Inferência estatística

Murillo F. Rodrigues

Estimação

O que é?

- Objetivo: fazer generalizações sobre uma população
- Parâmetros populacionais: média, proporção, …
- Exemplos:

μ - média da característica da população:

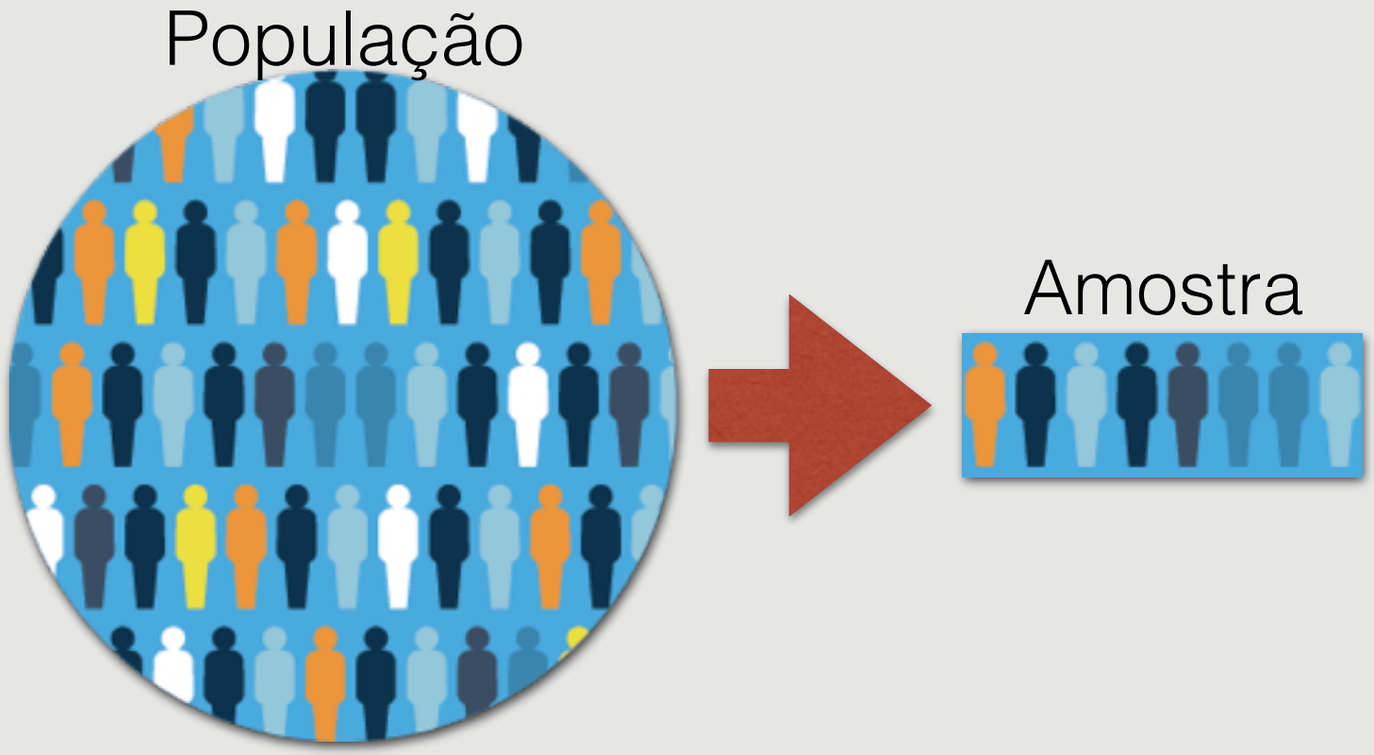
μ: taxa média de glicose de mulheres com idade superior a 60 anos, em certa localidade;

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

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

O que é?

- Variável de interesse X



- Com os elementos da amostra podemos estimar uma característica ou parâmetro populacional

- Estimador = estatística = função dos elementos da amostra que representa a característica de interesse

O que é?

- Estimador = estatística = função dos elementos da amostra que representa a característica de interesse

- Exemplo:

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

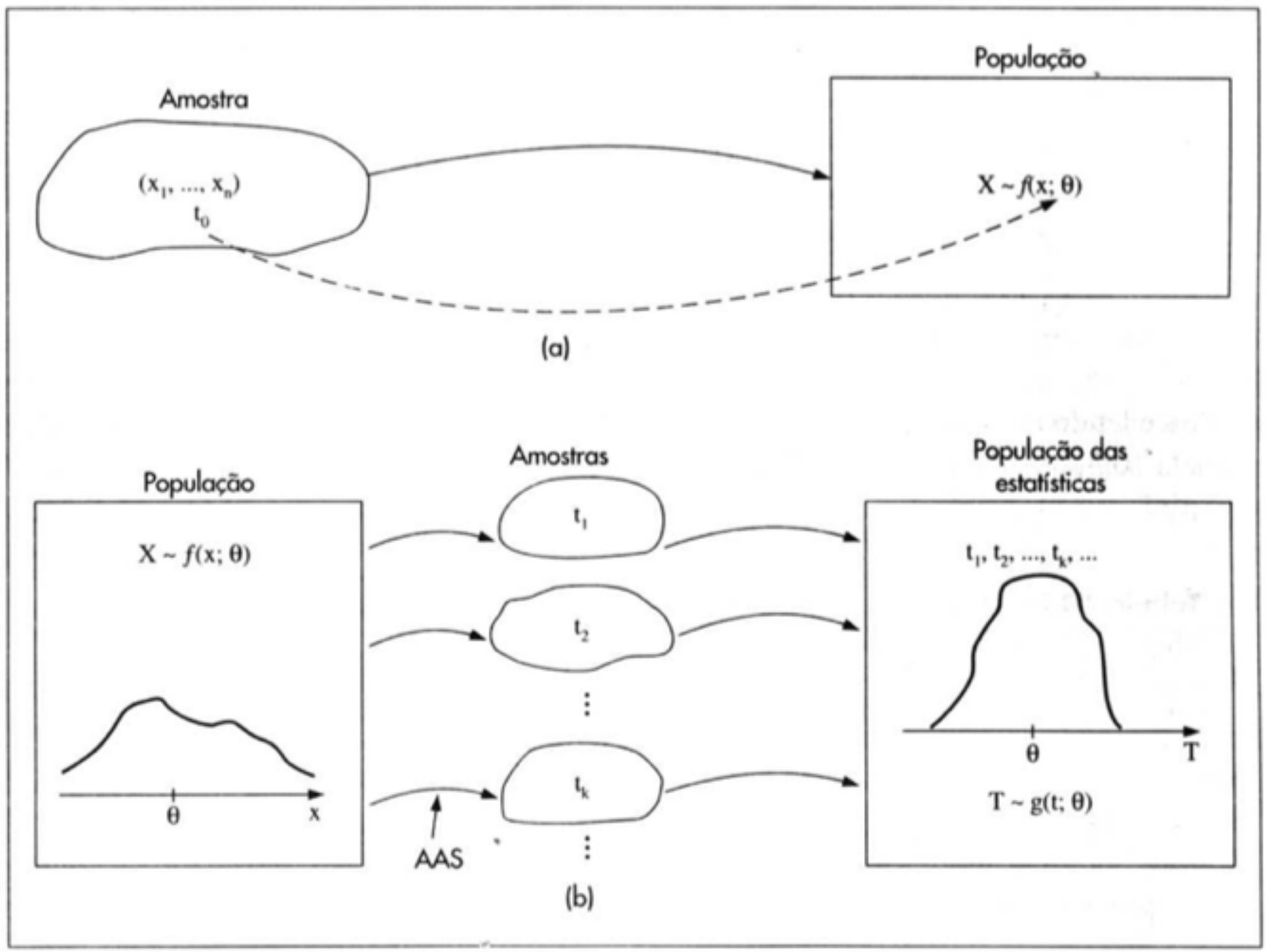
$\hat{p}$: proporção amostral (estimador da proporção p populacional).

- Estimativa é o valor assumido pelo estimador para a sua amostra

$\overline{x}$ é o valor de $\overline{X}$ para a amostra observada.

Distribuição amostral do estimador

Figura 10.1: (a) Esquema de inferência sobre θ.
(b) Distribuição amostral da estatística T.



Estimação da média

Estimação da média

- Objetivo: estimar a média populacional de uma variável X, a partir de uma amostra de valores de X
- Possíveis procedimentos:
  - Estimação pontual
  - Estimação intervalar

Estimação da média

- Estimador pontual para média populacional

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

- Estimativa pontual

$$\bar{x} = \frac{x_1 + \ldots + x_n}{n}$$

Estimação da média

- Estimativa intervalar ou intervalo de confiança
  - Estimadores pontuais são variáveis aleatórias e possuem **distribuições de probabilidade**
  - **Podemos incorporar essa incerteza na nossa estimativa?**

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

- **Como é a distribuição de probabilidade de uma média???**

Parênteses

- Erro padrão da média amostral

$$SE = \sqrt{\frac{\sigma^2}{n}}$$

$$SE = \sqrt{\frac{S^2}{n}}$$

Parênteses

- Z-score

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

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

Estimação da média

- Estimativa intervalar ou intervalo de confiança

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

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

Parênteses

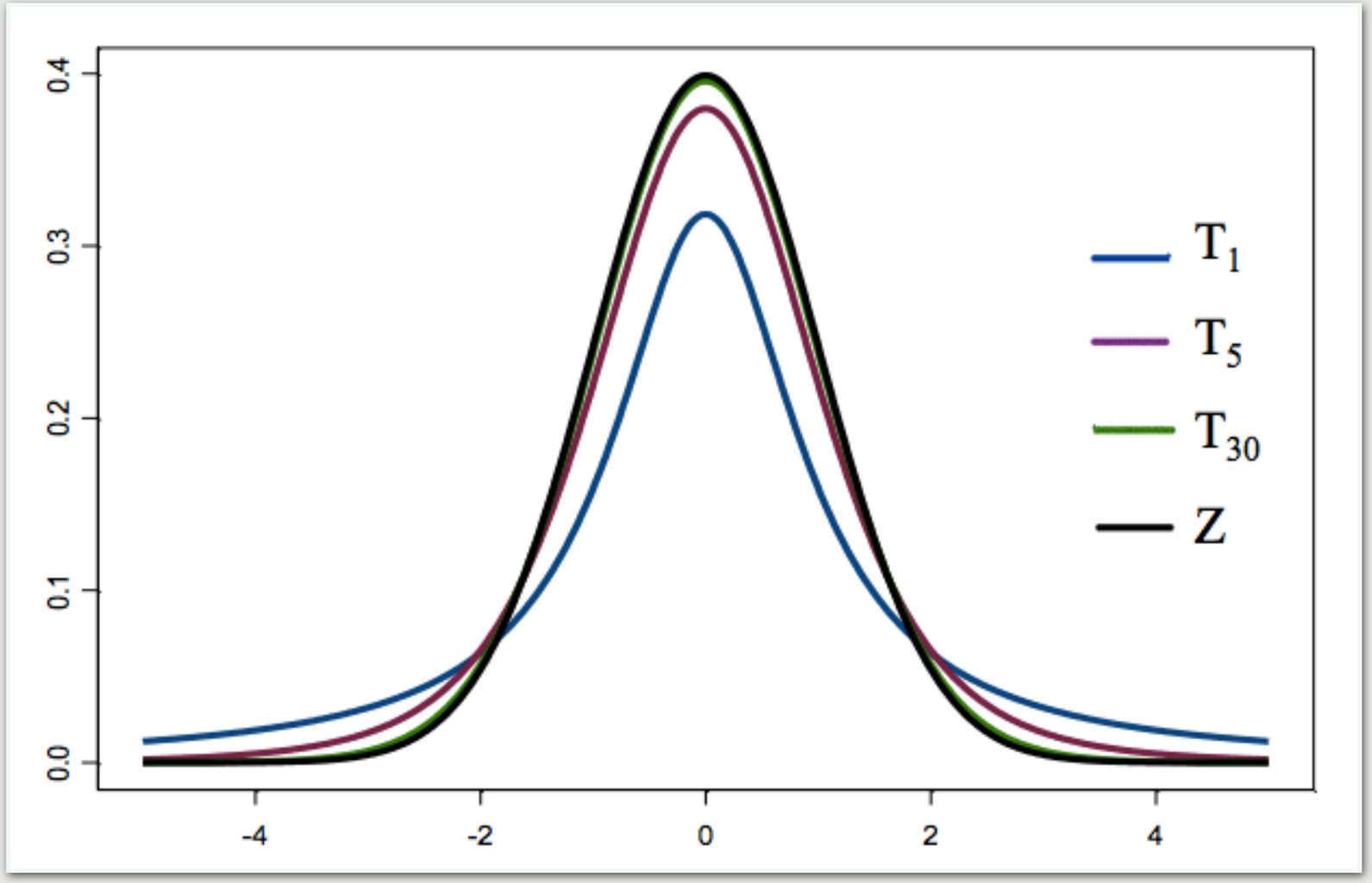
- Variância desconhecida = Estatística T

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

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

Parênteses

- Distribuição T



Estimação da média

- Estimativa intervalar ou intervalo de confiança

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

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

$$\varepsilon = t^{c}_{n-1}\sqrt{\frac{S^2}{n}}$$

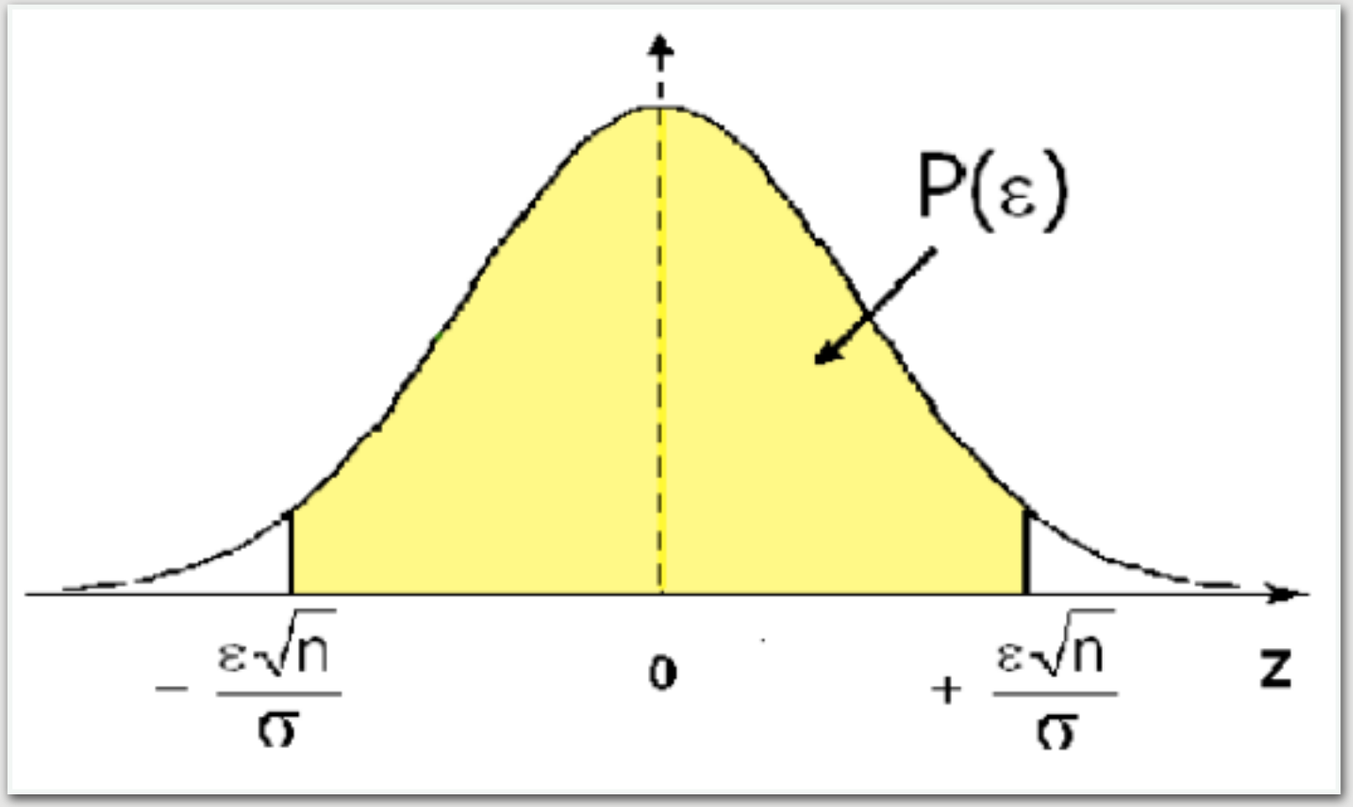
$$IC(\mu;\gamma) = \left[\bar{X} - t^{c}_{n-1}\sqrt{\frac{S^2}{n}}; \bar{X} + t^{c}_{n-1}\sqrt{\frac{S^2}{n}}\right]$$

Estimação da média

- Estimativa intervalar ou intervalo de confiança

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

$$\varepsilon = t^{c}_{n-1}\sqrt{\frac{S^2}{n}}$$



Estimação da média

- Estimativa intervalar ou intervalo de confiança

Como interpretar ???

Estimação da média

- Estimativa intervalar ou intervalo de confiança

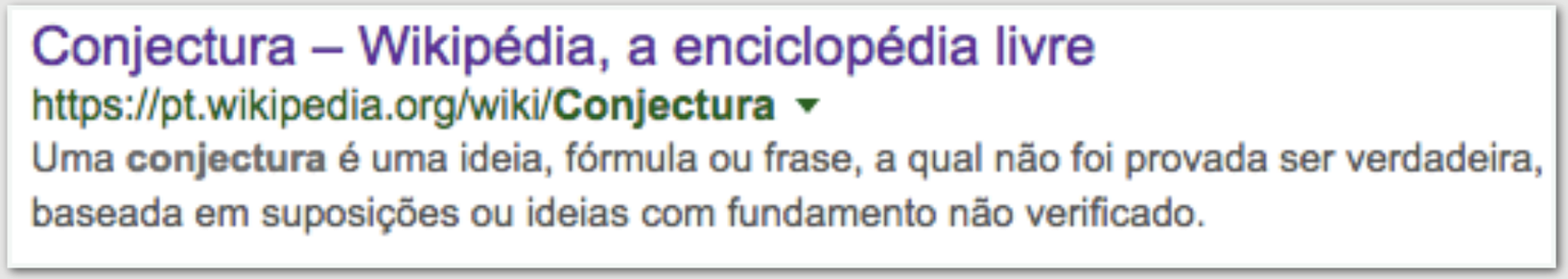
Como interpretar ???

Vamos usar o R para ilustrar a interpretação!

Teste de Hipóteses

O que é uma hipótese?

- É uma conjectura sobre um <u>parâmetro populacional</u>



- Exemplo:

- *"Eu acho que homens têm altura média maior que mulheres"*

Exemplo de hipótese

- Hipótese nula: homens e mulheres têm mesma altura média

- Hipótese alternativa: homens são, em média, mais altos que mulheres

- Seja X a variável altura dos homens e Y altura das mulhers:

  - $H_0 : \mu_X = \mu_Y$

  $H_1 : \mu_X > \mu_Y$

Exemplo de hipótese

- Seja X a variável altura dos homens e Y altura das mulhers:
  - $H_0 : \mu_X - \mu_Y = 0$
  - $H_1 : \mu_X - \mu_Y > 0$

- <u>Vamos ver se a altura média dos homens e mulheres dessa sala diferem!</u>

Como eu decido?

- Estatística T

$$T = \frac{\bar{X} - \mu}{\frac{S}{\sqrt{n}}}$$

$$T = \frac{\bar{X} - \bar{Y}}{\sqrt{S_p^2\left(\frac{1}{n} + \frac{1}{m}\right)}}$$

$$s_p^2 = \frac{(n-1)s_X^2 + (m-1)s_Y^2}{n+m-2}.$$

- Nível de significância $\alpha = 0.05$
- Evidência amostral $(n, \bar{x}, s)$

P-valor

- Determinar o nível descritivo. Como?

$$H_0 : \mu_X - \mu_Y = 0$$

$$H_1 : \mu_X - \mu_Y > 0$$

$$P = P(T >= T_{obs})$$

- Decisão: se P > alpha, então não rejeitamos H0!

RESUMO
Teste de hipóteses para a média populacional μ
(via nível descritivo)

(0) Descrever o **parâmetro** de interesse μ.

(1) Estabelecer as **hipóteses**:

H_0: $\mu = \mu_0$ contra uma das alternativas

H_1: $\mu \neq \mu_0$, H_1: $\mu > \mu_0$ ou H_1: $\mu < \mu_0$.

(2) Escolher a **Estatística de teste:**

$$Z = \sqrt{n}\,\frac{\bar{X} - \mu}{\sigma} \quad \text{ou} \quad T = \sqrt{n}\,\frac{\bar{X} - \mu}{S}$$

(3) Escolher um **nível de significância** α.

(4) Selecionar uma **amostra** casual simples de tamanho n
$\Rightarrow$ determinar a média amostral $\bar{x}_{obs}$ e o desvio padrão (populacional σ ou amostral s) .

(5B) Determinar o **nível descritivo** P

Se $H_1: \mu > \mu_0$, $P = P(Z \geq z_{obs})$ ou $P(T \geq t_{obs})$

Se $H_1: \mu < \mu_0$, $P = P(Z \leq z_{obs})$ ou $P(T \leq t_{obs})$

Se $H_1: \mu \neq \mu_0$, $P = 2 \times P(Z \geq |z_{obs}|)$ ou $2 \times P(T \geq |t_{obs}|)$

(6) **Decidir**, comparando P com o nível de significância α, e **concluir.**

Se $P \leq \alpha \Rightarrow$ rejeitamos H_0
Se $P > \alpha \Rightarrow$ não rejeitamos H_0

ANOVA

- Análise de Variância (Analysis of Variance)
- Comparação simultânea de médias de vários grupos

$$H_0 : \mu_1 = ... = \mu_k$$

$$H_1 : \mu_i \neq \mu_j$$

ANOVA

- Supondo *k* grupos, teríamos:

Grupo 1	Grupo 2	. . .	Grupo *k*	
y_{11}	y_{21}		y_{k1}	
y_{12}	y_{22}		y_{k2}	$\bar{y}_{k.}$
.	.		.	
.	.		.	
.	.		.	
y_{1n1}	y_{2n2}		y_{knk}	$s_{k.}$

- Termos importantes:
  - fator: critério de classificação (tratamentos)
  - nível: cada classificação ou grupo

ANOVA

- **Modelo estatístico:**

OBSERVAÇÃO = SISTEMÁTICA + ALEATÓRIA

- **Componente sistemático (previsível) : incorpora conhecimento que o pesquisador tem sobre o fenômeno**

- **Componente aleatório: representa as variações individuais que não são explicadas pela parte sistemática do modelo. Também conhecido como erro aleatório ou resíduo**

ANOVA

Modelo estatístico (1):

$$Y_{ij} = \mu_i + e_{ij}, \quad i = 1,\ldots, k, j = 1,\ldots, n_i,$$

com μ_i: média de Y no nível i (efeito do nível i),

e_{ij}: efeito aleatório do j-ésimo indivíduo do nível i,

Y_{ij}: variável resposta do j-ésimo indivíduo do nível i.

Suposição:

e_{ij} ~ normais, independentes com média 0 e variância σ^2.

Se a hipótese H_0 **for verdadeira**, o modelo pode ser reescrito:

Modelo estatístico (0):

$$Y_{ij} = \mu + e^*_{ij}, \quad i = 1,\ldots, k, j = 1,\ldots, n_i .$$

ANOVA

Variabilidade **Total**	=	Variabilidade **entre grupos**	+	Variabilidade **dentro dos grupos**

ANOVA

- Como calculamos aquelas variabilidades?

$$SQD = \sum_{i=1}^{k} \sum_{j=1}^{n_i} (y_{ij} - \bar{y}_{i.})^2$$

$$QMD = \frac{SQD}{n-k}$$

$$SQE = \sum_{i=1}^{k} n_i (\bar{y}_{i.} - \bar{y}_{..})^2$$

$$QME = \frac{SQE}{k-1}$$

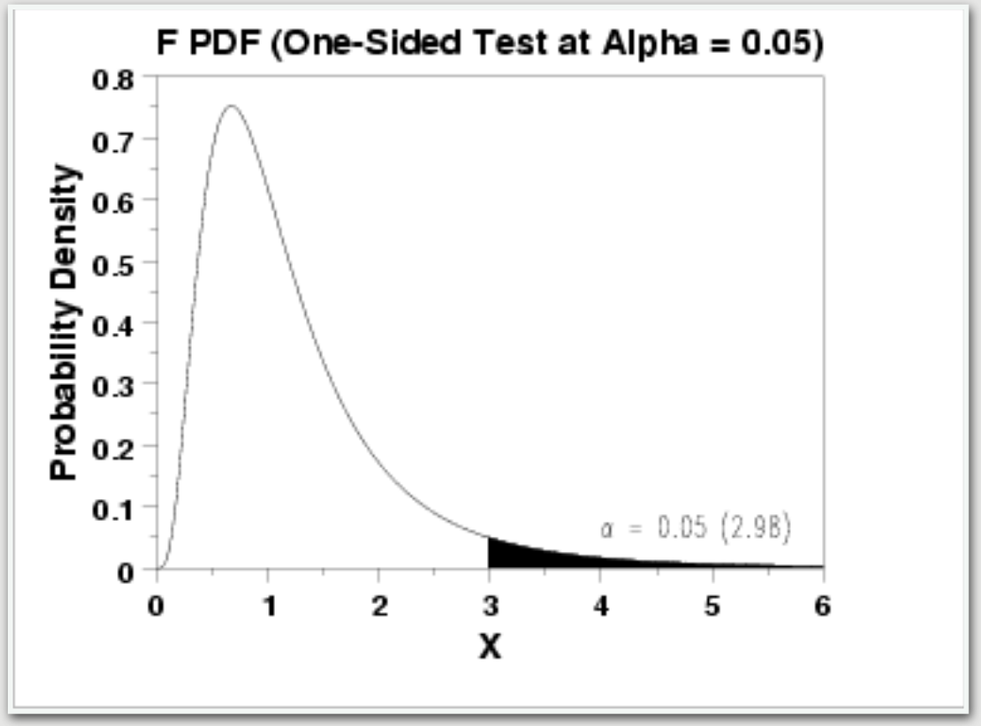
ANOVA

- Mas… como eu decido?

- Queremos comparar se a variação entre grupos é maior que a variação dentre grupos!

- Estatística F

$$F = \frac{QME}{QMD}$$



ANOVA

- Tudo resumido na tabela de ANOVA

Fonte de Variação	*g.l.*	Soma de Quadrados	Quadrado Médio	Teste F
Entre grupos	k - 1	SQE	$QME = SQE/(k - 1)$	$F = QME/QMD$
Dentro de grupos	$n - k$	SQD	$QMD = SQD/(n - k)$	
Total	n - 1	SQT		

com $F \sim F_{(k-1,\ n-k)}$

$$SQE = \sum_{i=1}^{k} n_i (\bar{y}_i - \bar{y})^2$$

$$SQD = \sum_{i=1}^{k} \sum_{j=1}^{n_i} (y_{ij} - \bar{y}_i)^2$$

$$SQT = \sum_{i=1}^{k} \sum_{j=1}^{n_i} (y_{ij} - \bar{y})^2$$

OBS.: $SQT = SQE + SQD$

ANOVA

- Tudo resumido na tabela de ANOVA

Fonte de Variação	*g.l.*	Soma de Quadrados	Quadrado Médio	Teste F
Entre grupos	k - 1	SQE	$QME = SQE/(k - 1)$	$F = QME/QMD$
Dentro de grupos	$n - k$	SQD	$QMD = SQD/(n - k)$	
Total	n - 1	SQT		

com $F \sim F_{(k-1,\ n-k)}$

$$SQE = \sum_{i=1}^{k} n_i (\bar{y}_i - \bar{y})^2$$

$$SQD = \sum_{i=1}^{k} \sum_{j=1}^{n_i} (y_{ij} - \bar{y}_i)^2$$

$$SQT = \sum_{i=1}^{k} \sum_{j=1}^{n_i} (y_{ij} - \bar{y})^2$$

OBS.: $SQT = SQE + SQD$

$$R^2 = \frac{SQE}{SQT}$$

Análise de Regressão Linear

Regressão Linear

- É uma ANOVA quando meu fator é uma variável contínua!

ANOVA

- Supondo *k* grupos, teríamos:

Grupo 1	Grupo 2	...	Grupo *k*	
y_{11}	y_{21}		y_{k1}	$\bar{y}_{k.}$
y_{12}	y_{22}		y_{k2}	$s_{k.}$
.	.		.	
.	.		.	
.	.		.	
y_{1n1}	y_{2n2}		y_{knk}	

- Termos importantes:
  - fator: critério de classificação (tratamentos)
  - nível: cada classificação ou grupo

Regressão Linear

- Modelo

O modelo estatístico para esta situação seria:

$$Y_i = \beta_0 + \beta_1 X_i + e_i$$

em que:

Y_i = valor observado para a variável dependente Y no i-ésimo nível da variável independente X.

β_0 = constante de regressão. Representa o intercepto da reta com o eixo dos Y.

β_1 = coeficiente de regressão. Representa a variação de Y em função da variação de uma unidade da variável X.

X_i = i-ésimo nível da variável independente X $(i = 1,2,\ldots,n)$

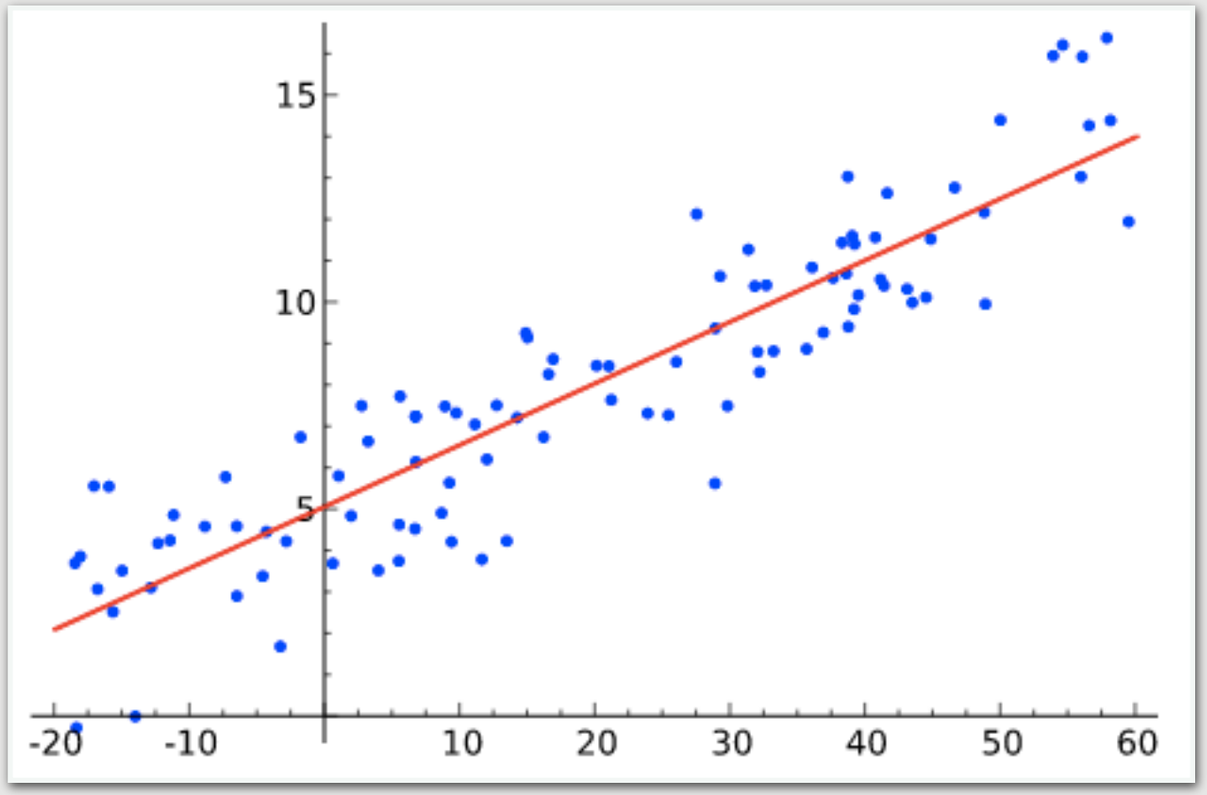
e_i = é o erro que está associado à distância entre o valor observado Y_i e o correspondente ponto na curva, do modelo proposto, para o mesmo nível i de X.

Regressão Linear

- Estimando os parâmetros

$$\hat{\alpha} = \bar{y} - \hat{\beta}\bar{x}$$

$$\hat{\beta} = \frac{cov(x, y)}{var(x)}$$



Regressão Linear

- Tabela de ANOVA

Fonte de Variação	*g.l.*	Soma de Quadrados	Quadrado Médio	Teste *F*
Regressão	1	*SQReg*	*QMReg = SQReg*	*F= QMReg/QMRes*
Resíduo	*n - 2*	*SQRes*	*QMRes = SQRes/(n - 2)*	
Total	*n - 1*	*SQT*		

$$SQT = \sum_{i=1}^{n} (y_i - \bar{y})^2$$

$$SQRes = \sum_{i=1}^{n} \left[y_i - (\hat{\alpha} + \hat{\beta} x_i) \right]^2 = \sum_{i=1}^{n} (y_i - \hat{y}_i)$$

Teste Qui-quadrado

Aderência

Qui-quadrado de aderência

Categorias	Frequência observada	Frequência esperada, sob H_0
1	O_1	E_1
2	O_2	E_2
3	O_3	E_3
⋮	⋮	⋮
k	O_k	E_k
Total	n	n

Qui-quadrado de aderência

Categorias	Frequência observada	Frequência esperada, sob H_0
1	O_1	E_1
2	O_2	E_2
3	O_3	E_3
⋮	⋮	⋮
k	O_k	E_k
Total	n	n

$$\chi^2 = \sum_{i=1}^{k} \frac{(O_i - E_i)^2}{E_i}$$

Qui-quadrado de aderência

Supondo H_0 verdadeira,

$$\chi^2 = \sum_{i=1}^{k} \frac{(O_i - E_i)^2}{E_i} \sim \chi^2_q, \quad \text{aproximadamente,}$$

sendo que $q = k - 1$ representa o número de graus de liberdade.