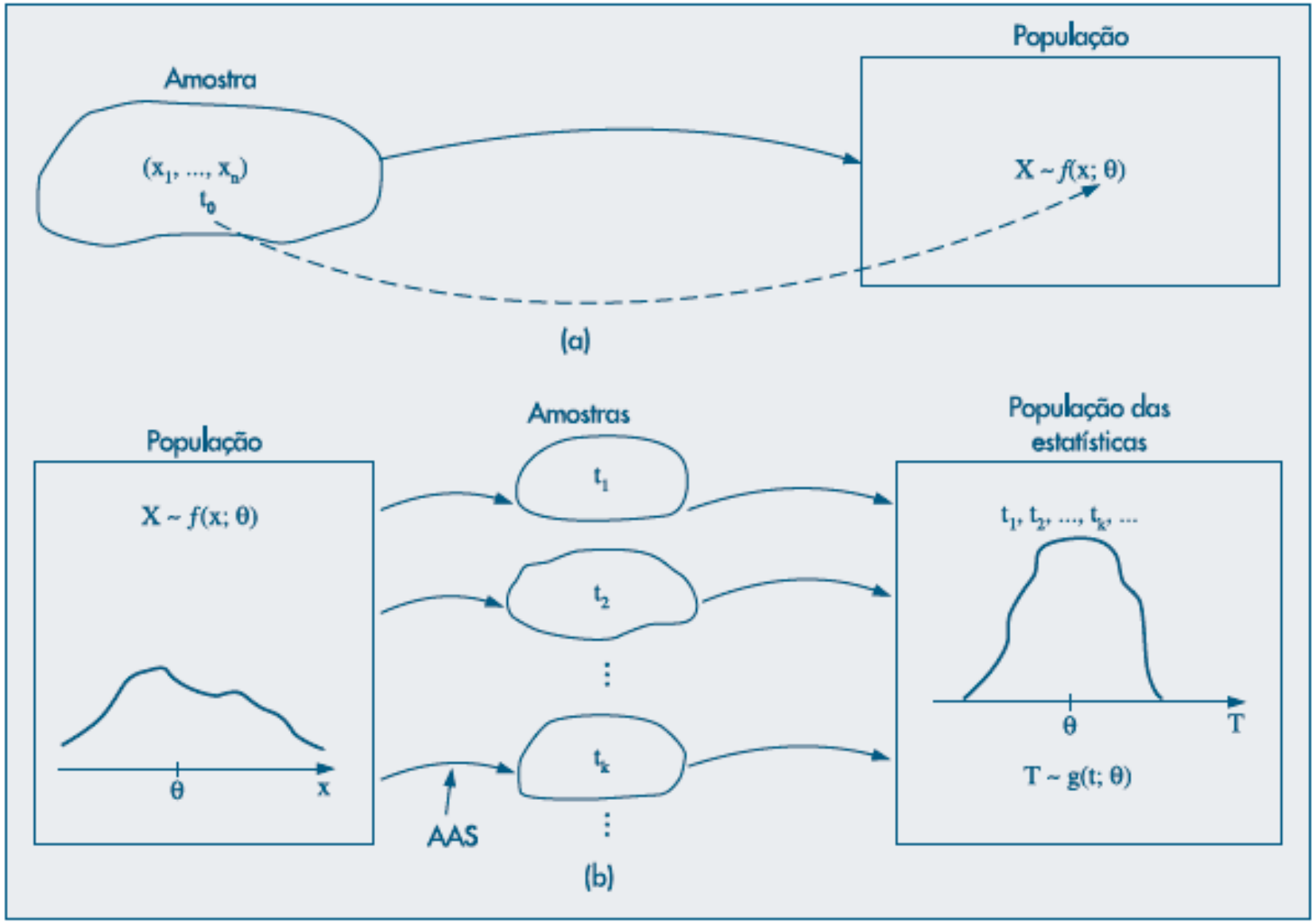
# Distribuições Amostrais

**O problema da inferência estatística: fazer uma afirmação sobre os parâmetros da população θ (média, variância, etc) através da amostra.**

**Usaremos uma AAS de n elementos sorteados dessa população, nossa decisão será baseada na estatística $T = f(X_1, ..., X_n)$.**

**Colhida essa amostra, observamos um particular valor de T: $t_0$, e baseados nesse valor é que faremos a afirmação sobre o parâmetro populacional θ.**

**Figura 10.1:** (a) Esquema de inferência sobre $\theta$.
(b) Distribuição amostral da estatística $T$.



- **Vamos observar *n* elementos, extraídos ao acaso e com reposição da população;**
- **Para cada elemento selecionado, observamos o valor da variável X de interesse.**

**Obtemos, então, uma amostra aleatória de tamanho *n* de X, que representamos por**

$X_1, X_2, \ldots, X_n$.

# Distribuição amostral da média

**Exemplo 1: Considere uma população em que uma variável X assume um dos valores do conjunto {1, 3, 5, 5, 7}. A distribuição de probabilidade de X é dada por**

| x | 1 | 3 | 5 | 7 |
|---|---|---|---|---|
| P(X=x) | 1/5 | 1/5 | 2/5 | 1/5 |

**É fácil ver que** $\mu_x = E(X) = 4{,}2$ ,

$\sigma_x^2 = Var(X) = 4{,}16.$

**Vamos relacionar <u>todas</u> as amostras possíveis de tamanho *n* = 2, selecionadas ao acaso e com reposição dessa população, e encontrar a distribuição da média amostral**

$$\overline{X} = \frac{X_1 + X_2}{2},$$

**sendo**

**$X_1$: valor selecionado na primeira extração; e**

**$X_2$: valor selecionado na segunda extração.**

| Amostra ($X_1$,$X_2$) | Probabilidade | Média Amostral |
|---|---|---|
| (1,1) | 1/25 | 1 |
| (1,3) | 1/25 | 2 |
| (1,5) | 2/25 | 3 |
| (1,7) | 1/25 | 4 |
| (3,1) | 1/25 | 2 |
| (3,3) | 1/25 | 3 |
| (3,5) | 2/25 | 4 |
| (3,7) | 1/25 | 5 |
| (5,1) | 2/25 | 3 |
| (5,3) | 2/25 | 4 |
| (5,5) | 4/25 | 5 |
| (5,7) | 2/25 | 6 |
| (7,1) | 1/25 | 4 |
| (7,3) | 1/25 | 5 |
| (7,5) | 2/25 | 6 |
| (7,7) | 1/25 | 7 |
| | 1 | |

**A distribuição de probabilidade de $\bar{X}$ para $n = 2$ é**

| $\bar{x}$ | 1 | 2 | 3 | 4 | 5 | 6 | 7 |
|---|---|---|---|---|---|---|---|
| $P(\bar{X} = \bar{x})$ | 1/25 | 2/25 | 5/25 | 6/25 | 6/25 | 4/25 | 1/25 |

**Neste caso,** $E(\bar{X}) = 4{,}2 = \mu_x$ e

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

**Repetindo o mesmo procedimento, para amostras de tamanho *n* = 3, temos a seguinte distribuição de probabilidade de $\overline{X}$,**

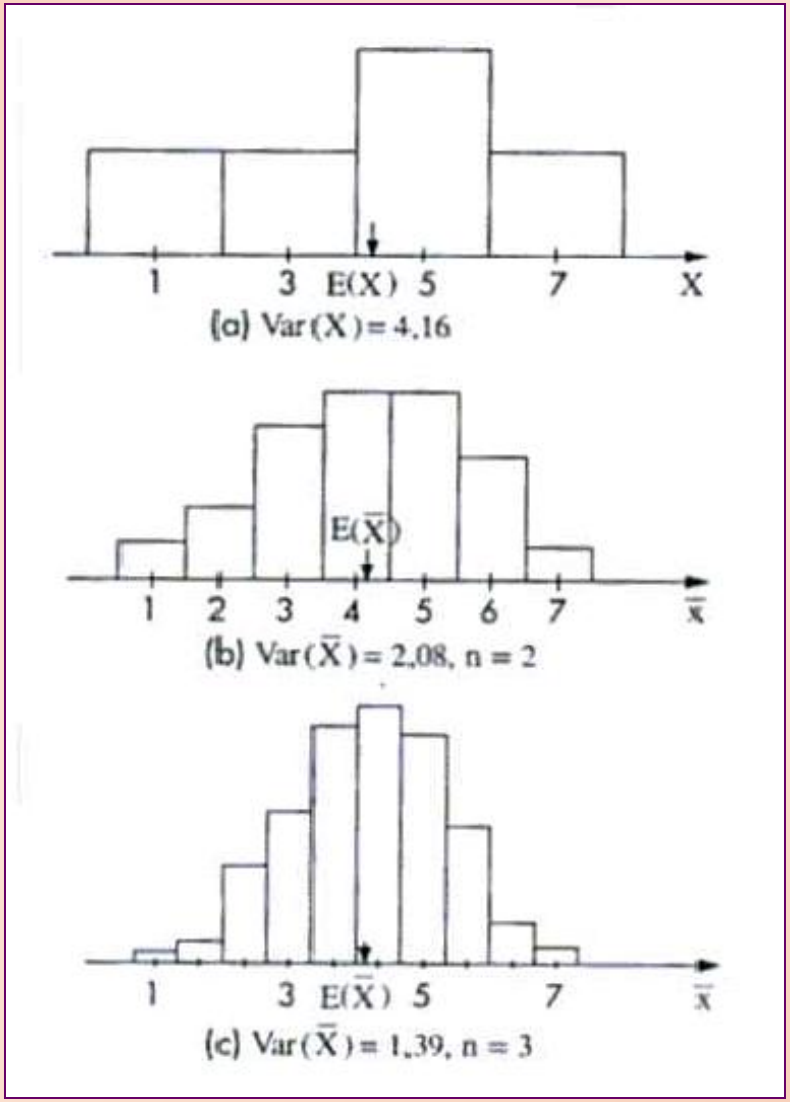
| $\overline{x}$ | $P(\overline{X} = \overline{x})$ |
|---|---|
| 1 | 1/125 |
| 5/3 | 3/125 |
| 7/3 | 9/125 |
| 3 | 16/125 |
| 11/3 | 24/125 |
| 13/3 | 27/125 |
| 5 | 23/125 |
| 17/3 | 15/125 |
| 19/3 | 6/125 |
| 7 | 1/125 |

**Neste caso,**

$$E(\overline{X}) = 4{,}2 = \mu_x \quad \text{e}$$

$$Var(\overline{X}) = 1{,}39 = \frac{\sigma_x^2}{3}.$$

**Figura 1: Histogramas correspondentes às distribuições de X e de $\overline{X}$, para amostras de {1,3,5,5,7}.**



**Dos histogramas, observamos que**

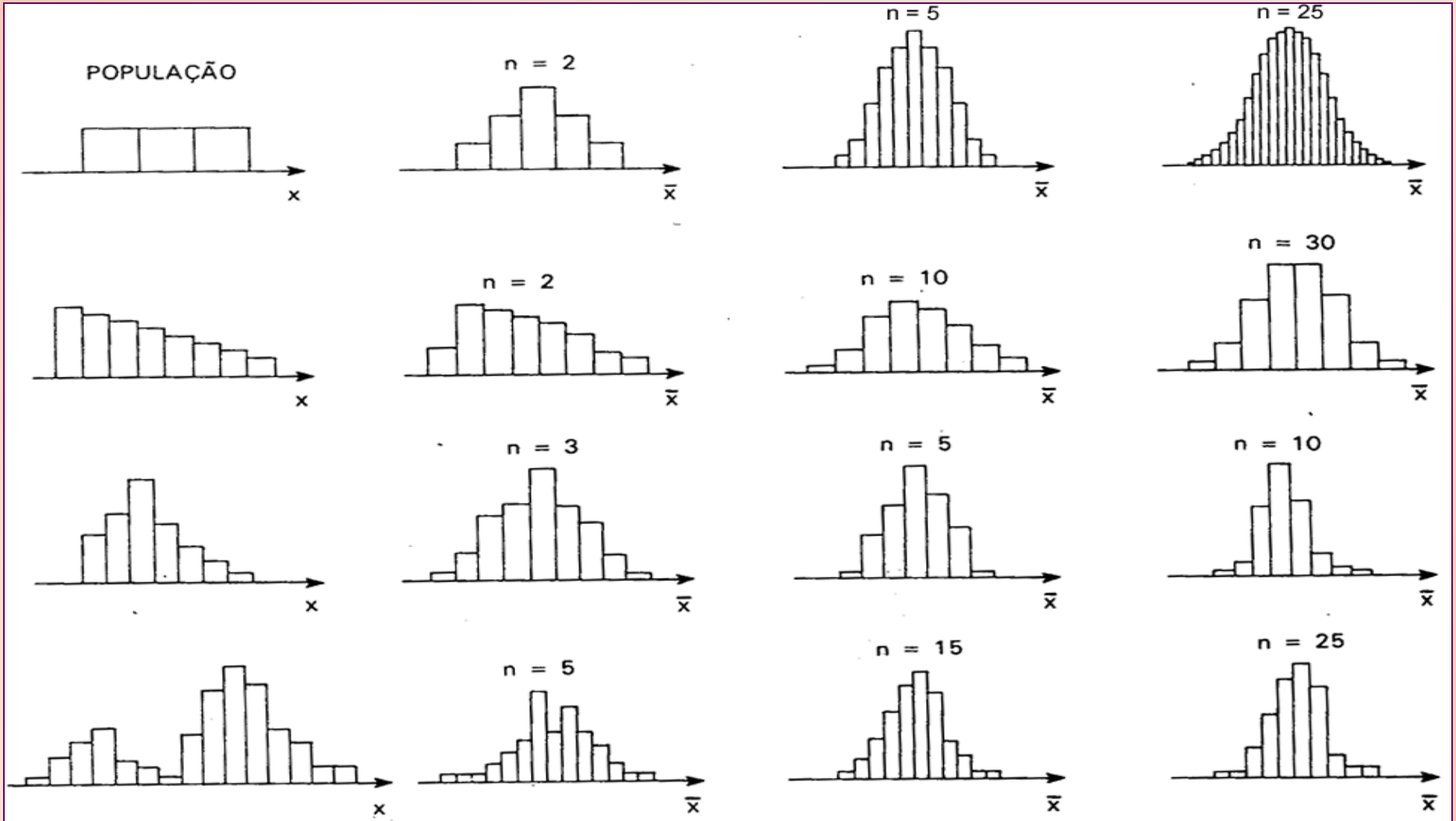
- **conforme *n* aumenta, os valores de $\bar{x}$ tendem a se concentrar cada vez mais em torno de**

$$E(\bar{x}) = 4{,}2 = \mu_x \text{ ,}$$

**uma vez que a variância vai diminuindo;**

- **os casos extremos passam a ter pequena probabilidade de ocorrência;**
- **para *n* suficientemente grande, a forma do histograma *aproxima-se de uma distribuição normal.***

# Figura 2: Histogramas correspondentes às *distribuições de* $\overline{X}$ para amostras de algumas populações.




Fonte: BUSSAB & MORETTIN. ***Estatística Básica***. São Paulo, Atual, 4ª edição, 2002, pp. 273.

**Esses gráficos sugerem que,**

**quando n aumenta, *independentemente da forma da distribuição* de X , a distribuição de probabilidade da média amostral $\bar{X}$ *aproxima-se* de uma distribuição normal.**

# Teorema do Limite Central

**Seja X uma v. a. que tem média $\mu$ e variância $\sigma^2$. Para uma amostra $x_1, x_2, ..., x_n$ , retirada ao acaso e com reposição de X, a distribuição de probabilidade da média amostral $\bar{X}$ *aproxima-se*, *para n grande*, de uma distribuição normal, com média $\mu$ e variância $\sigma^2 / n$ , ou seja,**

$$\bar{X} \sim N\left(\mu, \frac{\sigma^2}{n}\right), \text{para n grande, aproximadamente.}$$

## Comentários:

- **Se a distribuição de X é normal, então $\bar{X}$ tem distribuição normal <u>exata, para todo *n*.</u>**

- **O desvio padrão $\sqrt{\frac{\sigma^2}{n}} = \frac{\sigma}{\sqrt{n}}$ , que é**

**o desvio padrão da média amostral, também é denominado <u>erro padrão.</u>**

## Exemplo 2:

Uma maquina enchia pacotes de café cujos pesos seguiam uma distribuição N(500, 100). Colhendo-se uma amostra de n=100 pacotes e pesando-os. Se a máquina estiver regulada, qual a probabilidade de encontrarmos a média de 100 pacotes diferindo de 500g de menos de 2 gramas?

**Exemplo 3:**

Suponha que p = 30% dos estudantes de uma escola sejam mulheres. Colhemos uma AAS de n=10 estudantes e calculamos proporção de mulheres na amostra. Qual a probabilidade de que a proporção amostral difira de p em menos de 0,01?

# Distribuição amostral de uma Proporção

X = 1, se o indivíduo for portador da característica;
= 0, caso contrário;

**É fácil ver que** $\mu_x = E(X) = p$ ,

$\sigma_x^2 = Var(X) = p(1-p)$

Retirada de uma AAS(amostragem aleatória simples) dessa população, e indicando por $Y_n$ o total de indivíduos portadores da característica na amostra, então

$Y_n \sim b(n, p)$

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

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

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

# Outras Distribuições Amostrais

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

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

a mediana amostral, *md*, e o estimador

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

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

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

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

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

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

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

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

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

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

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

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

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

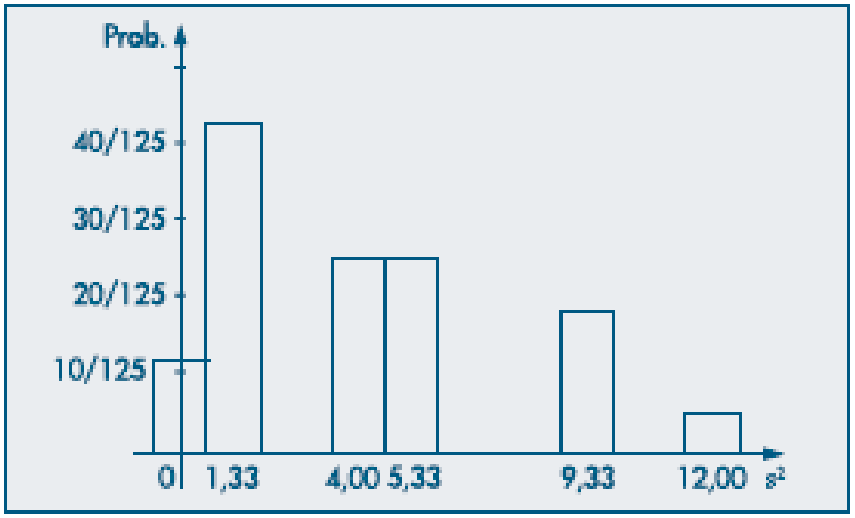



Figura 10.7: Distribuição amostral de md para amostras de tamanho $n = 3$ de $\{1, 3, 5, 5, 7\}$.

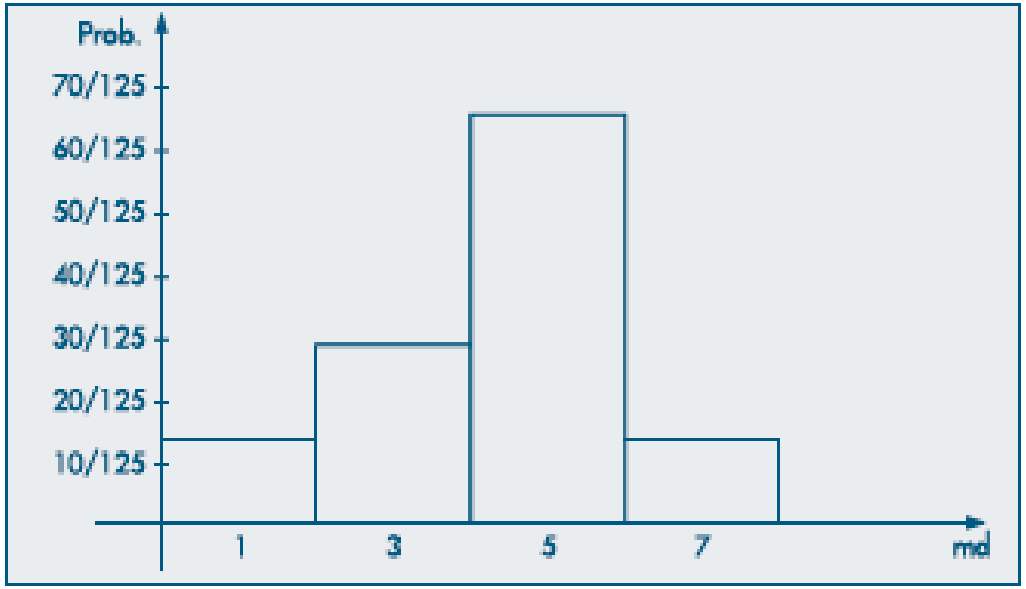
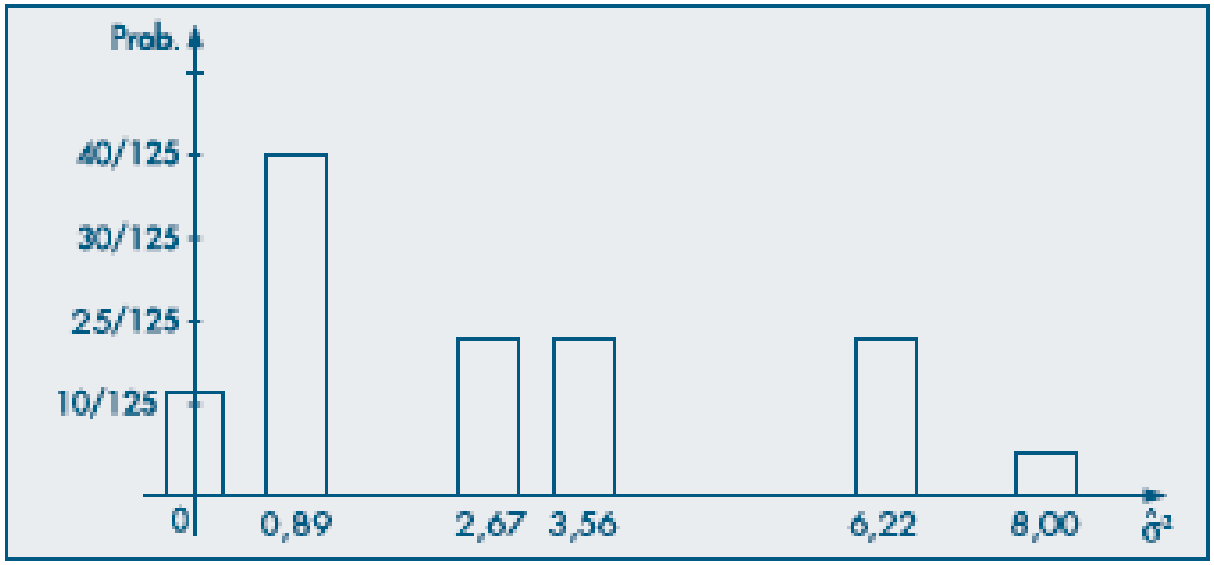


Figura 10.8: Distribuição amostral de $\hat{\sigma}^2$ para amostras de tamanho $n = 3$ extraídas de $\{1, 3, 5, 5, 7\}$.




Por exemplo, note que $E(S^2) = 4{,}16 = \sigma^2$, logo $S^2$ satisfaz uma propriedade análoga a $E(\overline{X}) = \mu$; dizemos que $\overline{X}$ e $S^2$ são estimadores *não-viesados* dos respectivos parâmetros $\mu$ e $\sigma^2$. 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.

# Dimensionamento da amostra

**Seja $P(\varepsilon) = \gamma$, a probabilidade da média amostral $\bar{X}$ estar a uma distância de, no máximo $\varepsilon$, da média populacional $\mu$ (desconhecida), ou seja,**

$$\gamma = P\left(\left|\bar{X} - \mu\right| \leq \varepsilon\right) = P\left(\mu - \varepsilon \leq \bar{X} \leq \mu + \varepsilon\right)$$

$$= P\left(\frac{-\varepsilon}{\sigma/\sqrt{n}} \leq \frac{\bar{X} - \mu}{\sigma/\sqrt{n}} \leq \frac{\varepsilon}{\sigma/\sqrt{n}}\right) \cong P\left(\frac{-\varepsilon\sqrt{n}}{\sigma} \leq Z \leq \frac{\varepsilon\sqrt{n}}{\sigma}\right),$$

**sendo $Z \sim N(0,1)$ .**

# Dimensionamento da amostra

A partir da relação $\varepsilon = z\frac{\sigma}{\sqrt{n}}$,

**o tamanho da amostra *n* é determinado por**

$$n = \left(\frac{z}{\varepsilon}\right)^2 \sigma^2,$$

**conhecendo-se o desvio padrão $\sigma$ de X, o erro $\varepsilon$ da estimativa e o coeficiente de confiança $\gamma$ do intervalo, sendo z tal que**

$\gamma = P(-z \leq Z \leq z)$ **e** $Z \sim N(0,1)$.

## Exemplo 4:

**A renda per-capita domiciliar numa certa região tem distribuição normal com desvio padrão $\sigma$ = 250 reais e média $\mu$ desconhecida. Se desejamos estimar a renda média $\mu$ com erro $\varepsilon$ = 50 reais e com uma confiança $\gamma$ = 95%, quantos domicílios devemos consultar?**

**X : renda per-capita domiciliar na região**

**$X \sim N(\mu ; 250^2)$**

**n = ?? tal que $\varepsilon$ = 50 reais,**

$$\gamma = 0{,}95 \Rightarrow z = 1{,}96$$

**Então,**

$$n = \left(\frac{z}{\varepsilon}\right)^2 \sigma^2$$

$$= \left(\frac{1{,}96}{50}\right)^2 (250)^2$$

$$= 96{,}04$$

**Aproximadamente 97 domicílios devem ser consultados.**

## Exemplo 5:

**A quantidade de colesterol X no sangue das alunas de uma universidade segue uma distribuição de probabilidades com desvio padrão $\sigma$ = 50 mg/dl e média $\mu$ desconhecida. Se desejamos estimar a quantidade média $\mu$ de colesterol com erro $\varepsilon$ = 20 mg/dl e confiança de 90%, quantas alunas devem realizar o exame de sangue?**

**X: quantidade de colesterol no sangue das alunas da universidade**

**$\sigma$ = 50 mg/dl**

**n = ?? tal que $\varepsilon$ = 20 mg/dl**

**$\gamma$ = 0,90 $\Rightarrow$ z = 1,65**

**Supondo que o tamanho da amostra a ser selecionada é suficientemente grande, pelo Teorema do Limite Central temos:**

$$n = \left(\frac{z}{\varepsilon}\right)^2 \sigma^2$$

$$= \left(\frac{1{,}65}{20}\right)^2 (50)^2$$

$$= 17{,}02$$

**Assim, aproximadamente 18 alunas devem realizar o exame de sangue.**

**No caso de proporção, usando a aproximação normal para proporção amostral, temos**

$$n = \left(\frac{z}{\varepsilon}\right)^2 p(1-p)$$

**Como não conhecemos p, podemos usar o fato de que p(1-p)<=1/4, para todos p.**

# Dimensionamento da amostra

Da relação $\varepsilon = z\sqrt{\dfrac{p(1-p)}{n}}$ ,

segue que o **tamanho amostral *n***, dados $\gamma$ e a margem de erro $\varepsilon$, tem a forma

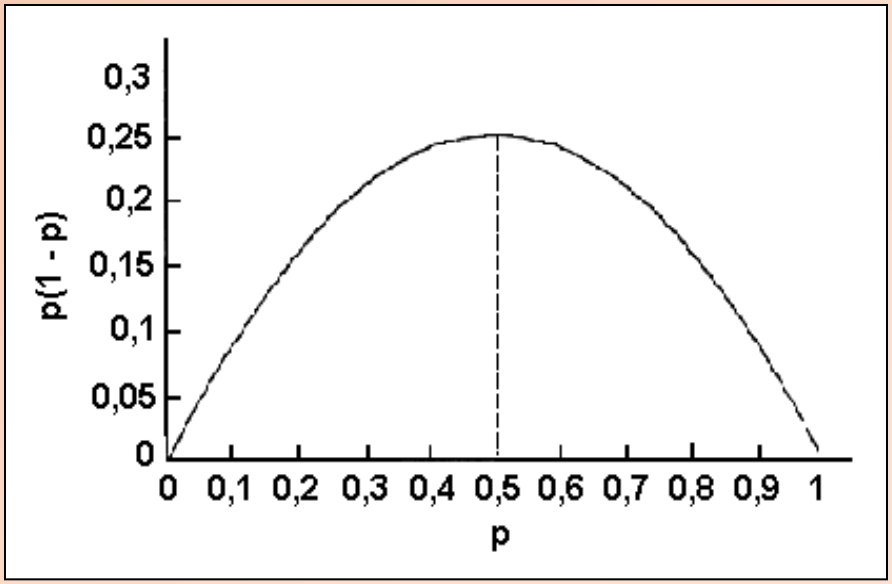
$$n = \left(\frac{z}{\varepsilon}\right)^2 p(1-p),$$

onde *z* é tal que $\gamma = P(-z \leq Z \leq z)$ e $Z \sim N(0,1)$.

Entretanto, nesta expressão, *n* depende de *p(1-p)*, que é desconhecido.

- ***Como calcular o valor de n?***

## Gráfico da função $p(1-p)$, para $0 \leq p \leq 1$.




**Pela figura observamos que:**

- **a função *p(1-p)* é uma parábola simétrica em torno de *p* = 0,5;**
- **o máximo de *p(1-p)* é 0,25, alcançado quando *p* = 0,5.**

**Assim, na prática, substituímos *p(1-p)* por seu valor máximo, obtendo**

$$n = \left(\frac{z}{\varepsilon}\right)^2 0{,}25\ ,$$

**que pode fornecer um valor de *n* maior do que o necessário.**

## Exemplo 6:

**No exemplo da USP (Exemplo 1) suponha que nenhuma amostra foi coletada. Quantos estudantes precisamos consultar de modo que a estimativa pontual esteja, no máximo, a 0,02 da proporção verdadeira *p*, com uma probabilidade de 0,95?**

**Dados do problema:**

**$\varepsilon = 0{,}02$ (erro da estimativa);**

**$P(\varepsilon) = \gamma = 0{,}95 \Rightarrow z = 1{,}96.$**

$$n = \left(\frac{1{,}96}{0{,}02}\right)^2 p(1-p) \leq \left(\frac{1{,}96}{0{,}02}\right)^2 0{,}25 = 2401 \text{ estudantes.}$$

**Pergunta:** ***É possível reduzir o tamanho da amostra quando temos alguma informação a respeito de p?***

**Por exemplo, sabemos que:**

- ***p*** **não é superior a 0,30, ou**
- ***p*** **é pelo menos 0,80, ou**
- ***p*** **está entre 0,30 e 0,60.**

**Resposta:** ***Depende do tipo de informação sobre p.***

**Em alguns casos, podemos substituir a informação** ***p(1-p)*****, que aparece na expressão de** ***n*****, por um valor menor que 0,25.**

# Redução do tamanho da amostra

**Vimos que, se nada sabemos sobre o valor de *p*, no cálculo de *n*, substituímos *p(1-p)* por seu valor máximo, e calculamos**

$$n = \left(\frac{z}{\varepsilon}\right)^2 0{,}25\,.$$

**Se temos a informação de que *p é no máximo 0,30* (*p* ≤ 0,30), então o valor máximo de *p(1-p)* será dado por 0,3x0,7 = 0,21.**
**Logo, reduzimos o valor de *n* para**

$$n = \left(\frac{z}{\varepsilon}\right)^2 0{,}21.$$

**Agora, se *p é pelo menos 0,80* ($p \geq 0,80$), então o máximo de *p(1-p)* é 0,8 x 0,2 = 0,16 e temos**

$$n = \left(\frac{z}{\varepsilon}\right)^2 0,16 .$$

**Mas, se *0,30 ≤ p ≤ 0,60* , o máximo de *p(1-p)* é 0,5x0,5 = 0,25 e, neste caso, não há redução, ou seja,**

$$n = \left(\frac{z}{\varepsilon}\right)^2 0,25.$$

## Exemplo 7:

**No Exemplo 6, suponha que temos a informação de que no máximo 30% dos alunos da USP foram ao teatro no último mês. Portanto, temos que $p \leq 0{,}30$ e, como vimos, o máximo de $p(1-p)$ neste caso é 0,21.**

**Assim, precisamos amostrar**

$$n = \left(\frac{z}{\varepsilon}\right)^2 0{,}21 = \left(\frac{1{,}96}{0{,}02}\right)^2 0{,}21 = 2017 \text{ estudantes,}$$

**conseguindo uma redução de 2401 – 2017 = 384 estudantes.**