ESTIMAÇÃO PARA A 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.

Exemplos:

μ: peso médio de homens na faixa etária de 20 a 30 anos, em uma certa localidade;

μ: idade média dos habitantes do sexo feminino na cidade de Santos, em 1990;

μ: salário médio dos empregados da indústria metalúrgica em São Bernardo do Campo, em 2001;

μ: taxa média de glicose em indivíduos do sexo feminino com idade superior a 60 anos, em determinada localidade;

μ: comprimento médio de jacarés adultos de uma certa raça.

• 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

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

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

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

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* $\overline{X}$.

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 $\bar{X}$,

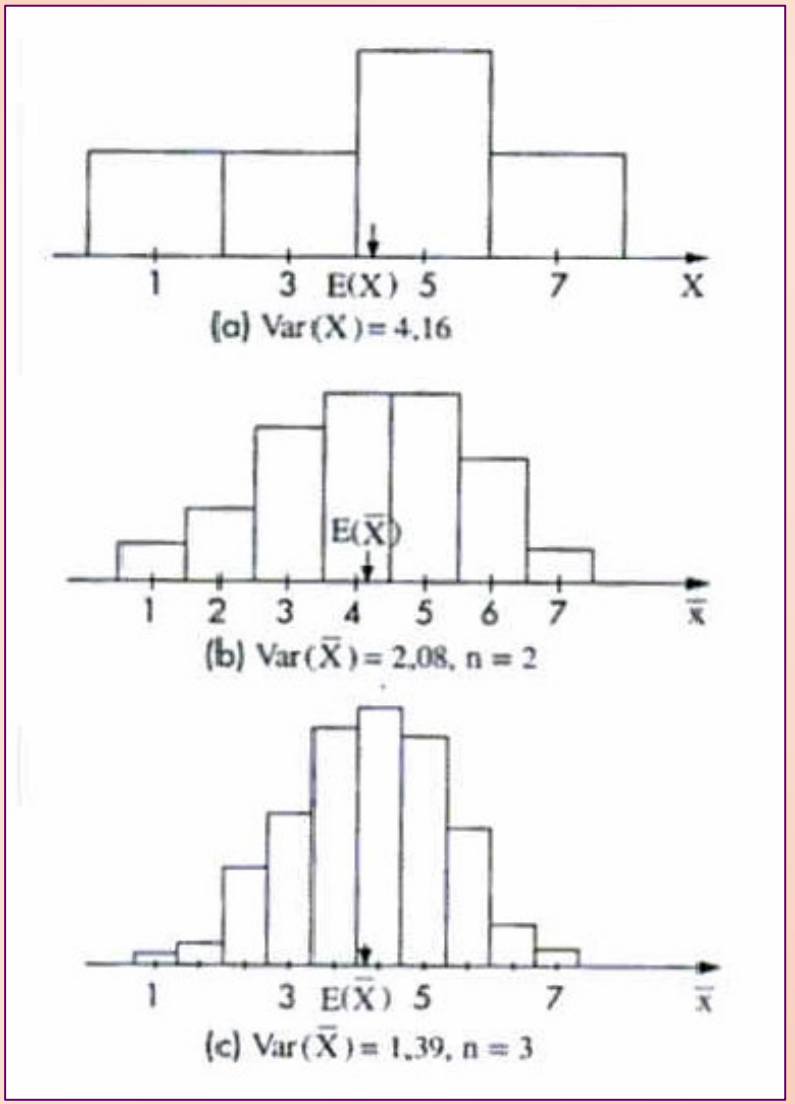
$\bar{x}$	$P(\bar{X}=\bar{x})$
1	1/125
5/3	3/125
7/3	9/125
3	16/125
11/3	24/125
13/3	27/125
5	23/125
17/3	15/125
19/3	6/125
7	1/125

Neste caso,

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

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

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



(a) Var(X) = 4.16

(b) Var($\overline{X}$) = 2.08, n = 2

(c) Var($\overline{X}$) = 1.39, n = 3

Dos histogramas, observamos que

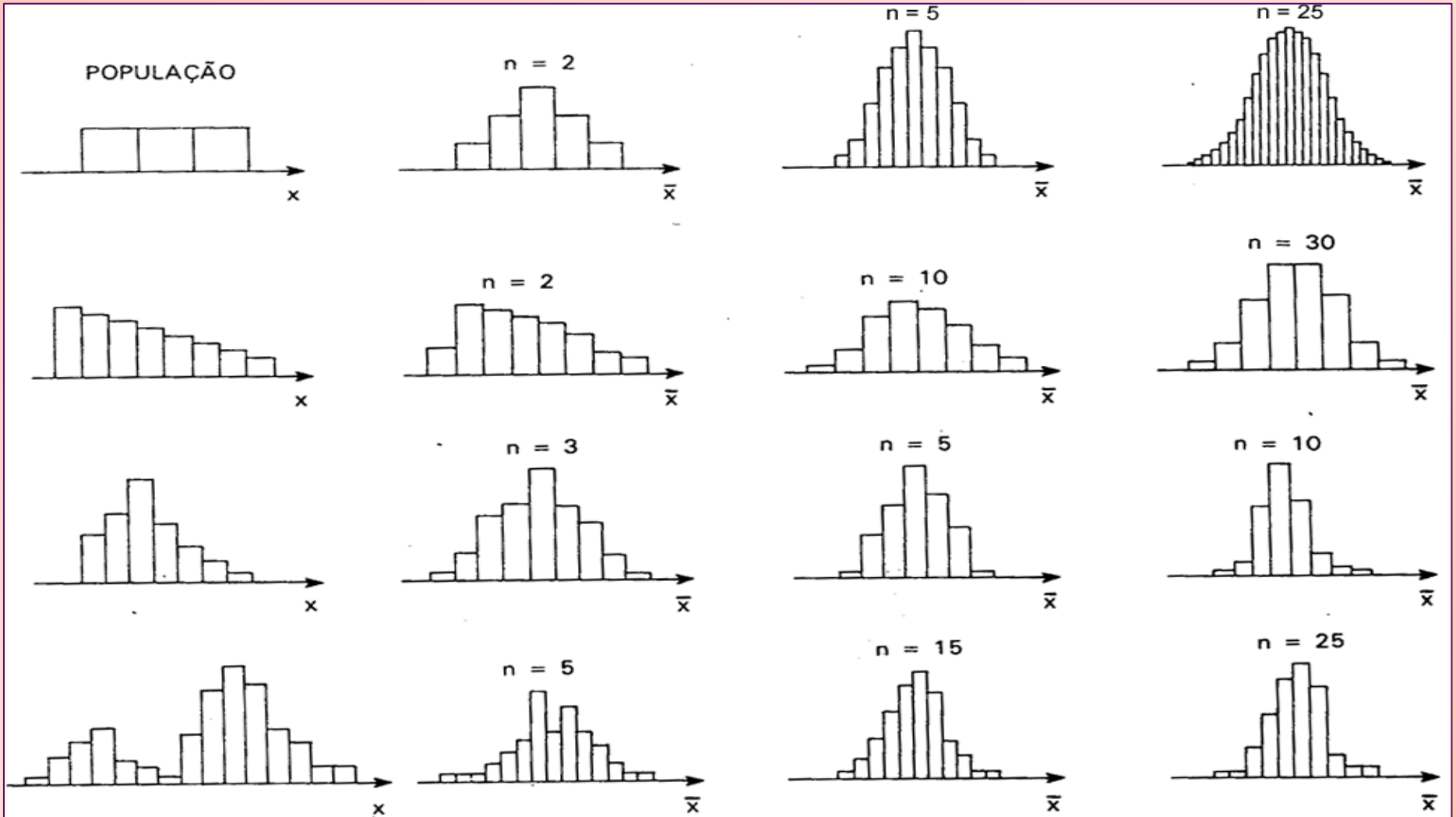
- **conforme *n* aumenta, os valores de $\bar{X}$ tendem a se concentrar cada vez mais em torno de**

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

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 μ e variância σ^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 $\overline{X}$ *aproxima-se, para n grande*, de uma distribuição normal, com média μ e variância σ^2 / n , ou seja,

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

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>

Intervalo de Confiança

Como vimos, o estimador por intervalo para a média μ tem a forma

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

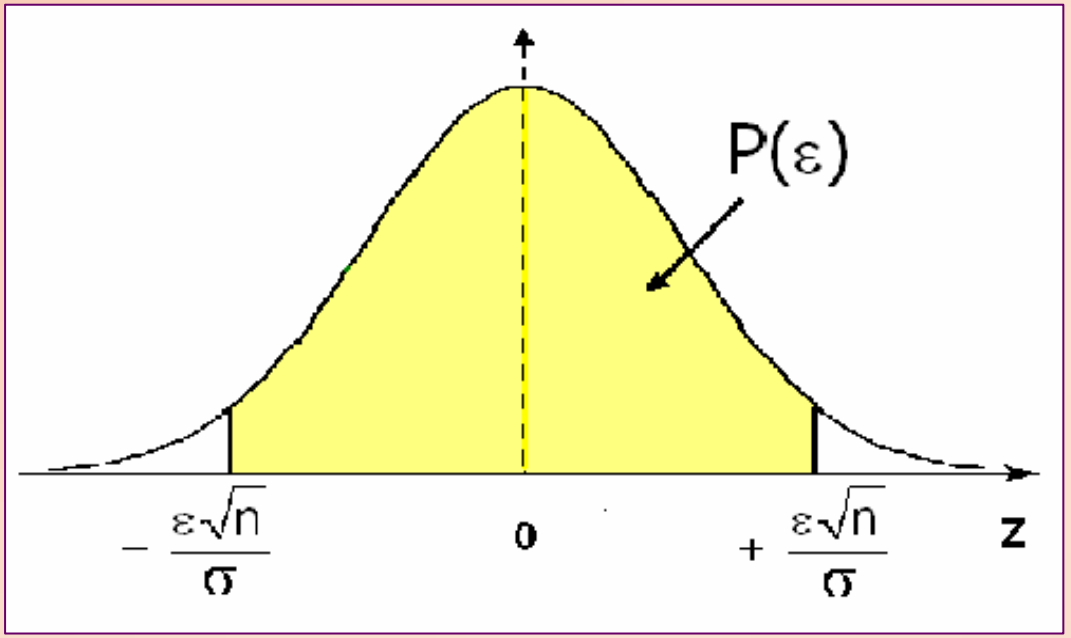
Pergunta: ***Como determinar*** ε ?

Seja $P(\varepsilon) = \gamma$, a probabilidade da média amostral $\bar{X}$ estar a uma distância de, no máximo ε, da média populacional μ (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 $Z \sim N(0,1)$.



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 2:

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

$\sigma = 10$ km/l

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

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

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 3:

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 σ = 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 $X \sim N(\mu ; 2{,}6^2)$

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

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

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

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

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

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

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

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 σ de X, o erro ε da estimativa e o coeficiente de confiança γ 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 σ = 250 reais e média μ desconhecida. Se desejamos estimar a renda média μ com erro ε = 50 reais e com uma confiança γ = 95%, quantos domicílios devemos consultar?

X : renda per-capita domiciliar na região
$X \sim N(\mu ; 250^2)$

n = ?? tal que ε = 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 σ = 50 mg/dl e média μ desconhecida. Se desejamos estimar a quantidade média μ de colesterol com erro ε = 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

σ = 50 mg/dl

n = ?? tal que ε = 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:

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

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

$$= 206{,}25$$

Assim, aproximadamente 207 alunas devem realizar o exame de sangue.

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

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

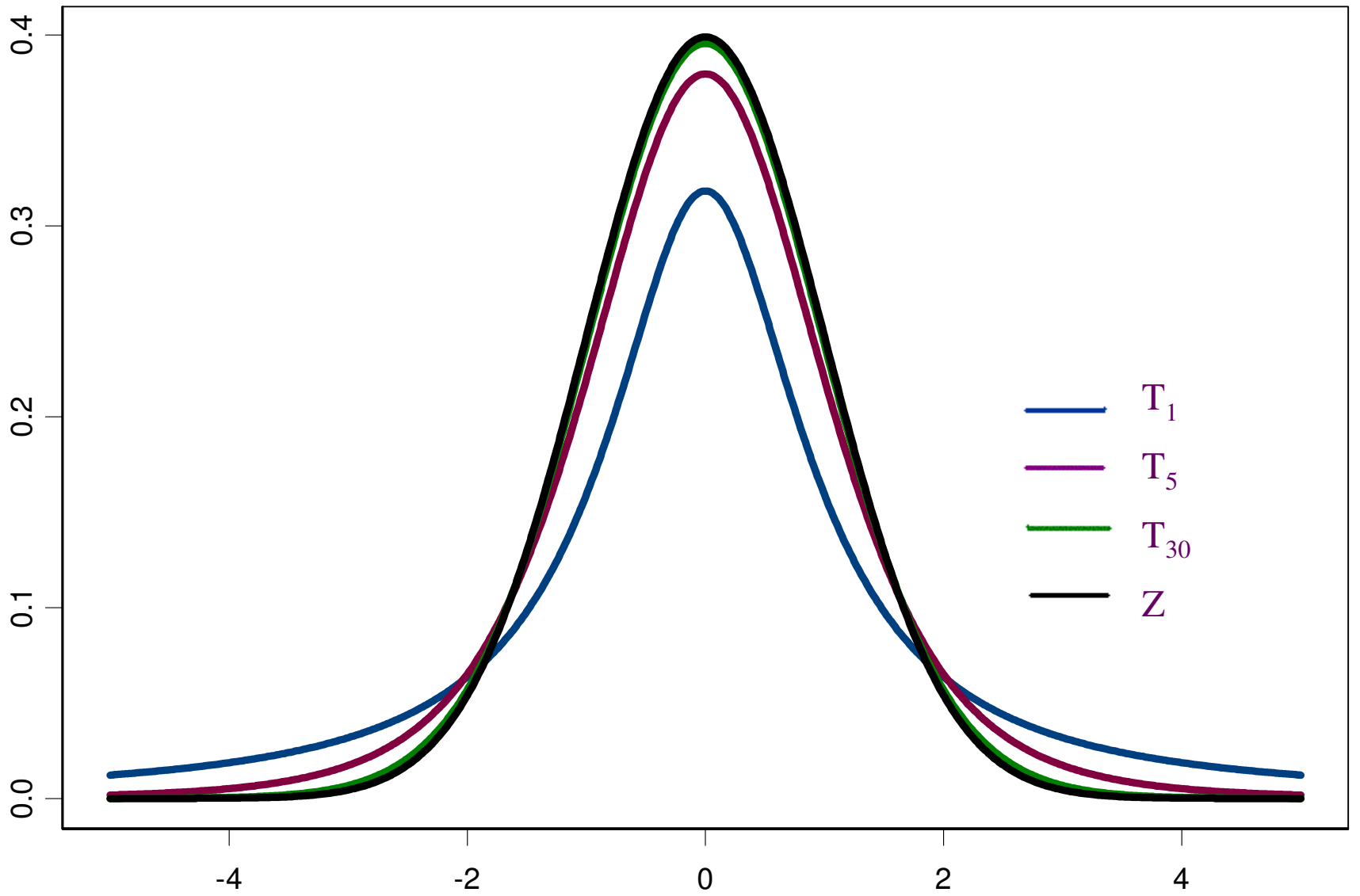
- **Se a variável na <u>população</u> tem distribuição <u>normal</u>, então**

Z **tem distribuição** $N(0,1)$

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

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

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


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

Assim, uma estimativa intervalar *aproximada* para a média populacional μ , quando o tamanho da amostra é grande e σ é desconhecido , 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 6:

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

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

$\gamma = 0,9 \Rightarrow z = 1,65$

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

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

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

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

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