Intervalo de Confianca

X

para a media de uma variavel aleatoria qualquer $\mu = EX$

para uma proporcao $p = EX$ quando $X \sim Bernoulli(p)$

parâmetros μ e p

Na aula passada foi discutida a ideia de estimador e algumas de suas propriedades.

Variavel aleatoria

Amostra

X_1
X_2
$\vdots$
X_n

$\to T(X_1, X_2, \ldots, X_n)$

Fornece uma estimativa pontual para um parametro.

Lembrando que temos

Estimadores e

Parametros

***Parâmetro*: quantidade desconhecida de uma característica <u>da população</u> e sobre a qual temos interesse.**

Exemplos: μ - *média da característica da população*:

μ : taxa média de glicose de mulheres com idade superior a 60 anos, em certa localidade;

p – *proporção de "indivíduos" em uma população com determinada característica.*

p: proporção de pacientes com menos de 40 anos diagnosticados com câncer nos pulmões.

X: variável de interesse : ***Renda***

Vamos observar n elementos, extraídos ao acaso da população, de forma independente;

Para cada elemento selecionado, observamos o valor da variável X de interesse.

$X_{(\text{João})}$

Obtemos, então, uma **amostra aleatória** ($a.a.$) de tamanho n de X, que representamos por

$$X_1, X_2, \ldots, X_n, \qquad i.i.d.$$

sendo X_i a variável de interesse para o i-ésimo indivíduo da amostra.

Uma vez selecionada a amostra saberemos a renda de *João* (x_1)

Estimador: função dos elementos da amostra, construída com a finalidade de representar, ou estimar, um parâmetro da característica de interesse X na população.

→ **Estimador** (ou estatística) $\Rightarrow \boldsymbol{f(X_1, X_2,...,X_n)}$.

$T(X_1, X_2, \ldots, X_n)$

Ex.: $\bar{X}$: média amostral (estimador da média μ da característica X da população).

$\hat{p}$: proporção amostral (estimador da proporção p populacional).

Vamos discutir estes dois exemplos mais importantes: $\bar{X}$ *e* $\hat{p}$

Estimativa: valor numérico assumido pelo estimador, para a amostra selecionada.

Suponha que queremos estimar $\mu = EX$ onde X e' a variavel aleatoria de interesse.

Digamos X = "renda de um individuo escolhido ao acaso de certa populacao"

Usamos o estimador $\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$

Pelo TLC temos para a diferenca entre $\bar{X}$ e μ

$$e = (\bar{X} - \mu) \underset{\text{aprox.}}{\sim} N(0; \sigma^2_{\bar{X}})$$

"erro"

$$\sigma^2_{\bar{X}} = \frac{\sigma^2_X}{n}$$

$$\frac{\bar{X} - \mu}{\sigma_{\bar{X}}} = \frac{e}{\sigma_{\bar{X}}} \underset{\text{Aprox}}{\sim} N(0;1)$$

O erro $e = \overline{X} - \mu$

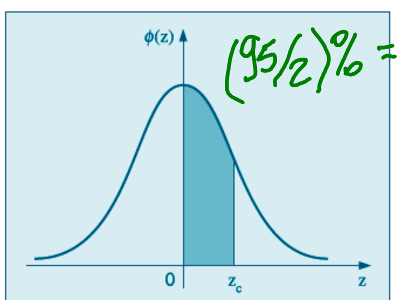
e' uma variavel aleatoria. Pelo TLC podemos estimar a chance dele ser "grande".

Por exemplo, temos

$$P(|e| < 1{,}96\,\sigma_{\overline{X}}) \cong 0{,}95$$

"aprox"

Figura 7.16: $P(0 \leq Z \leq z_c)$ fornecido pela Tabela III.



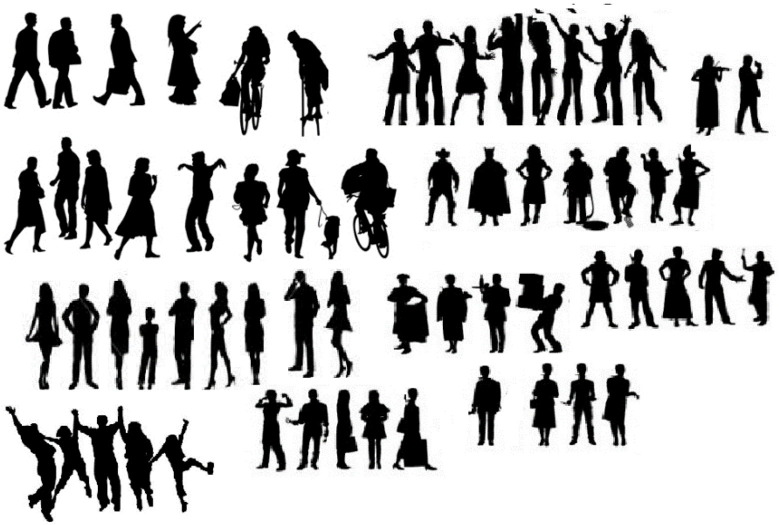
$(95/2)\% = 47{,}5\%$

1,96

$Z \sim N(0,1)$

$$P(-1{,}96 < Z < 1{,}96) = 0{,}95$$

95%



X - variável de interesse: ***Renda***

Por exemplo, obter a distribuição amostral da Média

Amostra 1

$\bar{x}_1$

Amostra 2

$\bar{x}_2$

...

...

Amostra *k*

$\bar{x}_k$

...

...

População das médias de amostras de tamanho n

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

Média amostral

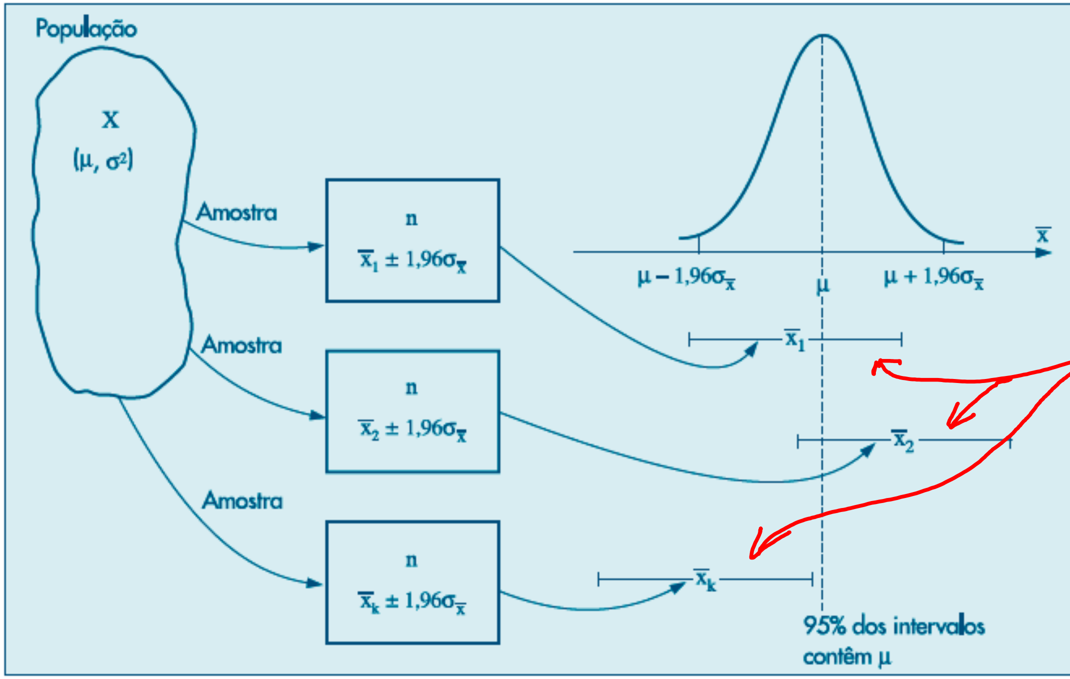
=

estimador de μ

$\bar{X}-\varepsilon$ $\bar{X}+\varepsilon$

$\bar{X}$

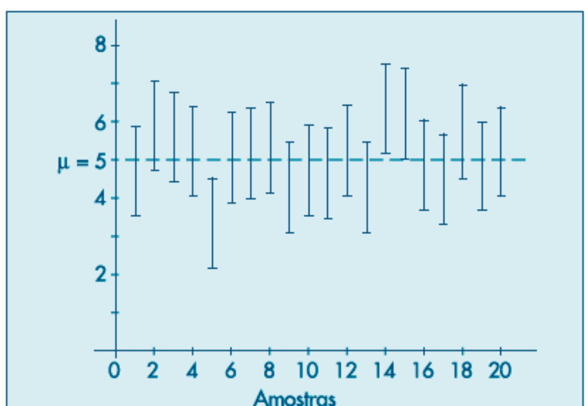
Figura 11.3: Significado de um IC para μ, com $\gamma = 0{,}95$ e σ^2 conhecido.



"sao
intervalos aleatorios"

Para um certo ε
e um certo n

Figura 11.4: Intervalos de confiança para a média de uma N(5, 9), para 20 amostras de tamanho n = 25.



qual e' a probabilidade
deste intervalo
conter o parametro?

γ (letra gama)

Estimativa por intervalo ou intervalo de confiança

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

• Os estimadores pontuais são variáveis aleatórias e, portanto, possuem uma distribuição de probabilidade, em geral, denominada ***distribuição amostral do estimador.***

Ideia: 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*.**

Lembrando:

(Teorema Limite Central)

Se a variável aleatória $\boldsymbol{X}$, na <u>população,</u> tem distribuição com média μ e variância σ^2, então, para uma amostra aleatória de tamanho n de X,

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

Aproximadamente, para n grande

e, portanto

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

Aproximadamente, para n grande

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

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

"a distância entre $\overline{X}$ e μ é menor ou igual a ε"

"Intervalo contem μ"

A probabilidade $P(\varepsilon)$ é também denominada coeficiente de confiança do intervalo, que denotamos por γ (gama).

$\overline{X} \geq \mu - \varepsilon$

$\overline{X} - \varepsilon \leq \mu \leq \overline{X} + \varepsilon$

$\overline{X} \leq \mu + \varepsilon$

$\overline{X}-\varepsilon \quad \overline{X}+\varepsilon \quad \overline{X}$

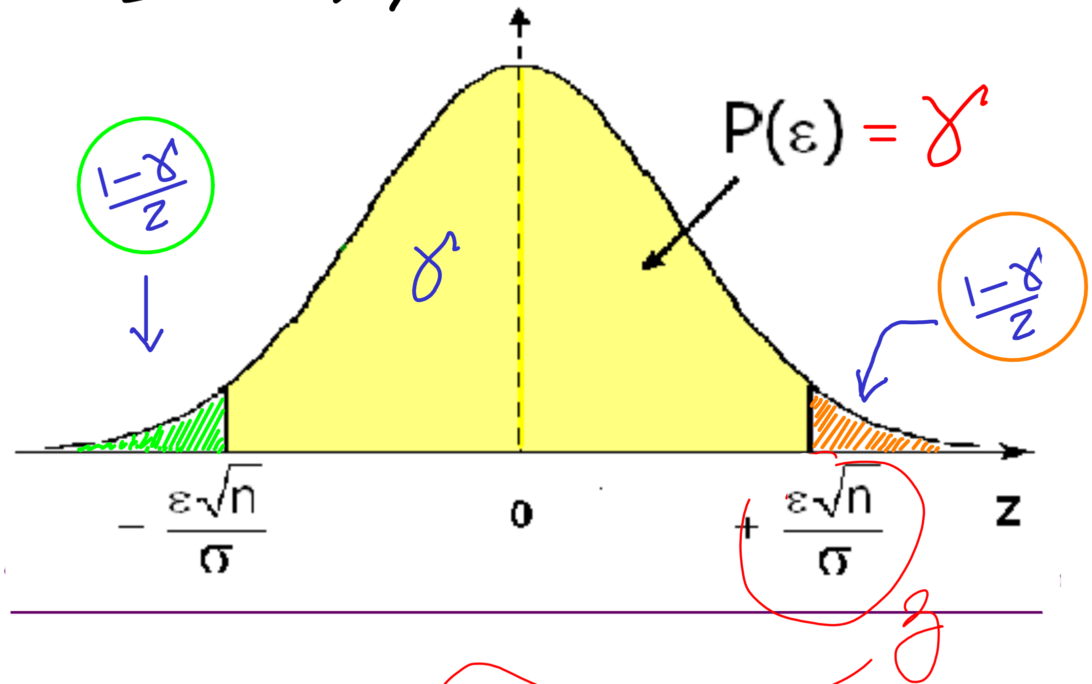
Desse modo, temos

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

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

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

$Z \sim N(0;1)$



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

Determinado a partir da "Tabela da Normal-Padrão"

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

$A(z) = \gamma + \frac{(1-\gamma)}{2}$

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

Erro na estimativa intervalar

Da igualdade $z = \varepsilon \dfrac{\sqrt{n}}{\sigma}$, segue que a margem de erro ε é dada por

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

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

O intervalo de confiança para a média μ, com coeficiente de confiança γ está dado por

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

sendo σ o desvio padrão (conhecido) de X.

"amosta aleatória"

Antes de selecionarmos uma $a.a.$, a probabilidade de que o intervalo

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

contenha a média μ verdadeira da população é γ.

Para o valor observado $\bar{x}$ de $\overline{X}$, o intervalo de 95% de confiança será $[\bar{x} - 1{,}96\frac{\sigma}{\sqrt{n}} \,;\, \bar{x} + 1{,}96\frac{\sigma}{\sqrt{n}}]$;

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

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 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 na população.

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

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

γ = 0,90 $\Rightarrow$ z = 1,65

A estimativa intervalar com 90% de confiança é dada por:

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

$$\left[10{,}5 - 1{,}65\frac{2{,}6}{\sqrt{25}}\,;10{,}5 + 1{,}65\frac{2{,}6}{\sqrt{25}}\right] =$$

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

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

Exemplo um pouco melhor de "media desconhecida mas variancia conhecida"

Exemplo 11.13. Uma máquina enche pacotes de café com uma variância igual a 100 g^2. Ela estava regulada para encher os pacotes com 500 g, em média. Agora, ela se desregulou, e queremos saber qual a nova média μ. Uma amostra de 25 pacotes apresentou uma média igual a 485 g. Vamos construir um intervalo de confiança com 95% de confiança para μ.

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

então

$$IC(\mu; 0{,}95) = 485 \pm 1{,}96 \times 2,$$

ou seja,

$$IC(\mu; 0{,}95) =]481,\ 489[,$$

pois $\sigma_{\bar{x}} = \sigma/\sqrt{n} = 10/5 = 2$g.

Dimensionamento da amostra

"Qual é o n?"

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

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

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

onde ε é a margem de erro, z é tal que

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

e σ é o desvio padrão (conhecido) de X.

Exemplo

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

$\varepsilon = 50$

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

n = ??

Então,

$$n = \left(\frac{z}{\varepsilon}\right)^2 \sigma^2 = \left(\frac{1{,}96}{50}\right)^2 \times (250)^2 = 96{,}04$$

Aproximadamente, 97 domicílios devem ser consultados.

Na prática, em geral, não conhecemos a variância populacional σ^2.

(Ou seja, tanto μ quanto σ são desconhecidos)

RESULTADO

Se X tem distribuição normal, com média μ e variância σ^2, então, para uma amostra aleatória de tamanho n de X, temos que

t_{n-1}

Distribuição "t de Student"

$$T = \frac{\overline{X} - \mu}{\sqrt{\frac{S^2}{n}}} \sim t_{n-1},$$

A distribuição t depende de n, MAS

$T \sim N(0,1)$

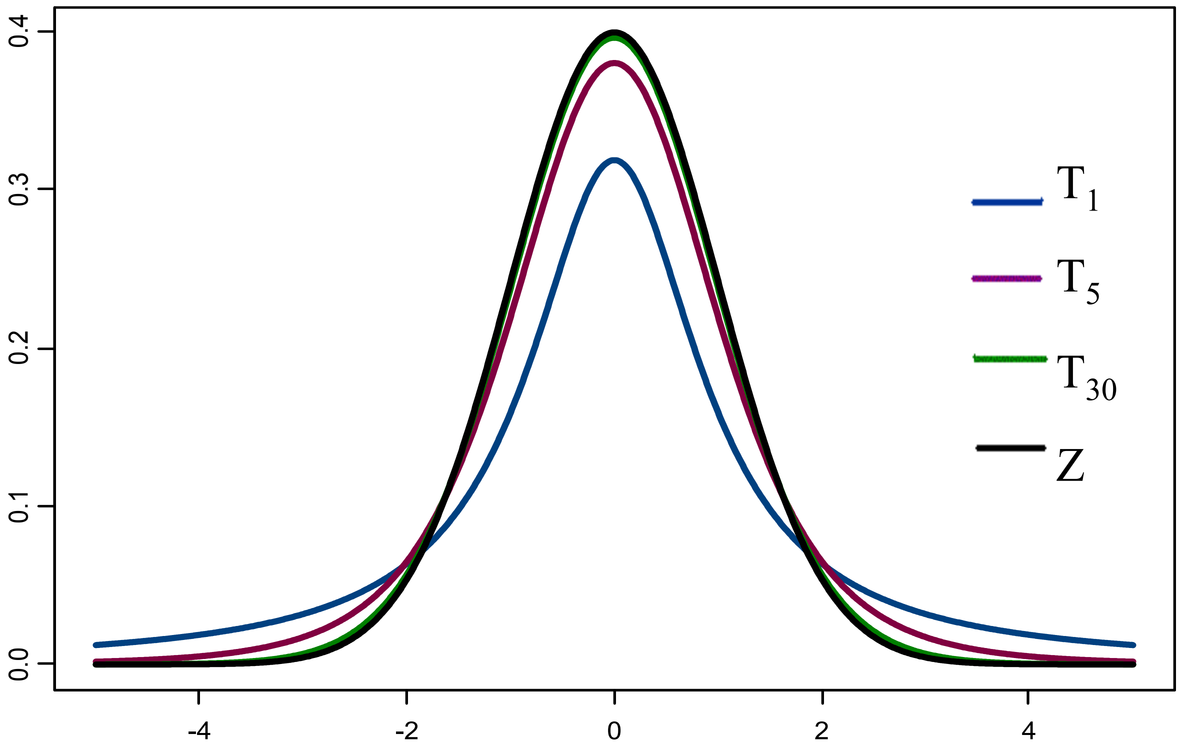
"aproximadamente se n for grande"

Estimador da variância σ^2:

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

é a variância amostral.

Distribuição *t* de *Student*



• tem caudas mais densas do que a distribuição normal;

• valores extremos são mais prováveis de ocorrer com a distribuição *t* do que com a normal padrão;

Vamos considerar agora o caso de "n grande" quando ha' pouca diferenca entre as duas distribuicoes.

Essa questao sera' retomada em "Testes de Hipoteses".

Exemplo

A quantidade de colesterol X no sangue das alunas de uma universidade tem uma distribuição aproximadamente normal com média e desvio padrão desconhecidos.

Para estimar a quantidade média de colesterol μ é selecionada uma amostra de 60 alunas. A média e o desvio padrão amostrais encontrados são 182 mg/dl e 50 mg/dl, respectivamente. Determine um intervalo de confiança com coeficiente de confiança de 90% para μ.

X: quantidade de colesterol no sangue das alunas da universidade $\Rightarrow X \sim N(\mu, \sigma^2)$

$n = 60 \Rightarrow \bar{x} = 182$ e $s = 50$

$\gamma = 0{,}90 \Rightarrow$ ***z=1,65, usando a tabela da Normal Padrão***

Substituindo os valores:

$$IC(\mu; 90\%) = (182 - 1{,}65\,\frac{50}{\sqrt{60}}\ ;\ 182 + 1{,}65\,\frac{50}{\sqrt{60}}) = \boldsymbol{(\,171{,}35\ ;\ 192{,}25\,)}$$

Estimador para uma proporcao

Sumonha que $X \sim Bernoulli(p)$

$p = EX$ parâmetro de interesse

Estimador

$$\hat{p} = \frac{1}{n}\sum_{i=1}^{n} X_i = \text{"proporção amostral"}$$

onde $X_1, X_2, \ldots, X_n$ são os valores amostrais.

Pelo TLC

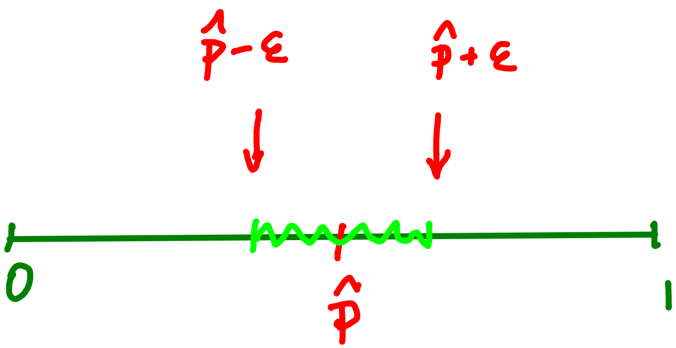
$$\hat{p} \sim N\left(p; \frac{p(1-p)}{n}\right)$$

aprox

ou

$$e = \hat{p} - p \sim N\left(0; \frac{p(1-p)}{n}\right) = \text{"erro"}$$

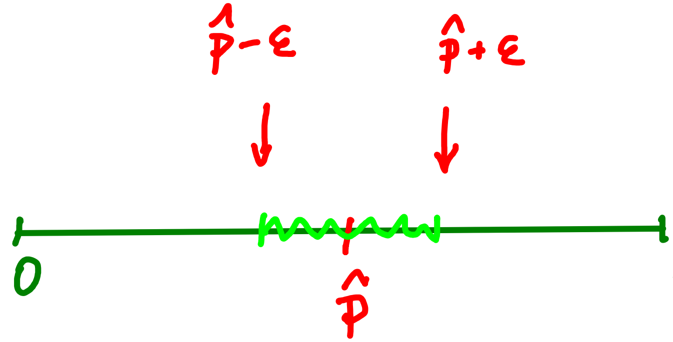
Intervalo de confianca para p com coeficiente de confianca γ



com "Probabilidade do intervalo conter p é γ"

$$IC[p;\gamma] = [\hat{p} - \varepsilon;\ \hat{p} + \varepsilon]$$

"Probabilidade do intervalo conter p é γ"



$$\left.\begin{array}{l} \hat{p}-\varepsilon < p \Rightarrow \hat{p} < p+\varepsilon \\ \quad\text{e} \\ \hat{p}+\varepsilon > p \Rightarrow \hat{p} > p-\varepsilon \end{array}\right\} \quad p-\varepsilon < \hat{p} < p+\varepsilon$$

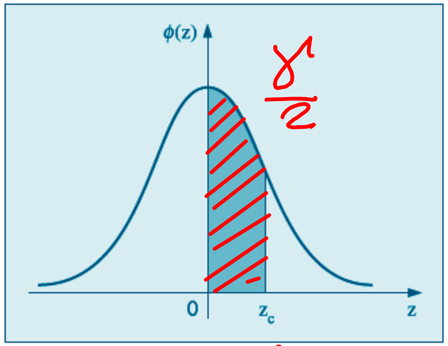
$$P(p-\varepsilon < \hat{p} < p+\varepsilon) = \gamma$$

$$P\left(\frac{p-\varepsilon-p}{\sqrt{\frac{p(1-p)}{n}}} < \frac{\hat{p}-p}{\sqrt{\frac{p(1-p)}{n}}} < \frac{p+\varepsilon-p}{\sqrt{\frac{p(1-p)}{n}}} \right) = \gamma$$

$$P\left(\frac{-\varepsilon\sqrt{n}}{\sqrt{p(1-p)}} < Z < \frac{\varepsilon\sqrt{n}}{\sqrt{p(1-p)}} \right) = \gamma$$

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

Figura 7.16: $P(0 \leqslant Z \leqslant z_c)$ fornecido pela Tabela III.



z_γ

"Tabela" $\Rightarrow$ $$z_\gamma = \frac{\varepsilon\sqrt{n}}{\sqrt{p(1-p)}}$$

$$\Downarrow$$

$$\varepsilon = z_\gamma \sqrt{\frac{p(1-p)}{n}}$$

"Entao" o intervalo de confianca para p (desconhecido) e'

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

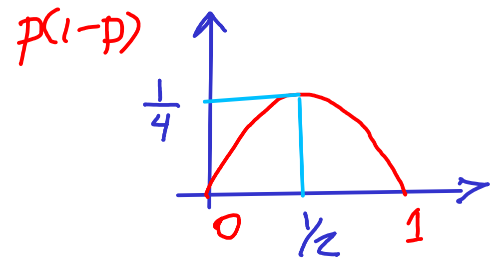
?????

Duas solucoes: "Conservadora" e "Otimista"

Conservadora: tomo um intervalo que garante

"Probabilidade do intervalo conter p é pelo menos γ"

$$P(p-\varepsilon < \hat{p} < p+\varepsilon) \geq \gamma$$



Considero a maior variancia possivel:

$$\sigma_{\bar{X}} = \sigma_{\hat{p}} = \sqrt{\frac{p(1-p)}{n}} \leq \sqrt{\frac{1/4}{n}} = \frac{1}{2\sqrt{n}}$$

$$\varepsilon = z_\gamma \sqrt{\frac{p(1-p)}{n}} \leq \frac{z_\gamma}{2\sqrt{n}}$$

$$IC[p;\gamma] = \left[\hat{p} - \frac{z_\gamma}{2\sqrt{n}} \,;\, \hat{p} + \frac{z_\gamma}{2\sqrt{n}}\right]$$

"Solucao Otimista": uso a estimativa de $\sigma_{\bar{X}} = \sigma_{\hat{p}} = \sqrt{\hat{p}(1-\hat{p})}$

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

Exemplo 11.15. Numa pesquisa de mercado, $n = 400$ pessoas foram entrevistadas sobre determinado produto, e 60% delas preferiram a marca A. Aqui, $\hat{p} = 0{,}6$ e um intervalo de confiança para p com coeficiente de confiança $\gamma = 0{,}95$ será

$$0{,}6 \pm (1{,}96)\ 1/\sqrt{1600} = 0{,}6 \pm 0{,}049,$$

ou seja

$$\mathrm{IC}(p;\, 0{,}95) = \,]0{,}551;\, 0{,}649[.$$

intervalo "conservador" ou "cauteloso"

pois pode ser bem maior que o necessario se p for "pequeno" ou "grande"

Na pratica, o mais frequente e' usar o resultado matematico

$$\frac{\hat{p} - p}{\sqrt{\dfrac{\hat{p}(1-\hat{p})}{n}}} \sim N(0,1)$$

aprox.

e considerar o "intervalo otimista"

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

Exemplo 11.16. Suponha que em $n = 400$ provas obtemos $k = 80$ sucessos. Vamos obter um intervalo de confiança para p com $\gamma = 0{,}90$. Como $\hat{p} = 80/400 = 0{,}2$ e $\hat{q} = 1 - \hat{p} = 0{,}8$, então

$$0{,}2 \pm (1{,}645)\sqrt{(0{,}2)(0{,}8)/400} = 0{,}2 \pm 0{,}033,$$

ou seja,

$$IC(p;\, 0{,}90) =]0{,}167;\, 0{,}233[.$$

("otimista")

Usando (11.40) o intervalo conservador é

$$IC(p;\, 0{,}90) =]0{,}159;\, 0{,}241[.$$

compare!