos retirar todas as possíveis AAS de tamanho n dessa população, e para cada uma calcular a média $\overline{X}$. Em seguida, consideremos a distribuição amostral e estudemos suas propriedades. Voltemos a considerar, a título de ilustração, o Exemplo 10.7.

Exemplo 10.10. A população {1, 3, 5, 5, 7} tem média $\mu = 4{,}2$ e variância $\sigma^2 = 4{,}16$. A distribuição amostral de $\overline{X}$ está na Tabela 10.3, da qual obtemos

$$E(\overline{X}) = \sum_i \overline{x}_i\, p_i = 1 \times \frac{1}{25} + 2 \times \frac{2}{25} + 3 \times \frac{5}{25} + 4 \times \frac{6}{25} + 5 \times \frac{6}{25}$$

$$+ 6 \times \frac{4}{25} + 7 \times \frac{1}{25} = 4{,}2.$$

De modo análogo, encontramos

$$\text{Var}(\overline{X}) = 2{,}08.$$

Verificamos, aqui, dois fatos: primeiro, a média das médias amostrais coincide com a média populacional; segundo, a variância de $\overline{X}$ é igual à variância de X, dividida por $n = 2$. Estes dois fatos não são casos isolados. Na realidade, temos o seguinte resultado.

Teorema 10.1. Seja X uma v.a. com média μ e variância σ^2, e seja $(X_1, ..., X_n)$ uma AAS de X. Então,

$$E(\overline{X}) = \mu \quad \text{e} \quad \text{Var}(\overline{X}) = \frac{\sigma^2}{n}.$$

Prova. Pelas propriedades vistas no Capítulo 8, temos:

$$E(\overline{X}) = (1/n)\ \{E(X_1) + ... + E(X_n)\}$$

$$= (1/n)\ \{\mu + \mu + ... + \mu\} = n\mu/n = \mu.$$

De modo análogo, e pelo fato de $X_1, ..., X_n$ serem independentes, temos

$$\text{Var}(\overline{X}) = (1/n^2)\ \{\text{Var}(X_1) + ... + \text{Var}(X_n)\}$$

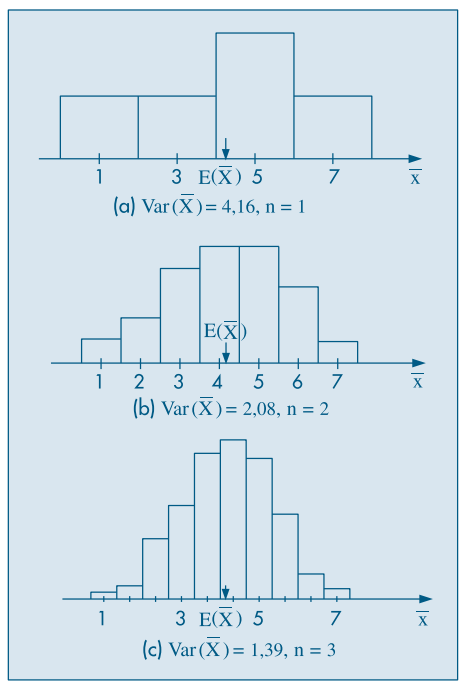
$$= (1/n^2)\ \{\sigma^2 + ... + \sigma^2\} = n\sigma^2/n^2 = \sigma^2/n.$$

Determinamos, então, a média e a variância da distribuição amostral de $\overline{X}$. Vejamos, agora, como obter informação sobre a forma da distribuição dessa estatística.

Exemplo 10.10. (continuação) Para a população {1, 3, 5, 5, 7}, vamos construir os histogramas das distribuições de $\overline{X}$ para n = 1, 2 e 3.

(i) Para n = 1, vemos que a distribuição de $\overline{X}$ coincide com a distribuição de X, com $E(\overline{X}) = E(X) = 4{,}2$ e $\text{Var}(\overline{X}) = \text{Var}(X) = 4{,}16$ (Figura 10.4(a)).

Figura 10.4: Distribuição de $\overline{X}$ para amostras de {1, 3, 5, 5, 7}.



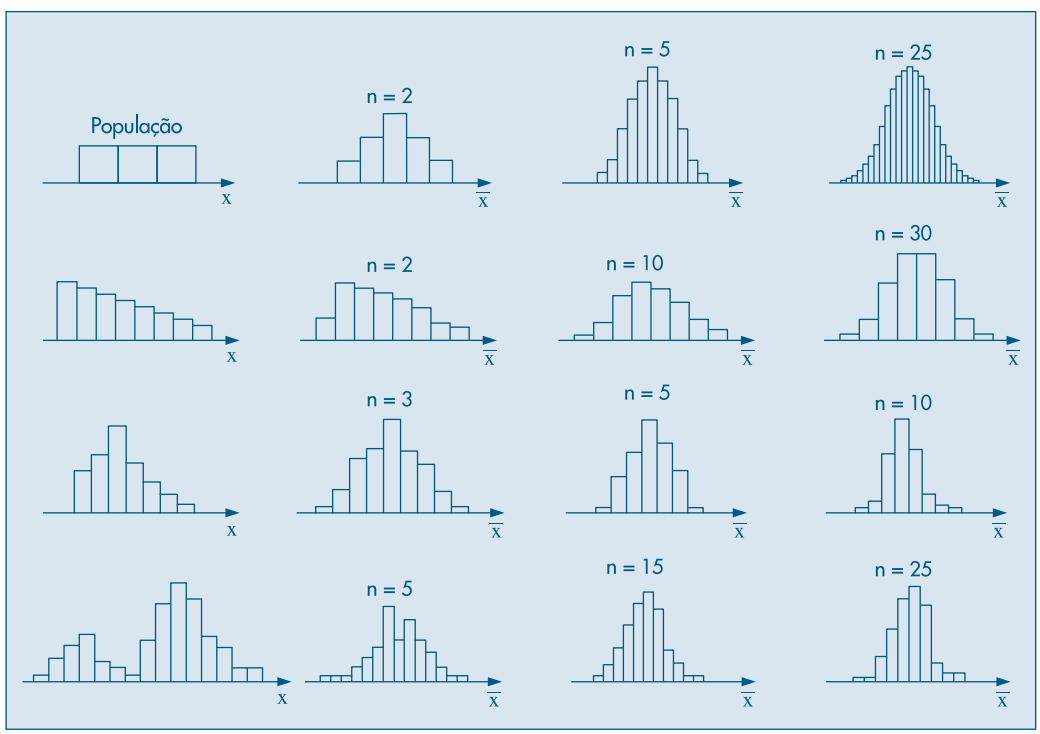
(ii) Para $n = 2$, baseados na Tabela 10.3, temos a distribuição de $\overline{X}$ dada na Figura 10.4(b), com $E(\overline{X}) = 4{,}2$ e $\text{Var}(\overline{X}) = 2{,}08$.

(iii) Finalmente, para $n = 3$, com os dados da Tabela 10.6, temos a distribuição de $\overline{X}$ na Figura 10.4 (c), com $E(\overline{X}) = 4{,}2$ e $\text{Var}(\overline{X}) = 1{,}39$.

Observe que, conforme n vai aumentando, o histograma tende a se concentrar cada vez mais em torno de $E(\overline{X}) = E(X) = 4{,}2$, já que a variância vai diminuindo. Os casos extremos passam a ter pequena probabilidade de ocorrência. Quando n for suficientemente grande, o histograma alisado aproxima-se de uma distribuição normal. Essa aproximação pode ser verificada analisando-se os gráficos da Figura 10.5, que mostram o comportamento do histograma de $\overline{X}$ para várias formas da distribuição da população e vários valores do tamanho da amostra n.

Esses exemplos sugerem que, quando o tamanho da amostra aumenta, independentemente da forma da distribuição da população, a distribuição amostral de $\overline{X}$ aproxima-se cada vez mais de uma distribuição normal. Esse resultado, fundamental na teoria da Inferência Estatística, é conhecido como *Teorema Limite Central* (TLC).

Figura 10.5: Histogramas correspondentes às distribuições amostrais de $\overline{X}$ para amostras extraídas de algumas populações.



Teorema 10.2. (TLC) Para amostras aleatórias simples $(X_1, ..., X_n)$, retiradas de uma população com média μ e variância σ^2 finita, a distribuição amostral da média $\overline{X}$ aproxima-se, para n grande, de uma distribuição normal, com média μ e variância σ^2/n.

A demonstração completa desse teorema exigiria recursos dos quais não dispomos, portanto não será dada, mas o importante é sabermos como esse resultado pode ser usado.

Observemos que, se a população for normal, então $\overline{X}$ terá distribuição *exata* normal. Esse resultado segue do fato de que a distribuição de uma combinação linear de v.a.'s normais independentes tem ainda distribuição normal. No caso da $\overline{X}$, a média e variância dessa normal serão dadas pelo Teorema 10.1. A prova dessa propriedade depende do conceito de função geradora de momentos, que não será objeto deste livro. O leitor interessado pode consultar Meyer (1965), por exemplo.

Exemplo 10.11. Voltemos ao Exemplo 10.4, onde uma máquina enchia pacotes cujos pesos seguiam uma distribuição $N(500, 100)$. Colhendo-se um amostra de $n = 100$ pacotes e pesando-os, pelo que foi dito acima, $\overline{X}$ terá uma distribuição normal com média 500 e variância 100/100 = 1. Logo, se a máquina estiver regulada, a probabilidade de encontrarmos a média de 100 pacotes diferindo de 500 g de menos de 2 gramas será

$$P(|\overline{X} - 500| < 2) = P(498 < \overline{X} < 502) = P(-2 < Z < 2) \approx 95\%.$$

Ou seja, dificilmente 100 pacotes terão uma média fora do intervalo (498, 502). Caso 100 pacotes apresentem uma média fora desse intervalo, podemos considerar como um evento raro, e será razoável supor que a máquina esteja desregulada.

Outra maneira de apresentar o TLC é por meio do

Corolário 10.1. Se $(X_1, ..., X_n)$ for uma amostra aleatória simples da população X, com média μ e variância σ^2 finita, e $\overline{X} = (X_1 + ... + X_n)/n$, então

$$Z = \frac{\overline{X} - \mu}{\sigma/\sqrt{n}} \sim N(0, 1). \tag{10.2}$$

Basta notar que se usou a transformação usual de reduzir a distribuição de $\overline{X}$ a uma normal padrão. Observe, também, que (10.2) pode ser escrita como

$$Z = \frac{\sqrt{n}\,(\overline{X} - \mu)}{\sigma} \sim N(0, 1). \tag{10.3}$$

Chamemos de e a v.a. que mede a diferença entre a estatística $\overline{X}$ e o parâmetro μ, isto é, $e = \overline{X} - \mu$; e é chamado o *erro amostral da média*. Então, temos o

Corolário 10.2. A distribuição de e aproxima-se de uma distribuição normal com média 0 e variância σ^2/n, isto é,

$$\frac{\sqrt{n}\,e}{\sigma} \sim N(0, 1). \tag{10.4}$$

O TLC afirma que $\overline{X}$ aproxima-se de uma normal quando n tende para o infinito, e a rapidez dessa convergência (veja a Figura 10.5) depende da distribuição da popula-

ção da qual a amostra é retirada. Se a população original tem uma distribuição próxima da normal, a convergência é rápida; se a população original se afasta muito de uma normal, a convergência é mais lenta, ou seja, necessitamos de uma amostra maior para que $\overline{X}$ tenha uma distribuição aproximadamente normal. Para amostras da ordem de 30 ou 50 elementos, a aproximação pode ser considerada boa.

Problemas

7. Uma v.a. X tem distribuição normal, com média 100 e desvio padrão 10.
   (a) Qual a $P(90 < X < 110)$?
   (b) Se $\overline{X}$ for a média de uma amostra de 16 elementos retirados dessa população, calcule $P(90 < \overline{X} < 110)$.
   (c) Represente, num único gráfico, as distribuições de X e $\overline{X}$.
   (d) Que tamanho deveria ter a amostra para que $P(90 < \overline{X} < 110) = 0{,}95$?

8. A máquina de empacotar um determinado produto o faz segundo uma distribuição normal, com média μ e desvio padrão 10 g.
   (a) Em quanto deve ser regulado o peso médio μ para que apenas 10% dos pacotes tenham menos do que 500 g?
   (b) Com a máquina assim regulada, qual a probabilidade de que o peso total de 4 pacotes escolhidos ao acaso seja inferior a 2 kg?

9. No exemplo anterior, e após a máquina estar regulada, programou-se uma carta de controle de qualidade. De hora em hora, será retirada uma amostra de quatro pacotes e esses serão pesados. Se a média da amostra for inferior a 495 g ou superior a 520 g, encerra-se a produção para reajustar a máquina, isto é, reajustar o peso médio.
   (a) Qual é a probabilidade de ser feita uma parada desnecessária?
   (b) Se o peso médio da máquina desregulou-se para 500 g, qual é a probabilidade de continuar a produção fora dos padrões desejados?

10. A capacidade máxima de um elevador é de 500 kg. Se a distribuição X dos pesos dos usuários for suposta $N(70, 100)$:
    (a) Qual é a probabilidade de sete passageiros ultrapassarem esse limite?
    (b) E seis passageiros?

10.9 Distribuição Amostral de uma Proporção

Vamos considerar uma população em que a proporção de elementos portadores de certa característica é p. Logo, podemos definir uma v.a. X, da seguinte maneira:

$$X = \begin{cases} 1, & \text{se o indivíduo for portador da característica} \\ 0, & \text{se o indivíduo não for portador da característica,} \end{cases}$$

logo,

$$\mu = E(X) = p, \quad \sigma^2 = \text{Var}(X) = p(1 - p).$$

Retirada uma AAS dessa população, e indicando por Y_n o total de indivíduos portadores da característica na amostra, já vimos que

$$Y_n \sim b(n, p).$$

Vamos definir por $\hat{p}$ a proporção de indivíduos portadores da característica na amostra, isto é,

$$\hat{p} = \frac{Y_n}{n}.$$

Então,

$$P(Y_n = k) = P(Y_n/n = k/n) = P(\hat{p} = k/n),$$

ou seja, a distribuição amostral de $\hat{p}$ é obtida da distribuição de Y_n.

Vimos na seção 7.5 que a distribuição binomial pode ser aproximada pela distribuição normal. Vamos mostrar que a justificativa desse fato está no TLC. Inicialmente, observe que

$$Y_n = X_1 + X_2 + \ldots + X_n,$$

onde cada X_i tem distribuição de Bernoulli, com média $\mu = p$ e variância $\sigma^2 = p(1 - p)$, e são duas a duas independentes. Podemos escrever que

$$Y_n = n\overline{X},$$

mas pelo TLC, $\overline{X}$ terá distribuição aproximadamente normal, com média p e variância $\frac{p(1-p)}{n}$, ou seja,

$$\overline{X} \sim N\left(p, \frac{p(1-p)}{n}\right).$$

Logo, a transformação $Y_n = n\overline{X}$ terá a distribuição

$$Y_n \sim N(np, np(1 - p)),$$

que foi a aproximação adotada na seção 7.5.

Observe que $\overline{X}$, na expressão acima, é a própria variável $\hat{p}$ e, desse modo, para n grande podemos considerar a distribuição amostral de p como aproximadamente normal:

$$\hat{p} \sim N\left(p, \frac{p(1-p)}{n}\right).$$

Exemplo 10.12. Suponha que $p = 30\%$ dos estudantes de uma escola sejam mulheres. Colhemos uma AAS de $n = 10$ estudantes e calculamos $\hat{p}$ = proporção de mulheres na amostra. Qual a probabilidade de que $\hat{p}$ difira de p em menos de 0,01? Temos que essa probabilidade é dada por

$$P(|\hat{p} - p| < 0{,}01) = P(-0{,}01 < \hat{p} - p < 0{,}01).$$

Mas, $\hat{p} - p \sim N\left(0, \frac{p(1-p)}{n}\right)$, e como $p = 0{,}3$, temos que

$$\text{Var}(\hat{p}) = (0,\ 3)(0,\ 7)/10 = 0{,}021,$$

e, portanto, a probabilidade pedida é igual a

$$P\left(\frac{-0{,}01}{\sqrt{0{,}021}} < Z < \frac{0{,}01}{\sqrt{0{,}021}}\right) = P(-0{,}07 < Z < 0{,}07) = 0{,}056.$$

Problemas

11. Sabe-se que 20% das peças de um lote são defeituosas. Sorteiam-se oito peças, com reposição, e calcula-se a proporção $\hat{p}$ de peças defeituosas na amostra.
    (a) Construa a distribuição exata de $\hat{p}$ (use a tábua da distribuição binomial).
    (b) Construa a aproximação normal à binomial.
    (c) Você pensa que a segunda distribuição é uma boa aproximação da primeira?
    (d) Já sabemos que, para dado p fixo, a aproximação melhora à medida que n aumenta. Agora, se n for fixo, para qual valor de p a aproximação é melhor?

12. Um procedimento de controle de qualidade foi planejado para garantir um máximo de 10% de itens defeituosos na produção. A cada 6 horas sorteia-se uma amostra de 20 peças e, havendo mais de 15% de defeituosas, encerra-se a produção para verificação do processo. Qual a probabilidade de uma parada desnecessária?

13. Supondo que a produção do exemplo anterior esteja sob controle, isto é, $p = 10\%$, e que os itens sejam vendidos em caixas com 100 unidades, qual a probabilidade de que uma caixa:
    (a) tenha mais do que 10% de defeituosos?
    (b) não tenha itens defeituosos?

10.10 Outras Distribuições Amostrais

Do mesmo modo que estudamos a distribuição amostral de $\bar{X}$, podemos, em princípio, estudar a distribuição amostral de qualquer estatística $T = f(X_1, \ldots, X_n)$. Mas, quanto mais complexa for essa relação f, mais difícil será a derivação matemática das propriedades dessa estatística. Vejamos alguns exemplos.

Exemplo 10.13. Na Tabela 10.6 apresentamos a distribuição de três outras estatísticas; a variância da amostra,

$$S^2 = \frac{1}{(n-1)} \sum_{i=1}^{n} (X_i - \bar{X})^2,$$

a mediana amostral, *md*, e o estimador

$$\hat{\sigma}^2 = \frac{1}{n} \sum_{i=1}^{n} (X_i - \bar{X})^2,$$

que difere de S^2 apenas no denominador, e que foi estudado no Capítulo 3. Desta tabela, obtemos as distribuições amostrais apresentadas nas Tabelas 10.7, 10.8 e 10.9.

Tabela 10.6: Distribuição amostral de algumas estatísticas obtidas de amostra de tamanho $n = 3$, retiradas da população $\{1, 3, 5, 5, 7\}$ ($\mu = 4{,}2$, $\sigma^2 = 4{,}16$ e $Md = 5$).

Tipo de amostra	Freqüência (prob. × 125)	Soma	Soma dos quadrados	Média $\bar{x}$	Mediana md	Variância	
						s^2	$\hat{\sigma}^2$
111	1	3	3	1,00	1	0	0
113	3	5	11	1,67	1	4/3	8/9
115	6	7	27	2,33	1	16/3	32/9
117	3	9	51	3,00	1	12	8
133	3	7	19	2,33	3	4/3	8/9
135	12	9	35	3,00	3	4	8/3
137	6	11	59	3,67	3	28/3	56/9
155	12	11	51	3,67	5	16/3	32/9
157	12	13	75	4,33	5	28/3	56/9
177	3	15	99	5,00	7	12	8
333	1	9	27	3,00	3	0	0
335	6	11	43	3,67	3	4/3	8/9
337	3	13	67	4,33	3	16/3	32/9
355	12	13	59	4,33	5	4/3	8/9
357	12	15	83	5,00	5	4	8/3
377	3	17	107	5,67	7	16/3	32/9
555	8	15	75	5,00	5	0	0
557	12	17	99	5,67	5	4/3	8/9
577	6	19	123	6,33	7	4/3	8/9
777	1	21	147	7,00	7	0	0
Total	125						

Tabela 10.7: Distribuição amostral da variância S^2, para amostras de tamanho 3, retiradas da população $\{1, 3, 5, 5, 7\}$.

s^2	0,00	1,33	4,00	5,33	9,33	12,00
$P(S^2 = s^2)$	11/125	42/125	24/125	24/125	18/125	6/125

$E(S^2) = 4{,}16, \quad \text{Var}(S^2) = 11{,}28.$

Tabela 10.8: Distribuição amostral da mediana da amostra md para amostras de tamanho 3, retiradas da população $\{1, 3, 5, 5, 7\}$.

md	1	3	5	7
Prob.	13/125	31/125	68/125	13/125

$E(md) = 4{,}30, \quad \text{Var}(md) = 2{,}54.$

Tabela 10.9: Distribuição amostral da variância $\hat{\sigma}^2$, para amostras de tamanho 3, retiradas da população {1, 3, 5, 5, 7}.

$\hat{\sigma}^2$	0,00	0,89	2,67	3,56	6,22	8,00
Prob.	11/125	42/125	24/125	24/125	18/125	6/125

$$E(\hat{\sigma}^2) = 2{,}77, \quad \text{Var}(\hat{\sigma}^2) = 5{,}04.$$

Os gráficos das funções de probabilidade estão nas Figuras 10.6, 10.7 e 10.8. A obtenção das propriedades dessas estatísticas, de modo geral, não é uma tarefa fácil, e os modelos de probabilidade resultantes correspondem a distribuições mais complexas.

Figura 10.6: Distribuição amostral de S^2 para amostras de tamanho $n = 3$ extraídas de {1, 3, 5, 5, 7}.

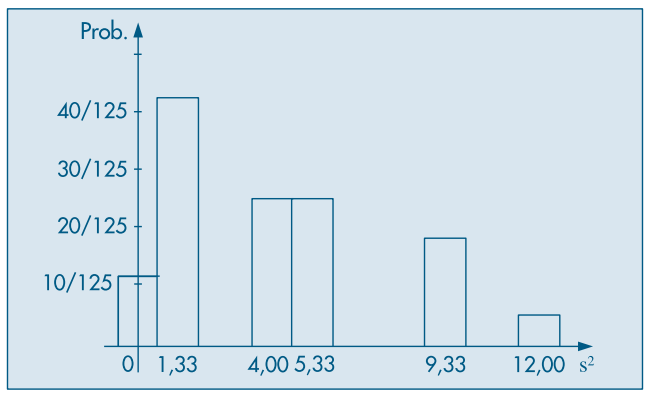


Figura 10.7: Distribuição amostral de md para amostras de tamanho $n = 3$ de {1, 3, 5, 5, 7}.

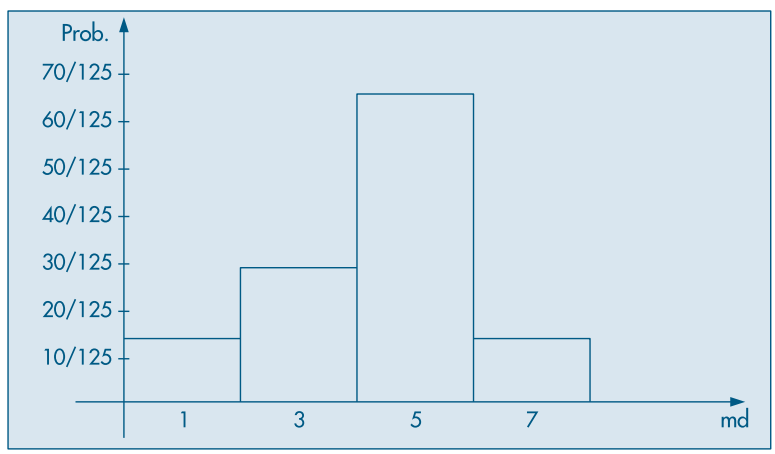
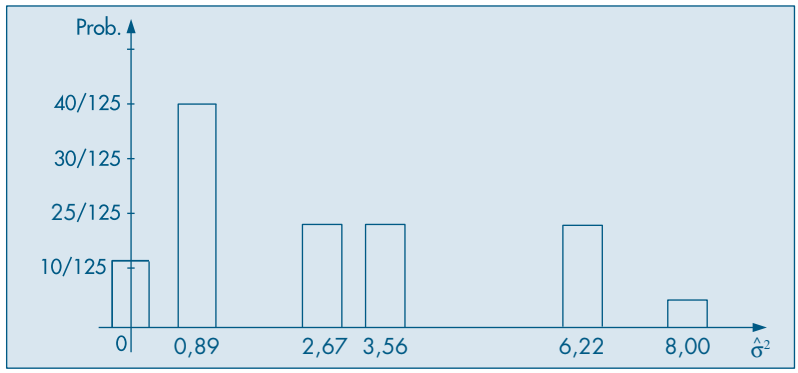


Figura 10.8: Distribuição amostral de $\hat{\sigma}^2$ para amostras de tamanho $n = 3$ extraídas de $\{1, 3, 5, 5, 7\}$.



Por exemplo, note que $E(S^2) = 4{,}16 = \sigma^2$, logo S^2 satisfaz uma propriedade análoga a $E(\bar{X}) = \mu$; dizemos que $\bar{X}$ e S^2 são estimadores *não-viesados* dos respectivos parâmetros μ e σ^2. Esta propriedade já não vale para md e $\hat{\sigma}^2$, pois $E(md) = 4{,}3$, enquanto $Md = 5{,}0$ e $E(\hat{\sigma}^2) = 2{,}77$ e não 4,16. Vemos que $\hat{\sigma}^2$ sub-estima a verdadeira variância.

Também pode-se demonstrar que S^2 segue uma distribuição que é um múltiplo de uma distribuição qui-quadrado (χ^2), quando a população tem distribuição normal. Ver a seção 11.9. Já a mediana *md*, obtida de amostras de uma população simétrica, com média μ e variância σ^2, segue aproximadamente uma distribuição normal, com média $E(md) = \mu$ e $\text{Var}(md) = (\pi\sigma^2)/(2n)$. Note que se exigem mais suposições do que aquelas mencionada no TLC. Nos Capítulos 11 e 12 voltaremos a discutir algumas distribuições amostrais e suas aplicações.

Problemas

14. Usando os dados da Tabela 10.2:
    (a) construa a distribuição amostral de $\hat{\sigma}^2$ e compare com a distribuição amostral de S^2 (Tabela 10.5). Você notou alguma propriedade de S^2 que seja "melhor" do que de $\hat{\sigma}^2$?
    (b) seja U a média de elementos distintos de amostras de tamanho $n = 3$. Por exemplo, se a amostra observada for (1, 1, 3), então $u = (1 + 3)/2 = 2$. Construa a distribuição amostral de U;
    (c) compare as distribuições amostrais de U e $\bar{X}$.

15. Na tabela abaixo tem-se a distribuição dos salários da Secretaria A.

Classes de salários	Freqüência relativa
4,5 ⊢ 7,5	0,10
7,5 ⊢ 10,5	0,20
10,5 ⊢ 13,5	0,40
13,5 ⊢ 16,5	0,20
16,5 ⊢ 19,5	0,10

(a) Calcule a média, a variância e a mediana dos salários nessa população.
(b) Construa a distribuição amostral da média e da mediana para amostras de tamanho 2, retiradas dessa população.
(c) Mostre que a média $\overline{X}$ e a mediana md da amostra são estimadores não-viesados da mediana Md da população, no sentido que $E(\overline{X}) = E(\text{md}) = \text{Md}$.
(d) Qual dos dois estimadores não-viesados você usaria para estimar Md nesse caso? Por quê?
(e) Baseado na distribuição amostral da média, encontre a distribuição amostral da estatística

$$Z = \frac{\overline{X} - \mu}{\sigma}\sqrt{n},$$

para $n = 2$.
(f) Quais são os valores de $E(Z)$ e $\text{Var}(Z)$?
(g) Construa a distribuição amostral da estatística

$$S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \overline{X})^2,$$

e faça o seu histograma.
(h) Calcule a média e variância de S^2.
(i) Baseando-se nas distribuições amostrais anteriores, determine a distribuição amostral da estatística

$$t = \frac{\overline{X} - \mu}{S}\sqrt{n},$$

e construa seu histograma. Qual o problema encontrado?
(j) Calcule a média e variância de t, quando possível.
(k) Calcule a $P(|t| < 2)$ e $P(|t| < 4{,}30)$.

16. Tente esboçar como ficariam os histogramas das estatísticas abaixo, para amostras de tamanho grande.
(a) S^2 (faça o histograma da distribuição da Tabela 10.5)
(b) $Z = \frac{\overline{X} - \mu}{\sigma}\sqrt{n}$ (Veja o Teorema Limite Central)
(c) $t = \frac{\overline{X} - \mu}{S}\sqrt{n}$, definida no problema anterior (compare com a expressão e o resultado obtido em (b)).

10.11 Determinação do Tamanho de uma Amostra

Em nossas considerações anteriores fizemos a suposição que o tamanho da amostra, n, era conhecido e fixo. Podemos, em certas ocasiões, querer determinar o tamanho da amostra a ser escolhida de uma população, de modo a obter um erro de estimação previamente estipulado, com determinado grau de confiança.

Por exemplo, suponha que estejamos estimando a média μ populacional e para tanto usaremos a média amostral, $\overline{X}$, baseada numa amostra de tamanho n. Suponha que se queira determinar o valor de n de modo que

$$P(|\overline{X} - \mu| \leq \varepsilon) \geq \gamma, \tag{10.5}$$

com $0 < \gamma < 1$ e ε é o *erro amostral* máximo que podemos suportar, ambos valores fixados.

Sabemos que $\overline{X} \sim N(\mu, \sigma^2/n)$, logo $\overline{X} - \mu \sim N(0, \sigma^2/n)$ e portanto (10.5) pode ser escrita

$$P(-\varepsilon \leq \overline{X} - \mu \leq \varepsilon) = P\left(\frac{-\sqrt{n}\varepsilon}{\sigma} \leq Z \leq \frac{\sqrt{n}\varepsilon}{\sigma}\right) \approx \gamma,$$

com $Z = (\overline{X} - \mu)\sqrt{n}/\sigma$. Dado γ, podemos obter z_γ da $N(0,1)$, tal que $P(-z_\gamma < Z < z_\gamma) = \gamma$, de modo que

$$\frac{\sqrt{n}\,\varepsilon}{\sigma} = z_\gamma,$$

do que obtemos finalmente

$$n = \frac{\sigma^2 z_\gamma^2}{\varepsilon^2}. \tag{10.6}$$

Note que em (10.6) conhecemos z_γ e ε, mas σ^2 é a variância desconhecida da população. Para podermos ter uma idéia sobre n devemos ter alguma informação prévia sobre σ^2 ou, então, usar uma pequena amostra piloto para estimar σ^2.

Exemplo 10.13. (continuação) Suponha que uma pequena amostra piloto de $n = 10$, extraída de uma população, forneceu os valores $\overline{X} = 15$ e $S^2 = 16$. Fixando-se $\varepsilon = 0{,}5$ e $\gamma = 0{,}95$, temos

$$n = \frac{16 \times (1{,}96)^2}{(0{,}5)^2} \approx 245.$$

No caso de proporções, usando a aproximação normal da seção 10.9 para $\hat{p}$, é fácil ver que (10.6) resulta

$$n = \frac{z_\gamma^2 p(1-p)}{\varepsilon^2}. \tag{10.7}$$

Como não conhecemos p, a verdadeira proporção populacional, podemos usar o fato de que $p(1-p) \leq 1/4$, para todo p, e (10.7) fica

$$n \approx \frac{z_\gamma^2}{4\varepsilon^2}. \tag{10.8}$$

Por outro lado, se tivermos alguma informação sobre p ou pudermos estimá-lo usando uma amostra piloto, basta substituir esse valor estimado em (10.7).

Exemplo 10.14. Suponha que numa pesquisa de mercado estima-se que no mínimo 60% das pessoas entrevistadas preferirão a marca A de um produto. Essa informação é baseada em pesquisas anteriores. Se quisermos que o erro amostral de $\hat{p}$ seja menor do que $\mathcal{E} = 0{,}03$, com probabilidade $\gamma = 0{,}95$, teremos

$$n \approx \frac{(1{,}96)^2(0{,}6)(0{,}4)}{(0{,}03)^2} = 1.024,$$

na qual usamos o fato de que $p \geqslant 0{,}60$. Veja também os Problemas 19, 20 e 41.

Problemas

17. Suponha que uma indústria farmacêutica deseja saber a quantos voluntários se deva aplicar uma vacina, de modo que a proporção de indivíduos imunizados na amostra difira de menos de 2% da proporção verdadeira de imunizados na população, com probabilidade 90%. Qual o tamanho da amostra a escolher? Use (10.8).

18. No problema anterior, suponha que a indústria tenha a informação de que a proporção de imunizados pela vacina seja $p \geqslant 0{,}80$. Qual o novo tamanho de amostra a escolher? Houve redução?

19. Seja o tamanho de amostra dado por (10.7) e n_0 dado por (10.8). Prove que, para todo p, temos $n \leqslant n_0$. (Use a função $f(p) = p(1-p)$ para sua resposta.)

20. Suponha que haja a informação $p \leqslant p_0 < 0{,}5$, com p_0 conhecida. Se $n_1 = z_\gamma^2 p_0(1-p_0)/\mathcal{E}^2$, mostre que $n \leqslant n_1 < n_0$. Mostre que essa mesma relação vale se soubermos que $p \geqslant p_0 > 0{,}5$.
[Sugestão: note que $f(p) = p(1-p)$ é crescente em $[0; 0{,}5]$, atinge o máximo em 0,5 e depois é decrescente em $[0{,}5; 1]$.]

10.12 Exemplos Computacionais

Vimos, no Exemplo 10.7, como escolher todas as possíveis amostras de tamanho $n = 2$, com reposição, da população $\{1, 3, 5, 5, 7\}$. Obtemos $5^2 = 25$ amostras. Como já salientamos em seções anteriores, ao escolher uma amostra de uma população, estamos na realidade gerando valores de uma v.a. com determinada distribuição de probabilidades, supostamente conhecida. No exemplo, podemos pensar na v.a. X, assumindo os valores $x_1 = 1$, $x_2 = 3$, $x_3 = 5$, $x_4 = 5$, $x_5 = 7$, com probabilidades todas iguais a 0,2. Portanto, para escolher uma amostra de tamanho $n = 2$, basta gerar dois valores dessa distribuição, como aprendemos no Capítulo 9.

Os programas Excel, SPlus e Minitab têm comandos apropriados para gerar amostras de uma população especificada.

Exemplo 10.15. O Excel usa a opção *Amostragem*, dentro de "Análise de Dados" do menu "Ferramentas". Na coluna G do quadro do Exemplo 9.5, temos uma amostra aleatória simples (com reposição), de tamanho $n = 5$ da população $\mathcal{P} = \{1, 2, ..., 10\}$, que está na coluna F.

Exemplo 10.16. O SPlus usa o comando *sample(x,n)* para gerar uma amostra *sem* reposição de tamanho n do conjunto x e o comando *sample(x,n,replace*=T) para gerar uma amostra *com* reposição. O Quadro 10.1 mostra como obter amostras de tamanho $n = 7$ do conjunto $x = \{1, 2, 3, ..., 15\}$, sem e com reposição.

Quadro 10.1: Geração de amostras. SPlus.

```
> x<–c (1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15)
>
>
> sample (x, 7)
[1]  6  7  4  2  3  10  5
>
>
> sample (x, 7, replace=T)
[1] 12 14 11 10 15 4 11
```

Exemplo 10.17. O Minitab usa os comandos Sample e Replace para obter amostras. Temos, no Quadro 10.2, amostras de tamanho $n = 5$ obtidas do conjunto $\{1, 2, ..., 10\}$ (na coluna C1). Na coluna C2 temos uma amostra sem reposição e na coluna C3 uma amostra com reposição.

Quadro 10.2: Geração de amostras. Minitab.

```
      C1   C2   C3
 1     1   10    8
 2     2    1    3
 3     3    8    8      MTB > Sample 5 C1 C2.
 4     4    2    6      MTB >
 5     5    7    4      MTB > Sample 5 C1 C3;
 6     6                SUBC> Replace.
 7     7                MTB >
 8     8
 9     9
10    10
```

10.13 Problemas e Complementos

21. Uma v.a. X tem distribuição normal com média 10 e desvio padrão 4. Aos participantes de um jogo é permitido observar uma amostra de qualquer tamanho e calcular a média amostral. Ganha um prêmio aquele cuja média amostral for maior que 12.
    (a) Se um participante escolher uma amostra de tamanho 16, qual é a probabilidade de ele ganhar um prêmio?
    (b) Escolha um tamanho de amostra diferente de 16 para participar do jogo. Qual é a probabilidade de você ganhar um prêmio?
    (c) Baseado nos resultados acima, qual o melhor tamanho de amostra para participar do jogo?

22. Se uma amostra com 36 observações for tomada de uma população, qual deve ser o tamanho de uma outra amostra para que o desvio padrão dessa amostra seja 2/3 do desvio padrão da média da primeira?

23. Definimos a variável $e = \overline{X} - \mu$ como sendo o erro amostral de média. Suponha que a variância dos salários de uma certa região seja 400 reais^2.
    (a) Determine a média e a variância de e.
    (b) Que proporção das amostras de tamanho 25 terão erro amostral absoluto maior do que 2 reais?
    (c) E qual a proporção das amostras de tamanho 100?
    (d) Nesse último caso, qual o valor de d, tal que $P(|e| > d) = 1\%$?
    (e) Qual deve ser o tamanho da amostra para que 95% dos erros amostrais absolutos sejam inferiores a um real?

24. A distribuição dos comprimentos dos elos da corrente de bicicleta é normal, com média 2 cm e variância 0,01 cm^2. Para que uma corrente se ajuste à bicicleta, deve ter comprimento total entre 58 e 61 cm.
    (a) Qual é a probabilidade de uma corrente com 30 elos não se ajustar à bicicleta?
    (b) E para uma corrente com 29 elos?
    [Observação: suponha que os elos sejam selecionados ao acaso para compor a corrente, de modo que se tenha independência.]

25. Cada seção usada para a construção de um oleoduto tem um comprimento médio de 5 m e desvio padrão de 20 cm. O comprimento total do oleoduto será de 8 km.
    (a) Se a firma construtora do oleoduto encomendar 1.600 seções, qual é a probabilidade de ela ter de comprar mais do que uma seção adicional (isto é, de as 1.600 seções somarem menos do que 7.995 m)?
    (b) Qual é a probabilidade do uso exato de 1.599 seções, isto é, a soma das 1.599 seções estar entre 8.000 m e 8.005 m?

26. Um professor dá um teste rápido, constante de 20 questões do tipo certo-errado. Para testar a hipótese de o estudante estar adivinhando a resposta, ele adota a seguinte regra de decisão: "Se 13 ou mais questões estiverem corretas, ele não está adivinhando". Qual é a probabilidade de rejeitarmos a hipótese, sendo que na realidade ela é verdadeira?

27. Um distribuidor de sementes determina, por meio de testes, que 5% das sementes não germinam. Ele vende pacotes com 200 sementes com garantia de 90% de germinação. Qual é a probabilidade de que um pacote não satisfaça à garantia?

28. Uma empresa fabrica cilindros com 50 mm de diâmetro, sendo o desvio padrão 2,5 mm. Os diâmetros de uma amostra de quatro cilindros são medidos a cada hora. A média da amostra é usada para decidir se o processo de fabricação está operando satisfatoriamente. Aplica-se a seguinte regra de decisão: "Se o diâmetro médio de amostra de quatro cilindros for maior ou igual a 53,7 mm, ou menor ou igual a 46,3 mm, deve-se parar o processo. Se o diâmetro médio estiver entre 46,3 e 53,7 mm, o processo continua.
    (a) Qual é a probabilidade de se parar o processo se a média dos diâmetros permanecer em 50 mm?
    (b) Qual é a probabilidade de o processo continuar se a média dos diâmetros se deslocar para 53,7 mm?

29. O CD-Veículos traz os preços de 30 carros nacionais e importados, extraídos da população de todos os carros vendidos no mercado. Supondo que o desvio padrão dessa amostra seja um bom representante do verdadeiro desvio padrão da população, qual será o tamanho de uma outra amostra a ser escolhida, de modo que, com probabilidade 90%, a média amostral difira da verdadeira média de menos de 0,02?

30. Tabela de Números Aleatórios. Para sortear AAS, costuma-se usar tabelas de números aleatórios, que são coleções de dígitos construídos aleatoriamente e que simulam o processo de sorteio. Na Tabela VII, apresentamos um pequeno conjunto de números aleatórios. Podem ser usados do seguinte modo: se quisermos selecionar dez nomes de uma lista de 90 pessoas, devemos começar numerando-os 01, 02, ..., 90. Em seguida, escolhemos duas colunas, digamos as duas primeiras, e tomamos os dez primeiros números; no caso, serão:
61, 94, 50, 51, 25, 63, 12, 38, 22, 07, 61.
Observe que o 94 foi eliminado, pois não existe esse número na população, e o 61 deverá aparecer repetido. Para outras explicações e tabelas maiores, consultar Pereira e Bussab (1974).

31. Como você usaria uma tabela (ou um gerador) de números aleatórios para sortear uma amostra nas seguintes situações:
   (a) 5 alunos de sua classe;
   (b) 10 alunos de sua escola;
   (c) 15 domicílios de seu bairro;
   (d) 20 ações negociadas na Bolsa de São Paulo;
   (e) 5 números de uma população cujos elementos são numerados de 1 a 115. Existe algum modo de "apressar" o sorteio?
   (f) 5 números de uma população de 115 nomes, cujos números vão de 612 a 726;
   (g) 5 números de uma população de 115 nomes, cuja numeração não é seqüencial, mas está compreendida entre os números 300 e 599.

32. Distribuição amostral da diferença de duas médias. Consideremos duas populações X com parâmetros μ_1 e σ_1^2 e Y com parâmetros μ^2 e σ_2^2. Sorteiam-se duas amostras independentes: a da primeira população de tamanho n e a da segunda de tamanho m. Calculam-se as médias amostrais $\overline{X}$ e $\overline{Y}$.
   (a) Qual a distribuição amostral de $\overline{X}$? E de $\overline{Y}$?
   (b) Defina $D = \overline{X} - \overline{Y}$. O que você entende por distribuição amostral de D?
   (c) Calcule $E(D)$ e $\text{Var}(D)$.
   (d) Como você acha que será a distribuição de D? Por quê?

33. A distribuição dos salários (em salários mínimos) de operários do sexo masculino de uma grande fábrica é $N(5{,}4;\ 1{,}69)$, e a de operários do sexo feminino é $N(5{,}4;\ 2{,}25)$.
Sorteiam-se duas amostras, uma com 16 homens e outra com 16 mulheres. Se D for a diferença entre o salário médio dos homens e das mulheres:
   (a) Calcule $P(|D| > 0{,}5)$.
   (b) Qual o valor de d tal que $P(|D| > d) = 0{,}05$?
   (c) Que tamanho comum deveriam ter ambas as amostras para que $P(|D| > 0{,}4) = 0{,}05$?

34. Numa escola A, os alunos submetidos a um teste obtiveram média 70, com desvio padrão 10. Em outra escola B, os alunos submetidos ao mesmo teste obtiveram média 65 e desvio padrão 15. Se colhermos na escola A uma amostra de 36 alunos e na B, uma de 49 alunos, qual é a probabilidade de que a diferença entre as médias seja superior a 6 unidades?

35. Distribuição amostral da diferença de duas proporções. Usando os resultados do Problema 32, qual seria a distribuição de $\hat{p}_1 - \hat{p}_2$, a diferença entre as proporções de amostras independentes retiradas de populações com parâmetros p_1 e p_2?

36. Amostras sem reposição de populações finitas. Suponha uma população com N elementos. Vimos que se extrairmos uma amostra de tamanho n, *com reposição*, e calcularmos a média amostral $\bar{X}$, então $E(\bar{X}) = \mu$ e $\text{Var}(\bar{X}) = \sigma^2/n$, onde μ e σ^2 são a média e a variância da população, respectivamente. No entanto, se a amostragem for feita *sem reposição*, então $E(\bar{X}) = \mu$ continua a valer, mas

$$\text{Var}(\bar{X}) = \frac{\sigma^2}{n}\,\frac{N-n}{N-1}.$$

O fator $(N - n)/(N - 1)$ é chamado *fator de correção para populações finitas*. Note que se n for muito menor que N, então esse fator é aproximadamente igual a um, e amostras com ou sem reposição são praticamente equivalentes.

Considere, agora, uma população $\mathcal{P} = \{1, 3, 5, 5, 7\}$, logo $N = 5$. Retire amostras de tamanho $n = 2$, *sem reposição*, e construa a distribuição amostral de $\bar{X} = (X_1 + X_2)/2$. Obtenha $E(\bar{X})$ e $\text{Var}(\bar{X})$ e verifique que esta é dada pela fórmula acima.

37. Planos probabilísticos. Existem vários planos probabilísticos que são utilizados em situações práticas. Vamos descrever brevemente alguns deles.

(a) *Amostragem Aleatória Simples (AAS).* Nesse plano as n unidades que compõem a amostra são selecionadas de tal forma que todas as possíveis amostras têm a mesma probabilidade de serem escolhidas. Podemos ter AAS com e sem reposição. No Exemplo 9.6 cada amostra com reposição tem probabilidade 1/25 de ser escolhida.

(b) *Amostragem Aleatória Estratificada.* Nesse procedimento, a população é dividida em subpopulações ou estratos, usualmente de acordo com os valores (ou categorias) de uma variável, e depois AAS é utilizada na seleção de uma amostra de cada estrato. Por exemplo, considere uma população de $N = 10$ estudantes, para os quais definimos as variáveis renda familiar (X_1) e classe social (X_2), categorizada como A, B ou C. Então, $\mathcal{P} = \{1, 2, ..., 10\}$ e suponha que a matriz de dados seja

$$\mathbf{D} = \begin{bmatrix} 10 & 8 & 15 & 6 & 22 & 12 & 7 & 16 & 13 & 11 \\ B & C & A & C & A & B & C & A & B & B \end{bmatrix}.$$

Podemos considerar três estratos, determinados pela variável X_2:

$$\mathcal{P}_A = \{3, 5, 8\}, \quad \mathcal{P}_B = \{1, 6, 9, 10\}, \quad \mathcal{P}_C = \{2, 4, 7\}.$$

Um dos objetivos da estratificação é homogeneizar a variância dentro de cada estrato, relativamente à principal variável de interesse.

(c) *Amostragem Aleatória por Conglomerados.* Como no item (b), a população é dividida em grupos (subpopulações) distintos, chamados conglomerados. Por exemplo, podemos dividir uma

cidade em bairros ou quadras. Usamos AAS para selecionar uma amostra de conglomerados e depois todos os indivíduos dos conglomerados selecionados são analisados.

(d) *Amostragem em Dois Estágios*. A população é dividida em grupos, como em (c). Num primeiro estágio, através de AAS, selecionamos algumas subpopulações. Num segundo estágio, usando novamente AAS, retiramos amostras das subpopulações selecionadas na primeiro estágio.

(e) *Amostragem Sistemática*. Nesse plano, supõe-se que temos uma listagem das unidades populacionais. Para k fixado, sorteamos um elemento entre os k primeiros da listagem. Depois observamos, sistematicamente, indivíduos separados por k unidades. Por exemplo, se $k = 10$ e sorteamos o oitavo elemento, observamos depois o décimo oitavo, vigésimo oitavo etc.

38. Distribuição do máximo de uma amostra. Considere M o máximo de uma AAS $X_1, \ldots, X_n$, escolhida de uma população com densidade $f(x)$ e f.d.a. $F(x)$. Seja $F_M(m)$ a f.d.a. de M. Então, $F_M(m) = P(M \leq m)$. Agora, o evento $\{M \leq m\}$ é equivalente ao evento $\{X_i \leq m$, para todo $1 \leq i \leq n\}$. Como as v.a. X_i são independentes, teremos

$$F_M(m) = P(M \leq m) = P(X_1 \leq m, \ldots, X_n \leq m) = P(X_1 \leq m) \ldots P(X_n \leq m) = [F(m)]^n.$$

Portanto, a densidade de M é dada por

$$f_M(m) = F'_M(m) = n[F(m)]^{n-1} f(m).$$

39. Obtenha a densidade de M para o caso de uma amostra de uma distribuição uniforme no intervalo $(0, \theta)$.

40. Suponha que temos a população $X \sim N(167; 25)$. Gere 100 amostras de tamanho 5 dessa população, usando algum programa de geração de valores de uma distribuição normal, como o Excel ou Minitab.

(a) Esboce a distribuição amostral de $\overline{X}$ (histograma) e calcule as principais medidas-resumo; faça *box plots* e ramos-e-folhas.

(b) Mesma questão para md = mediana da amostra.

(c) Compare as duas distribuições, ressaltando as principais diferenças.

(d) Estude a distribuição da estatística "variância da amostra".

41. Tamanho de uma amostra. Na prática, não conhecemos a distribuição de v.a. X e retiramos uma amostra a fim de estimar algum parâmetro dessa distribuição. Suponha, agora, que nosso interesse esteja na média $\mu = E(X)$. Para estimá-la, colhemos uma amostra $X_1, X_2, \ldots, X_n$ de X. Logo, as v.a. X_i são independentes, cada uma delas tem a mesma distribuição que X e $E(X_i) = \mu$, $\forall i = 1, \ldots, n$. Para estimar μ consideramos a média amostral $\overline{X}$.

Um problema que se apresenta é determinar o tamanho da amostra a colher. Isso pode ser feito usando a TLC, como vimos na seção 10.11.

Agora, vamos ver um procedimento diferente, também baseado no TLC, mas que envolve uma *regra de parada* para determinar o número de dados a colher. Esse procedimento foi sugerido por Ross (1997). Pelo TLC podemos escrever

$$P(|\overline{X} - \mu| > c\,\sigma/\sqrt{n}) \approx P(|Z| > c) = 2[1 - \Phi(c)], \tag{10.9}$$

para qualquer constante $c > 0$, onde $Z \sim N(0, 1)$ e $\Phi(\cdot)$ denota a f.d.a. de Z. Por exemplo, se $c = 1{,}96$, a probabilidade acima é $0{,}05$.

Suponha que, em vez de colher uma pequena amostra piloto para estimar σ, tenhamos informação suficiente para escolher um valor aceitável, digamos d, para o desvio padrão de $\bar{X}$, que é dado por $\sigma/\sqrt{n}$.

Por (10.9), podemos escrever, por exemplo,

$$P(|\bar{X}-\mu| \leq 1{,}96d) \approx 0{,}95.$$

Segue-se que podemos amostrar seqüencialmente de X até que $S/\sqrt{n} < d$, em que calculamos S com os valores até então escolhidos.

O seguinte algoritmo pode, então, ser adotado:

(1) Escolha um valor aceitável d para $\sigma/\sqrt{n}$.

(2) Gere pelo menos 30 dados (para obter uma estimativa razoável de σ).

(3) Continue a gerar dados, parando quando, com n dados, $S/\sqrt{n} < d$, com

$$S^2 = \sum (X_i - \bar{X})^2/(n-1).$$

(4) Estime μ por $\bar{X} = \sum X_i/n$.

Esse método implica podermos calcular $\bar{X}$ e S^2 *recursivamente*. Isso pode ser feito por meio das seguintes fórmulas, facilmente verificáveis:

$$\bar{X}_j = \frac{1}{j}\sum_{i=1}^{j} X_i, \quad S_j^2 = \frac{1}{j-1}\sum_{i=1}^{j}(X_i - \bar{X}_j)^2, \quad j \geq 2,$$

$$S_1^2 = 0,$$

$$\bar{X}_0 = 0,$$

$$\bar{X}_{j+1} = \bar{X}_j + \frac{X_{j+1} - \bar{X}_j}{j+1},$$

$$S_{j+1}^2 = \left(1 - \frac{1}{j}\right)S_j^2 + (j+1)(\bar{X}_{j+1} - \bar{X}_j)^2.$$

Suponha $x_1 = 3$, $x_2 = 5$, $x_3 = 2$, $x_4 = 6$, $x_5 = 4$. Então, usando as fórmulas acima, obtenha, recursivamente, $\bar{X}_i$, S_i^2, $i = 1, 2, 3, 4, 5$.

42. Suponha uma população $\mathcal{P} = \{1, 2, \ldots, N\}$ e a v.a. X definida sobre $\mathcal{P}$. Então, $T = \sum_{i=1}^{N} X_i$ é chamado *total populacional*. A média populacional é $\mu = T/N$ e a variância populacional é $\sigma^2 = \sum_{i=1}^{N}(X_i - \mu)^2/N$. Considere uma AAS de tamanho n extraída de $\mathcal{P}$ e $\bar{X}$ a média amostral. Considere o estimador $\hat{T} = N\bar{X}$. Mostre que $E(\hat{T}) = T$ e $\mathrm{Var}(\hat{T}) = N^2\sigma^2/n$.

43. Suponha que queiramos retirar uma amostra de uma distribuição de Bernoulli com parâmetro p. Escolhidos k dados $x_1, x_2, \ldots, x_k$, temos que $\bar{x}_k = \sum_i x_i/k$ é um estimador de p. Então um estimador natural da variância $\sigma^2 = p(1-p)$ da população é $\bar{x}_k(1-\bar{x}_k)$. Como ficaria o algoritmo descrito no Problema 41 para essa situação?