Probabilidade e Inferência Estatística com R - Módulo 2

Profª. Angélica Maria Tortola Ribeiro

October 20, 2020

UTFPRUNIVERSIDADE QUETRANSFORMA

- **População:** É o conjunto de todos os elementos sob investigação com pelo menos uma característica em comum.
- **Amostra:** É qualquer subconjunto da população.

 Exemplos:

 1. P1: Todos os moradores de Curitiba.
 2. P2: Todos os estudantes da UTFPR.
 3. P3: Todos os celulares produzidos por uma fábrica.
 4. P4: Todos os consumidores de algum produto.

- **População:** É o conjunto de todos os elementos sob investigação com pelo menos uma característica em comum.
- **Amostra:** É qualquer subconjunto da população.

  Exemplos:
  1. P1: Todos os moradores de Curitiba.
  2. P2: Todos os estudantes da UTFPR.
  3. P3: Todos os celulares produzidos por uma fábrica.
  4. P4: Todos os consumidores de algum produto.

- **População:** É o conjunto de todos os elementos sob investigação com pelo menos uma característica em comum.
- **Amostra:** É qualquer subconjunto da população.

  Exemplos:

  1. P1: Todos os moradores de Curitiba.
  2. P2: Todos os estudantes da UTFPR.
  3. P3: Todos os celulares produzidos por uma fábrica.
  4. P4: Todos os consumidores de algum produto.

- **População:** É o conjunto de todos os elementos sob investigação com pelo menos uma característica em comum.
- **Amostra:** É qualquer subconjunto da população.

  Exemplos:

  1. P1: Todos os moradores de Curitiba.
  2. P2: Todos os estudantes da UTFPR.
  3. P3: Todos os celulares produzidos por uma fábrica.
  4. P4: Todos os consumidores de algum produto.

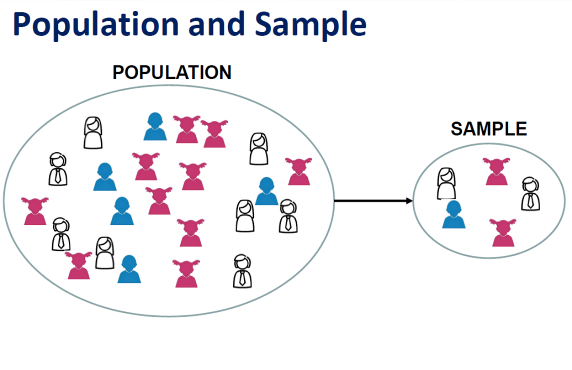


Figure: Representação de população e amostra.

- **Variáveis:** São as características de interesse de uma população.

  Exemplos:

  1. **P1:** altura, peso, nº de filhos, renda familiar, grau de escolaridade (F, M, S), tem ensino superior (sim ou não), classe social (baixa, média, alta), sexo (masculino, feminino).
  2. **P3:** possíveis variáveis são: peso do aparelho, nº de defeitos por aparelho, tem defeito (sim ou não), tamanho do aparelho (pequeno, médio, grande), nº de aparelhos produzidos por hora.

- **Variáveis:** São as características de interesse de uma população.

  Exemplos:

  1. **P1:** altura, peso, nº de filhos, renda familiar, grau de escolaridade (F, M, S), tem ensino superior (sim ou não), classe social (baixa, média, alta), sexo (masculino, feminino).
  2. **P3:** possíveis variáveis são: peso do aparelho, nº de defeitos por aparelho, tem defeito (sim ou não), tamanho do aparelho (pequeno, médio, grande), nº de aparelhos produzidos por hora.

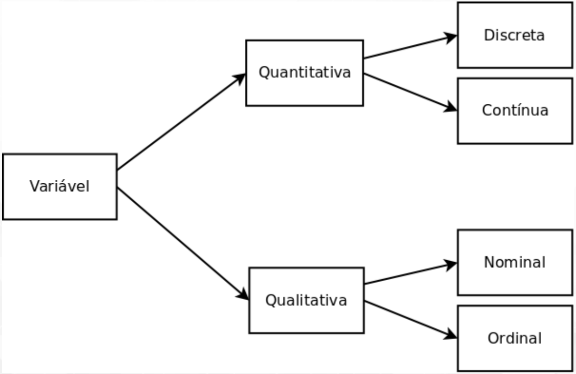


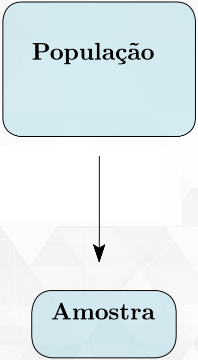
FIGURE: Classificação das variáveis.

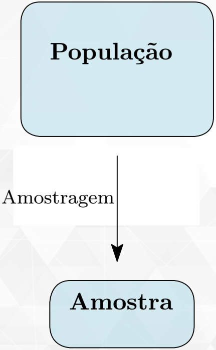
Table: Classificação das variáveis

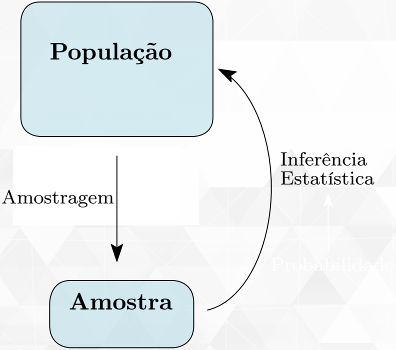
Discretas	Contínuas	Ordinais	Nominais
nº de filhos	peso	tamanho	sexo
nº de defeitos	altura	classe social	tem defeito
nº de aparelhos	renda familiar	Grau de escolaridade	tem ensino superior

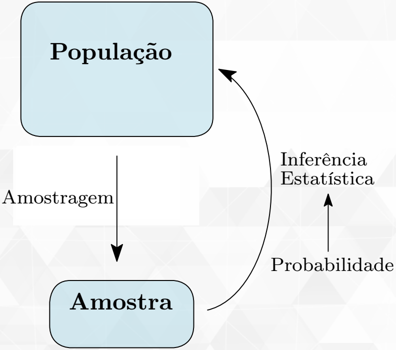
Exercício: Apresentar 2 exemplos de variáveis para cada uma das classificações da Tabela 1, considerando as populações **P2** e **P4**.

População


População
Amostra


População
Amostragem
Amostra


População
Amostragem
Amostra
Inferência
Estatística


População
Amostragem
Amostra
Inferência
Estatística
Probabilidade

Inferência Estatística

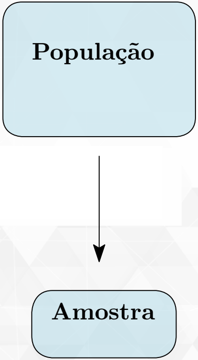
Conjunto de técnicas utilizadas para tirar conclusões sobre determinada(s) característica(s) da população, a partir de informações colhidas de uma amostra.

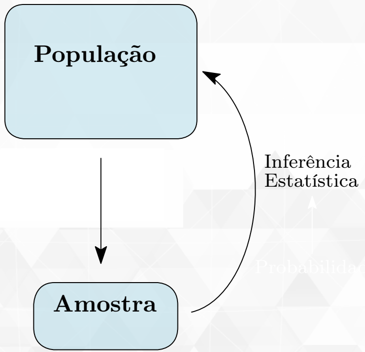
Amostra Aleatória (a.a.)

Uma a.a. de tamanho n de uma população X, com dada distribuição, é o conjunto de n variáveis aleatórias independentes $X_1, X_2, ..., X_n$, cada uma com a mesma distribuição de X.

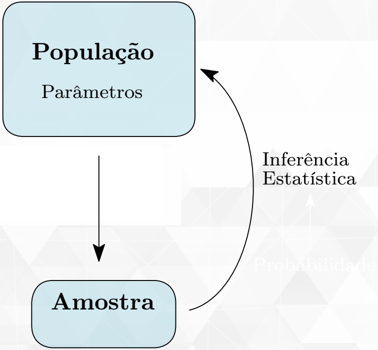
População

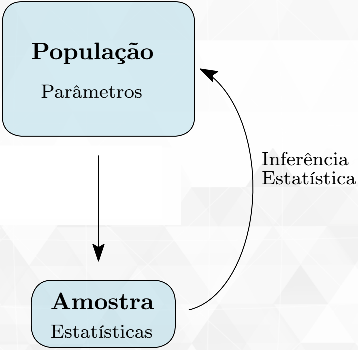
Parâmetros e Estatísticas




População
Amostra
Inferência
Estatística

Parâmetros e Estatísticas




População
Parâmetros
Inferência Estatística
Amostra
Estatísticas

Parâmetros

São medidas utilizadas para descrever características da população.

Estatísticas

São características da amostra, ou seja, uma estatística T é uma função de $X_1, X_2, ..., X_n$.

P1: Moradores de Curitiba

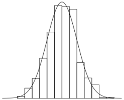
Variável X: Altura

Parâmetros e Estatísticas

P1: Moradores de Curitiba

Variável X: Altura

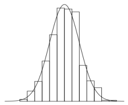
$$X \sim N(\mu, \sigma^2)$$



P1: Moradores de Curitiba

Variável X: Altura

$$X \sim N(\underbrace{\mu, \sigma^2})$$

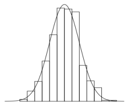


Parâmetros

P1: Moradores de Curitiba

Variável X: Altura

$$X \sim N\underbrace{(\mu, \sigma^2)}$$



Parâmetros

Amostra

$$\left.\begin{array}{l} \bar{X} = \frac{X_1 + X_2 + \ldots + X_n}{n} \\ S^2 = \sum_{i=1}^{n} \frac{(X_i - \bar{X})^2}{n-1} \\ S = \sqrt{S^2} \end{array}\right\}$$

Estatísticas

P1: Moradores de Curitiba

Variável X: Altura

$$X \sim N(\underbrace{\mu, \sigma^2}_{})$$

Parâmetros

Amostra

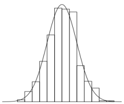
$$\left.\begin{aligned} \bar{X} &= \frac{X_1+X_2+\ldots+X_n}{n} \\ S^2 &= \sum_{i=1}^{n} \frac{(X_i-\bar{X})^2}{n-1} \\ S &= \sqrt{S^2} \end{aligned}\right\}$$

Estatísticas

P1: Moradores de Curitiba

Variável X: Altura

$$X \sim N\underbrace{(\mu, \sigma^2)}_{\text{Parâmetros}}$$



Amostra

$$\left.\begin{array}{l} \bar{X} = \frac{X_1 + X_2 + \ldots + X_n}{n} \\ S^2 = \sum_{i=1}^{n} \frac{(X_i - \bar{X})^2}{n-1} \\ S = \sqrt{S^2} \end{array}\right\} \longrightarrow \text{Estatísticas}$$

P1: Moradores de Curitiba

Variável X: Tem filhos (S, N)

P1: Moradores de Curitiba
Variável X: Tem filhos (S, N)

$$X \sim Bernoulli(p)$$

Parâmetros e Estatísticas

P1: Moradores de Curitiba

Variável X: Tem filhos (S, N)

$$X \sim Bernoulli(p)$$

$$X = \begin{cases} 0, & \text{se Fracasso} \\ 1, & \text{se Sucesso} \end{cases}$$

P1: Moradores de Curitiba

Variável X: Tem filhos (S, N)

$$X \sim Bernoulli(p)$$

$$X = \begin{cases} 0, & \text{se Fracasso} \\ 1, & \text{se Sucesso} \end{cases}$$

Parâmetro

P1: Moradores de Curitiba

Variável X: Tem filhos (S, N)

$$X \sim Bernoulli(p)$$

$$X = \begin{cases} 0, & \text{se Fracasso} \\ 1, & \text{se Sucesso} \end{cases}$$

Parâmetro

Amostra

Parâmetros e Estatísticas

P1: Moradores de Curitiba

Variável X: Tem filhos (S, N)

$$X \sim Bernoulli(p)$$

$$X = \begin{cases} 0, & \text{se Fracasso} \\ 1, & \text{se Sucesso} \end{cases}$$

Parâmetro

Amostra

$$\hat{p} = \frac{X_1 + X_2 + \ldots + X_n}{n}$$

Estatística

P1: Moradores de Curitiba

Variável X: Tem filhos (S, N)

$$X \sim Bernoulli(p)$$

$$X = \begin{cases} 0, & \text{se Fracasso} \\ 1, & \text{se Sucesso} \end{cases}$$

$p \rightarrow$ Parâmetro

Amostra $\rightarrow \hat{p} = \frac{X_1 + X_2 + \ldots + X_n}{n} \rightarrow$ Estatística

- **Exemplos de parâmetros:** média populacional, variância populacional, desvio-padrão populacional, proporção populacional;
- **Exemplos de estatísticas:** média amostral, variância amostral, desvio-padrão amostral, proporção amostral;

Table: Parâmetros e Estatísticas.

Medidas	Parâmetros	Estatísticas
Média	$\mu = \dfrac{X_1 + X_2 + \ldots + X_N}{N}$	$\bar{X} = \dfrac{X_1 + X_2 + \ldots + X_n}{n}$
Variância	$\sigma^2 = \sum_{i=1}^{N} \dfrac{(X_i - \mu)^2}{N}$	$S^2 = \sum_{i=1}^{n} \dfrac{(X_i - \bar{X})^2}{n-1}$ ou $\hat{\sigma}^2 = \sum_{i=1}^{n} \dfrac{(X_i - \bar{X})^2}{n}$
Desvio-padrão	$\sigma = \sqrt{\sigma^2}$	$S = \sqrt{(S^2)}$
Proporção	$p = \dfrac{X_1 + X_2 + \ldots + X_N}{N}$	$\hat{p} = \dfrac{X_1 + X_2 + \ldots + X_n}{n}$

Exemplo 1 no R: Parâmetros e Estatísticas

Exercícios no R:

1. Simular 5000 pesos de uma população com distribuição Normal com parâmetros $\mu = 90$ e $\sigma^2 = 9$;
2. Construir um histograma para esta variável, considerando as densidades de frequência, e adicionar a curva teórica da Normal;
3. Selecionar uma amostra, com reposição de tamanho $n = 100$ desta população;
4. Calcular as estatísticas: média, variância e desvio-padrão da amostra.

Exercícios no R:

1. Simular 5000 pesos de uma população com distribuição Normal com parâmetros $\mu = 90$ e $\sigma^2 = 9$;
2. Construir um histograma para esta variável, considerando as densidades de frequência, e adicionar a curva teórica da Normal;
3. Selecionar uma amostra, com reposição de tamanho $n = 100$ desta população;
4. Calcular as estatísticas: média, variância e desvio-padrão da amostra.

Exercícios no R:

1. Simular 5000 pesos de uma população com distribuição Normal com parâmetros $\mu = 90$ e $\sigma^2 = 9$;
2. Construir um histograma para esta variável, considerando as densidades de frequência, e adicionar a curva teórica da Normal;
3. Selecionar uma amostra, com reposição de tamanho $n = 100$ desta população;
4. Calcular as estatísticas: média, variância e desvio-padrão da amostra.

Exercícios no R:

1. Simular 5000 pesos de uma população com distribuição Normal com parâmetros $\mu = 90$ e $\sigma^2 = 9$;
2. Construir um histograma para esta variável, considerando as densidades de frequência, e adicionar a curva teórica da Normal;
3. Selecionar uma amostra, com reposição de tamanho $n = 100$ desta população;
4. Calcular as estatísticas: média, variância e desvio-padrão da amostra.

Estimador:

São estatísticas utilizadas para estimar os parâmetros. Aos valores observados dos estimadores dá-se o nome de **estimativas**.

Table: Parâmetros e Estimadores.

Parâmetros	Estimadores	Propriedades
μ	$\bar{X} = \dfrac{X_1 + X_2 + \ldots + X_n}{n}$	Não viciado e consistente
σ^2	$S^2 = \sum_{i=1}^{n} \dfrac{(X_i - \bar{X})^2}{n-1}$	Não viciado e consistente
p	$\hat{p} = \dfrac{X_1 + X_2 + \ldots + X_n}{n}$	Não viciado e consistente

População

$$X \sim N(\mu, \sigma^2)$$

População

$X \sim N(\mu, \sigma^2)$

Amostra 1

$\bar{x}_1$

População

$X \sim N(\mu, \sigma^2)$

Amostra 1
$\bar{x}_1$

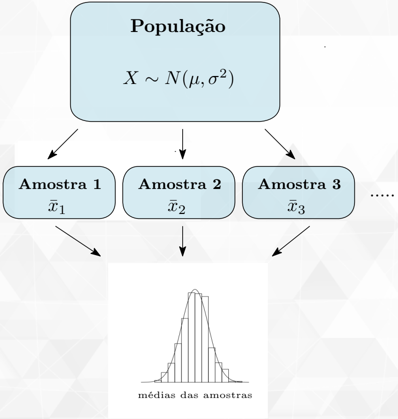
Amostra 2
$\bar{x}_2$

População

$$X \sim N(\mu, \sigma^2)$$

Amostra 1	**Amostra 2**	**Amostra 3**	.....
$\bar{x}_1$	$\bar{x}_2$	$\bar{x}_3$	

Distribuição Amostral



Distribuição amostral de um Estimador:

É o comportamento probabilístico do estimador, isto é, sua distribuição de probabilidade, caso todas as possíveis amostras fossem retiradas.

Exemplo 2 no R: Distribuição amostral da média: População Normal.

População

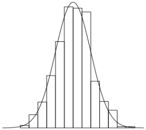
$$X \sim N(\mu, \sigma^2)$$

Distribuição amostral da média: População Normal

População

$$X \sim N(\mu, \sigma^2)$$

$\downarrow \forall n$

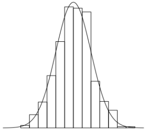


médias das amostras

População

$$X \sim N(\mu, \sigma^2)$$

$\forall n$



médias das amostras

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

Exemplo 3 no R: Distribuição amostral da média: População não Normal.

População

$$X \sim f(x;\theta) \quad \text{ou}$$
$$X \sim P(X = x;\theta)$$

População

$$X \sim f(x;\theta) \quad \text{ou}$$
$$X \sim P(X = x;\theta)$$

$$\begin{cases} E(X) \\ Var(X) \end{cases}$$

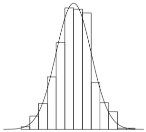
População

$X \sim f(x;\theta)$ ou

$X \sim P(X = x;\theta)$

$$\begin{cases} E(X) \\ Var(X) \end{cases}$$

$n \to \infty$



médias das amostras

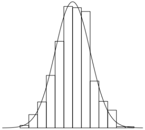
Distribuição amostral da média: População não Normal

População

$$X \sim f(x;\theta) \quad \text{ou}$$
$$X \sim P(X = x;\theta)$$

$$\begin{cases} E(X) \\ Var(X) \end{cases}$$

$n \to \infty$



médias das amostras

$$\bar{X} \sim N\left(E(X), \frac{Var(X)}{n}\right), n \to \infty$$

UTFPRUNIVERSIDADE QUETRANSFORMA

Exemplo 4 no R: Distribuição amostral da proporção

População

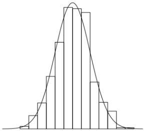
$$X \sim Bernoulli(p)$$

Distribuição amostral da proporção

População

$X \sim Bernoulli(p)$

$n \to \infty$

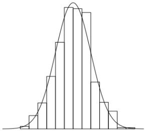


proporções das amostras

População

$$X \sim Bernoulli(p)$$

$$\downarrow n \to \infty$$



proporções das amostras

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

Exemplo 5 no R:

1. Os salários de funcionários de uma empresa seguem distribuição Normal, com média 1500 e desvio padrão 250. Calcule a probabilidade de que os salários médios das amostras de tamanhos 10 estejam entre 1400 e 1600.

$$X \sim N(1500, 250^2)$$

$$X \sim N(1500, 250^2)$$

$$\bar{X} \sim N\left(1500, \frac{250^2}{10}\right)$$

$$P(1400 \leq \bar{X} \leq 1600)$$

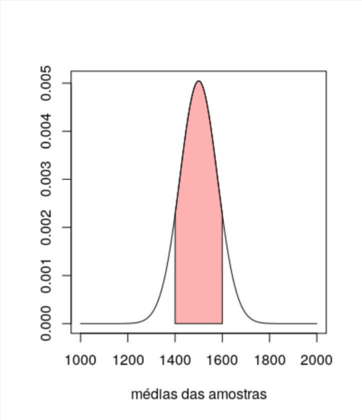
$$X \sim N(1500, 250^2)$$

$$\downarrow$$

$$\bar{X} \sim N\left(1500, \frac{250^2}{10}\right)$$

$$\downarrow$$

$$P(1400 \leq \bar{X} \leq 1600)$$


0.000
0.001
0.002
0.003
0.004
0.005
1000
1200
1400
1600
1800
2000
médias das amostras

Exemplo 6 no R:

1. Suponha que estamos interessados em estimar a proporção de consumidores de certo produto. Considerando a seleção de uma amostra aleatória de 300 pessoas, calcule a probabilidade de que a proporção amostral de consumidores seja de, no mínimo, 35%, sabendo que a verdadeira proporção é de 40%.

$$X \sim Bernoulli(0.4)$$

$$P(\hat{p} \geq 0.35)$$

$$X \sim Bernoulli(0.4)$$

$$\downarrow$$

$$\hat{p} \sim N\left(0.4, \frac{0.4(1-0.4)}{300}\right)$$

$$\downarrow$$

$$P(\hat{p} \geq 0.35)$$

$$X \sim Bernoulli(0.4)$$

$$\downarrow$$

$$\hat{p} \sim N\left(0.4, \frac{0.4(1-0.4)}{300}\right)$$

$$\downarrow$$

$$P(\hat{p} \geq 0.35)$$

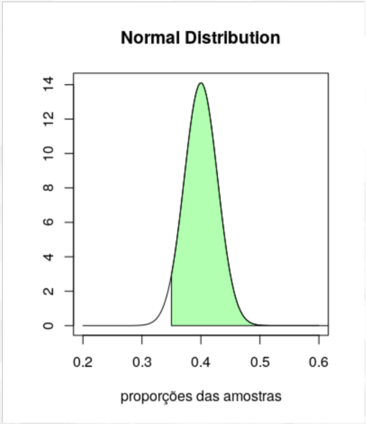
UTFPR UNIVERSIDADE QUE TRANSFORMA

DAEST DEPARTAMENTO ACADÊMICO DE ESTATÍSTICA

UTFPR UNIVERSIDADE TECNOLÓGICA FEDERAL DO PARANÁ

MINISTÉRIO DA EDUCAÇÃO

PÁTRIA AMADA BRASIL GOVERNO FEDERAL


Normal Distribution
0
2
4
6
8
10
12
14
0.2
0.3
0.4
0.5
0.6
proporções das amostras

Exercícios no R:

1. Admite-se que as alturas dos estudantes de uma universidade seguem distribuição Normal com média 172.72 cm e desvio padrão 7.62 cm. Calcule a probabilidade de que a altura média amostral se encontre entre 169.67 cm e 173.48 cm, considerando amostras aleatórias de 25 estudantes. **R: 0.6683**.
2. Suponha que 60% da população de uma certa cidade seja a favor da criação de um fundo público para fins de criação de áreas de lazer. Se 150 pessoas forem selecionadas aleatoriamente e entrevistadas, qual a probabilidade de que a proporção amostral de pessoas favoráveis seja menor que 0.52? **R:0.0227**

Tipos de estimação:

- Estimação pontual;
- Estimação por intervalo.

$$\bar{X} \rightarrow \mu$$

$$\hat{p} \rightarrow p$$

$$S^2 \rightarrow \sigma^2$$

$$X \sim N(\mu, \sigma^2)$$

$$\Big\downarrow \forall n$$

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

$$\Big\downarrow$$

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

$$X \sim N(\mu, \sigma^2)$$

$$\downarrow$$

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

$$X \sim Bernoulli(p)$$

$$\Big\downarrow \; n \to \infty$$

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

$$\Big\downarrow$$

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

- Considera além da estimação pontual, o erro associado à amostra utilizada no processo inferencial;
- Define limites de confiança para um parâmetro, com uma probabilidade (nível de confiança) conhecida;
- Para estimadores com distribuição de probabilidade simétrica, o IC para o parâmetro é simétrico em torno do seu estimador.

- De modo geral, seja θ o parâmetro de interesse, então o intervalo de confiança para θ com probabilidade $1-\alpha$ (nível de confiança) é:

$$IC(\theta,(1-\alpha)) = (Li; Ls)$$

ou seja,

$$P(Li \leq \theta \leq Ls) = 1-\alpha$$

- $Li \rightarrow$ limite inferior do intervalo;
- $Ls \rightarrow$ limite superior do intervalo;

- Supondo $X \sim N(\mu, \sigma^2) \rightarrow \bar{X} \sim N\left(\mu, \frac{\sigma^2}{n}\right)$, com σ^2 conhecido.
- E sabendo que $Z = \frac{(\bar{X} - \mu)}{\sigma/\sqrt{n}}$, o IC para μ pode ser obtido por:

$$P(-z_{\frac{\alpha}{2}} \leq Z \leq z_{\frac{\alpha}{2}}) = 1 - \alpha$$
$$P(-z_{\frac{\alpha}{2}} \leq \frac{(\bar{X} - \mu)}{\sigma/\sqrt{n}} \leq z_{\frac{\alpha}{2}}) = 1 - \alpha$$
$$P(\bar{X} - z_{\frac{\alpha}{2}} \frac{\sigma}{\sqrt{n}} \leq \mu \leq \bar{X} + z_{\frac{\alpha}{2}} \frac{\sigma}{\sqrt{n}}) = 1 - \alpha$$

$$IC(\mu,(1-\alpha)) = \left(\bar{X} - e; \bar{X} + e\right)$$

$$\text{onde} \quad e = z_{\frac{\alpha}{2}} \frac{\sigma}{\sqrt{n}}$$

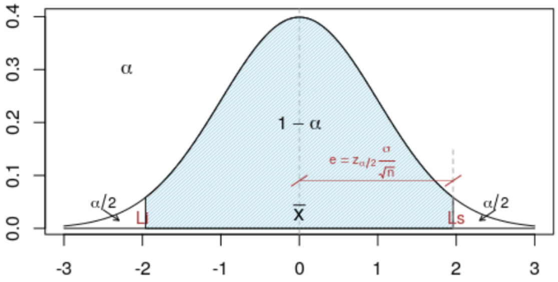
Intervalo de confiança para a média μ - σ^2 conhecida

Obs: O resultado acima também é válido mesmo que a população não seja Normal, mas a amostra seja suficientemente grande (TLC). O valor $z_{\alpha/2}$ é tal que $P(Z \geq z_{\alpha/2}) = \alpha/2$.

- Da relação anterior, pode-se encontrar o tamanho de amostra necessária, fazendo-se

$$n = \left(\frac{z_{\frac{\alpha}{2}}\sigma}{e}\right)^2$$

Intervalo de confiança para a média - σ^2 conhecida



Exemplo 7 no R:

1. Uma empresa deseja estudar seus layouts na Web. Assim, coletaram quanto tempo as pessoas passam em cada uma das página A ou B. Construir um intervalo para o tempo médio de permanencia na página A, com 95% e 99% de confiança, sabendo que os tempos médios de permanência na página A seguem distribuição Normal com desvio-padrão 1.
2. No item anterior, qual o tamanho de amostra necessário para se obter um erro de 0.17 com 99% de confiança?

Exercícios no R:

1. Construir um IC para o tempo médio de permanência na página B, com 95% e 99% de confiança, sabendo que os tempos médios de permanencia na página B seguem distribuição Normal com desvio-padrão populacional de 0.5. Qual o tamanho de amostra necessário para se obter um erro de 0.15 com 99% de confiança?
   R: $IC(\mu, 0.95) = (1.8134, 2.1712)$, $IC(\mu, 0.99) = (1.7571, 2.2274)$, $n = 74$.
2. Considera-se que o tempo de permanência de recém-formados no 1º emprego, em anos, é Normal com variância populacional de 1.96. Para uma amostra de 15 recém-formados, a média obtida foi de 2.7 anos. Construa um IC com 90% de confiança para o tempo médio populacional de permanência de recém-formados no 1º emprego.
   R: $IC(\mu, 0.90) = (2.105421, 3.294579)$.

- Supondo $X \sim N(\mu, \sigma^2)$, com σ^2 desconhecido, devemos estimá-lo pelo S, o que ontroduz uma incerteza maior na estimação. Assim,

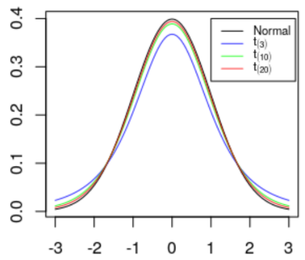
$$\frac{(\bar{X} - \mu)}{S/\sqrt{n}} \sim t_{n-1}$$

Distribuição t-student com parâmetro ν:

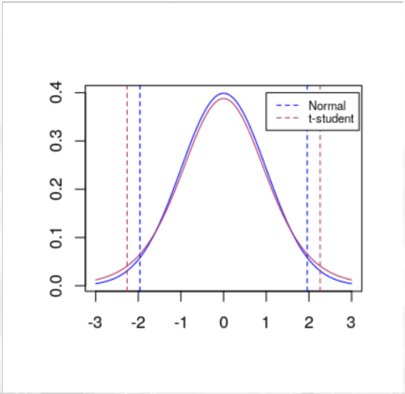
$$E(t) = 0$$
$$Var(t) = \frac{\nu}{\nu - 2}$$

Obs: Quando $\nu \to \infty$ então

$$t \to Normal$$



No R: dt(x, df) calcula a densidade da distribuição t com parâmetro df.



Assim, quando σ^2 desconhecida, o IC para μ pode ser obtido por:

$$P(-t_{((n-1),\frac{\alpha}{2})} \leq t \leq t_{((n-1),\frac{\alpha}{2})}) = 1 - \alpha$$

$$P(-t_{((n-1),\frac{\alpha}{2})} \leq \frac{(\bar{X} - \mu)}{S/\sqrt{n}} \leq t_{((n-1),\frac{\alpha}{2})}) = 1 - \alpha$$

$$P(\bar{X} - t_{((n-1),\frac{\alpha}{2})}\frac{S}{\sqrt{n}} \leq \mu \leq \bar{X} + t_{((n-1),\frac{\alpha}{2})}\frac{S}{\sqrt{n}}) = 1 - \alpha$$

$$IC(\mu, (1-\alpha)) = \left(\bar{X} - e; \bar{X} + e\right)$$

$$\text{onde} \quad e = t_{((n-1), \frac{\alpha}{2})} \frac{S}{\sqrt{n}}$$

Intervalo de confiança para a média μ - σ^2 desconhecida

Exemplo 8 no R:

O conjunto de dados income800.csv (disponível na página do DAEST), consiste de dados de renda anual de candidatos a empréstimos pessoais de uma empresa de empréstimo. Construa um IC com 95% de confiança para o verdadeiro rendimento anual médio.

Exercícios no R:

1. Das válvulas fabricadas por uma companhia retira-se uma amostra de 400 válvulas, e obtém-se a vida média de 800 horas e o desvio-padrão de 100 horas. Encontre o intervalo com 99% de confiança para a vida média das válvulas da população. **R:** $IC(\mu, 0.99) = (787.059, 812.941)$.
2. A resistência de determinado material está em estudo e sabe-se que essa variável é normalmente distribuída. Considerando a amostra abaixo, determine o IC para a resistencia média populacional com 90% de confiança. **R:** $IC(\mu, 0.90) = (5.5021, 6.9422)$.

4.9	7.0	8.1	4.5	5.6	6.8	7.2	5.7	6.2

Intervalo de confiança para a variância

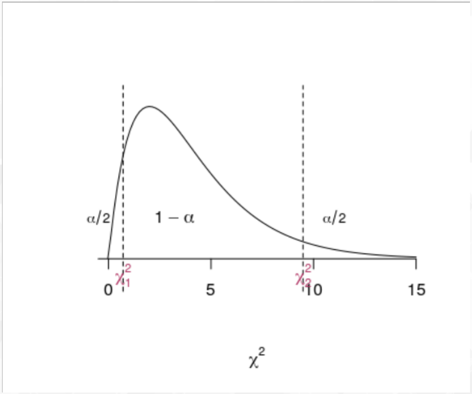
- Supondo $X \sim N(\mu, \sigma^2)$, com média e variância desconhecidos, o IC para a variância populacional é dado por:

$$IC(\sigma^2, (1-\alpha)) = \left(\frac{(n-1)S^2}{\chi_2^2}, \frac{(n-1)S^2}{\chi_1^2} \right)$$

onde, os valores χ_2^2 e χ_1^2 são tais que

$$P(\chi \leq \chi_1^2) = P(\chi \geq \chi_2^2) = \alpha/2$$

Intervalo de confiança para a variância



Figure

Exemplo 9 no R:

Sabe-se que as vendas diárias de um produto, em reais, seguem distribuição Normal. Uma amostra de dias de vendas, apresentou os valores:

275	258	242	229	225	134	227	249	259	191

Construa um intervalo com 90% de confiança para a variância populacional.

- Sabe-se que se $n \to \infty$, então $\hat{p} \sim N\left(p, \dfrac{p(1-p)}{n}\right)$
- Assim, o IC para a proporção fica:

$$IC(p, (1-\alpha)) = (\hat{p} - e; \hat{p} + e)$$

onde, $e = z_{\frac{\alpha}{2}} \sqrt{\dfrac{\hat{p}(1-\hat{p})}{n}}$

Exemplo 10 no R:

Considere o conjunto de dados "Melanoma" disponível no pacote "MASS" do R, que consiste de dados amostrais de 205 pacientes com melanoma maligno na Dinamarca. Considerando a variável sexo, encontre o IC para a proporção de pacientes do sexo feminino com 90% de confiança.

Exercícios no R:

1. Considerando ainda o conjunto de dados "Melanoma" do pacote "MASS", encontre o IC para a proporção de pacientes do sexo masculino com 99% de confiança.
   R: $IC(p, 0.99) = (0.2978, 0.4729)$.
2. Considerando a variável "ulcer" do conjunto de dados "Melanoma" do pacote "MASS", encontre o IC para a proporção de pacientes com presença de úlcera com 95% de confiança.
   R: $IC(p, 0.99) = (0.3710, 0.5069)$.

- **Hipótese Estatística:** é uma afirmação sobre um ou mais parâmetros da população;
- **Teste de Hipóteses:** é um conjunto de ferramentas estatísticas que nos permitem verificar se os dados amostrais trazem evidências que apoiem ou não a hipótese formulada.

- Seja μ a altura média de uma população (parâmetro) sobre o qual queremos verificar uma afirmação (testar uma hipótese).
- Queremos verificar se μ é igual a uma quantidade fixada, por exemplo, se μ é igual a 180. Então, formulamos as hipóteses:
  1. **Hipótese nula** (H_0): é a hipótese que estamos colocando à prova.
  2. **Hipótese alternativa** (H_1): é a hipótese que será considerada aceitável, caso a hipótese nula seja rejeitada.

Hipótese Bilateral:

$$H_0 : \mu = 180$$
$$H_1 : \mu \neq 180$$

Hipótese Unilateral à esquerda:

$$H_0 : \mu \geq 180$$
$$H_1 : \mu < 180$$

Hipótese Unilateral à direita:

$$H_0 : \mu \leq 180$$
$$H_1 : \mu > 180$$

- Qualquer que seja a decisão tomada, estamos sujeitos a cometer erros.
- **Erro de tipo I:** Rejeitar a hipótese nula quando essa é verdadeira.

$$\alpha = P(\text{Erro Tipo I}) = P(\text{Rejeitar } H_0 | H_0 \text{ é verdadeira})$$

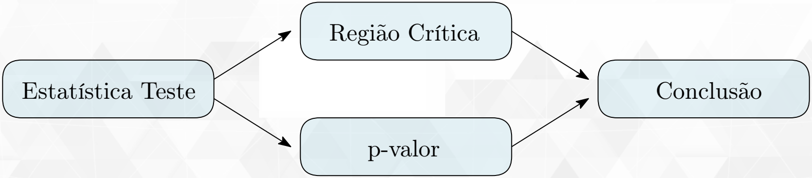
- **Erro de tipo II:** Não rejeitar a hipótese nula quando essa é falsa.

$$\beta = P(\text{Erro Tipo II}) = P(\text{Não rejeitar } H_0 | H_0 \text{ é falsa})$$

- Considerando um tamanho fixo da amostra, não é possível controlar simulataneamente ambos os erros.
- Estabeleceu-se em fixar α. Em seguida, calcular o tamanho da amostra que reduza a β.

Situação real	Conclusão do teste	
	Rejeitar H_0	Não rejeitar H_0
H_0 verdadeira	Erro tipo I	decisão correta
H_0 falsa	decisão correta	Erro tipo II

- Geralmente, fixa-se α (nível de significância) em 10%, 5%, 1%.


Estatística Teste
Região Crítica
p-valor
Conclusão

Hipótese Bilateral:

$$H_0 : \mu = \mu_0$$
$$H_1 : \mu \neq \mu_0$$

Hipótese Unilateral à esquerda:

$$H_0 : \mu \geq \mu_0$$
$$H_1 : \mu < \mu_0$$

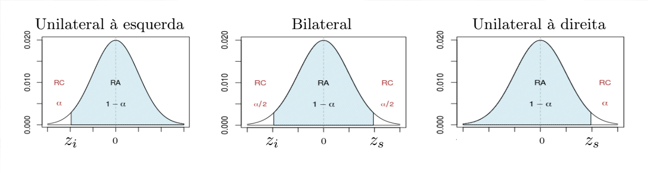
Hipótese Unilateral à direita:

$$H_0 : \mu \leq \mu_0$$
$$H_1 : \mu > \mu_0$$

$$X \sim N(\mu, \sigma^2) \longrightarrow \sigma^2 \text{ conhecida}$$

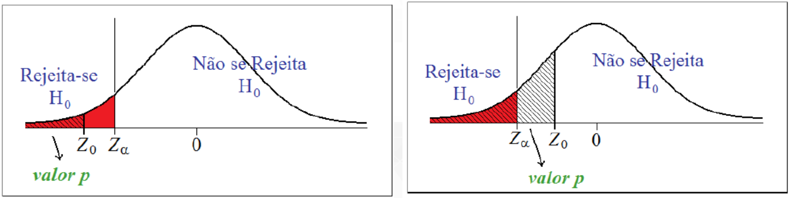
$$\downarrow$$

$$Z_0 = \frac{\bar{X} - \mu_0}{\sigma/\sqrt{n}} \sim N(0,1)$$



Se $Z_0 \in RC$, rejeitamos H_0, ao nível α de significância.

- **p-valor:** é a probabilidade de se observar valores mais extremos do que o que foi observado, supondo que a hipótese H_0 é verdadeira.
- **Teste unilateral à direita:** p-valor $= P(Z > Z_0)$.
- **Teste unilateral à esquerda:** p-valor $= P(Z < Z_0)$.
- **Teste bilateral:** p-valor $= 2P(Z > |Z_0|)$.

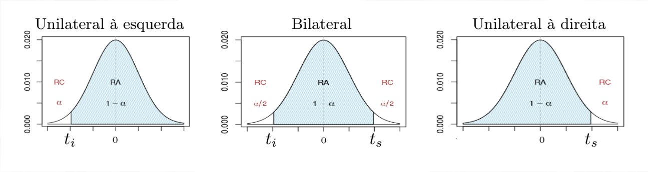


Se p-valor $\leq \alpha$, rejeitamos H_0, ao nível α de significância.

$$X \sim N(\mu, \sigma^2) \longrightarrow \sigma^2 \text{ desconhecida}$$

$$\downarrow$$

$$T_0 = \frac{\bar{X} - \mu_0}{S/\sqrt{n}} \sim t_{(n-1)}$$



Se $T_0 \in RC$, rejeitamos H_0, ao nível α de significância.

- **Teste unilateral à direita:** p-valor $= P(T > T_0)$.
- **Teste unilateral à esquerda:** p-valor $= P(T < T_0)$.
- **Teste bilateral:** p-valor $= 2P(T > |T_0|)$.

UTFPR UNIVERSIDADE QUE TRANSFORMA

DAEST DEPARTAMENTO ACADÊMICO DE ESTATÍSTICA

UTFPR UNIVERSIDADE TECNOLÓGICA FEDERAL DO PARANÁ

MINISTÉRIO DA EDUCAÇÃO

PÁTRIA AMADA BRASIL GOVERNO FEDERAL

Exemplo 11 no R:

Considere o conjunto de dados income800.csv (disponível na página do DAEST), que consiste de dados de renda anual de candidatos a empréstimos pessoais de uma empresa de empréstimo. Testar se o redimento médio populacional é de 70000 ao nível de 5% de significância.

Hipótese Bilateral:

$$H_0 : \sigma^2 = \sigma_0^2$$
$$H_1 : \sigma^2 \neq \sigma_0^2$$

Hipótese Unilateral à esquerda:

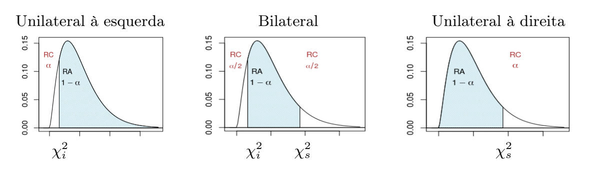
$$H_0 : \sigma^2 \geq \sigma_0^2$$
$$H_1 : \sigma^2 < \sigma_0^2$$

Hipótese Unilateral à direita:

$$H_0 : \sigma^2 \leq \sigma_0^2$$
$$H_1 : \sigma^2 > \sigma_0^2$$

$$X \sim N(\mu, \sigma^2)$$

$$\chi_0^2 = \frac{(n-1)S^2}{\sigma_0^2} \sim \chi^2_{(n-1)}$$



Se $\chi_0^2 \in RC$, rejeitamos H_0, ao nível α de significância.

EXEMPLO 12 NO R:

Considere novamente o conjunto de dados income800.csv (disponível na página do DAEST). Testar se a variabilidade populacional é de 10^9 ao nível de 5% de significância.

Exercícios no R:

Considere o conjunto de dados da web (disponível na página página do DAEST), supondo agora variâncias populacionais desconhecidas.

1. Testar se o tempo médio de permanência na página A é de 1.5 ao nível de 1% de significância. **R:** $t_0 = 3.1265$, p-value $= 0.0029$, Rejeita H_0.
2. Testar se a variabilidade do tempo de permanência na página A é menor ou igual a 0.6 ao nível de 5% de significância. **R:** $\chi_0 = 61.406$, p-value $= 0.1099$. Não rejeita H_0.

Teste de Hipótese para comparar dois grupos

População 1

$X_1 \sim N(\mu_1, \sigma_1^2)$

n_1

Amostra de P1

$\bar{X}_1 \quad S_1^2$

População 2

$X_2 \sim N(\mu_2, \sigma_2^2)$

n_2

Amostra de P2

$S_2^2 \quad \bar{X}_2$

independentes

UTFPRUNIVERSIDADE QUETRANSFORMA

DAEST DEPARTAMENTO ACADÊMICO DE ESTATÍSTICA

UTFPR UNIVERSIDADE TECNOLÓGICA FEDERAL DO PARANÁ

MINISTÉRIO DA EDUCAÇÃO

PÁTRIA AMADA BRASIL GOVERNO FEDERAL

- Teste para comparar as variâncias.
- Teste para comparar as médias, quando as variâncias são iguais.
- Teste para comparar as médias, quando as variâncias são diferentes.

Hipótese Bilateral:

$$H_0 : \sigma_1^2 = \sigma_2^2$$
$$H_1 : \sigma_1^2 \neq \sigma_2^2$$

Hipótese Unilateral à esquerda:

$$H_0 : \sigma_1^2 \geq \sigma_2^2$$
$$H_1 : \sigma_1^2 < \sigma_2^2$$

Hipótese Unilateral à direita:

$$H_0 : \sigma_1^2 \leq \sigma_2^2$$
$$H_1 : \sigma_1^2 > \sigma_2^2$$

População 1

$X_1 \sim N(\mu_1, \sigma_1^2)$

n_1

Amostra de P1

$S_1^2 \quad \bar{X}_1$

População 2

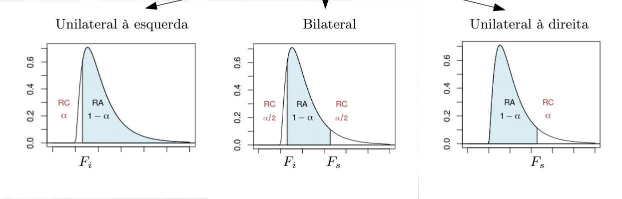
$X_2 \sim N(\mu_2, \sigma_2^2)$

n_2

Amostra de P2

$S_2^2 \quad \bar{X}_2$

$$F_0 = \frac{S_1^2}{S_2^2} \sim F_{(n_1-1, n_2-1)}$$



Hipótese Bilateral:

$$H_0 : \mu_1 = \mu_2$$
$$H_1 : \mu_1 \neq \mu_2$$

Hipótese Unilateral à esquerda:

$$H_0 : \mu_1 \geq \mu_2$$
$$H_1 : \mu_1 < \mu_2$$

Hipótese Unilateral à direita:

$$H_0 : \mu_1 \leq \mu_2$$
$$H_1 : \mu_1 > \mu_2$$

População 1

$X_1 \sim N(\mu_1, \sigma^2)$

n_1

Amostra de P1

$\bar{X}_1 \quad S_1^2$

População 2

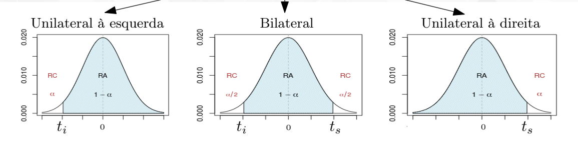
$X_2 \sim N(\mu_2, \sigma^2)$

n_2

Amostra de P2

$S_2^2 \quad \bar{X}_2$

$$t_0 = \frac{\bar{X}_1 - \bar{X}_2}{S_p\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} \sim t_{(n_1+n_2-2)} \longrightarrow S_p = \sqrt{\frac{(n_1-1)S_1^2 + (n_2-1)S_2^2}{n_1+n_2-2}}$$



População 1

$X_1 \sim N(\mu_1, \sigma_1^2)$

n_1

Amostra de P1

$\bar{X}_1 \quad S_1^2$

População 2

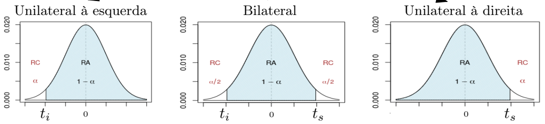
$X_2 \sim N(\mu_2, \sigma_2^2)$

n_2

Amostra de P2

$S_2^2 \quad \bar{X}_2$

$$t_0 = \frac{\bar{X}_1 - \bar{X}_2}{\sqrt{\dfrac{S_1^2}{n_1} + \dfrac{S_2^2}{n_2}}} \sim t_{(v)} \longrightarrow v = \frac{\left(\dfrac{S_1^2}{n_1} + \dfrac{S_2^2}{n_2}\right)^2}{\dfrac{\left(\frac{S_1^2}{n_1}\right)^2}{n_1 - 1} + \dfrac{\left(\frac{S_2^2}{n_2}\right)^2}{n_2 - 1}}$$



Exemplo 13 no R:

Considere o conjunto de dados durabilidade.txt (disponível na página do DAEST), que consiste da durabilidade em dias de aparelhos eletrônicos produzidos por duas empresas A e B.

1. Testar se as variâncias populacionais são iguais ao nível de 5% de significância.
2. Testar se as durabilidades médias populacionais são iguais ao nível de 5% de significância.

Exercícios no R:

Considere o conjunto de dados notas.txt (disponível na página do DAEST), que consiste das notas estatística de duas turmas 1 e 2.

1. Testar se as variâncias populacionais são iguais ao nível de 5% de significância. **R:** $F_0 = 1.9642$, p-value = 2.898e-06.
2. Testar se as notas médias populacionais são iguais ao nível de 5% de significância. **R:** $t_0 = 3.314$, df = 268.21, p-value = 0.001046.

Hipótese Bilateral:

$$H_0 : p = p_0$$
$$H_1 : p \neq p_0$$

Hipótese Unilateral à esquerda:

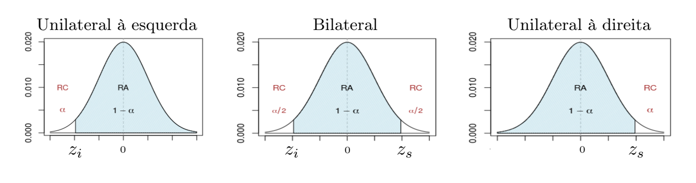
$$H_0 : p \geq p_0$$
$$H_1 : p < p_0$$

Hipótese Unilateral à direita:

$$H_0 : p \leq p_0$$
$$H_1 : p > p_0$$

$$X \sim Bernoulli(p)$$

$$Z_0 = \frac{\hat{p} - p_0}{\sqrt{\frac{p_0(1-p_0)}{n}}} \sim N(0,1)$$



EXEMPLO 14 NO R:

Considere a variável sexo do conjunto de dados notas.txt (disponível na página do DAEST). Testar ao nível de 5% de significância se a proporção de mulheres é igual a 60%.