Estimadores de Momentos

A **média** populacional é um caso particular daquilo que chamamos de momento. Na realidade, ela é o ***primeiro momento.***

Se X *for uma v.a. contínua, com densidade* $f(x;\theta_1, ..., \theta_r)$*, dependendo de r parâmetros, então:*

$$\mu_1 = E(X) = \int_{-\infty}^{\infty} x f(x;\, \theta_1,\, \ldots,\, \theta_r)dx.$$

Por exemplo, suponha que X *tenha distribuição normal, com parâmetros* μ *e* σ^2.

Aqui, $\theta_1 = \mu$, $\theta_2 = \sigma^2$ e $r = 2$. *Temos, nesse caso, que* $E(X) = \mu$.

Podemos, em geral, definir o *k-ésimo momento* de X por

$$\mu_k = E(X^k) = \int_{-\infty}^{\infty} x^k f(x;\, \theta_1,\, \ldots,\, \theta_r)dx, \quad k = 1, 2, \ldots$$

Assim, para $k = 2$, obtemos o segundo momento

$$E(X^2) = \int_{-\infty}^{\infty} \mathbf{x}^2 f(x;\, \theta_1,\, \ldots,\, \theta_r)dx.$$

$$E(X^2) = \mathrm{Var}(X) + [E(X)]^2 = \sigma^2 + \mu^2.$$

agora, que colhemos uma amostra de tamanho n da população $(X_1, ..., X_n)$. Definimos o chamado *k-ésimo momento amostral* por

$$m_k = \frac{1}{n}\sum_{i=1}^{n} X_i^k, \quad k = 1, 2, ... \tag{11.23}$$

Temos, portanto, que $m_1 = \bar{X}$ e $m_2 = \sum_{i=1}^{m} X_i^2/n$.

Definição. Dizemos que $\hat{\theta}_1, ..., \hat{\theta}_r$ são estimadores obtidos pelo método dos momentos se eles forem soluções das equações

$$m_k = \mu_k, \quad k = 1, 2, ..., r. \tag{11.24}$$

Exemplo 11.8. Se X tem média μ e variância σ^2, teremos as seguintes relações válidas para os dois primeiros momentos populacionais:

$$E(X) = \mu, \quad E(X^2) = \sigma^2 + \mu^2,$$

do que obtemos

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

Temos, também, os dois primeiros momentos amostrais:

$$m_1 = \frac{1}{n}\sum_{i=1}^{n} X_i = \overline{X},$$

$$m_2 = \frac{1}{n}\sum_{i=1}^{n} X_i^2.$$

Os estimadores obtidos pelo método dos momentos serão

$$\hat{\mu}_M = m_1 = \overline{X},$$

$$\hat{\sigma}_M^2 = m_2 - m_1^2 = \frac{1}{n}\sum_{i=1}^{n} X_i^2 - \overline{X}^2 = \hat{\sigma}^2.$$

Estimadores de Máxima Verossimilhança

Exemplo. Suponha que temos *n provas de Bernoulli com P (sucesso) = p*, *0 < p < 1* e *X = número de sucessos.*

*Estimar **p**?*

Uma amostra, por exemplo, *n = 3 e obtemos dois sucessos e um fracasso.*

***A função* de verossimilhança** é
L(p) = P(2 sucessos e 1 fracasso) = p2(1 – p).

Maximizando essa função em relação a p, *obtemos*
$L'(p) = 2p(1 - p) - p2 = 0 \Rightarrow p(2 - 3p) = 0$,

$$p = 0 \text{ ou } p = 2/3.$$

o ponto máximo é $\hat{p} = 2/3$,

que é o estimador de **máxima verossimilhança (EMV) de *p*.**

X ~ binomial (n, p), a função de verossimilhança nesse caso é

$$L(p) = p^x (1 - p)^{n-x},$$

que é a probabilidade de se obter x sucessos e $n - x$ fracassos. O máximo dessa função ocorre no mesmo ponto que $\ell(p) = \log_e L(p)$. Denotando o logaritmo natural simplesmente por log, temos

$$\ell(p) = x \log p + (n - x) \log(1 - p).$$

Derivando e igualando a zero obtemos $\hat{p}_{MV} = x/n$.

No caso discreto, a função de verossimilhança pode ser escrita na forma

$$L(\theta; x_1, \ldots, x_n) = P(X_1 = x_1|\theta) \ldots P(X_n = x_n|\theta).$$

No caso de variáveis contínuas, a função de verossimilhança é definida da seguinte maneira. Suponha que a v.a. X tenha densidade $f(x; \theta)$, onde destacamos a dependência do parâmetro θ desconhecido. Retiramos uma amostra de X, de tamanho n, $(X_1, ..., X_n)$, e sejam $(x_1, ..., x_n)$ os valores efetivamente observados.

Definição. A função de verossimilhança é definida por

$$L(\theta; x_1, ..., x_n) = f(x_1; \theta) \dots f(x_n; \theta), \tag{11.30}$$

que deve ser encarada como uma função de θ. O estimador de máxima verossimilhança de θ é o valor $\hat{\theta}_{MV}$ que maximiza $L(\theta; x_1, ..., x_n)$.

Exemplo 11.11. Suponha que a v.a. X tenha distribuição exponencial, com parâmetro $\alpha > 0$, desconhecido, e queremos obter o EMV desse parâmetro. A densidade de X é dada por (7.26):

$$f(x; \alpha) = \begin{cases} \dfrac{1}{\alpha} e^{-x/\alpha}, & \text{se } x \geqslant 0 \\ 0, & \text{se } x < 0. \end{cases}$$

Então, a verossimilhança é dada por

$$L(\alpha|x) = (1/\alpha)^n e^{-\Sigma x_i/\alpha}$$

e a log-verossimilhança fica

$$\ell(\alpha|x) = -n \log \alpha - \sum_{i=1}^{n} x_i/\alpha.$$

Derivando e igualando a zero obtemos que o EMV de α é

$$\hat{\alpha}_{MV} = \frac{\sum_{i=1}^{n} x_i}{n}, \tag{11.31}$$

Intervalo de Confiança
Para média

Objetivo

Estimar a média μ de uma variável aleatória X, que representa uma característica de interesse de uma população, a partir de uma amostra.

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

Obtemos, então, uma <u>amostra aleatória de tamanho *n*</u> de X, que representamos por

$X_1, X_2, ..., X_n$.

Uma ***estimador pontual*** para μ é dado pela ***média amostral***,

$$\overline{X} = \frac{X_1 + X_2 + \ldots + X_n}{n} = \sum_{i=1}^{n} \frac{X_i}{n} .$$

Uma ***estimador intervalar*** ou *intervalo de confiança* para μ tem a forma

$$\left[\overline{X} - \varepsilon \; ; \; \overline{X} + \varepsilon\right],$$

sendo ε o **erro amostral** (margem de erro) calculado a partir da *distribuição de probabilidade de* $\overline{X}$.

Teorema do Limite Central

Seja X uma v. a. que tem média μ e variância σ^2. Para uma amostra $X_1, X_2, \ldots, X_n$, 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 μ e variância σ^2 / n , ou seja,

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

Intervalo de Confiança

Como vimos, o <u>estimador por intervalo para a média μ</u> tem a forma

$$\left[\bar{X} - \varepsilon \; ; \; \bar{X} + \varepsilon\right]$$

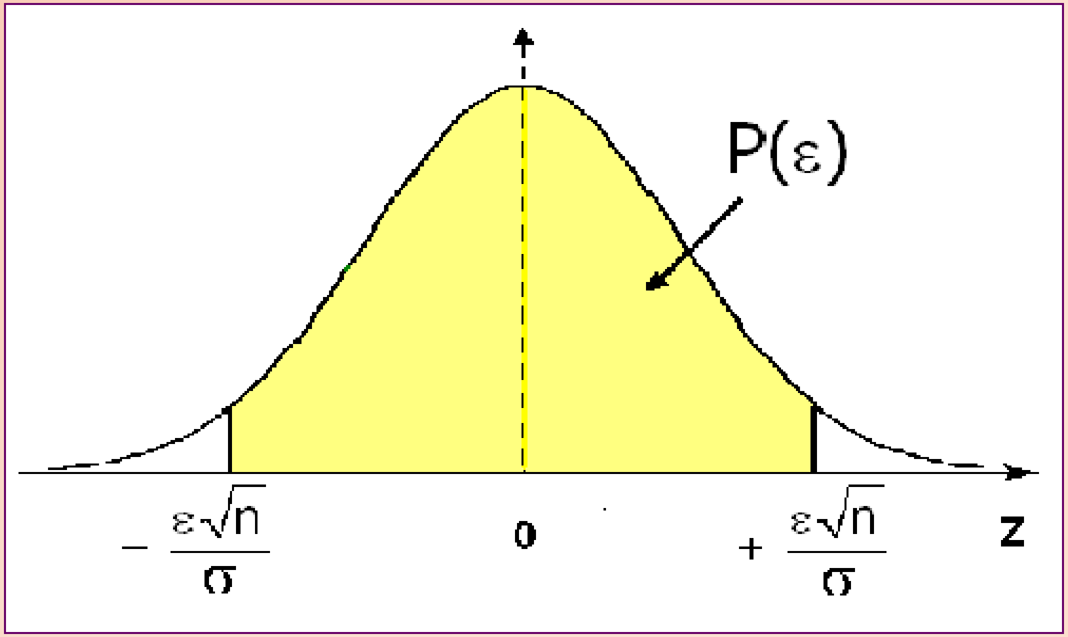
Pergunta: ***Como determinar*** ε **?**

Seja P(ε) = γ, a probabilidade da média amostral $\overline{X}$ estar a uma distância de, no máximo ε, da média populacional μ (desconhecida), ou seja,

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

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

sendo Z ~ N(0,1) .



Denotando $\frac{\varepsilon\sqrt{n}}{\sigma} = z$ **, temos que**

$$\gamma = P(-z \leq Z \leq z).$$

Assim, conhecendo-se o coeficiente de confiança γ obtemos z.

Erro na estimativa intervalar

Da igualdade $z = \dfrac{\varepsilon\sqrt{n}}{\sigma}$, segue que

o erro amostral ε é dado por

$$\varepsilon = z\frac{\sigma}{\sqrt{n}},$$

sendo z tal que $\gamma = P(-z \le Z \le z)$, com $Z \sim N(0,1)$.

O **intervalo de confiança para a média** μ, **com coeficiente de confiança** γ fica, então, dado por

$$\left[\overline{X} - z\frac{\sigma}{\sqrt{n}}\,;\ \overline{X} + z\frac{\sigma}{\sqrt{n}}\right],$$

sendo σ o desvio padrão de X.

Exemplo:

Não se conhece o consumo médio de combustível de automóveis da marca T. Sabe-se, no entanto, que o desvio padrão do consumo de combustível de automóveis dessa marca é 10 km/l. Na análise de 100 automóveis da marca T, obteve-se consumo médio de combustível de 8 km/l. Encontre um intervalo de confiança para o consumo médio de combustível dessa marca de carro. Adote um coeficiente de confiança igual a 95%.

X: consumo de combustível de automóveis da marca T

σ = 10 km/l

n = 100 $\Rightarrow$ $\bar{x}$ (média amostral) = 8 km/l

γ = 0,95 $\Rightarrow$ z = 1,96

Pelo Teorema do Limite Central, o intervalo de confiança de 95% é dado, aproximadamente, por

$$\left[\bar{X} - z\frac{\sigma}{\sqrt{n}}\ ;\ \bar{X} + z\frac{\sigma}{\sqrt{n}}\right] =$$

$$\left[8 - 1{,}96\frac{10}{\sqrt{100}}\ ;\ 8 + 1{,}96\frac{10}{\sqrt{100}}\right] =$$

$$\left[8 - 1{,}96\ ;\ 8 + 1{,}96\right]$$

$$\left[6{,}04\ ;\ 9{,}96\right]$$

Observe que o erro amostral ε é 1,96 km/l.

Exemplo:

Deseja-se estimar o tempo médio de estudo (em anos) da população adulta de um município. Sabe-se que o tempo de estudo tem distribuição normal com desvio padrão $\sigma = 2{,}6$ anos. Foram entrevistados $n = 25$ indivíduos, obtendo-se para essa amostra, um tempo médio de estudo igual a 10,5 anos. Obter um intervalo de 90% de confiança para o tempo médio de estudo populacional.

X : tempo de estudo, em anos e $X \sim N(\mu; 2{,}6^2)$

$n = 25 \quad \Rightarrow \quad \bar{x} = 10{,}5$ anos

$\gamma = 0{,}90 \quad \Rightarrow \quad z = 1{,}65$

A estimativa intervalar com 90% de confiança é dada por (em anos):

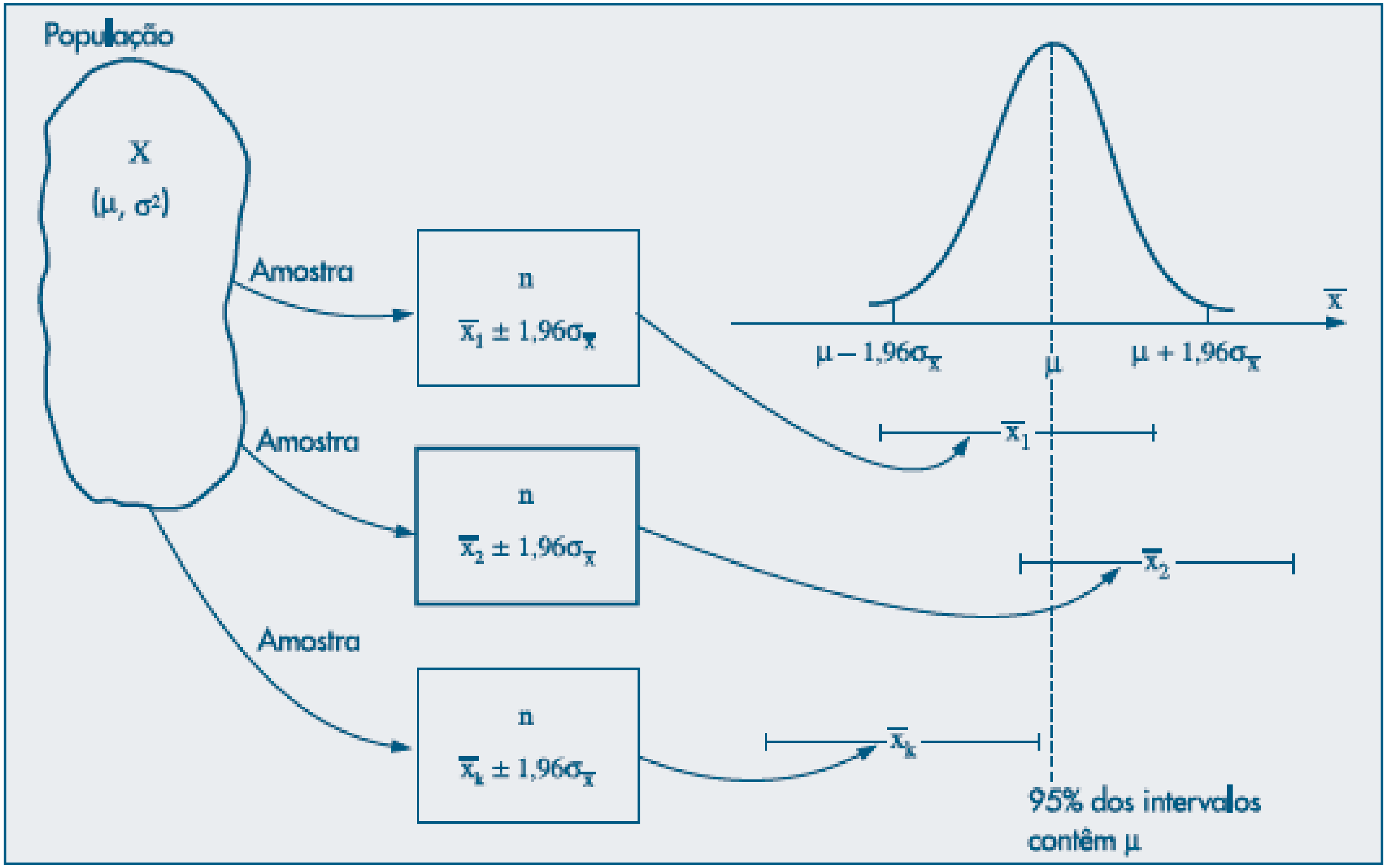
$$\left[\bar{x} - z\frac{\sigma}{\sqrt{n}}\ ;\ \bar{x} + z\frac{\sigma}{\sqrt{n}}\right] =$$

$$\left[10{,}5 - 1{,}65\frac{2{,}5}{\sqrt{25}}\ ;\ 10{,}5 + 1{,}65\frac{2{,}5}{\sqrt{25}}\right] =$$

$$\left[10{,}5 - 0{,}86\ ;\ 10{,}5 + 0{,}86\right]$$

$$\left[9{,}64\ ;\ 11{,}36\right].$$

Figura 11.3: Significado de um IC para μ, com $\gamma = 0{,}95$ e σ^2 conhecido.



Na prática, a variância populacional σ^2 é <u>desconhecida</u> e é substituída por sua estimativa,

$$s^2 = \frac{1}{n-1}\sum_{i=1}^{n}\left(x_i - \bar{x}\right)^2 .$$

A estimativa amostral do desvio padrão σ é $s = \sqrt{s^2}$.

Temos duas opções ao padronizar a variável $\overline{X}$

Se σ, o desvio padrão populacional, for conhecido, usamos

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

Se σ for desconhecido, usamos seu estimador, o desvio padrão amostral S, e consideramos a seguinte variável padronizada

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

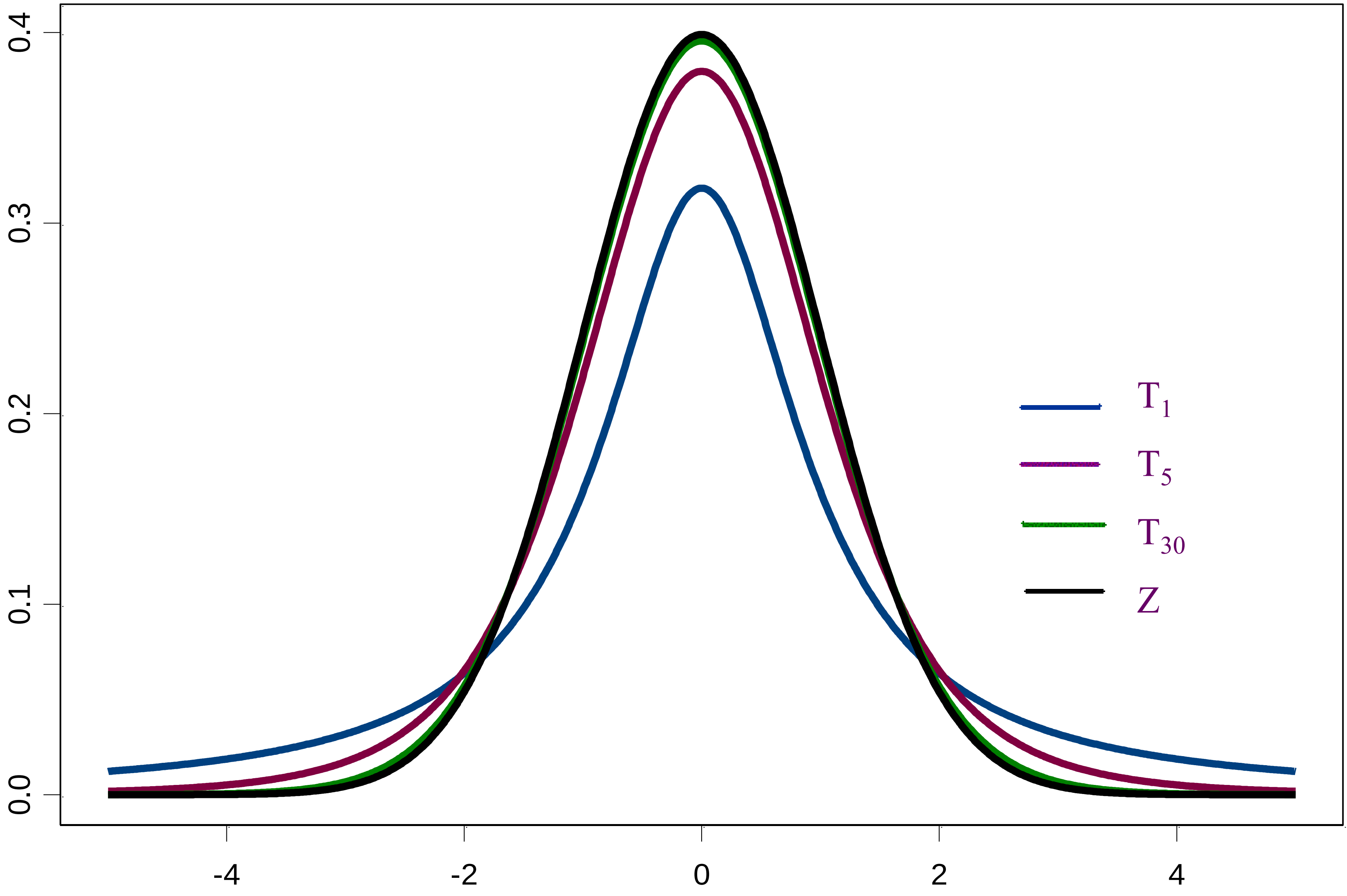
- **Se a variável na <u>população</u> tem distribuição <u>normal</u>, então**

Z **tem distribuição** $N(0,1)$

e T **tem distribuição** t **de Student com** $n\text{-}1$ **graus de liberdade.**

- **Se o tamanho** n **da amostra é grande, então**

Z **e** T **têm distribuição aproximadamente** $N(0,1)$.



0.0
0.1
0.2
0.3
0.4
-4
-2
0
2
4
T_1
T_5
T_{30}
Z

Assim, uma estimativa intervalar *aproximada* para a média populacional μ , quando o <u>tamanho da amostra é grande</u> e <u>σ é desconhecido</u> , é

$$\left[\bar{x} - z\frac{s}{\sqrt{n}} , \bar{x} + z\frac{s}{\sqrt{n}}\right],$$

sendo *s* o desvio padrão amostral e *z* tal que $\gamma = P(-z \leq Z \leq z)$ com $Z \sim N(0,1)$.

Exemplo:

Para estimar a renda semanal média de garçons de restaurantes em uma grande cidade, é colhida uma amostra da renda semanal de 75 garçons. A média e o desvio padrão amostrais encontrados são R$ 227 e R$ 15 respectivamente. Determine um intervalo de confiança, com coeficiente de confiança de 90%, para a renda média semanal.

X: renda semanal de garçons da cidade

$n = 75 \quad \Rightarrow \quad \bar{x} = 227 \;\text{ e }\; s = 15$

$\gamma = 0{,}9 \quad \Rightarrow \quad z = 1{,}65$

O intervalo de 90% de confiança é dado, aproximadamente, por (em reais).

$$\left[\bar{x} - z\frac{s}{\sqrt{n}}\ ,\ \bar{x} + z\frac{s}{\sqrt{n}}\right] =$$

$$\left[227 - 1{,}65\frac{15}{\sqrt{75}}\ ;\ 227 + 1{,}65\frac{15}{\sqrt{75}}\right] =$$

$$[227 - 2{,}86\ ;\ 227 + 2{,}86] =$$

$$[224{,}14\ ;\ 229{,}86]$$

Intervalo de Confiança
Para proporção

Objetivo

Estimar uma proporção *p* (desconhecida) de elementos em uma população, apresentando certa característica de interesse, a partir da informação fornecida por uma amostra.

Dois possíveis procedimentos de estimação:

- **Estimação pontual**
- **Estimação intervalar**

- Vamos observar *n* elementos, extraídos ao acaso e com reposição da população;

- Para cada elemento selecionado, verificamos a presença (sucesso) ou não (fracasso) da característica de interesse.

Estimador pontual

O ***estimador pontual para p***, também denominado ***proporção amostral***, é definido como

$$\hat{p} = \frac{X}{n},$$

sendo que

X denota o número de elementos na amostra que apresentam a característica;

n denota o tamanho da amostra coletada.

Se observamos o valor *k* da v. a. X, obtemos $\hat{p} = k / n$ que denominamos ***estimativa pontual para p.***

Exemplo: **Sejam,**

p**: proporção de alunos da USP que foram ao teatro pelo menos uma vez no último mês, e**

X: número de estudantes que respondem "sim" em uma pesquisa com *n* entrevistados.

Suponha que foram entrevistados $n = 500$ estudantes e que, desses, $k = 100$ teriam afirmado que foram ao teatro pelo menos uma vez no último mês.

A ***estimativa pontual*** (***proporção amostral***) ***para p*** é dada por:

$$\hat{p} = \frac{k}{n} = \frac{100}{500} = 0{,}20\,,$$

ou seja, 20% dos estudantes *entrevistados* afirmaram que foram ao teatro pelo menos uma vez no último mês.

Note que, outra amostra de mesmo tamanho pode levar a uma outra estimativa pontual para *p*.

Estimativa intervalar ou intervalo de confiança

•**Para uma amostra observada, os estimadores pontuais fornecem como estimativa um <u>único valor numérico</u> para o parâmetro.**

•**Os estimadores pontuais são <u>variáveis aleatórias</u> e, portanto, possuem uma distribuição de probabilidade, em geral, denominada *distribuição amostral*.**

Idéia: construir intervalos de confiança, que incorporem à estimativa pontual informações a respeito de sua variabilidade (erro amostral).

Intervalos de confiança são obtidos por meio da *distribuição amostral do estimador pontual*.

A ***estimativa intervalar*** corresponde a um intervalo determinado da seguinte maneira:

$$\left[\hat{p} - \varepsilon; \hat{p} + \varepsilon\right],$$

sendo ε o erro amostral ou margem de erro.

Pergunta: ***Como encontrar ε ?***

Seja P(ε) a probabilidade da estimativa pontual estar a uma distância de, no máximo, ε da proporção verdadeira *p*, ou seja,

$$P(\varepsilon) = P\left(\,|\hat{p} - p| \leq \varepsilon\,\right).$$

A probabilidade P(ε) é também denominada coeficiente de confiança do intervalo, que denotamos pela letra grega γ (gama).

Afirma-se ainda que a estimativa intervalar tem coeficiente de confiança γ = P(ε).

Formalmente,

$$P(\varepsilon) = P(|\hat{p} - p| \leq \varepsilon) = P\left(\left|\frac{X}{n} - p\right| \leq \varepsilon\right)$$

$$= P\left(p - \varepsilon \leq \frac{X}{n} \leq p + \varepsilon\right) = P(np - n\varepsilon \leq X \leq np + n\varepsilon)$$

$$= P\left\{-\frac{n\varepsilon}{\sqrt{np(1-p)}} \leq \frac{X - np}{\sqrt{np(1-p)}} \leq \frac{n\varepsilon}{\sqrt{np(1-p)}}\right\}.$$

Como $X \sim b(n,p)$ temos que, para n grande, a variável aleatória

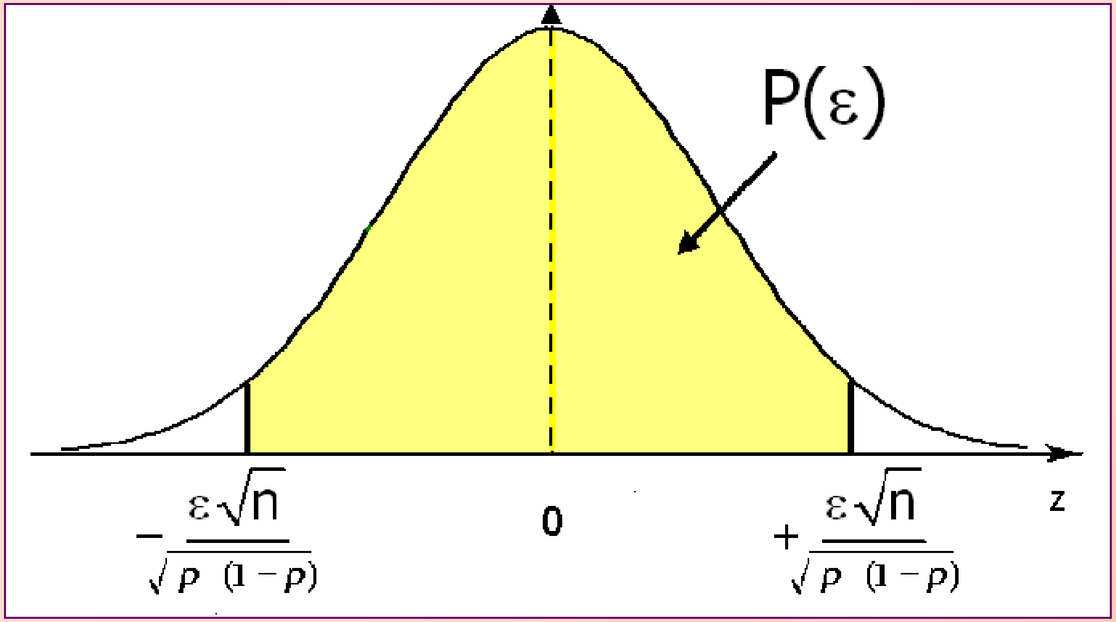
$$Z = \frac{X - np}{\sqrt{np(1-p)}}$$

tem distribuição N(0,1).

Deste modo, para *n* grande,

$$P(\varepsilon) \cong P\left\{-\frac{\varepsilon\sqrt{n}}{\sqrt{p(1-p)}} \le Z \le \frac{\varepsilon\sqrt{n}}{\sqrt{p(1-p)}}\right\},$$

onde Z ~ N(0,1).

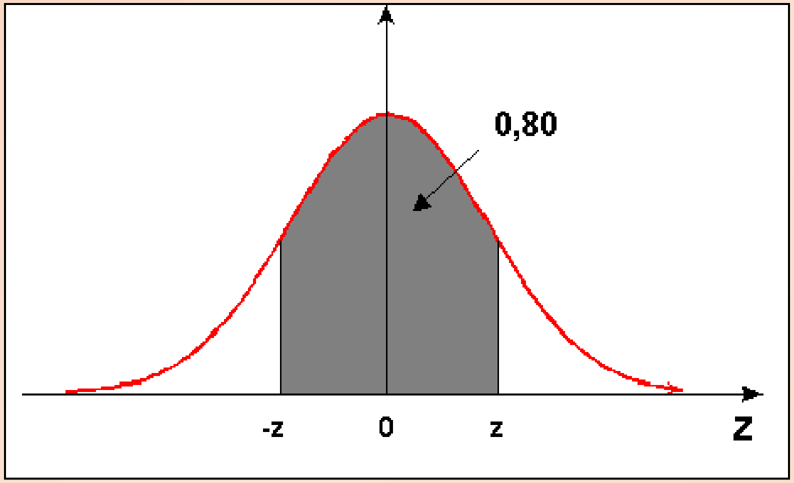


Denotando $\dfrac{\varepsilon\sqrt{n}}{\sqrt{p(1-p)}} = z$, temos que

$P(\varepsilon) = \gamma = P(-z \leq Z \leq z).$

Assim, podemos obter *z* conhecendo-se γ (ou $P(\varepsilon)$).

Por exemplo, considere $\gamma = 0{,}80$.



z é tal que A(*z*) = 0,90.

Pela tabela, temos *z* = 1,28.

Erro da estimativa intervalar

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

é imediato mostrar que o <u>erro amostral ε</u> é dado por

$$\varepsilon = z\sqrt{\frac{p(1-p)}{n}}\ ,$$

onde *z* é tal que γ = P(-z ≤ Z ≤ z), com Z~N(0,1).

Intervalo de confiança para *p*

Vimos que a estimativa intervalar para ***p*** tem a forma:

$$\left[\, \hat{p} - \varepsilon\, ;\; \hat{p} + \varepsilon \,\right],$$

com $\varepsilon = z\sqrt{\dfrac{p(1-p)}{n}}$ e *z* tal que γ = P(-z ≤ Z ≤ z) na N(0,1).

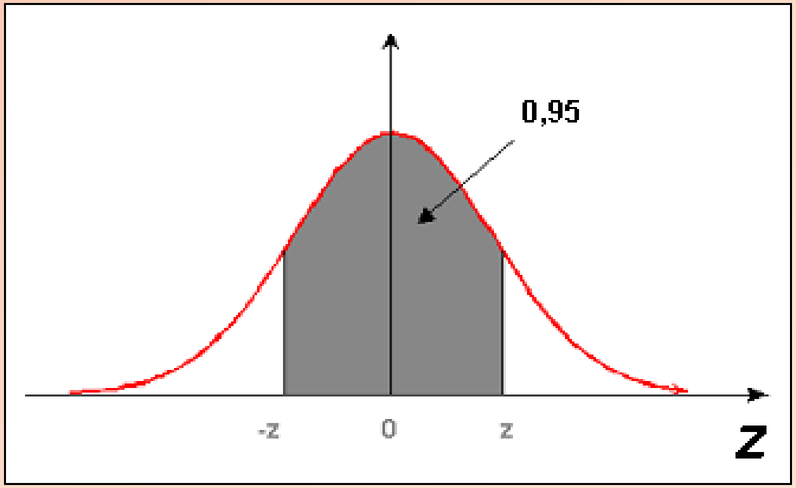
Na prática, substituímos a proporção desconhecida p pela proporção amostral $\hat{p}$, obtendo o seguinte ***intervalo de confiança com coeficiente de confiança*** γ*:*

$$IC(p\,;\gamma) = \left[\hat{p} - z\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}\;\;;\;\; \hat{p} + z\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}\right]$$

Exemplo:

No exemplo da USP, temos n = 500 e $\hat{p}$ = 0,20.

Construir um intervalo de confiança para *p* com coeficiente de confiança γ = 0,95.



Como γ = 0,95 fornece *z* = 1,96, o intervalo é dado por:

$$\left[\hat{p} - z\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}\ ;\ \hat{p} + z\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}\right]$$

$$= \left[0,20 - 1,96\sqrt{\frac{0,20 \times 0,80}{500}}\ ;\ 0,20 + 1,96\sqrt{\frac{0,20 \times 0,80}{500}}\right]$$

$$= [\,0,20 - 0,035\ ;\ 0,20 + 0,035\,] = [\,0,165\ ;\ 0,235\,].$$

Nesse intervalo (γ=0,95), a estimativa pontual para *p* é 0,20, com um erro amostral ε igual a 0,035.

Interpretação do IC com γ = 95%:

Se sortearmos 100 amostras de tamanho n=500 e construirmos os respectivos 100 intervalos de confiança, com coeficiente de confiança de 95%, esperamos que, aproximadamente, 95 destes intervalos contenham o verdadeiro valor de *p*.

Comentários:

Da expressão $\varepsilon = z\sqrt{\dfrac{p(1-p)}{n}}$ **, é possível concluir que:**

•para γ fixado, o erro diminui com o aumento de *n*.

•para *n* fixado, o erro aumenta com o aumento de γ.