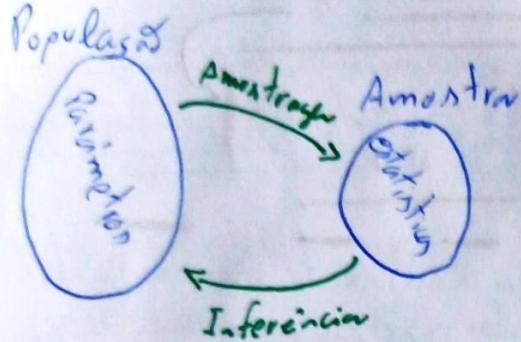


# Estatística Descritiva e Análise Exploratória de Dados

População: conjunto de elementos sobre o qual incide o estudo  
 ↳ tem parâmetros



Amostra: conjunto de elementos extraídos da população

Sondagem: estudo que incide na amostra

Censo: estudo que incide na população

Variável: característica em estudo

Amostragem ↳ Probabilística - aleatória  
 ↳ Não Probabilística - amostras por conveniência

Notações:

- População  $\boxed{X}$

- Amostra de  $n$  elementos

$$\hookrightarrow \boxed{(x_1, \dots, x_n)}$$

- Amostra de  $n$  elementos ordenada

$$\hookrightarrow \boxed{(x_{(1)}, \dots, x_{(n)})}$$

$$\boxed{(x_i), i=1, \dots, n}$$

Tipos de Dados/Variáveis ↳ Qualitativos - ex: sexo  
 ↳ Quantitativos - ex: peso

Tipos de Variáveis ↳ Discretas - valores inteiros  
 ↳ Contínuas - valores reais

Dados Qualitativos ↳ Nominais  
 ↳ Binários - ex: Macho/Fêmea  
 ↳ Ordinais - têm ordem

Dados Quantitativos ↳ Intervalares - não ex: 0 a 100  
 ↳ Relacionais - ex: tempo  
 ↳ Percentuais - ex: peso

Nota: Variáveis qualitativas são sempre discretas

Tabulacões

- Quando as variáveis são qualitativas organizam-se numa tabela de frequências

Frequência Absoluta - nº absoluto  $n_i / F_i$  de elementos

Frequência Relativa - Proporção  $f_i = \frac{n_i}{n} \times 100\%$

Frequência Acumulada ↳ Absolute  $\sum_{i=1}^n n_i / F_i$   
Relativa  $\bar{f}_i = \frac{\sum_{i=1}^n f_i}{n}$

↳ Somar/Accumular as frequências ordenadas

Moda - elemento mais frequente

Medidas de dispersão

Mediana - divide a amostra em partes iguais  
( $m$ )

$$M_z = \begin{cases} X\left(\frac{n+1}{2}\right) & , n \text{ é ímpar} \\ \frac{X(n_2) + X(\frac{n+1}{2})}{2} & , n \text{ é par} \end{cases}$$

Média - medida tendência

$$\bar{x} = \frac{\sum_{i=1}^n x_i}{n}$$

Quantil

$$Q_p = \begin{cases} X([np] + 1) & , np \notin \mathbb{N} \\ \frac{X(np) + X(np+1)}{2} & , np \in \mathbb{N} \end{cases}$$

$$\hookrightarrow [0 < p < 1]$$

Quartil

↳ dividem a distribuição de frequências em 4 partes iguais

Decil

↳ dividem em 10 partes iguais

Percentil

↳ dividem em 100 partes iguais

Variância ( $s^2$ )

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

$$s^2 = \frac{\sum_{i=1}^n x_i^2}{n-1} - \frac{n}{n-1} \cdot \bar{x}^2$$

Nota: muda a unidade de medida;  
mesma do quadrado

Desvio padrão ( $s$ )

$$s = \sqrt{s^2}$$

$s < \bar{x}$  → variabilidade reduzida  
 $s > \bar{x}$  → variabilidade elevada

Amplitude amostral ( $R$ )

$$R = \underbrace{X(n)}_{\text{máximo}} - \underbrace{X(1)}_{\text{mínimo}}$$

Amplitude interquartil (AIQ)

$$AIQ = Q_{\frac{3}{2}} - Q_{\frac{1}{2}}$$

Coeficiente de variação (CV)

$$CV = \frac{s}{\bar{x}} \times 100\%$$

\* Primeira folha

↳ medida de dispersão relativa  
para comparar a variabilidade  
de amostra em unidades distintas

## \* Continuação

	Variabilidade
<u>CV ≤ 15%</u>	<u>Fixa</u>
<u>15 &lt; CV ≤ 30%</u>	<u>Média</u>
<u>CV &gt; 30%</u>	<u>Elevada</u>

## Diagrama de espalhamento e quartílio

↳ Representação gráfica de variáveis quantitativas



## Barreira de outliers - intervalo de

valores que permite classificar uma observação como sendo outlier

$$B.I. = Q_{\frac{1}{4}} - 1,5 \times AIQ$$

$$B.S. = Q_{\frac{3}{4}} + 1,5 \times AIQ$$

$$[B.I., B.S.] \quad \swarrow \quad \searrow$$

todos os observações forem dentro da barreira são outliers

• Min e Max são os primeiros que não são outliers

## Tipos de amostras

→ → Simétrica

↳ Moda = Mediana = Média

→ → enviesada à direita

→ → enviesada à esquerda

## Dados agrupados (variável discreta)

### Média

$$\bar{x} = \frac{\sum c_i \cdot f_i}{n}$$

→  $\begin{cases} c_i - \text{classe} \\ k - n: \text{de } c_i \end{cases}$

↳ Moda: maior frequência absoluta

↳ Mediana: 1ª frequência relativa acumulada a exceder 50%

### Variância

$$s^2 = \frac{\sum c_i^2 \cdot f_i}{n-1} - \frac{\bar{x}}{n-1} \cdot \frac{\bar{x}^2}{n-1}$$

# Teoria das Probabilidades

## Acontecimento

- ↳ elementar: 1 resultado
- ↳ impossível: nunca acontece
- ↳ certo: acontece todos os dias
- ↳ simples: ex: sair face 2 no dado
- ↳ compõe: ex: sair face pai no dado

$A: \dots \rightarrow$  definir acontecimento

↳ letra maiuscula

## Espaço de resultados ( $\Omega$ )

- ↳ conjunto de resultados possíveis

## Probabilidade

$$P(A) = \frac{\# \text{casos favoráveis a } A}{\# \text{casos possíveis}}$$

↳ Nota: só acontece se os resultados  
são equiprováveis

$$P(A) = \lim_{n \rightarrow \infty} \frac{n_A}{n}$$

$$P(A) \geq 0$$

$P(S)=1 \rightarrow$  acontecimento certo

## Leis de De Morgan

$$\bar{A \cap B} = \bar{A} \cup \bar{B}$$

$$A \cup \bar{B} = \bar{A} \cap \bar{B}$$

## Axiomas

$$\rightarrow P(A) \geq 0, \forall A$$

$$\rightarrow P(\Omega) = 1$$

$$\rightarrow A \cap B = \emptyset \Rightarrow P(A \cup B) = P(A) + P(B)$$

## Consequências

$$\rightarrow P(\emptyset) = 0$$

$$\rightarrow \text{se } A \subseteq B, P(A) \leq P(B)$$

$$\rightarrow P(\bar{A}) = 1 - P(A)$$

$$\rightarrow P(A) \in [0, 1]$$

$$\rightarrow P(A - B) = P(A \cap \bar{B}) = P(A) - P(A \cap B)$$

$$\rightarrow P(A \cup B) = P(A) + P(B) - P(A \cap B)$$

## Probabilidade condicional

"Acontecendo B" ou "A dado B"

$$P(A|B) = \frac{P(A \cap B)}{P(B)}$$

$$P(A \cap B) = P(A|B) \times P(B)$$

Quando A e B são independentes:

$$P(A \cap B) = P(A) \times P(B)$$

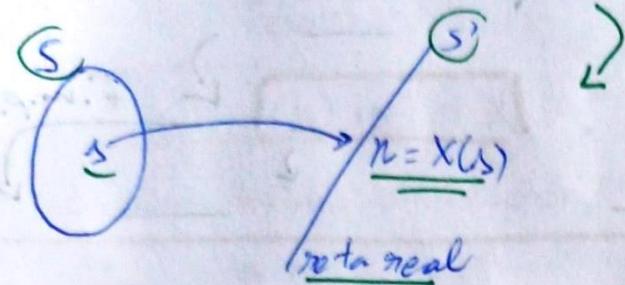
## Probabilidade total

$$P(A) = P(A \cap B) + P(A \cap \bar{B})$$

$$P(A) = \sum_{i \in N} P(A \cap B_i)$$

## Variável Aleatória (v.a.)

$X$  é uma função que associa a cada elemento do espaço de resultados  $S$ , um número real ( $\mathbb{R}$ )



Discreta - finito ou infinito numerável

Continua - infinito

## Função Massa de Probabilidade

→ retorna a  $P$  associada a cada valor que a v.a. discreta assume

$S \rightarrow$  espaço de resultados

$x_i \rightarrow$  v.a. discreta com valores  $x_i$

$$p(x_i) = p_i = P(X=x_i)$$

→ Probabilidade de  $x_i$

$$\begin{cases} P(X=x_i) > 0 \\ \sum_i P(X=x_i) = 1 \end{cases} \quad \left\{ \begin{array}{l} \text{Necessário para} \\ \text{ser considerada} \\ \text{função massa...} \end{array} \right.$$

## Representar

$X$	$(x_1 \dots x_n)$	Supórtete (valores que $X$ pode tomar)
$P(X=x)$	$P(X=x_1) \dots P(X=x_n)$	

## Função de distribuição acumulativa

$$F(x) = P(X \leq x)$$

$$F(-\infty) = 0 \quad \text{e} \quad F(+\infty) = 1$$

exemplo:

$$F(x) = P(X \leq x) = \begin{cases} 0, & x < 0 \\ \frac{1}{2}, & 0 \leq x < 1 \\ \frac{3}{4}, & 1 \leq x < 2 \\ 1, & x \geq 2 \end{cases}$$

## Válor médio

$$E(X) = \sum_i x_i \cdot p_i$$

### Propriedades:

1- Valor médio de uma constante é a própria constante

$$E(c) = c \quad E[E(X)] = E(X)$$

2- Valor médio de  $X^n$  é:

$$E(X^n) = \sum_i x_i^n \cdot p_i$$

3- Se  $a, b \in \mathbb{R}$ ,

$$E(ax+b) = aE(X) + b$$

4- O valor médio da soma é a soma dos valores médios

$$E(X+Y) = E(X) + E(Y)$$

## Variância de uma v.a.

$$V(x) = E(x^2) - E(x)^2$$

## Propriedades

1- Variância de uma constante é 0

2- Se  $Y = \alpha X + \beta$ ,  $\alpha, \beta \in \mathbb{R}$ :

$$V(Y) = \alpha^2 \cdot V(X)$$

3- Se X e Y forem v.a. independentes:

$$V(X+Y) = V(X-Y) = V(X) + V(Y)$$

Nota: A Covariância é a medida de variabilidade dijunta de duas v.a. em que:

$$\text{Cov}(x, y) = E\{[x-E(x)][y-E(y)]\}$$

4- Se X e Y forem dependentes:

$$V(x+y) = V(x) + V(y) + 2\text{Cov}(x, y)$$

$$V(x-y) = V(x) + V(y) - 2\text{Cov}(x, y)$$

## Derroio padrão de v.a.

$$\sigma(x) = \sqrt{V(x)}$$

## Coeficiente de Variação de v.a.

$$CV(x) = \frac{\sigma}{E(x)} \times 100\%$$

## Modelo de Bernoulli

↳ Seja x uma v.a. discreta que apenas toma dois valores:

- 0- intuscesso
- 1- succeso

$$X \sim \text{Ber}(p)$$

f.m.p.

↳ p probabilidade de succeso

$$f(x) = \begin{cases} p^x (1-p)^{1-x}, & x=0 \vee x=1 \\ 0, & x \neq 0, 1 \end{cases}$$

$$E[x] = p$$

$$\text{Var}[x] = p(1-p)$$

## Modelo Binomial

↳ Seja x uma v.a. que representa o nº de succeso em n provas de Bernoulli:

$$X \sim \text{Binomial}(n, p)$$

$n$ : nº de provas  
 $p$ : probabilidade de succeso [constante]

$$f(x) = \begin{cases} {}^n C_x p^x (1-p)^{n-x}, & x=0, \dots, n \\ 0, & x \neq 0, \dots, n \end{cases}$$

$$E[x] = np$$

$$\text{Var}[x] = np(1-p)$$

## Modelo Hipergeométrico

↳ Seja a v.a.  $X$  que representa o nº de elementos da amostra recolhida que possuem a característica A:

$$X \sim HG(N, n, M)$$

- $N$ : nº de elementos da população
- $n$ : dimensão da amostra
- $M$ : elementos da população com a característica A

$$f(x) = \frac{\binom{M}{x} \times \binom{N-M}{n-x}}{\binom{N}{n}}$$

f. imp.

$$E[X] = n \times \frac{M}{N}$$

$$\text{Var}[X] = \frac{n \times M \times (N-M) \times (N-n)}{N^2 \times (N-1)}$$

## Modelo de Poisson

↳ Seja  $X$  a v.a. que representa o nº de ocorrências de um determinado acontecimento num dado intervalo/periodo:

$$X \sim P(\lambda) \quad \lambda > 0$$

↳  $\lambda$  - valor médio

$$E[X] = \text{Var}[X] = \lambda$$

## Propriedades

↳ Probabilidade da ocorrência de um acontecimento é igual em intervalos da mesma amplitude

↳ A ocorrência de um acontecimento num dado intervalo de tempo é independente de outros intervalos

## Variável Aleatória Contínua

- ↪ no conjunto dos números ( $\mathbb{R}$ )
- ↪ a probabilidade no ponto é 0  
 $P(X = x_0) = 0, \forall x_0 \in \mathbb{R}$

↪ A função densidade de probabilidade caracteriza todos os variáveis e satisfaz as condições:

$$f(x) \geq 0, \forall x \in \mathbb{R}$$

$$\int_{-\infty}^{+\infty} f(x) dx = 1$$

↪ Seja  $x$  uma v.a. contínua, a F.D.P. de  $x$  é  $f(x)$  tal que:

$$P(a \leq x \leq b) = \int_a^b f(x) dx$$

↪ Representar num histograma

## Funções de Distribuição

$$F(x) = P(X \leq x) = \int_{-\infty}^x f(x) dx$$

↪  $F(x)$  é primitiva de  $f(x)$

↪  $0 \leq F(x) \leq 1$

$$\lim_{n \rightarrow -\infty} nF(n) = 0 \quad \lim_{n \rightarrow +\infty} F(n) = 1$$

↪ Função monótona não decrescente

↪ " continua em  $\mathbb{R}$

## Valor Médio

$$\hookrightarrow E(x) = \int_{-\infty}^{+\infty} xf(x) dx$$

## Variância

$$\hookrightarrow V(x) = E(x^2) - E(x)^2$$

$$E(x^2) = \int_{-\infty}^{+\infty} x^2 f(x) dx$$

## Modelo Uniforme

Uma v.a.  $x$  tem distribuição uniforme num intervalo  $[a, b]$  se a prob. de  $x$  pertencer a outro intervalo é proporcional ao comprimento do mesmo.

$$X \sim U[a, b]$$

$$f(x) = \begin{cases} \frac{1}{b-a}, & a \leq x \leq b \\ 0, & \text{c.c.} \end{cases}$$

$$F(x) = \begin{cases} 0, & x < a \\ \frac{x-a}{b-a}, & a \leq x \leq b \\ 1, & x > b \end{cases}$$

$$\hookrightarrow E(x) = \frac{a+b}{2}$$

$$\hookrightarrow V(x) = \frac{(b-a)^2}{12}$$

## Modelo Exponencial

↳ Utilizado para modelar o tempo que decorre entre dois acontecimentos consecutivos de Poisson ( $\lambda$ ) num intervalo de tempo unitário  $T$ .

↳  $X \sim \text{Exp}(\lambda)$

$$f(x) = \begin{cases} 0, & x \leq 0 \\ \lambda e^{-\lambda x}, & x > 0 \wedge \lambda > 0 \end{cases}$$

$$F(x) = \begin{cases} 0, & x \leq 0 \\ 1 - e^{-\lambda x}, & x > 0 \wedge \lambda > 0 \end{cases}$$

$$E(X) = \frac{1}{\lambda} \quad \text{e} \quad \text{Var}(X) = \frac{1}{\lambda^2}$$

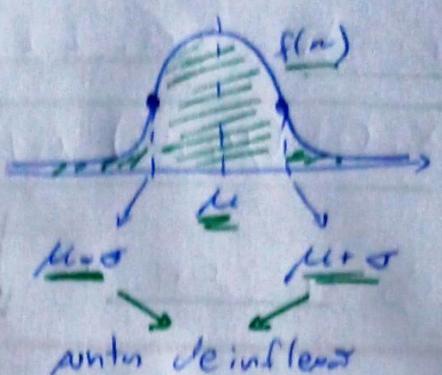
## Modelo Normal

↳ Seja  $X$  uma v.a. com distribuição normal, tem-se que esta variável é parametrizada por:  $\mu$  (valor médio) e  $\sigma$  (desvio padrão).

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

$$E(X) = \mu \quad \text{e} \quad \text{Var}(X) = \sigma^2$$

FDP  $\rightarrow$  Curva de Gauss  $\rightarrow f(n)$



↳ O caso tabulado corresponde à distribuição normal standard, em que:  $\mu = 0$  e  $\sigma = 1$

$X \sim N(\mu = 0, \sigma = 1)$

Para estandardizar:

$X \sim N(\mu, \sigma)$  estandardizar

$$z = \frac{X - \mu}{\sigma}, \quad z \sim N(0, 1)$$

$$F_z(z) = P(Z \leq z) = \Phi(z)$$

↳ Função de distribuição de  $z$

$$\Phi(-z) = 1 - \Phi(z)$$

$$\Phi(z) = a \Leftrightarrow z = \Phi^{-1}(a)$$

## Additividade da Normal

Sejam  $x_1, \dots, x_n$ , v.a. i.i.d.\* em que  $x_i \sim N(\mu_i, \sigma_i^2)$ ,  $i=1, \dots, n$

Considerando  $a_1, \dots, a_n$ , t.p.:

$$T = \sum_{i=1}^n x_i \sim N\left(\sum_{i=1}^n a_i \mu_i; \sqrt{\sum_{i=1}^n a_i^2 \sigma_i^2}\right)$$

Corolário: Se  $x_i \sim N(\mu, \sigma)$ :

$$T = \sum_{i=1}^n x_i \sim N(n\mu, \sigma\sqrt{n})$$

Consideremos uma amostra aleatória de dimensão  $n$ ,  $(x_1, \dots, x_n)$ , extraída de uma população  $x \sim N(\mu, \sigma)$

$$\bar{X} = \frac{1}{n} \cdot \sum_{i=1}^n x_i \rightarrow \text{Media}$$

$\bar{X}$  também é uma v.a.:

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

## Teorema do Limite Central

Sejam  $x_1, \dots, x_n$ , v.a. iid. com valor médio e variância finitas:

$$E(x_i) = \mu \quad \text{e} \quad \text{Var}(x_i) = \sigma^2$$

Então para um  $n > 30$ :

I

$$T = \sum_{i=1}^n x_i \stackrel{(a)}{\sim} N(n\mu, \sigma\sqrt{n})$$

\* se  $x_i$ s - ad - etc

II Seja  $\bar{x} = \frac{1}{n} \sum_{i=1}^n x_i$ , então:

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

\* i.i.d.: independentes e identicamente distribuídos

## Estimacão

↳ Tem o objetivo de arranjar funções da amostra para estimar os parâmetros da população

↳ Para  $\bar{x}$  ser um estimador:

①  $\bar{x}$  tem de ser centrado

$$E[\bar{x}] = \mu \quad \leftarrow$$

② tem a menor variabilidade possível

$$\text{Var}[\bar{x}] = \frac{\text{Var}[x]}{n} \rightarrow 0 \quad n \rightarrow \infty \quad \leftarrow$$

<u>Parâmetro</u>	<u>Estimador</u>
valor médio ( $\mu$ )	média ( $\bar{x}$ )
variancia ( $\sigma^2$ )	variancia ( $\sigma^2$ )
desvio padrão ( $\sigma$ )	desvio padrão ( $\sigma$ )
proporção ( $p$ )	média de sucesso ( $\hat{p}$ )

↳ O método mais utilizado para encontrar o melhor estimador para um parâmetro é o

## Método da Máxima Verossimilhança

↳ Existem duas formas de produzir estimativas para os parâmetros:

- estimativa pontual
- " " intervalar

## Estimativa Pontual

↳ Propõe-se operar um único (parâmetro) valor para o parâmetro

## Estimativa Intervalar

↳ Propõe-se um determinado intervalo com um certo grau de confiança que contenha o verdadeiro valor do parâmetro

### Intervalo de Confiança

$$\text{Grau de confiança} = 1 - \alpha$$

## Construir I.C.:

Dada uma amostra aleatória  $x_1, \dots, x_n$ , e seja  $\theta$  o parâmetro a estimar.

Considere-se a estatística  $T$ , função da amostra e de  $\theta$ , cuja distribuição é conhecida e não depende de  $\theta$ :

A esta estatística chamamos variável fatorial

O objetivo é encontrar o el:

$$P(a < T < b) = 1 - \alpha \quad \leftarrow$$

Encontrar o IC. para  $\theta$  constante em obter:

$$P(L_1 < \theta < L_0) = 1 - \alpha \quad \leftarrow$$

## Amplitude do I.C. =

$$= \text{Lim. Superior} - \text{Lim. Inferior}$$

$$\text{Erro/Precisão} = \frac{\text{Amplitude}}{2}$$

## Exercício

1. Definir Variável aleatória

2. Definir:

- Parâmetro a estimar
- Tipo de população
- Tamanho da Amostra
- Se  $\sigma$  é conhecido (para  $\mu$ )
- Nível de Confiança e  $\alpha$

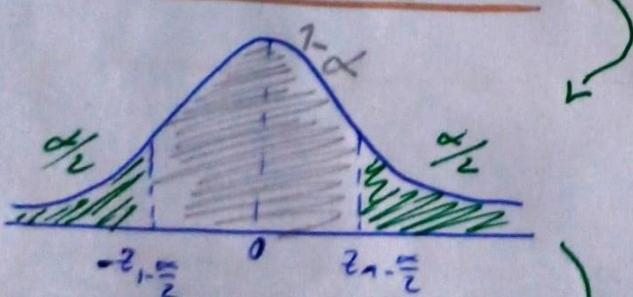
3. Escolher Variável Fatorial

4. Construir o intervalo de confiança

5. Concluir

## V.F. com Distribuição Normal

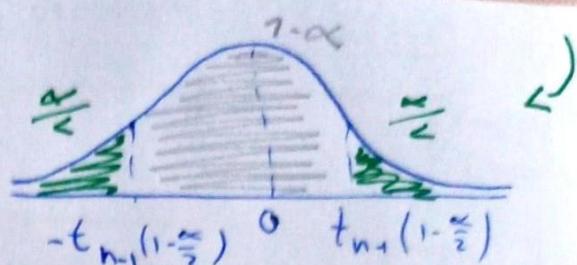
$$P[-z_{1-\frac{\alpha}{2}} < V.F. < z_{1-\frac{\alpha}{2}}] = 1-\alpha$$



$$z_{1-\frac{\alpha}{2}} = \Phi^{-1}(1-\frac{\alpha}{2})$$

## V.F. com Distribuição t-Student

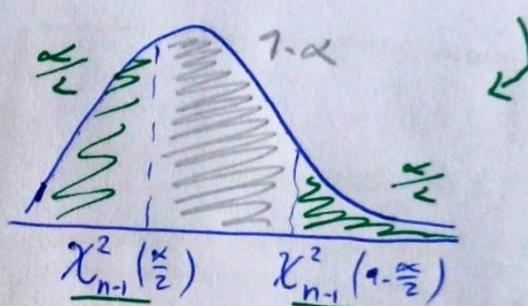
$$P[-t_{n-1}(1-\frac{\alpha}{2}) < V.F. < t_{n-1}(1-\frac{\alpha}{2})] = 1-\alpha$$



n-1 - graus de liberdade

## V.F. com Distribuição Qui-quadrado

$$P[\chi^2_{n-1}(\frac{\alpha}{2}) < V.F. < \chi^2_{n-1}(1-\frac{\alpha}{2})] = 1-\alpha$$



n-1 - graus de liberdade

## Testes de Hipóteses

↳ tomada de decisão

↳ conjetutar acerca do verdadeiro valor do parâmetro da população

↳  $H_0 \rightarrow$  hipótese nula

↳ assume como verdadeira

↳  $H_1 \rightarrow$  hipótese alternativa

↳ hipótese que se pretende testar

## Tipos de erros

Situação real	Decisão Rejeitar $H_0$	Não Rejeitar $H_0$
$H_0$ Verdadeira	Erro tipo I $\alpha$	✓
$H_0$ Falsa	✓	Erro tipo II $\beta$

$$\alpha = P(\text{Rejeitar } H_0 \mid H_0 \text{ Verdadeira})$$

$$\beta = P(\text{Não Rejeitar } H_0 \mid H_0 \text{ Falsa})$$

$\alpha \rightarrow$  nível de significância do teste

Nota: É mais grave cometer um erro do tipo I do que um do tipo II

Válores de  $\alpha$  comuns:

1%, 5%, 10%

Significa que há 5% de probabilidade de rejeitar  $H_0$ , sabendo que é verdadeira

## Definir Hipóteses

Parâmetro  $p$  (proporção) (exemplo)

$$(1) H_0: p \leq p_0 \quad vs \quad H_1: p > p_0$$

$$(2) H_0: p \geq p_0 \quad vs \quad H_1: p \leq p_0$$

$$(3) H_0: p = p_0 \quad vs \quad H_1: p \neq p_0$$

A estatística de teste é a variável aleatória do parâmetro a testar, com a definição da distribuição

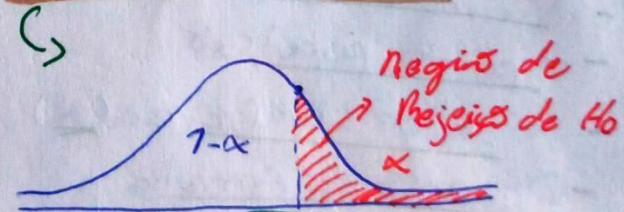
[p-value] \* → área produzida pelo valor observado

$$p = P(T \geq T_{obs}) \rightarrow \text{exemplo}$$

$$T_{obs} = \text{estatística de teste}$$

## Teste Unilateral à Direita

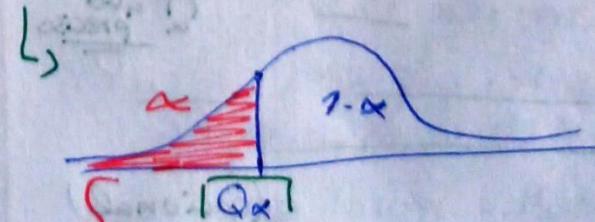
$$H_0: \theta \leq \theta_0 \quad vs \quad H_1: \theta > \theta_0$$



$$\text{A.R.: } [Q_{1-\alpha}, +\infty]$$

## Teste Unilateral à Esquerda

$$H_0: \theta \geq \theta_0 \quad vs \quad H_1: \theta < \theta_0$$

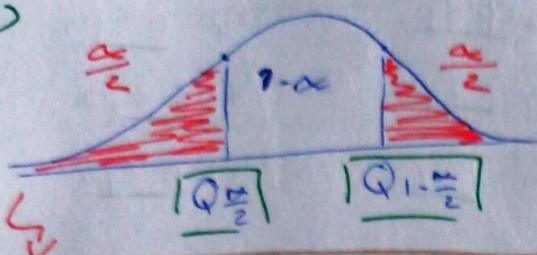


$$\text{A.R.: } ]-\infty, Q_\alpha]$$

## Teste Bilateral

$$H_0: \theta = \theta_0 \quad vs \quad H_1: \theta \neq \theta_0$$

↳



$$I.R.: ]-\infty; Q_{\frac{\alpha}{2}}] \cup [Q_{1-\frac{\alpha}{2}}, +\infty[$$

\* O p-value é o menor nível de significância ( $\alpha$ ) a partir do qual se começa a rejeitar  $H_0$ :

Se  $|x| \geq p\text{-value}$ , rejeitase  $H_0$

## Exercícios

1. Definir variáveis aleatórias

2. Definir:

- Parâmetro a testar
- Tipo de população
- Nível de significância ( $\alpha$ )
- Dimensão da amostra

\* 3. Especificar hipóteses a testar

4. Identificar Estatística de teste

5. Determinar V.E.T. ↳ meter  $H_0$  que é preciso

6. Desenhar gráficos e calcular a região de rejeição de  $H_0$

7. Calcular p-value (opcional)

8. Concluir

## Exercício do tipo:

↳ "Qual a probabilidade de ter tomado a decisão errada se o verdadeiro valor for  $\theta_0^*$ "

1. Dizer o erro a se pode cometer

2. Reescrever a região de rejeição em função do estimador (ex:  $\bar{X}, \hat{P}, S^2$ )

3. Representar o gráfico em função de  $(\bar{x}, \hat{P}, S^2)$

4. Calcular  $P[\bar{X} \leq \bar{x}_c] = (\alpha \text{ ou } 1 - \alpha)$   
e com isso calcular  $\bar{x}_c$  (ou  $\hat{P}_c, S_c^2$ )

5. Calcular  $P[\text{erros}]$

## Amostras Emparelhadas

↳ Medidas para o mesmo indivíduo para instantes distintos

↳ originam pares de observações

1. Definir variáveis aleatórias

2. Definir  $D$ :

$$D = X - Y \sim N(\mu_x - \mu_y, \sigma_D^2)$$

$$\sigma_D^2 = V(X) + V(Y) - 2 \text{Cov}(X, Y)$$

3. Parâmetro a testar: Amostra das diferenças

$$\mu_D = \mu_x - \mu_y$$

estimador  $\bar{d} = \frac{\sum_{i=1}^n d_i}{n}$  → cada observação

$$\sigma_D^2 = \frac{\sum_{i=1}^n d_i^2 - n \bar{d}^2}{n-1} = \frac{\sum_{i=1}^n (d_i - \bar{d})^2}{n-1}$$

## Testes de Ajustamento

↳ Testar a hipótese de que uma distribuição se ajusta a uma determinada população

↳ Teste Lilliefors - Kolmogorov - Smirnov

↳ Teste Shapiro - Wilk

### Teste Lilliefors - Kolmogorov - Smirnov

↳ quando os parâmetros não são conhecidos

↳  $n > 50$  (para mais precisa)

•  $D_{obs} \geq D_{\alpha}$   
•  $\alpha > p\text{-value}$

} Rejeita-se  $H_0$

### Teste Shapiro - Wilk

↳  $n \leq 50$

•  $W_{obs} \leq W_{\alpha}$   
•  $\alpha > p\text{-value}$

} Rejeita-se  $H_0$

### Hipóteses a testar:

$H_0: X \sim F(\dots) \text{ vs } H_1: X \not\sim F(\dots)$

↳ Normalmente testase a normalidade

## Regressão Linear

- ↳ Establecer relação entre a variável dependente e uma ou mais variáveis independentes

Com duas variáveis chama-se regressão linear simples

## Diagrama de dispersão

- ↳ Representar variáveis quantitativas

## Covariância Amostral

- ↳ medida de variabilidade conjunta de pares de observações

$$\text{cov}(x, y) = \frac{s_{xy}}{n-1}$$

$$s_{xy} = \sum_{i=1}^n x_i y_i - n \bar{x} \bar{y}$$

↳ Soma cruzada

Se:

- $\text{cov}(x, y) > 0 \rightarrow$  relação linear positiva  
↳ variam no mesmo sentido
- $\text{cov}(x, y) < 0 \rightarrow$  relação linear negativa  
↳ variam em sentido oposto
- $\text{cov}(x, y) = 0 \rightarrow$  não existe relação linear

## Coeficiente de Correlação Linear

- ↳ mede o grau de associação linear entre os valores amostrais

$$\rho_{xy} = \frac{\text{cov}(x, y)}{s_x s_y} = \frac{s_{xy}}{\sqrt{s_{xx} \cdot s_{yy}}}$$

$$s_{xy} = \sum_{i=1}^n x_i y_i - n \bar{x} \bar{y}$$

↳ soma das cruzadas

$$s_{xx} = \sum_{i=1}^n x_i^2 - n \bar{x}^2 = (n-1) s_x^2$$

↳ soma dos quadrados de x

$$s_{yy} = (\dots) \rightarrow$$

## Propriedades

- $-1 \leq \rho \leq 1$
- $\rho = -1$ : correlação linear negativa perfeita
- $\rho = 1$ : correlação linear positiva perfeita
- $\rho = 0$ : não existe correlação linear
- $|\rho| > 0$ : correlação linear aceitável

## Regressão Linear - Equação

$$Y_i = \beta_0 + \beta_1 x_i + \epsilon_i$$

Y - variável explicada / dependente

X - variável explicativa

E - erro aleatório  $\rightarrow E_i \sim N(0; \sigma^2)$

$\beta_0$  - ordenada na origem

$\beta_1$  - declive da reta

## Método de Regressão

### Interpretação dos coeficientes

$\beta_0$ : valor de  $y$  quando  $x$  é nulo

$\beta_1$ : por cada unidade de  $x$ ,  $y$  varia  $\beta_1$  unidades

### Método dos Mínimos Quadrados

Determinar os valores  $\beta_0$  e  $\beta_1$  que minimizam a soma dos quadrados dos erros (SQE)

$$\text{SQE} = \sum_{i=1}^n e_i^2 = \sum_{i=1}^n (y_i - \beta_0 - \beta_1 x_i)^2$$

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

## Estimadores de $\beta_0$ e $\beta_1$

$$\hat{\beta}_1 = \frac{s_{xy}}{s_{xx}} = \frac{\text{cov}(x, y)}{s_x^2}$$

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}$$

### Coeficiente de determinação

↳ Permite medir a qualidade do ajustamento feito pelo método dos mínimos quadrados

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

↳ Soma dos quadrados totais

$$SQE = \sum_{i=1}^n (y_i - \hat{y}_i)^2 = s_{yy} - \frac{s_{xy}^2}{s_{xx}}$$

↳ Soma de quadrados dos erros

$$SQR = \sum_{i=1}^n (\hat{y}_i - \bar{y})^2 = \frac{s_{xy}^2}{s_{xx}}$$

↳ Soma de quadrados de regressão

Variabilidade total do conjunto de observações de  $Y$

$$SQT = SQE + SQR$$

## Coeficiente de determinação (cont.)

↳ Determine o percentagem de variabilidade da variável dependente que é explicada à custa da variável independente

$$\Rightarrow R^2 = \frac{SQR}{SQT} = 1 - \frac{SQE}{SQT} = \frac{Sxy^2}{SxxSyy}$$

### Propriedades

↳ Quadrado do Coeficiente de Correlação Linear

- $0 \leq R^2 \leq 1$
- $R^2 = 1$ , ajuste perfeito
- $R^2 = 0$ ; ajuste inútil

### Sobre os Erros

$$|\epsilon_i \sim \text{Normal}(0; \sigma^2)|$$

- (1)  $E[\epsilon_i] = 0$
- (2)  $V[\epsilon_i] = \sigma^2$
- (3)  $\text{Cov}(\epsilon_i, \epsilon_j) = 0$ , para  $i \neq j$
- (4) Normalmente Distribuídos

### Erros observados:

$$\hookrightarrow |e_i = y_i - \hat{y}_i|$$

$$|Y = \beta_0 + \beta_1 X + E| \quad 2$$

Parâmetro	Estimador	Equação
$\beta_0$	$\hat{\alpha} / \hat{\beta}_0$	$a = \bar{y} - b\bar{x}$
$\beta_1$	$\hat{B} / \hat{\beta}_1$	$b = \frac{\sum xy}{\sum x^2}$
$y_o$	$\hat{Y}_o$	$\hat{y}_o = a + b x_o$