

Vamos caminhar neste semestre na estrada paramétrica do mundo estatístico. Como sempre o professor Maurício apresentará exemplos do livro dos professores Bussab& Morettin.

Este foi o livro inicial de vocês na primeira disciplina da nossa graduação e depois utilizado em Probabilidade I e Probabilidade II.

Com o professor Bussab fiz a disciplina de Amostragem no meu mestrado no IME-USP e com o professor Morettin fiz as disciplinas de Inferência Paramétrica e Métodos não Paramétricos.

Esta disciplina ofereço a estes grandes mestres.

Vamos iniciar nossa disciplina pelo capítulo 10 da nona edição. Precisamos ler com cuidado todo o texto. Explorar os exemplos e refazê-los no *R*. Fazer todos os exercícios.

Vamos mostrar a capa do livro da minha edição:

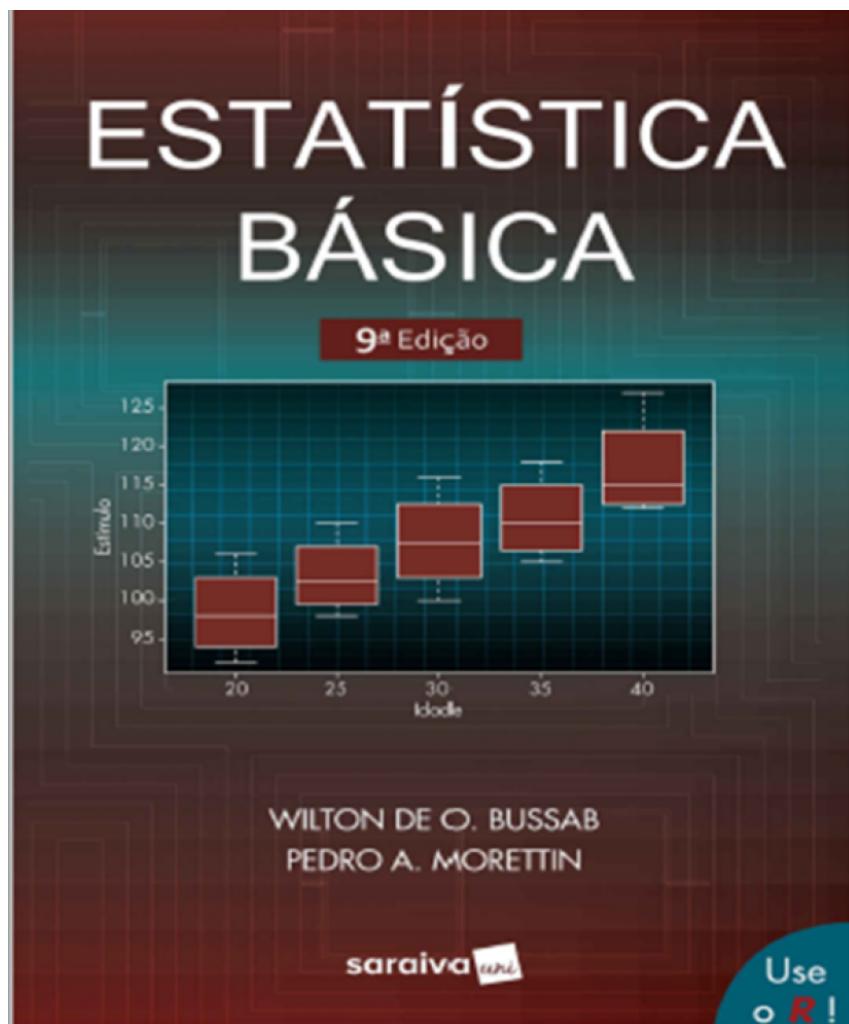


Figura 1:

Note que na capa aparece um diagrama box-plot que é um exemplo que será explorado

em nossa disciplina.

Serão estudados os capítulos: 10, 11, 12, 13, e 14.

Vamos comentar de leve os capítulos 15 e 16.

10.1: Introdução.

Vimos, na Parte I, como resumir descritivamente variáveis associadas a um ou mais conjuntos de dados.

Na Parte II, 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 está sendo 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, frequentemente o pesquisador tem alguma ideia 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 corres-

pondente; 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.

Se pudéssemos medir as alturas de todos os brasileiros adultos, teríamos meios de obter sua distribuição exata e, daí, produzir os correspondentes parâmetros. Mas, nessa situação, não teríamos necessidade de usar a inferência estatística!

Raramente se consegue obter a distribuição exata de alguma variável, ou porque isso é muito dispendioso, ou muito demorado ou, às vezes, porque consiste num processo destrutivo.

Por exemplo, se estivéssemos observando a durabilidade de lâmpadas e testássemos todas até queimarem, não restaria nenhuma para ser vendida. Assim, a solução é selecionar parte dos elementos (amostra), analisá-la e inferir propriedades para o todo (população).

Outras vezes, estamos interessados em explorar relações entre variáveis envolvendo experimentos mais complexos para a obtenção dos dados.

Por exemplo, gostaríamos de obter resposta para a seguinte indagação: a altura que um produto é colocado na gôndola de um supermercado afeta a sua venda?

Observe que para responder a questão precisamos obter dados de vendas com o produto oferecido em diferentes alturas, e que essas vendas sejam controladas para evitar interferências de outros fatores que não a altura.

Nesse caso, não existe claramente um conjunto de todos os elementos para os quais pudéssemos encontrar os parâmetros populacionais.

Recorrer a modelos para descrever o todo (população) facilita a identificação e solução do problema.

Nesse exemplo, supondo que as vendas V_h do produto oferecido na altura h ($h = 1$ representando baixo, $h = 2$ representando meio e $h = 3$ representando alto) segue uma distribuição próxima a normal, ou seja,

$$V_h \sim N(\mu_h, \sigma^2)$$

, o nosso problema passa a ser o de verificar, por meio de dados coletados do experimento (amostra), se existe evidência de igualdade das médias μ_1 , μ_2 e μ_3 .

$$H_0 : \mu_1 = \mu_2 = \mu_3 = \mu.$$

Note que, em nossa formulação do problema, consideraremos que as três situações de alturas resultam observações com a mesma variância σ^2 . Essa suposição poderia ser modificada.

Soluções de questões como as apresentadas acima são o objeto da inferência estatística.

Dois conceitos básicos são, portanto, necessários para o desenvolvimento da Inferência Estatística: população e amostra.

Definição. População é o conjunto de todos os elementos ou resultados sob investigação. Amostra é qualquer subconjunto da população.

Vejamos outros exemplos para melhor entender essas definições.

Vamos reservar a notação N para o tamanho populacional e n para o tamanho amostral.

Exemplo 10.1 Consideremos uma pesquisa para estudar os salários dos $N = 500$ funcionários da Companhia MB. Seleciona-se uma amostra de $n = 36$ indivíduos, e anotam-se os seus salários. A variável aleatória a ser observada é “salário”. A população é formada pelos 500 funcionários da companhia.

A amostra é constituída pelos 36 indivíduos selecionados. Na realidade, estamos interessados nos salários, portanto, para sermos mais precisos, devemos considerar como a população os 500 salários correspondentes aos 500 funcionários.

Consequentemente, a amostra será formada pelos 36 salários dos indivíduos selecionados. Podemos estudar a distribuição dos salários na amostra, e esperamos que esta reflita a distribuição de todos os salários, desde que a amostra tenha sido escolhida com cuidado.

Tabela 2.1 Informações sobre estado civil, grau de instrução, número de filhos, salário (expresso como fração do salário mínimo), idade (medida em anos e meses) e procedência de 36 empregados da seção de orçamentos da Companhia MB.

Nº	Estado civil	Grau de instrução	Nº de filhos	Salário (x sal. mín.)	Idade		Região de procedência
					anos	meses	
1	solteiro	ensino fundamental	—	4,00	26	03	interior
2	casado	ensino fundamental	1	4,56	32	10	capital
3	casado	ensino fundamental	2	5,25	36	05	capital
4	solteiro	ensino médio	—	5,73	20	10	outra
5	solteiro	ensino fundamental	—	6,26	40	07	outra
6	casado	ensino fundamental	0	6,66	28	00	interior
7	solteiro	ensino fundamental	—	6,86	41	00	interior
8	solteiro	ensino fundamental	—	7,39	43	04	capital
9	casado	ensino médio	1	7,59	34	10	capital
10	solteiro	ensino médio	—	7,44	23	06	outra
11	casado	ensino médio	2	8,12	33	06	interior
12	solteiro	ensino fundamental	—	8,46	27	11	capital
13	solteiro	ensino médio	—	8,74	37	05	outra
14	casado	ensino fundamental	3	8,95	44	02	outra
15	casado	ensino médio	0	9,13	30	05	interior
16	solteiro	ensino médio	—	9,35	38	08	outra
17	casado	ensino médio	1	9,77	31	07	capital
18	casado	ensino fundamental	2	9,80	39	07	outra
19	solteiro	superior	—	10,53	25	08	interior
20	solteiro	ensino médio	—	10,76	37	04	interior
21	casado	ensino médio	1	11,06	30	09	outra
22	solteiro	ensino médio	—	11,59	34	02	capital
23	solteiro	ensino fundamental	—	12,00	41	00	outra
24	casado	superior	0	12,79	26	01	outra
25	casado	ensino médio	2	13,23	32	05	interior
26	casado	ensino médio	2	13,60	35	00	outra
27	solteiro	ensino fundamental	—	13,85	46	07	outra
28	casado	ensino médio	0	14,69	29	08	interior
29	casado	ensino médio	5	14,71	40	06	interior
30	casado	ensino médio	2	15,99	35	10	capital
31	solteiro	superior	—	16,22	31	05	outra
32	casado	ensino médio	1	16,61	36	04	interior
33	casado	superior	3	17,26	43	07	capital
34	solteiro	superior	—	18,75	33	07	capital
35	casado	ensino médio	2	19,40	48	11	capital
36	casado	superior	3	23,30	42	02	interior

Fonte: Dados hipotéticos.

Figura 2:

Exemplo 10.2 Queremos estudar p , a proporção de indivíduos na cidade A que são favoráveis a certo projeto governamental. Uma amostra de 200 pessoas é sorteada, e a opinião de cada uma é registrada a favor ou contra o projeto.

A população consiste de todos os moradores da cidade, e a amostra é formada pelas 200 pessoas selecionadas. Podemos, como foi visto no Capítulo 5, definir a variável X , que toma o valor 1, se a resposta de um morador for favorável, e o valor 0, se a resposta for contrária ao projeto. Assim, nossa população pode ser reduzida à distribuição de

$$X \sim Ber(p),$$

com

$$P(X = x) = p^x(1 - p)^{1-x} I_A(x), \quad A = \{0, 1\}$$

com A o suporte da nossa variável e

$$\mu = p \quad e \quad \sigma^2 = p(1-p).$$

e a amostra X_1, X_2, \dots, X_{200} será constituída de uma sequência de 200 zeros e uns.

Seja S_n , o número de indivíduos na cidade A que são favoráveis ao projeto governamental na amostra. Assim se a amostragem é feita com reposição temos que:

$$S_n = \sum_{i=1}^n X_i \sim Bin(n, p).$$

Note que

$$P(S_n = s) = \binom{n}{s} p^s (1-p)^{n-s} I_{\{0,1,\dots,n\}}(s)$$

com

$$E(S_n) = np \quad e \quad V(S_n) = np(1-p).$$

Vamos estimar p por

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

No caso da amostragem sem reposição considere B , o número de indivíduos na cidade A que são favoráveis ao projeto governamental na população.

Assim

$$S_n \sim HG(N, B, n),$$

com

$$P(S_n = s) = \frac{\binom{B}{s} \binom{N-B}{n-s}}{\binom{N}{n}} I_A(s),$$

com

$$A = \{max(0, B + n - N), \dots, min(n, B)\}.$$

Por exemplo se $N = 13$, $B = 5$ e $n = 4$ temos:

$$L_i = max(0, B + n - N) = max(0, 5 + 4 - 13) = max(0, -4) = 0 \quad e \quad L_s = min(n, B) = min(4, 5) = 4$$

Assim o suporte será

$$A = \{0, 1, 2, 3, 4\}.$$

Por exemplo se $N = 13$, $B = 5$ e $n = 9$ temos:

$$L_i = max(0, B + n - N) = max(0, 5 + 9 - 13) = max(0, 1) = 1 \quad e \quad L_s = min(n, B) = min(9, 5) = 5$$

Assim o suporte será

$$A = \{1, 2, 3, 4, 5\}.$$

Além disso,

$$E(S_n) = np \quad e \quad V(S_n) = np(1-p) \frac{N-n}{N-1},$$

com

$$p = \frac{B}{N} \quad e \quad q = 1 - p = \frac{N-B}{N}.$$

A nossa população é descrita por apenas um parâmetro p e é importante saber como ele varia. Daí surge a noção de espaço paramétrico Θ . Neste caso

$$\Theta = [0, 1].$$

Exemplo 10.3 O interesse é investigar a duração de vida de um novo tipo de lâmpada, pois acreditamos que ela tenha uma duração maior do que as fabricadas atualmente. Então, 100 lâmpadas do novo tipo são deixadas acesas até queimarem.

A duração em horas de cada lâmpada é registrada. Aqui, a variável é a duração em horas de cada lâmpada. A população é formada por todas as lâmpadas fabricadas ou que venham a ser fabricadas por essa empresa, com o mesmo processo.

A amostra é formada pelas 100 lâmpadas selecionadas. Note-se que nesse caso não podemos observar a população, ou seja, a distribuição da duração de vida das lâmpadas na população, pois isso corresponderia a queimar todas as lâmpadas.

Assim, em alguns casos, não podemos observar a população toda, pois isso significaria danificar (ou destruir) todos os elementos da população.

Esse problema geralmente é contornado atribuindo-se um modelo teórico para a distribuição da variável populacional.

Por exemplo suponha que

$$X \sim Exp(\theta), \quad \theta > 0,$$

Aqui nosso espaço paramétrico é

$$\Theta = (0, \infty)$$

e o nosso suporte

$$A = (0, \infty).$$

A f.d.p. de X é dada por:

$$f(x; \theta) = \theta e^{-\theta x} I_A(x),$$

com

$$\mu = E(X) = \frac{1}{\theta} \quad e \quad \sigma^2 = \frac{1}{\theta^2}.$$

Como estimar θ usando a nossa amostra aleatória X_1, X_2, \dots, X_n ?

A ideia conhecida por Método dos Momentos é igualar o primeiro momento populacional $E(X)$ ao primeiro momento amostral \bar{X} .

$$E(X) = \bar{X} = \frac{S_n}{n}.$$

Logo

$$\frac{1}{\theta} = \frac{S_n}{n}.$$

Assim

$$\theta = \frac{n}{S_n} = \frac{1}{\bar{X}}.$$

Nosso estimador proposto será:

$$\hat{\theta} = T = \frac{1}{\bar{X}}.$$

Nosso estimador depende de $S_n = \sum_{i=1}^n X_i$. Precisamos achar a distribuição de S_n , a soma de n variáveis aleatórias independentes e identicamente distribuídas com a lei exponencial de parâmetro θ .

A função geradora de momentos de $X \sim Exp(\theta)$ é dada por:

$$M_X(t) = \frac{\theta}{\theta - t}, \quad t < \theta.$$

A função geradora de momentos de S_n é dada por:

$$M_{S_n}(t) = [M_X(t)]^n = \left[\frac{\theta}{\theta - t} \right]^n, \quad t < \theta$$

que é a fgm de uma gama de parâmetros $r = n$ e $\lambda = \theta$ e com f.d.p. dada por:

$$f_{S_n}(s) = \frac{\theta^n}{\Gamma(n)} s^{n-1} e^{-\theta s} I_A(s),$$

com

$$E(S_n) = \frac{n}{\theta} \quad e \quad V(S_n) = \frac{n}{\theta^2}.$$

Note que

$$\frac{1}{n} E(S_n) = E(\bar{X}) = \frac{1}{\theta},$$

assim \bar{X} é um estimador não viciado de $g(\theta) = \frac{1}{\theta}$.

Quem seria o estimador não viciado de θ ?

Nossa busca começa por achar a lei de

$$V = \bar{X}.$$

A f.g.m. de V é dada por:

$$M_V(t) = M_{\bar{X}}(t) = E(e^{tS_n/n}) = M_{S_n}(t/n)$$

$$M_V(t) = \left[\frac{\theta}{\theta - t/n} \right]^n, \quad t/n < \theta$$

$$M_V(t) = \left[\frac{n\theta}{n\theta - t} \right]^n, \quad t < n\theta$$

Assim

$$V = \bar{X} \sim Gama(r = n, \lambda = n\theta),$$

com

$$f_V(v) = \frac{n^n \theta^n}{\Gamma(n)} v^{n-1} e^{-n\theta v} I_A(v).$$

Vamos achar a esperança de $W = \frac{1}{V}$:

$$E\left(\frac{1}{V}\right) = \int_0^\infty \frac{1}{v} \frac{n^n \theta^n}{\Gamma(n)} v^{n-1} e^{-n\theta v} dv$$

$$= \frac{n^n \theta^n}{\Gamma(n)} \int_0^\infty v^{(n-1)-1} e^{-n\theta v} dv = \frac{n^n \theta^n}{\Gamma(n)} IGG(a = n-1 > 0, b = n\theta, c = 1).$$

$$= \frac{n^n \theta^n}{\Gamma(n)} \frac{\Gamma(n-1)}{n^{n-1} \theta^{n-1}}, \quad n > 1$$

$$E\left(\frac{1}{V}\right) = \frac{n\theta \Gamma(n-1)}{\Gamma(n)} = \frac{n\theta \Gamma(n-1)}{(n-1)\Gamma(n-1)}$$

$$E\left(\frac{1}{V}\right) = \frac{n\theta}{n-1}, n > 1.$$

Logo:

$$E\left(\frac{n-1}{nV}\right) = E\left(\frac{n-1}{n\bar{X}}\right) = E\left(\frac{n-1}{S_n}\right) = \theta.$$

Nosso estimador não viciado para θ seria

$$T = \frac{n-1}{S_n}, n > 1.$$

Usamos os seguintes resultados que não podem ser esquecidos pelos alunos:

Fato1: Se $a > 0$

$$\Gamma(a+1) = a \Gamma(a).$$

Fato2: Se $a > 0, b > 0, c > 0$

$$IGG(a, b, c) = \int_0^\infty v^{a-1} e^{-b v^c} dv = \frac{\Gamma(a/c)}{c b^{a/c}}.$$

Integral da função gamma generalizada

Exemplo 10.4 Em alguns casos, fazemos suposições mais precisas sobre a população (ou sobre a variável definida para os elementos da população).

Digamos que X represente o peso real de pacotes de café, enchidos automaticamente por uma máquina. Sabe-se que a distribuição de X pode ser representada por uma normal, com parâmetros μ e σ^2 desconhecidos.

Sorteamos 100 pacotes e medimos seus pesos.

A população será o conjunto de todos os pacotes enchidos ou que virão a ser enchidos pela máquina, e que pode ser suposta como normal.

A amostra será formada pelas 100 medidas obtidas dos pacotes selecionados, que pode ser pensada como constituída de 100 observações feitas de uma distribuição normal.

Veremos mais adiante como tal amostra pode ser obtida.

Neste caso temos um vetor de parâmetros

$$\theta = (\mu, \sigma^2)'.$$

que serão estimados a partir da amostra aleatória X_1, X_2, \dots, X_n .

Nosso espaço paramétrico é dado por:

$$\Theta = (-\infty, \infty) \times (0, \infty),$$

representando o fato que a média da normal é real e a variância positiva.

Nosso suporte é o conjunto dos números reais e a f.d.p. de X é dada por:

$$f(x, \mu, \sigma^2) = \frac{1}{\sqrt{2 \pi \sigma^2}} \exp\left(-\frac{(x - \mu)^2}{2\sigma^2}\right) I_A(x),$$

com

$$E(X) = \mu \quad e \quad V(X) = \sigma^2$$

A função geradora de momentos de X é dada por:

$$M_X(t) = e^{\mu t + \frac{1}{2} \sigma^2 t^2}.$$

Para estimar μ usamos

$$\bar{X} = \frac{\sum_{i=1}^n X_i}{n} = \frac{S_n}{n} \sim N\left(\mu, \frac{\sigma^2}{n}\right)$$

A função geradora de momentos de S_n é dada por:

$$M_{S_n}(t) = [M_X(t)]^n = \left[e^{\mu t + \frac{1}{2} \sigma^2 t^2} \right]^n = e^{n \mu t + \frac{1}{2} n \sigma^2 t^2},$$

que é a fgm de uma Normal de parâmetros $n \mu$ e $n \sigma^2$ e com f.d.p. dada por:

$$f_{S_n}(s, \mu, \sigma^2) = \frac{1}{\sqrt{2 \pi n} \sigma} \exp\left(-\frac{(s - n \mu)^2}{2n\sigma^2}\right) I_A(s),$$

com

$$E(S_n) = \mu \quad e \quad V(S_n) = n \sigma^2.$$

Seja $V = \bar{X}$. A fgm de V é dada por:

$$M_V(t) = M_{\bar{X}}(t) = E\left(e^{t S_n/n}\right) = M_{S_n}(t/n)$$

que é a fgm de uma Normal de parâmetros $n\mu$ e $n \sigma^2$

$$M_V(t) = \left[e^{n \mu \frac{t}{n} + \frac{1}{2} n \sigma^2 \frac{t^2}{n^2}} \right]^n$$

$$M_V(t) = \exp \left(\mu + \frac{1}{2} \frac{\sigma^2}{n} t^2 \right),$$

que é a fgm de uma Normal de parâmetros μ e $\frac{\sigma^2}{n}$.

Seja S^2 , a variância amostral, definida por:

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

Então

$$U = \frac{(n-1)S^2}{\sigma^2} = \frac{\sum_{i=1}^n (X_i - \bar{X})^2}{\sigma^2} \sim \chi^2(n-1).$$

Vamos provar mais adiante que:

$$E(S^2) = \sigma^2 \quad e \quad V(S^2) = \frac{2\sigma^4}{n-1}.$$

Além disso, em populações normais \bar{X} e S^2 são independentes.

Exemplo 10.5 Para investigar a “honestidade” de uma moeda, nós a lançamos 50 vezes e contamos o número de caras observadas.

A população, como no caso do Exemplo 10.2, pode ser considerada como tendo a distribuição da variável X , assumindo o valor 1, com probabilidade p , se ocorrer cara, e assumindo o valor 0, com probabilidade $q = 1 - p$, se ocorrer coroa. Ou seja, a população pode ser considerada como tendo distribuição de Bernoulli com parâmetro p .

A variável ficará completamente especificada quando conhecermos p . A amostra será uma sequência de 50 números zeros ou uns.

Na realidade testar a hipótese que a moeda é não viciada é testar:

$$H_0 : p = \frac{1}{2},$$

versus a hipótese

$$H_1 : p \neq \frac{1}{2}.$$

Exemplo 10.6 Há razões para supor que o tempo Y de reação a certo estímulo visual dependa da idade do indivíduo (esse exemplo será usado nos Capítulos 15 e 16). Suponha, ainda, que essa dependência seja linear.

Para verificarmos se essa suposição é verdadeira, obtiveram-se 20 dados da seguinte maneira: 20 pessoas foram selecionadas, sendo 10 homens e 10 mulheres. Dentro de cada grupo de homens e mulheres foram selecionadas duas pessoas das seguintes faixas de idade: 20, 25, 30, 35 e 40 anos. Cada pessoa foi submetida ao teste e seu tempo de reação y foi medido.

Veja a tabela 15.1:

Indivíduo	Y	W	X	Z
1	96	H	20	90
2	92	M	20	100
3	106	H	20	80
4	100	M	20	90
5	98	M	25	100
6	104	H	25	90
7	110	H	25	80
8	101	M	25	90
9	116	M	30	70
10	106	H	30	90
11	109	H	30	90
12	100	M	30	80
13	112	M	35	90
14	105	M	35	80
15	118	H	35	70
16	108	H	35	90
17	113	M	40	90
18	112	M	40	90
19	127	H	40	60
20	117	H	40	80

A população poderia ser considerada como formada por todas aquelas pessoas que viessem a ser submetidas ao teste, segundo o sexo e a idade. A amostra é formada pelas 20 medidas, que estão apresentadas na Tabela 15.1.

Observações.

(i) Os três últimos exemplos mostram uma ampliação do conceito definido de população, ou seja, designamos agora a população como a função probabilidade ou função densidade de probabilidade de uma v.a. X , modelando a característica de interesse. Esse artifício simplifica substancialmente o problema estatístico, exigindo no entanto uma proposta de modelo para a variável X . Nesses casos simplificaremos a linguagem, dizendo: “seja a população $f(x)$ ”. Por exemplo, “considere a população das alturas $X \sim (\mu, \sigma^2)$ ”.

(ii) Essa abordagem, por meio da distribuição de probabilidades, utiliza muitas vezes o conceito de população infinita contínua, exigindo um tratamento matemático mais cuidadoso. É mais fácil apresentar os problemas e soluções por meio de populações finitas. É o que faremos muitas vezes. Entretanto, é importante que o estudante aprenda a trabalhar com o conceito de modelo, explorando o caso de população $f(x)$.

10.3:Problemas de Inferência.

Como já dissemos anteriormente, o objetivo da Inferência Estatística é produzir afirmações sobre dada característica da população, na qual estamos interessados, a partir de informações colhidas de uma parte dessa população.

Essa característica população pode ser representada por uma variável aleatória.

Se tivéssemos informação completa sobre a função de probabilidade, no caso discreto, ou sobre a função densidade de probabilidade, no caso contínuo, da variável em questão, não teríamos necessidade de escolher uma amostra. Toda a informação desejada seria obtida por meio da distribuição da variável, usando-se a teoria estudada anteriormente.

Mas isso raramente acontece. Ou não temos qualquer informação a respeito da variável, ou ela é apenas parcial. Podemos admitir, como no exemplo das alturas de brasileiros adultos, que ela siga uma distribuição normal, mas desconhecemos os parâmetros que a caracterizam (média, variância).

Em outros casos, podemos ter uma ideia desses parâmetros, mas desconhecemos a forma da curva. Ou ainda, o que é muito frequente, não possuímos informações nem sobre os parâmetros, nem sobre a forma da curva. Em todos os casos, o uso de uma amostra nos ajudaria a formar uma opinião sobre o comportamento da variável (população).

Embora a identificação e a descrição da população sejam fundamentais no processo inferencial, é comum os pesquisadores dedicarem mais atenção em descrever a amostra do que a população para a qual serão feitas as afirmações. É imprescindível que se explique claramente a população investigada.

Neste livro, estaremos mais preocupados em trabalhar com populações descritas por modelos do que com populações finitas identificadas por elementos portadores de uma característica de interesse. Portanto, na maioria das vezes, iremos nos referir à “população X ”, significando que a variável de interesse X , definida sobre a população-alvo, segue uma distribuição $f(x)$. Nosso problema de interesse passaria a ser o de fazer afirmações sobre a forma da curva e seus parâmetros.

Alguns exemplos simples nos darão uma noção dos tipos de formulações e problemas que a inferência estatística pode nos ajudar a resolver.

Exemplo 10.5 (continuação) Voltemos ao exemplo da moeda. Indicando por X o número de caras obtidas depois de lançar a moeda 50 vezes, sabemos que, se tomados alguns cuidados quanto ao lançamento, X segue uma distribuição binomial, ou seja, $X \sim B(50, p)$.

Esse modelo é válido, admitindo-se ou não a “honestidade” da moeda, isto é, sendo ou não $p = 1/2$. Lançada a moeda, vamos supor que tenham ocorrido 36 caras. Esse resultado traz evidência de que a moeda seja “honesto”?

Para tomarmos uma decisão, podemos partir do princípio de que a moeda não favorece nem cara nem coroa, isto é, $p = 1/2$. Com essa informação e com o modelo binomial, podemos encontrar qual a probabilidade de se obterem 36 caras ou mais, e esse resultado nos ajudaria a tomar uma decisão. Suponha que a decisão foi rejeitar a “honestidade” da moeda: qual é a melhor estimativa para p , baseando-se no resultado observado?

Descrevemos aí os dois problemas básicos da Inferência Estatística: o primeiro é chamado teste de hipóteses, e o segundo, estimativa. Nos capítulos seguintes, esses problemas serão abordados com mais detalhes.

Exemplo 10.4 (continuação) Às vezes, o modelo teórico associado ao problema não é tão evidente. No caso da máquina de encher pacotes de café automaticamente, digamos que ela esteja regulada para encher os pacotes segundo uma distribuição normal com média 500 gramas e desvio padrão de 10 gramas, isto é, $X \sim N(500, 10^2)$.

Sabemos também que, às vezes, a máquina desregula-se e, quando isso acontece, o único parâmetro que se altera é a média, permanecendo a mesma variância. Para manter a produção sob controle, iremos colher uma amostra de 100 pacotes e pesá-los. Como essa amostra nos ajudará a tomar uma decisão?

Parece razoável, nesse caso, usarmos a média \bar{x} da amostra como informação pertinente para uma decisão. Mesmo que a máquina esteja regulada, dificilmente \bar{x} será igual a 500 gramas, dado que os pacotes apresentam certa variabilidade no peso. Mas se \bar{x} não se afastar muito de 500 gramas, não existirão razões para suspeitarmos da qualidade do procedimento de produção. Só iremos pedir uma revisão se $\bar{x} = 500$, em valor absoluto, for “muito grande”.

O problema que se apresenta agora é o de decidir o que é próximo ou distante de 500 gramas. Se o mesmo procedimento de colher a amostra de 100 pacotes fosse repetido um número muito grande de vezes, sob a condição de a máquina estar regulada, teríamos ideia do comportamento da v.a. \bar{X} , e saberíamos dizer se aquele valor observado é ou não um evento raro de ocorrer. Caso o seja, é mais fácil suspeitar da regulagem da máquina do que do acaso. Vemos, então, a importância nesse caso de se conhecer as propriedades da distribuição da variável \bar{X} .

Exemplo 10.6 (continuação) A descrição matemática da v.a. Y : tempo de reação ao estímulo é um pouco mais complexa. Podemos supor que esse tempo, para uma dada idade x , seja uma v.a. com distribuição normal, com média dependendo da idade x , ou seja, podemos escrever

$$Y \sim N(\mu(x), \sigma^2).$$

A linearidade expressa no problema pode ser incluída da seguinte maneira:

$$\mu(x) = \alpha + \beta x.$$

Note que agora nossa amostra seriam os pares ordenados:

$$(X_1, Y_1), (X_2, Y_2), \dots, (X_n, Y_n).$$

O nosso vetor de parâmetros é dado por:

$$\theta = (\alpha, \beta, \sigma^2),$$

com espaço paramétrico

$$\Theta = (-\infty, \infty) \times (-\infty, \infty) \times (0, \infty).$$

Seja $\hat{\beta}$ o estimador de mínimos quadrados de β dado por:

$$\hat{\beta} = \frac{\sum_{i=1}^n (X_i - \bar{X})(Y_i - \bar{Y})}{\sum_{i=1}^n (X_i - \bar{X})^2}.$$

Pode-se mostrar que

$$E(\hat{\beta}) = \beta \quad e \quad V(\hat{\beta}) = \frac{\sigma^2}{\sum_{i=1}^n (X_i - \bar{X})^2}.$$

O estimador de mínimos quadrados de α é dado por:

$$\hat{\alpha} = \bar{Y} - \hat{\beta} \bar{X}.$$

Pode-se mostrar que

$$E(\hat{\alpha}) = \alpha \quad e \quad V(\hat{\alpha}) = \left[\frac{1}{n} + \frac{\bar{X}^2}{\sum_{i=1}^n (X_i - \bar{X})^2} \right] \sigma^2.$$

Voltaremos a esse modelo no capítulo 16. Outra maneira de escrever as duas relações anteriores é

$$Y|x \sim N(\alpha + \beta x, \sigma^2).$$

Leia-se "Y dado x".

Exemplo 15.1, página 498 Um psicólogo está investigando a relação entre o tempo que indivíduo leva para reagir a um estímulo visual (Y) e alguns fatores como sexo (W), idade(X) e acuidade visual(Z , medida em percentagem).

Na tabela 15.1, página 438 temos os tempos para $n = 20$ indivíduos (valores da v.a. Y). o fator sexo tem dois níveis:

$i = 1$: sexo masculino (H) e $i = 2$ sexo feminino (M), com $n_1 = 10, n_2 = 10$.

O fator idade tem cinco níveis $i = 1$:indivíduos com 20 anos de idade, $i = 2$:indivíduos com 25 anos de idade, $i = 3$: indivíduos com 30 anos de idade, $i = 4$:indivíduos com 35 anos de idade e $i = 5$:indivíduos com 40 anos de idade. Aqui $n_1 = n_2 = n_3 = n_4 = n_5 = 4$.

A acuidade visual , como porcentagem da visão completa, também gera 5 níveis: $i = 1$:indivíduos com 100 % da visão, $i = 2$:indivíduos com 90 % da visão e assim por diante. Não foi possível controlar essa variável a priori como as outras duas , já que ela exige exames oftalmológicos para sua mensuração.. Daí o desbalanceamento dos tamanhos observados: $n_1 = 2, n_2 = 10, n_3 = 5, n_4 = 2, n_5 = 1$.

Indivíduo	Y	W	X	Z
1	96	H	20	90
2	92	M	20	100
3	106	H	20	80
4	100	M	20	90
5	98	M	25	100
6	104	H	25	90
7	110	H	25	80
8	101	M	25	90
9	116	M	30	70
10	106	H	30	90
11	109	H	30	90
12	100	M	30	80
13	112	M	35	90
14	105	M	35	80
15	118	H	35	70
16	108	H	35	90
17	113	M	40	90
18	112	M	40	90
19	127	H	40	60
20	117	H	40	80

Podemos, por exemplo, estimar os parâmetros α e β , baseados na amostra de 20 dados. Ou podemos querer investigar a possibilidade de β ser igual a zero, significando

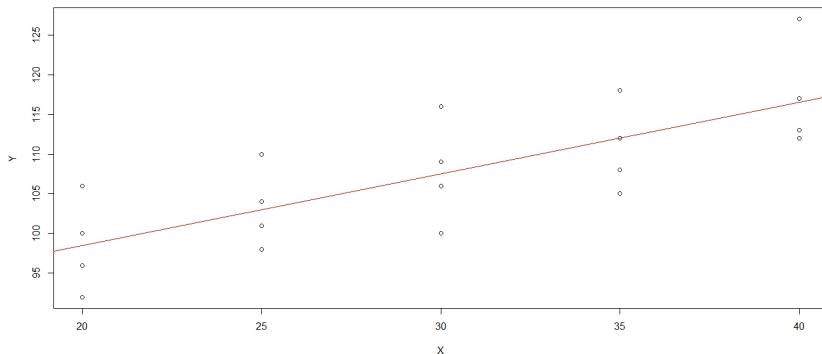


Figura 3:

que a idade não afeta o tempo de reação.

Novamente, os dois principais problemas de inferência aparecem aqui: estimativa e teste de uma hipótese. Um outro problema importante em inferência é o de previsão. Por exemplo, considerando um grupo de pessoas de 40 anos, poderemos prever com o modelo acima qual será o respectivo tempo de reação.

Repetir um mesmo experimento muitas vezes, sob as mesmas condições, nem sempre é possível, mas em determinadas condições é possível determinar teoricamente o comportamento de algumas medidas feitas na amostra, como por exemplo a média. Mas isso depende, em grande parte, do procedimento (plano) adotado para selecionar a amostra. Assim, em problemas envolvendo amostras, antes de tomarmos uma decisão, teríamos de responder a quatro perguntas:

- (a) Qual a população a ser amostrada?
- (b) Como obter os dados (a amostra)?
- (c) Que informações pertinentes (estatísticas) serão retiradas da amostra?
- (d) Como se comporta(m) a(s) estatística(s) quando o mesmo procedimento de escolher a amostra é usado numa população conhecida?

Nas seções e capítulos subsequentes, tentaremos responder a essas perguntas.

10.4 :Como Selecionar uma Amostra

As observações contidas em uma amostra são tanto mais informativas sobre a população quanto mais conhecimento explícito ou implícito tivermos dessa mesma população. Por

exemplo, a análise da quantidade de glóbulos brancos obtida de algumas gotas de sangue da ponta do dedo de um paciente dará uma ideia geral da quantidade dos glóbulos brancos no corpo todo, pois sabe-se que a distribuição dos glóbulos brancos é homogênea, e de qualquer lugar que se tivesse retirado a amostra ela seria “representativa”. Mas nem sempre a escolha de uma amostra adequada é imediata. Voltando ao Exemplo 10.2, para o qual queríamos obter uma amostra de habitantes para saber a opinião sobre um projeto governamental, escolhendo intencionalmente uma amostra de 200 indivíduos moradores de certa região beneficiada pelo projeto, saberemos de antemão que o resultado conterá um viés de seleção. Isto é, na amostra, a proporção de pessoas favoráveis ao projeto deverá ser maior do que no todo, donde a importância da adoção de procedimentos científicos que permitam fazer inferências adequadas sobre a população.

A maneira de se obter a amostra é tão importante, e existem tantos modos de fazê-lo, que esses procedimentos constituem especialidades dentro da Estatística, sendo Amostragem e Planejamento de Experimentos as duas mais conhecidas. Poderíamos dividir os procedimentos científicos de obtenção de dados amostrais em três grandes grupos:

(a) Levantamentos Amostrais, nos quais a amostra é obtida de uma população bem definida, por meio de processos bem protocolados e controlados pelo pesquisador. Podemos, ainda, subdividi-los em dois subgrupos: levantamentos probabilísticos e não probabilísticos. O primeiro reúne todas aquelas técnicas que usam mecanismos aleatórios de seleção dos elementos de uma amostra, atribuindo a cada um deles uma probabilidade, conhecida a priori, de pertencer à amostra. No segundo grupo estão os demais procedimentos, tais como: amostras intencionais, nas quais os elementos são selecionados com o auxílio de especialistas, e amostras de voluntários, como ocorre em alguns testes sobre novos medicamentos e vacinas. Ambos os procedimentos têm suas vantagens e desvantagens. A grande vantagem das amostras probabilísticas é medir a precisão da amostra obtida, baseando-se no resultado contido na própria amostra. Tais medidas já são bem mais difíceis para os procedimentos do segundo grupo.

Estão nessa situação os Exemplos 10.1 (conhecer os salários da Cia. MB), 10.2 (identificar a proporção de indivíduos favoráveis ao projeto), 10.4 (pesos dos pacotes de café) etc.

(b) Planejamento de Experimentos, cujo principal objetivo é o de analisar o efeito de uma variável sobre outra. Requer, portanto, interferências do pesquisador sobre o ambiente em estudo (população), bem como o controle de fatores externos, com o intuito de medir o efeito desejado. Podemos citar como exemplos aquele já citado sobre a altura de um produto na gôndola de um supermercado afetar as vendas e o Exemplo 10.6. Em ensaios clínicos em medicina, esse tipo de estudo é bastante usado, como por exemplo para testar se um novo medicamento é eficaz ou não para curar certa doença.

(c) Levantamentos Observacionais, nos quais os dados são coletados sem que o pesquisador tenha controle sobre as informações obtidas, exceto eventualmente sobre possíveis erros grosseiros. As séries de dados temporais são exemplos típicos desses levantamentos.

Por exemplo, queremos prever as vendas de uma empresa em função de vendas passadas. O pesquisador não pode selecionar dados, esses são as vendas efetivamente ocorridas. Nesses casos, a especificação de um modelo desempenha um papel crucial na ligação entre dados e população.

No caso de uma série temporal, o modelo subjacente é o de processo estocástico; podemos pensar que a série efetivamente observada é uma das infinitas possíveis realizações desse processo. A população hipotética aqui seria o conjunto de todas essas realizações, e a série observada seria a amostra. Veja Morettin e Toloi (2006) para mais informações.

Neste livro, iremos nos concentrar principalmente em levantamentos amostrais e, mais ainda, num caso simples de amostragem probabilística, a amostragem aleatória simples, com reposição, a ser designada por AAS. O leitor poderá consultar Bussab e Bolfarine (2005) para obter mais detalhes sobre outros procedimentos amostrais. Um breve resumo sobre alguns planos é dado no Problema 37. Noções sobre planejamento de experimentos podem ser vistas em Peres e Saldiva (1982).

Problemas:

1. Dê sua opinião sobre os tipos de problemas que surgiriam nos seguintes planos amostrais:
 - (a) Para investigar a proporção dos operários de uma fábrica favoráveis à mudança do início das atividades das 7h para as 7h 30 min, decidiu-se entrevistar os 30 primeiros operários que chegasse à fábrica na quarta-feira.
 - (b) Mesmo procedimento, só que o objetivo é estimar a altura média dos operários.
 - (c) Para estimar a porcentagem média da receita municipal investida em lazer, enviaram-se questionários a todas as prefeituras, e a amostra foi formada pelas prefeituras que enviaram as respostas.
 - (d) Para verificar o fato de oferecer brindes nas vendas de sabão em pó, tomaram-se quatro supermercados na zona sul e quatro na zona norte de uma cidade. Nas quatro lojas da zona sul, o produto era vendido com brinde, enquanto nas outras quatro era vendido sem brinde. No fim do mês, compararam-se as vendas da zona sul com as da zona norte.
2. Refazer o Problema 7 do Capítulo 8.
Numa urna, há cinco tiras de papel, numeradas 1, 3, 5, 5, 7. Uma tira é sorteada e recolocada na urna; então, uma segunda tira é sorteada. Sejam X_1 e X_2 o primeiro e o segundo números sorteados.
 - (a) Determine a distribuição conjunta de X_1 e X_2 .

- (b) Determine as distribuições marginais de X_1 e X_2 . Elas são independentes?
- (c) Encontre a média e a variância de X_1 , X_2 e $\bar{X} = \frac{X_1+X_2}{2}$.
- (d) Como seriam as respostas anteriores se a primeira tira de papel sorteada não fosse devolvida à urna antes da segunda extração? Comente.

10.5: Amostragem Aleatória Simples

A amostragem aleatória simples é a maneira mais fácil para selecionarmos uma amostra probabilística de uma população. Além disso, o conhecimento adquirido com esse procedimento servirá de base para o aprendizado e desenvolvimento de outros procedimentos amostrais, planejamento de experimentos, estudos observacionais etc. Comecemos introduzindo o conceito de AAS de uma população finita, para a qual temos uma listagem de todas as N unidades elementares.

Podemos obter uma amostra nessas condições, escrevendo cada elemento da população num cartão, misturando-os numa urna e sorteando tantos cartões quantos desejarmos na amostra. Esse procedimento torna-se inviável quando a população é muito grande. Nesse caso, usa-se um processo alternativo, no qual os elementos são numerados e em seguida sorteados por meio de uma tabela de números aleatórios (veja a sua utilização em Problemas e Complementos) ou por meio do uso de computadores, que podem gerar números aleatórios (veja o Capítulo 9).

Utilizando-se um procedimento aleatório, sorteia-se um elemento da população, sendo que todos os elementos têm a mesma probabilidade de ser selecionados. Repete-se o procedimento até que sejam sorteadas as n unidades da amostra. Podemos ter uma AAS com reposição, se for permitido que uma unidade possa ser sorteada mais de uma vez, e sem reposição, se a unidade sorteada for removida da população. Do ponto de vista da quantidade de informação contida na amostra, amostrar sem reposição é mais adequado. Contudo, a amostragem com reposição conduz a um tratamento teórico mais simples, pois ela implica que tenhamos independência entre as unidades selecionadas. Essa independência facilita o desenvolvimento das propriedades dos estimadores que serão considerados.

Portanto, para o restante do livro, o plano amostral considerado será o de amostragem aleatória simples com reposição, que denotaremos simplesmente por AAS. Vejamos com algum detalhe o significado mais preciso de uma amostra.

Exemplo 10.7 Considere o Problema 2 acima, em que colhemos todas as amostras possíveis de tamanho 2, com reposição, da população $\{1, 3, 5, 5, 7\}$. Defina a variável X : valor assumido pelo elemento na população. Então, a distribuição de X é dada pela Tabela 10.1.

Tabela 1: Tabela 10.1 Distribuição da v.a. X para o problema 2

x	1	3	5	7
$P(X = x)$	$\frac{1}{5}$	$\frac{1}{5}$	$\frac{2}{5}$	$\frac{1}{5}$

Indicando por X_1 o número selecionado na primeira extração e por X_2 o número selecionado na segunda extração, vimos que era possível escrever a distribuição conjunta do par (X_1, X_2) . Veja também a Tabela 10.2.

Além disso, as distribuições marginais de X_1 e X_2 são independentes e iguais à distribuição de X . Desse modo, cada uma das 25 possíveis amostras de tamanho 2 que podemos extrair dessa população corresponde a observar uma particular realização da v.a. (X_1, X_2) , com X_1 e X_2 independentes e $P(X_1 = x) = P(X_2 = x) = P(X = x)$, para todo x . Essa é a caracterização de amostra casual simples que iremos usar neste livro.

Definição. Uma amostra aleatória simples de tamanho n de uma variável aleatória X , com dada distribuição, é o conjunto de n variáveis aleatórias independentes X_1, X_2, \dots, X_n , cada uma com a mesma distribuição de X . Ou seja, a amostra será a n -upla ordenada (X_1, X_2, \dots, X_n) , em que X_i indica a observação do i -ésimo elemento sorteado. Quando a população é caracterizada por uma distribuição de probabilidades, o modo mais simples para sortear uma AAS é usar os procedimentos de simulação estudados no Capítulo 9. O processo de simular uma observação de uma distribuição especificada por seus parâmetros nada mais é do que retirar uma AAS de tamanho um da população. Desse modo, para retirar uma AAS (com reposição) de n indivíduos da população X , basta gerar n números aleatórios independentes dessa distribuição.

Exemplo 10.8 Vamos retirar uma AAS de 5 alturas (em cm) de uma população de mulheres cujas alturas X seguem a distribuição $N(167; 25)$. Usando-se, por exemplo, o gerador de números aleatórios do Excel, fornecendo os parâmetros $\mu = 167$ e $\sigma = 5$, além do tamanho da amostra $n = 5$, obtemos os valores:

$$x_1 = 165, x_2 = 161, x_3 = 168, x_4 = 173, x_5 = 173.$$

Note que, se você for gerar uma tal amostra, poderá obter valores diferentes desses. Observe, também, que o primeiro elemento a ser observado pode ser qualquer valor da população simulada $N(167; 25)$. Desse modo, indicando por X_1 o valor observado na primeira extração, concluímos que $X_1 \sim N(167; 25)$. Como a geração do segundo número aleatório é feita independentemente do segundo, resulta que a v.a. X_2 , valor observado na segunda extração, também segue uma distribuição $N(167; 25)$, e assim por diante. Diante do exposto, vemos que continua válida a definição de AAS dada acima, quando a amostra é retirada de uma população referenciada pela sua distribuição de probabilidades. No caso de uma população X contínua, com f.d.p. $f(x)$, a f.d.p. conjunta da amostra (X_1, X_2, \dots, X_n) , segundo o que vimos no Capítulo 8, será dada por

$$f(x_1, x_2, \dots, x_n) = f_1(x_1)f_2(x_2)\dots f_n(x_n) = \prod_{i=1}^n f_i(x_i),$$

em que $f(x_i)$ denota a distribuição (marginal) de $X_i, i = 1, 2, \dots, n$.

No caso da AAS temos:

$$f(x_1, x_2, \dots, x_n) = \prod_{i=1}^n f(x_i).$$

Antes de prosseguirmos, seria interessante fazer uma comparação da inferência estatística com o processo de simulação da população. Podemos imaginar que qualquer característica X de interesse seja produzida por um “programa” (modelo) de gerador de números aleatórios, e que somente o “proprietário” (natureza) desse programa é que conhece a forma da distribuição de X , os valores dos parâmetros etc. relacionados ao programa. Quando “obtemos” a amostra, estamos apenas observando o resultado da simulação, não conhecemos nada do processo gerador dos dados. O objetivo da inferência estatística é fornecer critérios para nos ajudar a descobrir a forma da distribuição e/ou parâmetros usados pelo “proprietário”. Bons indicadores desses valores nos ajudam a entender melhor os fenômenos e fazer previsões para futuras observações. 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, \dots, X_n) definido acima.

Problemas

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

Número de filhos	Porcentagem
0	10
1	20
2	30
3	25
4	15
Total	100

- (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, \dots, X_n) , esta será dada por

$$\bar{X} = \frac{1}{n} (X_1 + X_2 + \dots + X_n) = \frac{\sum_{i=1}^n X_i}{n}.$$

É fácil verificar que \bar{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_2, \dots, X_n) .

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

As estatísticas mais comuns são:

A média da amostra :

$$\bar{X} = \frac{\sum_{i=1}^n X_i}{n}.$$

A variância da amostra:

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

O menor valor da amostra:

$$Y_1 = X_{(1)} = \text{Min}(X_1, X_2, \dots, X_n).$$

O maior valor da amostra:

$$Y_n = X_{(n)} = \text{Max}(X_1, X_2, \dots, X_n).$$

A amplitude total da amostra:

$$A = W = Y_n - Y_1 = X_{(n)} - X_{(1)}.$$

A i -ésima maior observação da amostra ou a i -ésima estatística de ordem da amostra:

$$Y_i = X_{(i)}, \quad i = 1, 2, \dots, n.$$

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

$$Y_1 = X_{(1)} \leq Y_2 = X_{(2)} \leq \dots \leq Y_n = X_{(n)},$$

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 $q_1 = q(0, 25)$, o primeiro quartil, $q_2 = q(0, 50)$, o segundo quartil ou mediana amostral, e $q_3 = q(0, 75)$, o terceiro quartil.

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 $\mu = E(X)$ e sua variância $\sigma^2 = \text{Var}(X)$.

Os símbolos mais comuns são dados na tabela a seguir.

Denominação	População	Amostra
Média	$\mu = E(X)$	$\bar{X} = \frac{\sum_{i=1}^n X_i}{n}$.
Mediana	$Md = Q_2$	$md = q_2$
Variância	σ^2	$S^2 = \frac{\sum_{i=1}^n (X_i - \bar{X})^2}{n-1}$.
Número de elementos	N	n
Proporção	P	p
Quantil	$Q(p)$	$q(p)$
Quartis	Q_1, Q_2, Q_3	q_1, q_2, q_3
Distância Interquartil	$d_Q = Q_3 - Q_1$	$d_q = q_3 - q_1$
Função densidade	$f(x)$	histograma
Função de Distribuição	$F(x)$	$F_e(x)$

em que $F_e(x)$ significa função de distribuição empírica definida por:

$$F_e(x) = \frac{N(x)}{n},$$

em que $N(x)$ é o número de observações da amostra menores ou iguais a 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 por meio 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, \dots, X_n) , ou seja, $T = h(X_1, X_2, \dots, X_n)$. 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_2, \dots, 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, em que 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 .

Vejamos alguns exemplos simples para aclarar um pouco mais o conceito de distribuição amostral de uma estatística. Nossa 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.

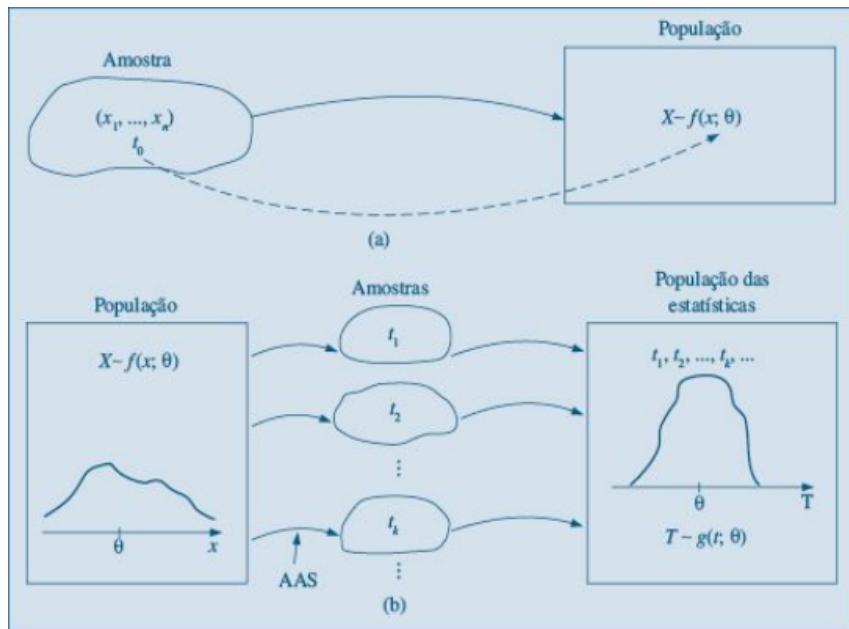


Figura 4:

X_2		1	3	5	7	Total
X_1	1	$\frac{1}{25}$	$\frac{1}{25}$	$\frac{2}{25}$	$\frac{1}{25}$	$\frac{1}{5}$
	3	$\frac{1}{25}$	$\frac{1}{25}$	$\frac{2}{25}$	$\frac{1}{25}$	$\frac{1}{5}$
5	$\frac{2}{25}$	$\frac{2}{25}$	$\frac{4}{25}$	$\frac{2}{25}$	$\frac{2}{25}$	$\frac{2}{5}$
7	$\frac{1}{25}$	$\frac{1}{25}$	$\frac{2}{25}$	$\frac{1}{25}$	$\frac{1}{25}$	$\frac{1}{5}$
Total	$\frac{1}{5}$	$\frac{1}{5}$	$\frac{2}{5}$	$\frac{1}{5}$	$\frac{1}{5}$	1

Vejamos qual é a distribuição da estatística

$$\bar{X} = \frac{X_1 + X_2}{2}. \quad (10.1)$$

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

$$P(\bar{X} = 1) = \frac{1}{25}.$$

Obteremos a média igual a 3 quando ocorrer o evento $A = \{(1, 5), (3, 3), (5, 1)\}$, logo

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

Procedendo de maneira análoga para os demais valores que \bar{X} pode assumir, obtemos a Tabela 10.3, que dá a distribuição da v.a. X .

Na Figura 10.2 temos as distribuições de X e de \bar{X} .

\bar{x}	1	2	3	4	5	6	7	Total
$P(X = \bar{x})$	$\frac{1}{25}$	$\frac{2}{25}$	$\frac{5}{25}$	$\frac{6}{25}$	$\frac{6}{25}$	$\frac{4}{25}$	$\frac{1}{25}$	1

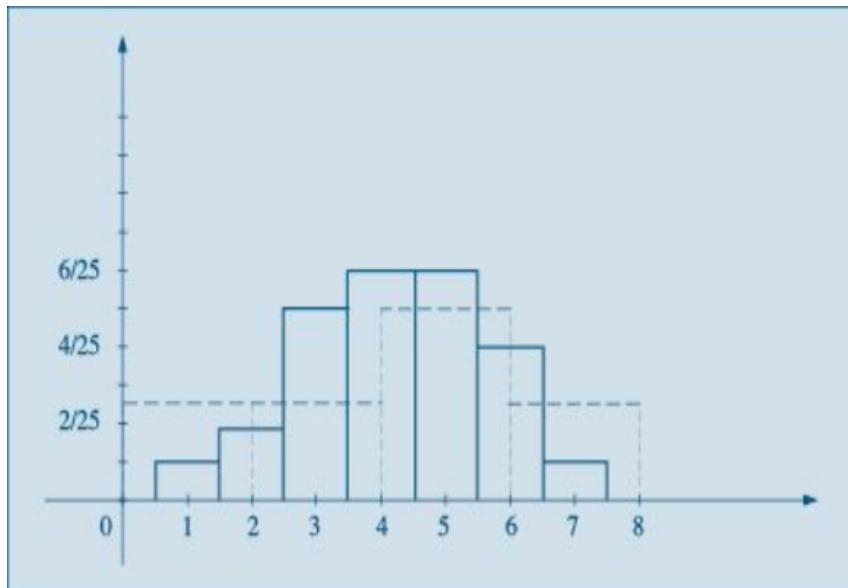


Figura 5:

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

$$W = Y_n - Y_1$$

a amplitude total e

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

respectivamente.

A distribuição amostral de W é dada por:

w	0	2	4	6	Total
$P(W = w)$	$\frac{7}{25}$	$\frac{10}{25}$	$\frac{6}{25}$	$\frac{2}{25}$	1

A distribuição amostral de S^2 é dada por:

s^2	0	2	8	18	Total
$P(S^2 = s^2)$	$\frac{7}{25}$	$\frac{10}{25}$	$\frac{6}{25}$	$\frac{2}{25}$	1

Para responder estas perguntas vamos construir um quadro auxiliar:

(x_1, x_2)	$P(X_1 = x_1, X_2 = x_2)$	\bar{x}	w	s^2
(1, 1)	0,04	1	0	0
(1, 3)	0,04	2	2	2
(1, 5)	0,08	3	4	8
(1, 7)	0,04	4	6	18
(3, 1)	0,04	2	2	2
(3, 3)	0,04	3	0	0
(3, 5)	0,08	4	2	2
(3, 7)	0,04	5	4	8
(5, 1)	0,08	3	4	8
(5, 3)	0,08	4	2	2
(5, 5)	0,16	5	0	0
(5, 7)	0,08	6	2	2
(7, 1)	0,04	6	2	18
(7, 3)	0,04	5	4	8
(7, 5)	0,08	6	2	2
(7, 7)	0,04	7	0	0

Com ele fica fácil calcular as distribuições amostrais pedidas.

Exemplo 10.5 (continuação) No caso do lançamento de uma moeda 50 vezes, usando como estatística T = 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 a 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

$$P(T \geq 36 | n = 50, p = 0,5) = 0,0013 = 0,13\%.$$

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 $p > 0,5$.

Vamos usar o **R** para calcular esta probabilidade. Vamos calcular a exata usando a binomial e a aproximada com e sem a correção de continuidade.

```
> ##X= número de caras em 50 lançamentos de uma moeda honesta
>
> n=50;p=1/2;n;p
[1] 50
[1] 0.5
>
>
> ##P=P(X>=36)=1-P(X<=35)
>
> ##P_e=probabilidade exata
>
> P_e=1-pbinom(35,n,p);P_e
```

```

[1] 0.001301086
> round(P_e,4)
[1] 0.0013
>
>
> ##P_as=probabilidade aproximada sem correção de continuidade
>
>
> mu=n*p;mu
[1] 25
> sigma2=n*p*(1-p);sigma2
[1] 12.5
> sigma=sqrt(sigma2);sigma
[1] 3.535534
> P_as=1-pnorm(35,mu,sigma);P_as
[1] 0.002338867
> round(P_as,4)
[1] 0.0023
>
>
> ##P_ac=probabilidade aproximada com correção de continuidade
>
>
> P_ac=1-pnorm(35.5,mu,sigma);P_ac
[1] 0.001489733
> round(P_ac,4)
[1] 0.0015
>

```

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 ideia 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:

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

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

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

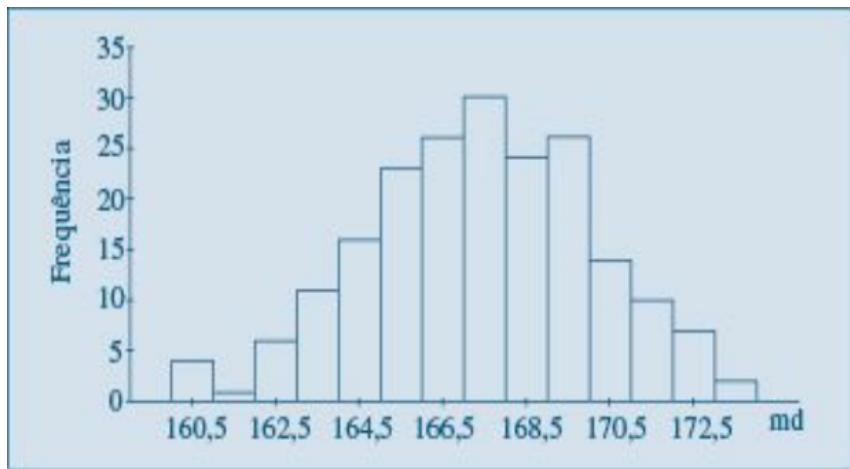


Figura 6:

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

Problemas:

4. Usando os dados da Tabela 10.2, construa a distribuição amostral da estatística

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

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 $Var(X_i)$, $i = 1, 2$;

(c) construa a distribuição amostral de

$$\bar{X} = \frac{X_1 + X_2}{2}.$$

(d) calcule $E(\bar{X})$ e $Var(\bar{X})$;

(e) faça num mesmo gráfico os histogramas de X e de \bar{X} ;

(f) construa as distribuições amostrais de

$$S^2 = \sum_{i=1}^2 (X_i - \bar{X})^2 = |X_1 - X_2| \quad \text{e} \quad \hat{\sigma}^2 = \frac{\sum_{i=1}^2 (X_i - \bar{X})^2}{2}.$$

(g) baseado no resultado de (f), qual dos dois estimadores você usaria para estimar a variância de X ? Por quê?

(h) calcule $P(|\bar{X} - \mu| > 1)$.

6. Ainda com os dados do Problema 3, e para amostras de tamanho 3:

(a) determine a distribuição amostral de \bar{X} e faça o histograma;

(b) calcule a média e variância de \bar{X} ;

(c) calcule $P(|\bar{X} - \mu| > 1)$.

(d) se as amostras fossem de tamanho 4, a $P(|\bar{X} - \mu| > 1)$ 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 \bar{X} , a média da amostra. Consideremos uma população identificada pela variável X , cujos parâmetros média populacional $\mu = E(X)$ e variância populacional $\sigma^2 = Var(X)$ são supostos conhecidos. Vamos retirar todas as possíveis AAS de tamanho n dessa população, e para cada uma calcular a média \bar{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-(Continuação): A população $\{1, 3, 5, 5, 7\}$ tem média $\mu = 4,2$ e variância $\sigma^2 = 4,16$.

A distribuição amostral de \bar{X} está na Tabela 10.3, da qual obtemos

$$E(\bar{X}) = \sum_i \bar{x}_i p_i = 1 \times \frac{1}{25} + 2 \times \frac{1}{25} + 3 \times \frac{1}{25} + 4 \times \frac{1}{25} + 5 \times \frac{2}{25} + 6 \times \frac{1}{25} + 7 \times \frac{1}{25} = 4,2$$

De modo análogo, encontramos

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

Verificamos, aqui, dois fatos: primeiro, a média das médias amostrais coincide com a média populacional; segundo, a variância de \bar{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, \dots, X_n) uma AAS de X . Então,

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

Prova. Pelas propriedades vistas no Capítulo 8, temos:

$$E(\bar{X}) = (1/n)(E(X_1) + E(X_2) + \dots + E(X_n)) = (1/n)(\mu + \mu + \dots + \mu) = \frac{n\mu}{n} = \mu.$$

De modo análogo, e pelo fato de X_1, X_2, \dots, X_n serem independentes, temos

$$\text{Var}(\bar{X}) = (1/n^2)(V(X_1) + V(X_2) + \dots + V(X_n)) = (1/n^2)(\sigma^2 + \sigma^2 + \dots + \sigma^2) = \frac{n\sigma^2}{n^2} = \frac{\sigma^2}{n}.$$

Determinamos, então, a média e a variância da distribuição amostral de \bar{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 \bar{X} para $n = 1, 2$ e 3 .

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

Figura 10.4 Distribuição de X para amostras de 1, 3, 5, 5, 7.

(ii) Para $n = 2$, baseados na Tabela 10.3, temos a distribuição de \bar{X} dada na Figura 10.4(b), com

$$E(\bar{X}) = 4,2 \quad \text{e} \quad \text{Var}(\bar{X}) = 2,08.$$

(iii) Finalmente, para $n = 3$, com os dados da Tabela 10.6, temos a distribuição de \bar{X} na Figura 10.4 (c), com

$$E(\bar{X}) = 4,2 \quad \text{e} \quad \text{Var}(\bar{X}) = 1,39.$$

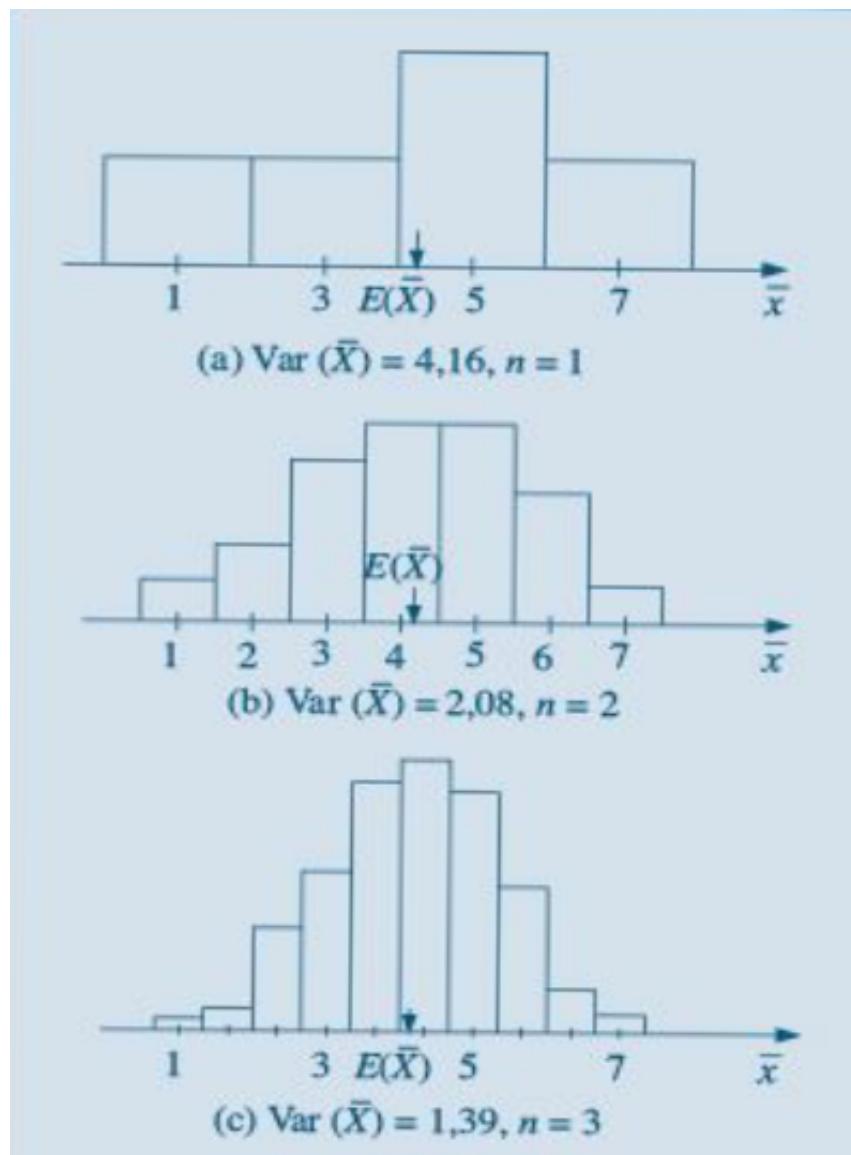


Figura 7:

Observe que, conforme n vai aumentando, o histograma tende a se concentrar cada vez mais em torno de $E(\bar{X}) = E(X) = 4,2$, 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 \bar{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 \bar{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).

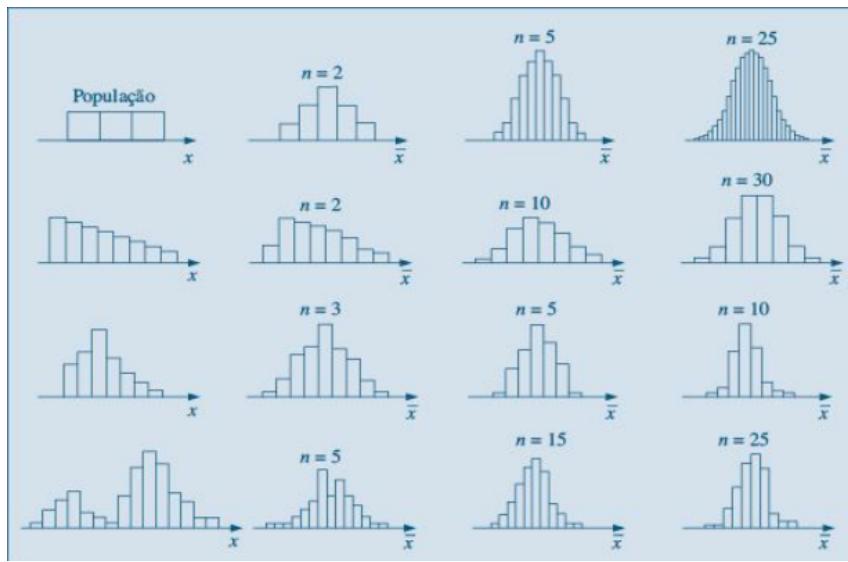


Figura 8:

Teorema 10.2. (TLC) Para amostras aleatórias simples (X_1, X_2, \dots, X_n) , retiradas de uma população com média μ e variância σ^2 finita, a distribuição amostral da média \bar{X} aproxima-se, para n grande, de uma distribuição normal, com média μ e variância $\frac{\sigma^2}{n}$. 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 \bar{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 \bar{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, em que 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, \bar{X} terá uma distribuição normal com média 500 e variância

$$\frac{\sigma^2}{n} = \frac{100}{100} = 1.$$

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

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

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.

Vamos resolver no **R**:

>

```
> #####X peso do pacote de café
>
> #####X ~N(500,100)
> mu=500;sigma2=100;sigma=sqrt(sigma2);mu;sigma
[1] 500
[1] 10
>
>
> ##X_1,X_2,...,X_100-amostra aleatória
>
> ##Xb=media da amostra é uma variável aleatória
> muXb=mu;muXb
[1] 500
>
> n=100;n
[1] 100
> mu
[1] 500
> sigma2Xb=sigma2/n;sigma2Xb
[1] 1
> sigmaXb=sqrt(sigma2Xb);sigmaXb
[1] 1
>
> ##p=P(|Xb-500|<2)=P(498<Xb<502)=P(Xb<502)-P(Xb<=498)=p_2-p_1
>
> p_2=pnorm(502,muXb,sigmaXb);p_2
[1] 0.9772499
>
> p_1=pnorm(498,muXb,sigmaXb);p_1
[1] 0.02275013
>
> p=p_2-p_1;p;round(p,2)
[1] 0.9544997
[1] 0.95
>
> ##Vamos calcular pela Normal Padrão
>
>
> z_2=(502-muXb)/sigmaXb;z_2
[1] 2
>
> z_1=(498-muXb)/sigmaXb;z_1
[1] -2
>
>
> ##p=P(498<Xb<502)=P(-2<Z<2)= P(-2<Z<0)+P(0<Z<2)= 2P(0<Z<2)=2*p_3
>
> p_3=pnorm(2) -pnorm(0);p_3
```

```

[1] 0.4772499
>
> ##Número de casas decimais da tabela III do livro
>
> round(p_3,5)
[1] 0.47725
>
>
> p=2*p_3;p;round(p,2)
[1] 0.9544997
[1] 0.95
>
>

```

Outra maneira de apresentar o TLC é por meio do

Corolário 10.1 Se (X_1, X_2, \dots, X_n) for uma amostra aleatória simples da população X , com média μ e variância σ^2 finita, e

$$\bar{X} = \frac{X_1 + X_2 + \dots + X_n}{n}$$

, então

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

(10.2) Basta notar que se usou a transformação usual de reduzir a distribuição de \bar{X} a uma normal padrão. Observe, também, que (10.2) pode ser escrita como (10.3)

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

Chamemos de E a v.a. que mede a diferença entre a estatística \bar{X} e o parâmetro μ , isto é,

$$E = \bar{X} - \mu.$$

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 $\frac{\sigma^2}{n}$, isto é,

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

O TLC afirma que 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 população

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 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 \bar{X} for a média de uma amostra de 16 elementos retirados dessa população, calcule $P(90 < \bar{X} < 110)$.
 - (c) Represente, num único gráfico, as distribuições de X e \bar{X} .
 - (d) Que tamanho deveria ter a amostra para que $P(90 < \bar{X} < 110) = 0,95$?
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, os quais 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?

Solução: Seja $X \sim N(70, 100)$ a distribuição do peso dos usuários do elevador. Seja X_i o peso do i -ésimo passageiro, $i = 1, 2, \dots, n$ e

$$S_n = \sum_{i=1}^n X_i,$$

o peso total dos n passageiros.

Sabemos que X_1, X_2, \dots, X_n são variáveis aleatórias independentes e identicamente distribuídas a X .

Sabemos que

$$E(S_n) = n\mu = n\mu, \quad V(S_n) = n\sigma^2.$$

Temos que:

$$S_n \sim N(n\mu, n\sigma^2),$$

$$Z = \frac{S_n - n\mu}{\sqrt{n\sigma^2}} \sim N(0, 1).$$

O problema pede

$$P(S_n > 500) = P\left(Z > \frac{500 - 70n}{\sqrt{100n}}\right).$$

Vamos resolver o item **a**. Para $n = 7$ temos:

$$p_a = P(S_7 > 500) = P\left(Z > \frac{500 - 490}{\sqrt{700}}\right).$$

$$p_a = P(S_7 > 500) = P(Z > 0,38) = 0,5 - P(0 < Z \leq 0,38),$$

$$p_a = 0,5 - 0,14803 = 0,35197 \approx 0,352.$$

Isto quer dizer que de cada 1000 viagens 352 terão problemas de excesso de peso.

Vamos resolver o item **b**. Para $n = 6$ temos:

$$p_b = P(S_6 > 500) = P\left(Z > \frac{500 - 420}{\sqrt{600}}\right).$$

$$p_b = P(S_6 > 500) = P(Z > 3,27) = 0,5 - P(0 < Z \leq 3,27),$$

$$p_b = 0,5 - 0,49946 = 0,00054 \approx 0,0005.$$

Isto quer dizer que de cada 10000 viagens somente 5 terão problemas de excesso de peso. Bem mais seguro.

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,

$$X \sim Ber(p)$$

com

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

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

$$S_n \sim Bin(n, p).$$

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

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

Então,

$$P(S_n = k) = P\left(\frac{S_n}{n} = \frac{k}{n}\right) = P\left(\hat{p} = \frac{k}{n}\right),$$

ou seja, a distribuição amostral de \hat{p} é obtida da distribuição de S_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

$$S_n = X_1 + X_2 + \dots + X_n = \sum_{i=1}^n X_i,$$

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

Podemos escrever que

$$S_n = n \bar{X},$$

mas pelo TLC, \bar{X} terá distribuição aproximadamente normal, com

$$E(\bar{X}) = p \quad e \quad V(\bar{X}) = \frac{p(1-p)}{n},$$

ou seja,

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

Logo, a transformação $S_n = n\bar{X}$ terá a distribuição

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

que foi a aproximação adotada na Seção 7.5.

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

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} a 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

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

e, portanto, a probabilidade pedida é igual a

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

Vamos fazer a solução exata.

Seja S o número de mulheres na amostra de tamanho $n = 10$. Assim

$$S \sim Bin(10, p = 0,3)$$

e a f.p. de S é dada por:

$$P(S = s) = \binom{10}{s} 0,3^s 0,7^{10-s} I_A(s), \quad A = \{0, 1, \dots, 10\}.$$

Seja

$$\hat{p} = \frac{S}{10}$$

e

$$\hat{p} - p = \frac{S}{10} - p = \frac{S - 10p}{10}.$$

Para $p = 0,3$ temos:

$$\hat{p} - p = \frac{S - 3}{10},$$

$$|\hat{p} - p| = \frac{|S - 3|}{10} < 0,01.$$

Assim

$$|S - 3| < 0,1 \quad \text{ou} \quad -0,1 < S - 3 < 0,1$$

ou

$$2,9 < S < 3,1.$$

O único ponto que satisfaz a desigualdade é $S = 3$.

Assim

$$P(|\hat{p} - p| < 0,01) = P(S = 3) = \binom{10}{3} 0,3^3 0,7^7 = 0,26,$$

bem distante da obtida pela aproximação da normal.

Vamos fazer pelo **R**

```
>
> ####X=número de mulheres na amostra de tamanho amostral .
>
> #####pc leia-se p chapéu
>
> #####E=|pc-p|, erro amostral absoluto.
>
> n=10;p=0.3
>
> x=0:10;x
[1] 0 1 2 3 4 5 6 7 8 9 10
> px=dbinom(x,n,p)
>
```

```

> pc=x/n
>
> E=abs(pc-p)
>
> tab=cbind(x,px,pc,E);tab
x          px      pc      E
[1,] 0 0.0282475249 0.0 0.3
[2,] 1 0.1210608210 0.1 0.2
[3,] 2 0.2334744405 0.2 0.1
[4,] 3 0.2668279320 0.3 0.0
[5,] 4 0.2001209490 0.4 0.1
[6,] 5 0.1029193452 0.5 0.2
[7,] 6 0.0367569090 0.6 0.3
[8,] 7 0.0090016920 0.7 0.4
[9,] 8 0.0014467005 0.8 0.5
[10,] 9 0.0001377810 0.9 0.6
[11,] 10 0.0000059049 1.0 0.7
>
> ###E <0,1 equivale a E=0 ou S=3
>
> ##Pode-se obter a distribuição amostral do erro absoluto na coluna E.
>

```

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.
- Construa a distribuição exata de \hat{p} (use a tábua da distribuição binomial).
 - Construa a aproximação normal à binomial.
 - Você pensa que a segunda distribuição é uma boa aproximação da primeira?
 - Já sabemos que, para dado p fixo, a aproximação melhora conforme 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 $T = h(X_1, X_2, \dots, X_n)$. Mas quanto mais complexa for essa relação h , 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,

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

a mediana amostral, md ,

$$md = \begin{cases} \frac{Y_{n/2} + Y_{(n/2+1)}}{2}, & \text{se } n \text{ é par.} \\ Y_{(n+1)/2}, & \text{se } n \text{ é ímpar..} \end{cases}$$

e o estimador

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

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.

Nossa população é representada pela variável X com a seguinte função de probabilidade:

$$f(x) = P(X = x) = \frac{1}{25} I_{\{1,3,7\}}(x) + \frac{2}{25} I_{\{5\}}(x).$$

Os parâmetros populacionais que serão analisados:

$$\mu = 4, 2 \quad , \quad \sigma^2 = 4, 16 \quad e \quad Md = 5.$$

Vamos construir um quadro auxiliar para os tipos de amostra de tamanho $n = 3$.

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

O suporte de \bar{X} é dado por:

$$A = \left\{ 1, \frac{5}{3}, 2, \frac{7}{3}, 3, \frac{11}{3}, \frac{13}{3}, 5, \frac{17}{3}, \frac{19}{3}, 7 \right\}$$

A função de probabilidade de \bar{X} é dada por:

\bar{x}	1	$\frac{5}{3}$	2	$\frac{7}{3}$	3	$\frac{11}{3}$	5	$\frac{17}{3}$	$\frac{19}{3}$	7
$P(\bar{X} = \bar{x})$	$\frac{1}{125}$	$\frac{3}{125}$	$\frac{9}{125}$	$\frac{16}{125}$	$\frac{30}{125}$	$\frac{24}{125}$	$\frac{23}{125}$	$\frac{15}{125}$	$\frac{6}{125}$	$\frac{1}{125}$

O valor esperado da média amostral é dado por:

$$E(\bar{X}) = 4,2 = \mu.$$

A variância da média amostral é dada por:

$$Var(\bar{X}) = \frac{4,16}{3} = \frac{\sigma^2}{n}.$$

```
> EX=sum(x*px);EX
[1] 4.2
> EX2=sum(x^2*px);EX2
[1] 21.8
> VX=EX2-EX^2;VX
```

```
[1] 4.16
>
> ###Valores da média amostral
>
> xb=c(1,5/3,7/3,3,11/3,13/3,5,17/3,19/3,7)
>
> p=c(1,3,9,16,24,27,23,15,6,1)/125
> sum(p)
[1] 1
>
> EXbar=sum(xb*p);EXbar
[1] 4.2
>
> EXbar2=sum(xb^2*p);EXbar2
[1] 19.02667
>
> VXbar=EXbar2-EXbar^2;VXbar
[1] 1.386667
> n=3;sigma2=4.16
>
> sigma2/n;VXbar
[1] 1.386667
[1] 1.386667
> require(MASS)
>
> fractions(sigma2/n);fractions(VXbar)
[1] 104/75
[1] 104/75
>
>
```



```
> md=c(1,3,5,7)
> pmd=c(13,31,68,13)/125
> sum(pmd)
[1] 1
>
> Emd=sum(md*pmd);Emd
[1] 4.296
>
> Emd2=sum(md^2*pmd);Emd2
[1] 21.032
> Vmd=Emd2-Emd^2;Vmd
[1] 2.576384
>
```

```
>
> s2=c(0,4/3,4,16/3,28/3,12)
> ps2=c(11,42,24,24,18,6)/125
> sum(ps2)
[1] 1
>
> ES2=sum(s2*ps2);ES2
[1] 4.16
>
> ES2==sigma2 ##S^2 é um estimador não viciado para sigma^2.
[1] TRUE
>
>
> ES22=sum(s2^2*ps2);ES22
[1] 28.58667
>
> VS2= ES22-ES2^2;VS2
[1] 11.28107
>
>
> ###0 quarto momento central de X é dado por:
>
> mu_4=sum((x-EX)^4*px);mu_4; mu_4/n
[1] 33.8432
[1] 11.28107
>

> ###Seja U estimador da variância com o denominador n.
>
>
> u=c(0,8/9,8/3,32/9,56/9,8)
> pu=c(11,42,24,24,18,6)/125
> sum(pu)
[1] 1
> EU=sum(u*pu);EU
[1] 2.773333
> EU==sigma2
[1] FALSE
>
> ##U é um estimador viciado para sigma^2.
>
> EU2=sum(u^2*pu);EU2
[1] 12.70519
> VU=EU2-EU^2;VU
[1] 5.013807
>
```

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.

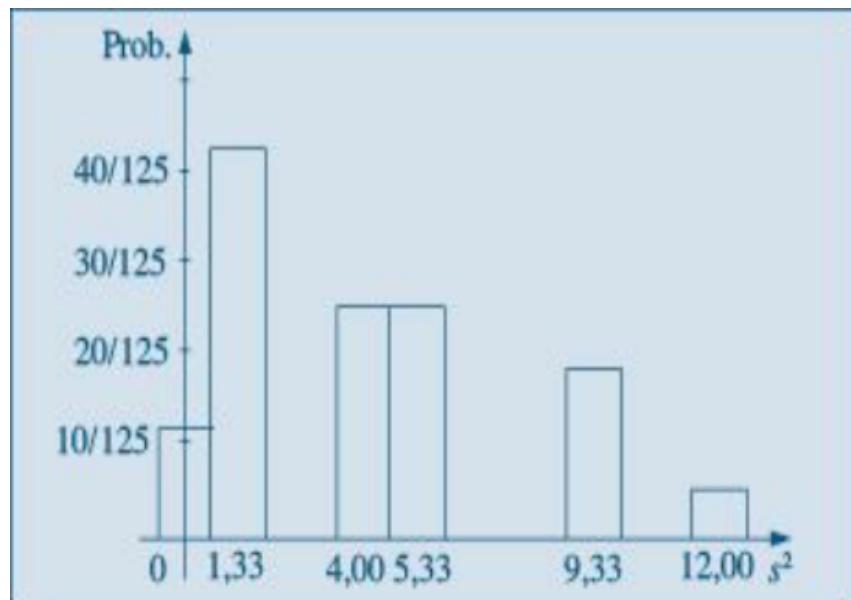


Figura 9:

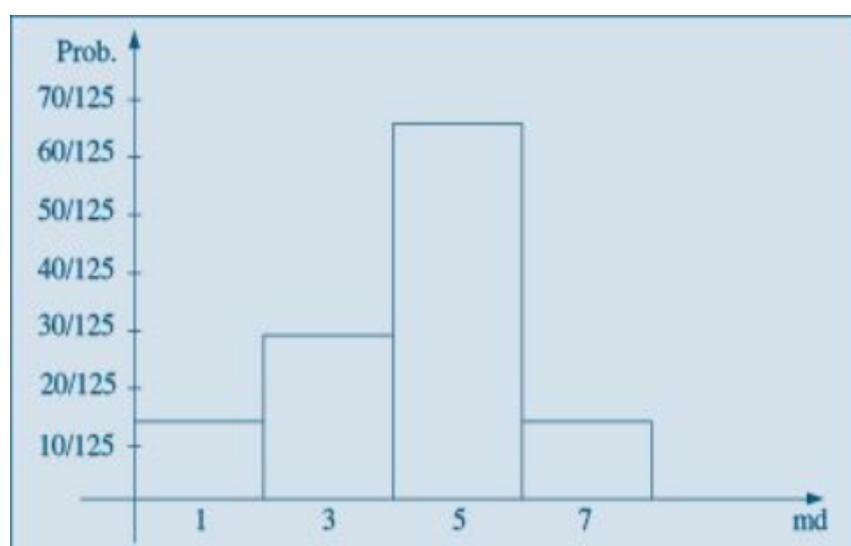


Figura 10:

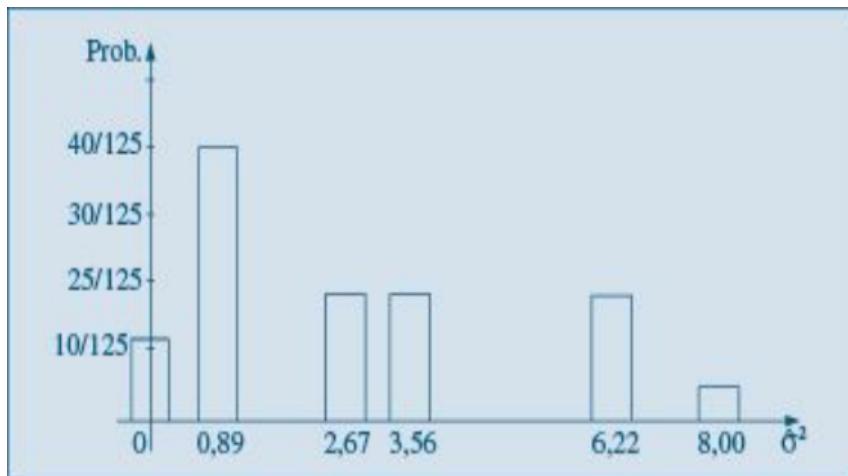


Figura 11:

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.

Na realidade a esperança de S^2

$$E(S^2) = \sigma^2.$$

Assim a variância amostral S^2 é um estimador não viciado para a variância populacional σ^2 .

Além disso

$$Var(S^2) = \frac{1}{n} \left(\mu_4 - \frac{n-3}{n-1} \sigma^4 \right),$$

em que

$$\mu_4 = E((X - \mu)^4)$$

é o quarto momento central de X .

Se a lei de X for normal com média μ e variância σ^2 e uma amostra aleatória (X_1, X_2, \dots, X_n) é retirada então:

$$\bar{X} \sim N \left(\mu, \frac{\sigma^2}{n} \right),$$

$$V = \frac{(n-1)S^2}{\sigma^2} \sim \chi^2(n-1),$$

e

$$S^2 \sim Gama \left(r = \frac{n-1}{2}, \lambda = \frac{n-1}{2\sigma^2} \right).$$

Neste caso temos:

$$E(S^2) = \frac{r}{\lambda} = \frac{n-1}{2} \times \frac{2\sigma^2}{n-1} = \sigma^2,$$

$$V(S^2) = \frac{r}{\lambda^2} = \frac{n-1}{2} \times \frac{4\sigma^4}{(n-1)^2} = \frac{2\sigma^4}{n-1}.$$

Vamos calcular a variância usando a fórmula geral.

Sabemos que a curtose da $N(\mu, \sigma^2)$ é 3. Assim

$$K_4 = \frac{E((X-\mu)^4)}{\sigma^4} = \frac{\mu_4}{\sigma^4} = 3$$

e

$$\mu_4 = 3\sigma^4.$$

$$Var(S^2) = \frac{1}{n} \left(\mu_4 - \frac{n-3}{n-1} \sigma^4 \right) = \frac{1}{n} \left(3\sigma^4 - \frac{n-3}{n-1} \sigma^4 \right),$$

$$Var(S^2) = \frac{\sigma^4}{n} \left(3 - \frac{n-3}{n-1} \right) = \frac{\sigma^4}{n} \left(\frac{3n-3-n+3}{n-1} \right),$$

$$Var(S^2) = \frac{\sigma^4}{n} \times \frac{2n}{n-1} = \frac{2\sigma^4}{n-1}.$$

Outro fato muito importante em populações normais é que a média amostral \bar{X} e a variância amostral S^2 são variáveis aleatórias independentes.

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

$$E(md) = \mu \quad e \quad Var(md) = \frac{\pi \sigma^2}{2n}.$$

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 $u = \frac{(1+3)}{2} = 2$. 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	Frequência 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
Total	1

(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 \bar{X} e a mediana md da amostra são estimadores não viesados da mediana Md da população, no sentido que $E(\bar{X}) = E(md) = Md$.

(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

$$Z = \frac{\sqrt{n}(\bar{X} - \mu)}{\sigma}$$

para $n = 2$.

(f) Quais são os valores de $E(Z)$ e $Var(Z)$?

(g) Construa a distribuição amostral da estatística

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

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

$$t = \frac{\sqrt{n}(\bar{X} - \mu)}{S}$$

e construa seu histograma. Qual é 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)

$$Z = \frac{\sqrt{n}(\bar{X} - \mu)}{\sigma}$$

(Veja o Teorema Limite Central)

(c)

$$t = \frac{\sqrt{n}(\bar{X} - \mu)}{S}$$

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 estimativa previamente estipulado, com determinado grau de confiança.

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

$$P(|\bar{X} - \mu| \leq \epsilon) \geq \gamma,$$

(10.5)

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

$$\bar{X} \sim N(\mu, \frac{\sigma^2}{n})$$

, logo

$$\bar{X} - \mu \sim N(0, \frac{\sigma^2}{n}).$$

e portanto (10.5) pode ser escrita

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

com

$$Z = \frac{\sqrt{n}(\bar{X} - \mu)}{\sigma}.$$

Dado γ , podemos obter z_γ da $N(0, 1)$, tal que

$$P(-z_\gamma < Z < z_\gamma) = \gamma$$

, de modo que

$$\frac{\sqrt{n}\epsilon}{\sigma} = z_\gamma$$

do que obtemos finalmente (10.6)

$$n = \frac{\sigma^2 z_\gamma^2}{\epsilon^2}.$$

Note que em (10.6) conhecemos z_γ e ϵ , mas σ^2 é a variância desconhecida da população. Para podermos ter uma ideia 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 $\bar{X} = 15$ e $S^2 = 16$. Fixando-se $\epsilon = 0,5$ e $\gamma = 0,95$, temos

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

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

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

(10.7)

Como não conhecemos p , a verdadeira proporção populacional, podemos usar o fato de que

$$p(1-p) \leq \frac{1}{4}, \quad \text{para todo } p,$$

e (10.7) fica

$$n = \frac{z_\gamma^2}{4 \epsilon^2}.$$

(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 $\epsilon = 0,03$, com probabilidade $\gamma = 0,95$, teremos

$$n \approx \frac{(1,96)^2 \times 0,6 \times 0,4}{(0,03)^2} = 1024.$$

na qual usamos o fato de que $p \geq 0,60$.

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 $p \geq 0,80$. 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 $n \leq n_0$. (Use a função $f(p) = p(1 - p)$ para sua resposta.)
20. Suponha que haja a informação $p \leq p_0 < 0,5$, com p_0 conhecida. Se mostre que $n \leq n_1 < n_0$. Mostre que essa mesma relação vale se soubermos que $p \geq p_0 > 0,5$.

[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, Minitab e R 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 $P = \{1, 2, \dots, 10\}$, que está na coluna F.

Exemplo 10.16 O R e o SPlus usam 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

$$x = \{1, 2, 3, \dots, 15\}$$

, sem e com reposição.

Quadro 10.1 Geração de amostras. R e 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 $\{1, 2, \dots, 10\}$ (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
4	4	2	6
5	5	7	4
6	6		
7	7		
8	8		
9	9		
10	10		

MTB > Sample 5 C1 C2.
 MTB >
 MTB > Sample 5 C1 C3;
 SUBC>Replace.
 MTB >

1 10.13 Problemas Suplementares

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 $E = X - \mu$ como sendo o erro amostral de média. Suponha que a variância dos salários de uma certa região seja 400 reais².

(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 $0,01 \text{ cm}^2$. 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?

Na página 518 item 7 temos estatísticas de vendas de 30 veículos novos segundo o tipo: Nacional (N) e Importado(I) em março de 1999. Preço em dólares, comprimento em metros e motor em CV.

Veículo	Preço	Comprimento	Motor	Tipo
Asia Towner	9.440	3,36	40	I
Audi A3	38.850	4,15	125	I
Chevrolet Astra	10.532	4,11	110	N
Chevrolet Blazer	16.346	4,60	106	N
Chevrolet Corsa	6.176	3,73	60	N
Chevrolet Tigra	12.890	3,92	100	I
Chevrolet Vectra	13.140	4,47	110	N
Chrysler Neon	31.640	4,36	115	I
Dodge Dakota	11.630	4,98	121	N
Fiat Fiorino	6.700	4,16	76	N
Fiat Marea	12.923	4,39	127	N
Fiat Uno Mille	5257	3,64	57	N
Fiat Palio	6.260	3,73	61	N
Fiat Siena	7.780	4,10	61	I
Ford Escort	10.767	4,20	115	I
Ford Fiesta	6.316	3,83	52	N
Ford Ka	5.680	3,62	54	N
Ford Mandeia	33.718	4,56	130	I
Honda Civic	14.460	4,45	106	N
Hyundai Accent	21.500	4,12	91	I
Peugeot 106	13.840	3,68	50	I
Renault Clio	13.700	3,70	74	I
Toyota Corolla	15.520	4,39	116	N
Toyota Perua	24.632	4,40	96	N
VW Gol	6.340	3,81	54	N
VW Golf	22.200	4,15	100	I
VW Parati	9.300	4,08	69	N
VW Polo	12.018	4,14	99	I
VW Santana	11.386	4,57	101	N
VW Saveiro	7.742	4,38	88	N

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).

Vamos apresentar a tabela VII do livro Bussab e Morettin.

Tabela VII – Números Aleatórios																			
61	09	26	29	85	11	95	77	79	04	57	00	91	29	59	83	53	87	02	02
94	47	40	99	93	82	13	22	40	33	19	72	55	69	82	16	94	21	66	39
50	40	50	55	79	00	58	17	26	30	38	11	54	89	04	13	69	17	35	48
51	01	75	76	54	43	11	28	32	75	33	09	04	78	74	91	56	79	43	39
25	45	79	30	63	56	44	70	05	04	31	81	46	02	92	32	06	71	12	48
63	94	61	14	24	60	27	00	00	95	54	31	59	00	79	94	46	32	61	90
12	95	04	73	06	72	76	88	55	62	38	79	18	68	10	31	93	58	66	92
38	06	78	00	85	42	57	29	28	34	79	91	93	58	82	97	37	07	64	67
22	69	28	18	25	08	90	93	53	17	54	12	21	03	56	30	88	53	46	82
07	95	63	14	76	53	62	10	21	57	55	74	57	68	22	38	84	55	57	49
61	41	81	16	97	55	19	65	08	62	26	38	74	32	30	44	64	64	91	80
97	15	71	92	40	28	33	35	23	32	75	36	18	98	41	10	50	93	75	95
39	81	34	84	33	83	42	77	35	00	51	42	82	63	30	47	01	98	96	73
58	35	04	52	06	81	24	32	74	53	28	82	43	35	01	73	34	47	05	76
52	85	30	59	37	00	49	88	07	43	08	04	00	48	36	23	31	88	80	88
41	92	93	01	94	13	33	63	32	35	38	91	18	89	71	67	46	73	42	47
88	51	22	59	99	51	20	74	13	55	30	41	25	99	10	26	01	33	24	13
11	12	32	28	25	67	22	97	11	73	55	24	09	23	47	12	93	44	80	47
33	02	06	80	29	39	78	49	81	21	42	00	99	80	44	56	33	83	46	16
03	67	08	29	16	04	92	31	62	03	94	53	02	60	55	72	46	68	25	93
41	54	93	90	86	52	14	58	90	34	83	00	73	38	14	50	77	58	08	94
18	84	83	61	42	96	82	86	02	30	40	16	65	55	63	20	40	24	79	80
06	15	93	11	72	17	32	31	84	89	53	66	01	99	53	75	79	92	20	61
12	74	92	15	60	93	84	37	29	62	24	96	78	93	28	34	41	69	04	51
79	13	36	81	55	51	46	66	68	85	07	73	35	42	52	61	29	21	02	34
01	78	33	32	06	16	45	94	09	18	40	14	73	03	61	80	69	79	52	95
90	73	28	21	38	57	39	36	24	33	31	99	64	86	19	61	55	50	65	14
44	10	20	96	70	32	41	46	22	97	08	22	02	47	43	57	15	87	76	59
52	47	00	27	41	43	70	17	52	44	51	26	94	73	17	72	16	51	81	77
23	03	84	44	29	43	57	05	46	59	89	00	65	01	20	27	32	66	34	56

Figura 12:

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 sequencial, 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 \bar{X} e \bar{Y} .

- (a) Qual a distribuição amostral de \bar{X} ? E de \bar{Y} ?
- (b) Defina $D = \bar{X} - \bar{Y}$. O que você entende por distribuição amostral de D ?
- (c) Calcule $E(D)$ e $Var(D)$.
- (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. Considere a população $\mathbb{P} = \{1, 3, 5, 5, 7\}$. Retire amostras de tamanho $n = 2$, sem reposição e construa a distribuição amostral de

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

Obtenha $E(X)$ e $Var(X)$ e verifique (10.9).

37. Obtenha a densidade de M , dada por (10.10), para o caso de uma amostra de uma distribuição uniforme no intervalo $(0, \theta)$.

38. Suponha que temos a população $X \sim (167; 25)$. 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 \bar{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”.

39. Suponha uma população $\mathbb{P} = \{1, 2, \dots, N\}$ e a v.a. X definida sobre \mathbb{P} . Então,

$$T = \sum_{i=1}^N X_i$$

é chamado total populacional.

A média populacional é

$$\mu = \frac{T}{N}$$

e a variância populacional é

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

Considere uma AAS de tamanho n extraída de \mathbb{P} e \bar{X} a média amostral. Considere o estimador

$$\hat{T} = N\bar{X}.$$

Mostre que

$$E(\hat{T}) = T \text{ e } Var(\hat{T}) = \frac{N^2 \sigma^2}{n}.$$

40. Suponha que queiramos retirar uma amostra de uma distribuição de Bernoulli com parâmetro p . Escolhidos k dados x_1, x_2, \dots, x_k , temos que

$$\bar{X}_k = \frac{\sum_{i=1}^k x_i}{k}$$

é um estimador de p .

Então um estimador natural da variância $\sigma^2 = p(1 - p)$ da população é

$$S^2 = \bar{X}_k \times (1 - \bar{X}_k).$$

Como ficaria o algoritmo descrito no CM-4 para essa situação?

2 10.14 Complementos Metodológicos.

1. Amostras sem reposição de populações finitas. Suponha uma população com N elementos e X_1, X_2, \dots, X_N valores de uma variável como idade com

$$\begin{aligned} \mu &= \frac{\sum_{i=1}^N X_i}{N} \quad e \\ N \sigma^2 &= \sum_{i=1}^N (X_i - \bar{X})^2 = \sum_{i=1}^N X_i^2 - N \bar{X}^2 \\ N \sigma^2 &= \sum_{i=1}^N X_i^2 - N \bar{X}^2 \\ N^2 \bar{X}^2 &= \left(\sum_{i=1}^N X_i \right)^2 = \sum_{i=1}^N X_i^2 - 2 \sum_{i=1}^{N-1} \sum_{j=i+1}^N X_i X_j \\ \bar{X}^2 &= \frac{1}{N^2} \sum_{i=1}^N X_i^2 - \frac{2}{N^2} \sum_{i=1}^{N-1} \sum_{j=i+1}^N X_i X_j \end{aligned}$$

Finalmente

$$\begin{aligned} N \sigma^2 &= \sum_{i=1}^N X_i^2 - N \bar{X}^2 \\ N \sigma^2 &= \sum_{i=1}^N X_i^2 - N \frac{1}{N^2} \sum_{i=1}^N X_i^2 + \frac{2}{N^2} \sum_{i=1}^{N-1} \sum_{j=i+1}^N X_i X_j \end{aligned}$$

$$\begin{aligned}
N\sigma^2 &= \sum_{i=1}^N X_i^2 - \frac{1}{N} \sum_{i=1}^N X_i^2 + \frac{2}{N} \sum_{i=1}^{N-1} \sum_{j=i+1}^N X_i X_j \\
N\sigma^2 &= \frac{N-1}{N} \sum_{i=1}^N X_i^2 + \frac{2}{N} \sum_{i=1}^{N-1} \sum_{j=i+1}^N X_i X_j \\
\frac{N^2}{N-1} \sigma^2 &= \left[\sum_{i=1}^N X_i^2 + \frac{2}{N-1} + \sum_{i=1}^{N-1} \sum_{j=i+1}^N X_i X_j \right]
\end{aligned}$$

. Vimos que se extraímos uma amostra de tamanho n , com reposição, e calcularmos a média amostral \bar{X} , então

$$E(\bar{X}) = \mu = \frac{\sum_{i=1}^N X_i}{N} \quad e \quad Var(\bar{X}) = \frac{\sigma^2}{n},$$

onde μ e σ^2 são a média e a variância da população, respectivamente. No entanto, se a amostragem for feita sem reposição, então continua a valer, mas

$$E(\bar{X}) = \mu \quad e \quad Var(\bar{X}) = \frac{\sigma^2}{n} \frac{N-n}{N-1}.$$

(10.9) 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.

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

Prova:

Para cada elemento vamos definir uma variável indicadora da inclusão desse elemento na amostra.

$$I_i = \begin{cases} 1 & \text{se o } i\text{-ésimo elemento pertence à amostra} \\ 0 & \text{caso contrário...} \end{cases}$$

A lei de

$$I_i \sim B(p = \frac{n}{N}),$$

pois

$$p = P(I_i = 1) = \frac{\binom{N-1}{n-1}}{\binom{N}{n}} = \frac{(N-1)!}{(n-1)!(N-n)!} \times \frac{n!(N-n)!}{N!} = \frac{n}{N}.$$

$$E(I_i) = p = \frac{n}{N} \quad e \quad V(I_i) = pq = \frac{n}{N} \frac{N-n}{N}.$$

Seja

$$S_n = \sum_{i=1}^N X_i E(I_i),$$

Logo

$$E(S_n) = \sum_{i=1}^N X_i \frac{n}{N} = n \frac{\sum_{i=1}^N X_i}{N} = n \mu.$$

Como

$$\bar{X} = \frac{S_n}{n}$$

temos que:

$$E(\bar{X}) = \frac{E(S_n)}{n} = \frac{n\mu}{n} = \mu.$$

Além disso

$$Var(\bar{X}) = Var\left(\frac{S_n}{n}\right) = \frac{1}{n^2} Var(S_n).$$

Sabemos que

$$\begin{aligned} Var(S_n) &= Var\left(\sum_{i=1}^N X_i I_i\right) = \sum_{i=1}^N V(X_i I_i) + 2 \sum_{i=1}^{N-1} \sum_{j=i+1}^N Cov(X_i I_i, X_j I_j), \\ Var(S_n) &= \sum_{i=1}^N X_i^2 V(I_i) + 2 \sum_{i=1}^{N-1} \sum_{j=i+1}^N X_i X_j Cov(I_i, I_j), \end{aligned}$$

Seja

$$\begin{aligned} A &= \sum_{i=1}^N X_i^2 V(I_i) = \sum_{i=1}^N X_i^2 \frac{n}{N} \frac{N-n}{N} = \frac{n}{N} \frac{N-n}{N} \sum_{i=1}^N X_i^2. \\ A &= \frac{n(N-n)}{N^2} \sum_{i=1}^N X_i^2. \end{aligned}$$

Sabemos que:

$$Cov(I_i, I_j) = E(I_i I_j) - E(I_i) E(I_j)$$

A lei de

$$I_i I_j \sim B(p = \frac{n(n-1)}{N(N-1)}),$$

pois

$$p = P(I_i I_j = 1) = \frac{\binom{N-2}{n-2}}{\binom{N}{n}} = \frac{(N-2)!}{(n-2)!(N-n)!} \times \frac{n!(N-n)!}{N!} = \frac{n(n-1)}{N(N-1)}.$$

$$E(I_i I_j) = p = \frac{n(n-1)}{N(n-1)}.$$

Assim,

$$Cov(I_i, I_j) = \frac{n(n-1)}{N(N-1)} - \frac{n^2}{N^2} = \frac{n}{N} \left(\frac{n-1}{N-1} - \frac{n}{N} \right)$$

$$Cov(I_i, I_j) = \frac{n}{N} \frac{(n-1)N - n(N-1)}{N(N-1)}$$

$$Cov(I_i, I_j) = \frac{n}{N} \frac{(nN - N - nN + n)}{N(N-1)} = -\frac{N-n}{N^2(N-1)}$$

Logo

$$\begin{aligned} B &= 2 \sum_{i=1}^{N-1} \sum_{j=i+1}^N X_i X_j Cov(I_i, I_j) = -2 \frac{N-n}{N^2(N-1)} \sum_{i=1}^{N-1} \sum_{j=i+1}^N X_i X_j \\ Var(S_n) &= A + B = \frac{n(N-n)}{N^2} \sum_{i=1}^N X_i^2 - \frac{N-n}{N^2(N-1)} 2 \sum_{i=1}^{N-1} \sum_{j=i+1}^N X_i X_j \\ Var(S_n) &= \frac{n(N-n)}{N^2} \left[\sum_{i=1}^N X_i^2 - \frac{2}{N-1} \sum_{i=1}^{N-1} \sum_{j=i+1}^N X_i X_j \right] \end{aligned}$$

Finalmente

$$Var(S_n) = \frac{n(N-n)}{N^2} \frac{N^2}{N-1} \sigma^2 = n \sigma^2 \frac{N-n}{N-1}.$$

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

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

Considere, agora, uma população

$$\mathbb{P} = \{1, 3, 5, 5, 7\}$$

, logo $N = 5$. Retire amostras de tamanho $n = 2$, sem reposição, e construa a distribuição amostral de

$$\bar{X} = \frac{X_1 + X_2}{2}$$

Obtenha $E(\bar{X})$ e $Var(\bar{X})$ e verifique que esta é dada pela fórmula acima.

solução: Já vimos que

$$\mu = E(X) = 4,2 \quad V(X) = \sigma^2 = 4,16.$$

A fração amostral par $n = 2$ e $N = 5$ vale

$$\frac{N-n}{N-1} = \frac{3}{4}.$$

$$Var(\bar{X}) = \frac{\sigma^2}{n} \frac{N-n}{N-1} = \frac{4,16}{2} \times \frac{3}{4} = 1,56.$$

As possíveis amostras sem reposição de tamanho $n = 2$:

(X_1, X_2)	$P(X_1 = x_1, X_2 = x_2)$	Média Amostral
(1,3)	0,1	2
(1,5)	0,2	3
(1,7)	0,1	4
(3,5)	0,2	4
(3,7)	0,1	5
(5,5)	0,1	5
(5,7)	0,2	6

A distribuição amostral de \bar{X} é dada por:

$$P(\bar{X} = x) = 0,1 I_{\{2\}}(x) + 0,2 I_{\{3,5,6\}}(x) + 0,3 I_{\{4\}}(x).$$

Assim,

$$E(\bar{X}) = 0,1 \times 2 + 0,2 \times (3 + 5 + 6) + 0,3 \times 4$$

$$E(\bar{X}) = 0,2 + 2,8 + 1,2 = 4,2 = \mu.$$

$$E(\bar{X}^2) = 0,1 \times 4 + 0,2 \times (9 + 25 + 36) + 0,3 \times 16$$

$$E(\bar{X}^2) = 0,4 + 14 + 4,8 = 19,2.$$

$$Var(\bar{X}) = 19,2 - 4,2^2 = 19,2 - 17,64 = 1,56.$$

2. 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 10.7, cada amostra com reposição tem probabilidade $1/25$ de ser escolhida.

Vamos usar o R para gerar uma amostra aleatória simples com e sem reposição de tamanho 3 de

$$\mathbb{P} = \{1, 3, 5, 5, 7\}.$$

```

> P=c(1,3,5,5,7)
> A1=sample(P,3,replace=T);A1
[1] 5 5 7
> A2=sample(P,3,replace=T);A2
[1] 5 1 7
> A1=sample(P,3,replace=T);A1
[1] 5 7 5
> A2=sample(P,3,replace=T);A2
[1] 3 5 3
> A3=sample(P,3,replace=T);A3
[1] 5 5 3
> A1=sample(P,3);A1
[1] 5 7 3
> A2=sample(P,3);A2
[1] 5 7 5
> A3=sample(P,3);A3
[1] 7 3 1

```

Usando esta população com elementos repetidos cada elemento tem probabilidade 0, 2.
Vamos usar que:

$$P(X = 1) = P(X = 3) = P(X = 7) = 0,2 \quad e \quad P(X = 5) = 0,4.$$

Agora a nossa população fica:

$$\mathbb{P} = \{1, 3, 5, 7\}.$$

Vamos usar o R :

```

>
> P=c(1,3,5,7)
>
>
> ###Com reposição:
>
>
> A1=sample(P,3,prob=c(1,1,2,1)/5,replace=T);A1
[1] 7 5 1
>
>
> A2=sample(P,3,prob=c(1,1,2,1)/5,replace=T);A2
[1] 5 7 5
>
>
> A3=sample(P,3,prob=c(1,1,2,1)/5,replace=T);A3
[1] 7 7 3
>
>
> ###Sem reposição:
>
> A1=sample(P,3,prob=c(1,1,2,1)/5);A1

```

```
[1] 1 5 7
>
> A2=sample(P,3,prob=c(1,1,2,1)/5);A2
[1] 5 7 3
>
> A3=sample(P,3,prob=c(1,1,2,1)/5);A3
[1] 7 3 5
>
```

Vamos agora simular o lançamento de uma honesta 100 vezes.

A nossa população

$$\mathbb{P} = \{0, 1\},$$

em que $X = 0$ significa que o lançamento resultou em coroa e $X = 1$ cara.

```
> n=100
> A=sample(P,n,replace=T);A
[1] 1 1 0 1 0 1 1 0 1 0 0 0 0 0 1 1 1 0 0 0 1 0 0 0 1 1 1 0 0 1 0 1 1 1 0 1 1
[38] 1 1 1 1 0 1 1 0 0 1 1 0 1 0 0 1 1 1 1 0 1 0 0 0 0 1 0 0 1 1 1 0 1 0 1 1 0 1 0
[75] 0 0 1 1 0 0 0 0 1 1 1 0 1 0 0 1 1 1 0 0 1 1 0 0 0
> table(A)
A
0 1
47 53
> table(A)/n
A
0 1
0.47 0.53
```

Na realidade sabemos que:

$$X \sim Bernoulli(1/2) = Bin(1, 1/2)$$

Assim a simulação fica mais fácil.

Veja como funciona:

```
> set.seed(32)
> A= rbinom(100,1,1/2);A
[1] 1 1 1 1 0 1 1 1 0 1 0 1 1 0 1 1 1 1 0 0 1 1 0 1 1 1 0 1 1 1 0 1 1 1 0 1 1
[38] 0 1 1 0 1 0 1 1 0 1 1 0 1 1 1 0 0 0 1 1 1 1 1 0 1 0 0 1 1 0 1 0 1 0 1 0 0 0
[75] 0 0 1 0 0 0 1 1 1 0 0 1 0 1 0 0 0 0 0 0 0 0 0 0 0 0
> table(A)
A
0 1
45 55
> table(A)/100
A
```

O aluno usará com semente aleatória o seu número de matrícula.

(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,

$$\mathbb{P} = \{1, 2, \dots, 10\}$$

e suponha que a matriz de dados seja

A matriz

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

Podemos considerar três estratos, determinados pela variável X_2 :

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

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

Seja $N = 10$ o tamanho da nossa população e N_h o tamanho populacional do estrato h , $h = 1, 2, 3$. Sejam ainda $E(X) = \mu$, a média populacional, e σ^2 , a variância populacional.

Sejam μ_h , a média populacional do h -ésimo estrato e σ_h^2 , a variância populacional do h -ésimo estrato.

A média populacional é dada por:

$$\mu = \frac{\sum_{i=1}^N X_i}{N} = \frac{10 + 8 + 15 + 6 + 22 + 12 + 7 + 16 + 13 + 11}{10}.$$

$$\mu = \frac{120}{10} = 12.$$

Note que

$$\mu = \frac{\sum_{i=1}^N X_i}{N} = \frac{\sum_{i \in A} X_i + \sum_{i \in B} X_i + \sum_{i \in C} X_i}{N},$$

$$\mu = \frac{N_1 \mu_1 + N_2 \mu_2 + N_3 \mu_3}{N} =$$

$$\mu = \sum_{h=1}^3 \frac{N_h}{N} \mu_h = \sum_{h=1}^3 W_h \mu_h.$$

Mas

$$W_1 = \frac{N_1}{N} = \frac{3}{10} = 0,3; W_2 = \frac{N_2}{N} = \frac{4}{10} = 0,4; W_3 = \frac{N_3}{N} = \frac{3}{10} = 0,3.$$

Note que:

$$W_1 + W_2 + W_3 = 1.$$

Vamos calcular as médias dos estratos:

$$\mu_1 = \frac{15 + 16 + 22}{3} = \frac{53}{3}.$$

$$\mu_2 = \frac{10 + 11 + 12 + 23}{3} = \frac{46}{4} = \frac{23}{2} = 11,5.$$

$$\mu_3 = \frac{6 + 7 + 8}{3} = \frac{21}{3} = 7.$$

$$\mu = \sum_{h=1}^3 W_h \mu_h. = 0,3 \times \frac{53}{3} + 0,4 \times 11,5 + 0,3 \times 7$$

$$\mu = 5,3 + 4,6 + 2,1 = 12.$$

A variância populacional dos salários é calculada por:

$$\sigma^2 = \frac{\sum_{i=1}^N (X_i - \mu)^2}{N} = \frac{\sum_{i=1}^N (X_i - 12)^2}{N} = \frac{208}{10} = 20,8.$$

Vamos explicar a variabilidade do numerador através da variabilidade dentro e entre estratos.

Assim,

$$\sum_{i=1}^N (X_i - \mu)^2 = \sum_{i \in A} (X_i - \mu)^2 + \sum_{i \in B} (X_i - \mu)^2 + \sum_{i \in C} (X_i - \mu)^2$$

Mas

$$\begin{aligned} \sum_{i \in A} (X_i - \mu)^2 &= \sum_{i \in A} (X_i - \mu_1 + \mu_1 - \mu)^2 \\ \sum_{i \in A} (X_i - \mu)^2 &= \sum_{i \in A} (X_i - \mu_1)^2 + 2(\mu_1 - \mu) \sum_{i \in A} (X_i - \mu_1) + \sum_{i \in A} (\mu_1 - \mu)^2 \\ \sum_{i \in A} (X_i - \mu)^2 &= \sum_{i \in A} (X_i - \mu_1)^2 + N_1(\mu_1 - \mu)^2. \\ \sum_{i \in A} (X_i - \mu)^2 &= N_1 \sigma_1^2 + N_1(\mu_1 - \mu)^2. \end{aligned}$$

De forma análoga temos:

$$\sum_{i \in B} (X_i - \mu)^2 = N_2 \sigma_2^2 + N_2(\mu_2 - \mu)^2.$$

$$\sum_{i \in C} (X_i - \mu)^2 = N_3 \sigma_3^2 + N_3(\mu_3 - \mu)^2.$$

Juntando temos:

$$\sum_{i=1}^N (X_i - \mu)^2 = \sum_{h=1}^3 N_h \sigma_h^2 + \sum_{h=1}^3 N_h (\mu_h - \mu)^2 = VD + VE.$$

Dividindo por N temos:

$$\begin{aligned}\sigma^2 &= \frac{\sum_{i=1}^N (X_i - \mu)^2}{N} = \frac{\sum_{h=1}^3 N_h \sigma_h^2}{N} + \frac{\sum_{h=1}^3 N_h (\mu_h - \mu)^2}{N} \\ \sigma^2 &= \sum_{h=1}^3 \frac{N_h}{N} \sigma_h^2 + \sum_{h=1}^3 \frac{N_h}{N} (\mu_h - \mu)^2, \\ \sigma^2 &= \sum_{h=1}^3 W_h \sigma_h^2 + \sum_{h=1}^3 W_h (\mu_h - \mu)^2.\end{aligned}$$

Vamos calcular as variâncias populacionais dos 3 estratos:

$$\begin{aligned}\sigma_1^2 &= \frac{\sum_{i \in A} (X_i - \mu_1)^2}{N_1} = \frac{\sum_{i \in A} X_i^2 - N_1 \mu_1^2}{N_1} \\ \sigma_1^2 &= \frac{\sum_{i \in A} X_i^2}{N_1} - \mu_1^2.\end{aligned}$$

$$\begin{aligned}\sigma_1^2 &= \frac{225 + 256 + 484}{3} - \frac{53^2}{9} = \frac{865}{3} - \frac{2809}{9}. \\ \sigma_1^2 &= \frac{965 \times 3 - 2809}{9} = \frac{2895 - 2809}{9} = \frac{86}{9}.\end{aligned}$$

$$\begin{aligned}\sigma_2^2 &= \frac{\sum_{i \in B} X_i^2}{N_2} - \mu_2^2. \\ \sigma_2^2 &= \frac{100 + 121 + 144 + 169}{4} - \frac{529}{4} = \frac{534}{4} - \frac{529}{4}.\end{aligned}$$

$$\sigma_2^2 = \frac{5}{4} = 1,25.$$

$$\begin{aligned}\sigma_3^2 &= \frac{\sum_{i \in C} X_i^2}{N_1} - \mu_3^2. \\ \sigma_3^2 &= \frac{36 + 49 + 64}{3} - 49 = \frac{149}{3} - 49 = \frac{2}{3}.\end{aligned}$$

A parte da variância explicada dentro os estratos é dada por:

$$\begin{aligned}\sigma_D^2 &= \sum_{h=1}^3 W_h \sigma_h^2 = 0,3 \times \frac{86}{9} + 0,4 \times \frac{5}{4} + 0,3 \times \frac{2}{3}. \\ \sigma_D^2 &= \frac{8,6}{3} + 0,5 + 0,2 = \frac{8,6}{3} + 0,7 = \frac{8,6 + 2,1}{3} = \frac{10,7}{3}\end{aligned}$$

A parte da variância explicada entre os estratos é dada por:

$$\begin{aligned}\sigma_E^2 &= \sum_{h=1}^3 W_h (\mu_h - \mu)^2. \\ \sigma_E^2 &= 0,3 \times \left(\frac{53}{3} - 12\right)^2 + 0,4 \times (11,5 - 12)^2 + 0,3 \times (7 - 12)^2 \\ \sigma_E^2 &= 0,3 \times \frac{289}{9} + 0,4 \times \frac{1}{4} + 0,3 \times 25 \\ \sigma_E^2 &= \frac{28,9}{3} + 0,1 + 7,5 = \frac{28,9}{3} + 7,6 = \frac{51,7}{3} \\ \sigma_D^2 + \sigma_E^2 &= \frac{10,7}{3} + \frac{51,7}{3} = \frac{62,4}{3} = 20,8 = \sigma^2.\end{aligned}$$

Vamos estimar a média populacional μ da nossa população estratificada. Isto é feito colhendo uma amostra aleatória simples de tamanho n_h dentro do estrato h .

Assim temos

$$n_1 + n_2 + n_3 = n$$

Vamos calcular a média e a variância amostral para cada estrato obtendo:

$$\bar{X}_h \quad e \quad S_h^2$$

com \bar{X}_h um estimador não viciado de μ_h .

O nosso estimador proposto é

$$T = \sum_{h=1}^3 W_h \bar{X}_h.$$

Vamos calcular a esperança de T :

$$E(T) = \sum_{h=1}^3 W_h E(\bar{X}_h) = \sum_{h=1}^3 W_h \mu_h = \mu.$$

A variância do nosso estimador é dada por:

$$V(T) = \sum_{h=1}^3 W_h^2 V(\bar{X}_h) = \sum_{h=1}^3 W_h^2 \frac{\sigma_h^2}{n_h},$$

que depende de σ_h^2 que é desconhecida.

Assim vamos usar:

$$\hat{V}(T) = \sum_{h=1}^3 W_h^2 \frac{S_h^2}{n_h}.$$

(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, por meio de AAS, selecionamos algumas subpopulações. Num segundo estágio, usando novamente AAS, retiramos amostras das subpopulações selecionadas no 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.

3. Distribuição do máximo de uma amostra. Considere M o máximo de uma AAS (X_1, X_2, \dots, 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

$$X_i \leq m, \text{ para todo } 1 \leq i \leq n.$$

Como as v.a. X_i são independentes, teremos

$$F_M(m) = P(M \leq m) = P(X_1 \leq m, X_2 \leq m, \dots, X_n \leq m),$$

Como as variáveis são independentes temos:

$$F_M(m) = P(X_1 \leq m)P(X_2 \leq m) \dots P(X_n \leq m) = [F(m)]^n.$$

Portanto, a densidade de M é dada por

$$f_M(m) = (F_M(m))'(m) = n[F(m)]^{n-1}f(m). \quad (10.10)$$

4. 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, \dots, X_n) de X . Logo, as v.a. X_i são independentes, cada uma delas tem a mesma distribuição que X e $E(X_i) = \mu, \forall i = 1, 2, \dots, n$. Para estimar μ consideraremos a média amostral \bar{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

$$P\left(\left|\bar{X} - \mu\right| > \frac{c\sigma}{\sqrt{n}}\right) = P(|Z| > c) = 2[1 - \Phi(c)] \quad (10.11)$$

para qualquer constante $c > 0$, em que $Z \sim N(0, 1)$ e $\Phi(z)$ 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 X , que é dado por $\frac{\sigma}{\sqrt{n}}$

Por (10.11), podemos escrever, por exemplo,

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

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

O seguinte algoritmo pode, então, ser adotado:

- (1) Escolha um valor aceitável d para $\frac{\sigma}{\sqrt{n}}$.
- (2) Gere pelo menos 30 dados (para obter uma estimativa razoável de σ).
- (3) Continue a gerar dados, parando quando, com n dados, $\frac{S}{\sqrt{n}} < d$ com

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

- (4) Estime μ por

$$\bar{X} = \frac{\sum X_i}{n}.$$

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

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

$$\bar{X}_0 = 0 \quad , \quad S_1^2 = 0,$$

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

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

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 , $i = 1, 2, 3, 4, 5$.