Variáveis Aleatórias Discretas

Função Probabilidade

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

Valor Esperado E(X)

$$\mu_x = E(X) = \sum_{i=1}^{n} x_i . P(X = x_i)$$

Variância

$$Var(X) = {\sigma_x}^2 = E(X^2) - [E(X)]^2 = E(X^2) - {\mu_x}^2$$

Onde: $E(X^2) = \sum_{i=1}^{n} {x_i}^2 . p_i$

O **desvio padrão de X** ($\boldsymbol{\sigma_x}$) é a raiz quadrada positiva da variância:

$$\boldsymbol{\sigma_x} = \sqrt{\boldsymbol{\sigma_x^2}}$$

Média

$$\mu_x = E(X) = \sum_{i=1}^{n} x_i . P(X = x_i)$$

Distribuição de Bernoulli

- Função de probabilidade: $\boldsymbol{P(X = x) = p^x . q^{1-x}}$

- Média:

$$E(X) = \mu_x = \sum_{i=0}^{1} x . P(X = x) = 0 . q + 1 . p = p$$

- Variância:

$$Var(X) = E(X^2) - [E(X)]^2$$

$$E(X^2) = \sum_{i=1}^{n} {x_i}^2 . p_i = 0^2 . q + 1^2 . p = p$$

$$Var\ (X) = p - p^2 = p . (1 - p) = \mathrm{p . q}$$

Distribuição Binomial

- Função de probabilidade:

$$P(X = x) = \binom{n}{x} . p^x . q^{n-x}, x = 1,2, \dots , n$$

Onde $\binom{n}{x} = \frac{n!}{x!(n-x)!}$

No qual:

- n é número de repetições dos experimentos;
- x é o número desejado de sucessos;
- $(n - x)$ é o número esperado de fracassos;
- p é a probabilidade de sucesso em um ensaio individual;
- q é a probabilidade de fracasso num ensaio individual.

- **Média:**

$$\mu_x = E(X) = n.p$$

- **Variância:**

$$Var(X) = n.p.q$$

Distribuição Poisson

- **Função de probabilidade:**

$$P(X = x) = \frac{e^{-\lambda}.\lambda^x}{x!}$$

Onde:

x=número de sucessos em um intervalo;

e=2,718 (número de Euler);

λ é a média de ocorrência de sucessos (tempo, área ou volume)

- **Média:**

$$E(X) = \mu_x = \lambda$$

- **Variância:**

$$V(X) = {\sigma_x}^2 = \lambda$$

Modelos Probabilísticos para v.a. Discretas		
BERNOULLI X ~ Ber(p)	BINOMIAL X ~ Bin(n,p)	POISSON X ~ Po(λ)
• Única realização de um experimento aleatório; • Sucesso (X=1) ou fracasso (X=0)	• n experimentos de Bernoulli em sequência; • Cada ensaio: sucesso ou fracasso; • A probabilidade de sucesso p é constante em cada ensaio.	• Interesse em contar o número de sucessos ocorridos em um experimento, em um intervalo de tempo, superfície ou volume.
$P(X=x)=p^x.q^{1-x}$	$P(X=x)=\binom{n}{x}.p^x.q^{n-x}$ $\binom{n}{x}=\frac{n!}{x!\,(n-x)!}$	$P(X=x)=\frac{e^{-\lambda}.\lambda^x}{x!}$
$E(X)=\mu_x=p$	$E(X)=\mu_x=n.p$	$E(X)=\mu_x=\lambda$
$Var(X)=\sigma_x{}^2=p.q$	$Var(X)=\sigma_x{}^2=n.p.q$	$V(X)=\sigma_x{}^2=\lambda$

Distribuição Normal

- Notação:
  $X \sim N(\mu,\sigma^2) \rightarrow X\ tem\ distibuição\ normal\ com\ média\ \mu=E(X)\ e\ variância\ \sigma^2=Var(x)$

- Suponha $X \sim N(\mu,\sigma^2)$ e deseja-se calcular P(a<X<b):

$$P(a<X<b)=\int_a^b f(x)dx=\int_a^b \frac{\mathbf{1}}{\boldsymbol{\sigma}\sqrt{\mathbf{2}\boldsymbol{\pi}}}\mathbf{exp}\left(-\frac{\mathbf{1}}{\mathbf{2}}\left(\frac{\boldsymbol{x}-\boldsymbol{\mu}}{\boldsymbol{\sigma}}\right)^{\mathbf{2}}\right)dx$$

NÃO PODE SER RESOLVIDA PELOS MÉTODOS TRADICIONAIS (SÓ PODE SER RESOLVIDA POR APROXIMAÇÃO NUMÉRICA)!!!

Padronização da distribuição Normal

Distribuição Normal Padrão

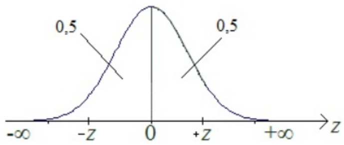
- Existe uma infinidade de distribuições normais.
- Então, utiliza-se a variável normal padronizada Z :

$$Z = \frac{X - \mu}{\sigma}$$

- Distribuição reduzida:

$$P(X \leq x) = P\left(Z \leq \frac{x - \mu}{\sigma}\right) = F(x)$$

- Z tem média zero e variância igual a 1;
- $Z \sim N(0{,}1)$: Z tem distribuição normal padronizada média 0 e variância 1;



TIPOS DE AMOSTRAGENS PROBABILÍSTICAS

Amostragem Aleatória Estratificada

1º CASO:

$$N_1 = N_2 = N_3 = \ldots = N_L \qquad n_i = \frac{n}{L}$$

Amostragem Aleatória Estratificada

2º CASO:

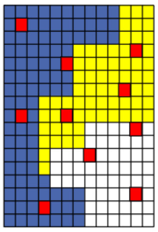
$$n_i = n\frac{N_i}{N}$$

Amostragem Aleatória Estratificada

- Sabendo que o tamanho da amostra é ***n***, como determinar o nº de indivíduos (n_i) a serem selecionados em cada um dos estratos?

- 3º CASO:
- Tamanho ótimo (considera a variabilidade)

Então: $$n_i = \left(\frac{N_i S_i}{\sum N_i S_i} \right) n$$



Teorema do Limite Central (TLC)

Considere uma população de tamanho *N* com média $\boldsymbol{\mu}$ e variância $\boldsymbol{\sigma}^2$. Se forem retiradas k amostras de tamanho *n* desta população, a média amostral ($\bar{\boldsymbol{x}}$) terá uma distribuição aproximadamente normal, com média das médias amostrais iguais a média da população (**amostragem com reposição**)

$$\boldsymbol{\mu_{\bar{x}} = \mu_x}$$

e com variância e desvio padrão das médias amostrais iguais a

$$\boldsymbol{\sigma_{\bar{x}}}^2 = \frac{\boldsymbol{\sigma_x}^2}{\boldsymbol{n}}$$

$$\boldsymbol{\sigma_{\bar{x}}} = \frac{\boldsymbol{\sigma_x}}{\sqrt{\boldsymbol{n}}}$$

Para distribuição amostral das médias: quando se conhece a variância da população ou a **amostra é grande (n ≥ 30),** utiliza-se a estatística Z da distribuição Normal, independente da distribuição da população.

$$z = \frac{\bar{x} - \mu_{\bar{x}}}{\sigma_{\bar{x}}}$$

- Distribuição amostral da média em pequenas amostras (**n<30**);
- **Não se conhece σ** (desvio padrão populacional);
- Conhece-se apenas a estimativa **s** do desvio padrão amostral;

$$t = \frac{\overline{x} - \mu_{\overline{x}}}{s_{\overline{x}}} = \frac{\overline{x} - \mu_{\overline{x}}}{\frac{s_x}{\sqrt{n}}}$$

Distribuição do Qui-Quadrado (χ^2)

- *É uma distribuição amostral de* ***variâncias;***
- *Ao retirar uma amostra de* ***n*** *elementos de uma população normal com média* μ *e variância* σ^2*, tem-se que a distribuição amostral da variância (****s²****) segue uma distribuição de* χ^2 ***(qui-quadrado)*** *com* ***n-1*** *graus liberdade :*

$$\boxed{\chi^2 = \frac{(n-1)s^2}{\sigma^2}}$$

- 1º caso: amostras grandes (n≥30)

a. Intervalo de confiança (IC):

$$IC(\mu)_{1-\alpha} : \bar{x} \pm e$$

b. Margem de erro:

$$z = \frac{\bar{x} - \mu_{\bar{x}}}{\sigma_{\bar{x}}} \qquad e = z_{\alpha/2} \cdot \frac{\sigma}{\sqrt{n}} \; ; \; e = z_{\alpha/2} \cdot \frac{s}{\sqrt{n}}$$

Intervalo de Confiança para Média

- 1º caso: amostras grandes (n≥30)

c. Tamanho da amostra

$$e = z\alpha/_2 \cdot \frac{\sigma}{\sqrt{n}}$$

$$\Downarrow$$

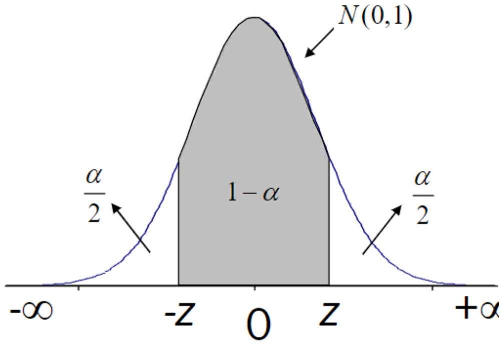
$$n = \left(\frac{z\alpha/_2 \cdot \sigma}{e}\right)^2$$

Intervalo de Confiança para Média

$X \sim N(\mu, \sigma^2)$ μ desconhecido, mas σ^2 conhecido

(Normal Padrão)

$$Z = \frac{\bar{X} - \mu}{\frac{\sigma}{\sqrt{n}}}$$



$$\boxed{P(\bar{X} - z\frac{\sigma}{\sqrt{n}} < \mu < \bar{X} + z\frac{\sigma}{\sqrt{n}}) = 1 - \alpha}$$

IC para média

$$P(|Z| > z) = \alpha$$

nível de significância

$$P(-z < Z < z) = 1 - \alpha$$

nível de confiança

Intervalo de Confiança para Média

- 2º caso: amostras pequenas (n<30)

a. Intervalo de confiança (IC):

$$IC(\mu)_{1-\alpha} : \bar{x} \pm e$$

b. Margem de erro:

$$t = \frac{\bar{x} - \mu_{\bar{x}}}{s} \qquad e = t_{\alpha/2} \cdot \frac{s}{\sqrt{n}}$$

Graus de liberdade (v)=n-1

Intervalo de Confiança para uma Proporção

a. Intervalo de confiança (IC):

$$\mathrm{IC}(\mathrm{p})_{1-\alpha} : \hat{\mathrm{p}} \pm \mathrm{e} \qquad \hat{p} = \frac{x}{n} = \frac{\textit{número de sucessos}}{\textit{tamanho da amostra}}$$

b. Margem de erro:

$$e = Z_{\alpha/2}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$$

c. Tamanho da amostra

$$e = Z_{\alpha/2}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$$

$$\Downarrow$$

$$n = \frac{z_{\alpha/2}^{2}\,\hat{p}(1-\hat{p})}{e^{2}}$$

Intervalo de Confiança para variância e desvio padrão

a. Intervalo de confiança (IC):

$$IC(\sigma^2)_{1-\alpha} = s^2 \pm e \qquad IC(\sigma)_{1-\alpha} = s \pm e$$

b. Margem de erro:

$$e = \frac{(n-1)s^2}{\chi^2} \qquad e = \sqrt{\frac{(n-1)s^2}{\chi^2}}$$

Teste de hipóteses

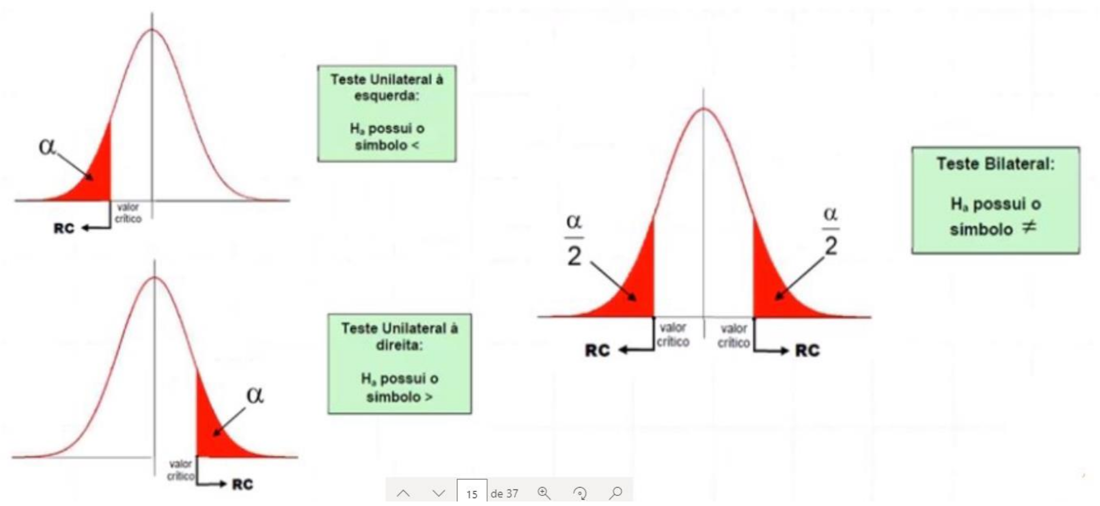
- **Passo a passo para a realização de um teste de hipóteses**

1. Escreva a **hipótese nula H_o** e a **hipótese alternativa H_1**

2. Calcule o **valor observado** (z_{obs}, t_{obs}...) utilizando a **fórmula** correspondente ao caso que está analisando.

Teste de hipóteses

3. Escolher o nível de significância (α) e estabelecer a **região crítica (RC)**



Teste de hipóteses

4. Obtenha o **valor crítico** (z_c, t_c...) do teste de acordo com o nível se significância do teste e com a **região crítica (RC)** utilizando a **tabela** de distribuição correspondente.

5. Marque o valor observado (z_{obs}, t_{obs}...) no gráfico

6. Conclua o teste:

   - se o **valor observado** $\in$ **RC**, então rejeita H_o

   - se o **valor observado** $\notin$ **RC**, então aceita H_o

- **População infinita, normal ou aproximadamente normal, variância populacional conhecida (amostra grande n≥30)**

Ho	H_1	R. CRITICA
$\mu \geq \mu_0$	$\mu < \mu_o$	$z < -z_\alpha$
$\mu \leq \mu_0$	$\mu > \mu_o$	$z > z_\alpha$
$\mu = \mu_0$	$\mu \neq \mu_o$	$z < -z_{\alpha/2}$ e $z > z_{\alpha/2}$

Estatística a ser utilizada =>

$$Z = \frac{\overline{X} - \mu_o}{\sigma \ / \sqrt{n}}$$

Teste de hipótese para média

- População infinita, normal ou aproximadamente normal, variância populacional desconhecida e amostra pequena (n<30).

Ho	H_1	R. CRITICA
$\mu \geq \mu_0$	$\mu < \mu_o$	$t < -t_\alpha$
$\mu \leq \mu_0$	$\mu > \mu_o$	$t > t_\alpha$
$\mu = \mu_0$	$\mu \neq \mu_o$	$t < -t_{\alpha/2}$ e $t > t_{\alpha/2}$

Estatística a ser utilizada =>

$$t = \frac{\overline{X} - \mu_O}{S \ / \sqrt{n}} \qquad \textbf{com v = n-1}$$

Teste de hipótese - diferença entre médias

- População infinita, normal ou aproximadamente normal, variância populacional conhecida (amostras grandes e pequenas)

Ho	H_1	R. CRITICA
$\mu_1 - \mu_2 \geq d_0$	$\mu_1 - \mu_2 < d_o$	$z < -z_\alpha$
$\mu_1 - \mu_2 \leq d_0$	$\mu_1 - \mu_2 > d_o$	$z > z_\alpha$
$\mu_1 - \mu_2 = d_0$	$\mu_1 - \mu_2 \neq d_o$	$z < -z_{\alpha/2}$ e $z > z_{\alpha/2}$

- Estatística a ser utilizada =>

$$Z = \frac{(\overline{X}_1 - \overline{X}_2) - (\mu_1 - \mu_2)}{\sqrt{\sigma_1^{\ 2} / n_1 + \sigma_2^{\ 2} / n_2}}$$

Teste de hipótese – uma proporção

Ho	H_1	R. CRITICA
$p \geq p_0$ $p \leq p_0$ $p = p_0$	$p < p_o$ $p > p_o$ $p \neq p_o$	$z<-z_\alpha$ $z>z_\alpha$ $z<-z_{\alpha/2}$ e $z>z_{\alpha/2}$

Estatística a ser utilizada =>

$$Z = \frac{\hat{p} - p_o}{\sqrt{\hat{p}\hat{q} / n}}$$

Teste de hipótese – diferença entre proporções

Ho	H_1	R. CRITICA
$p_1 - p_2 \geq p_0$ $p_1 - p_2 \leq p_0$ $p_1 - p_2 = p_0$	$p_1-p_2<p_o$ $p_1-p_2>p_o$ $p_1 - p_2 \neq p_o$	$z<-z_\alpha$ $z>z_\alpha$ $z<-z_{\alpha/2}$ e $z>z_{\alpha/2}$

Estatística a ser utilizada =>

$$Z = \frac{(\hat{p}_1 - \hat{p}_2) - (p_1 - p_2)}{\sqrt{(\hat{p}_1\hat{q}_1 / n_1) + (\hat{p}_2\hat{q}_2 / n_2)}}$$

TESTE DE QUI-QUADRADO (χ^2)

- A estatística do teste de Qui-Quadrado é dada por:

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

Onde:

fo_i = a frequência observada na classe i;

fe_i = a frequência esperada (calculada) para a classe i;

k = o número de classes.

TESTE DE QUI-QUADRADO (χ^2)

Os graus de liberdade associados ao teste de qui-quadrado são:

- Se o teste for de **aderência:** ***gl = k - r - 1***, (k = número de classes; r = número de parâmetros necessário para a determinação das fe_i).
- De uma forma geral, tem-se:
  - *r = 0 para distribuições uniformes ou polinomiais;*
  - *r = 1 para distribuições binomial, Poisson, exponencial, etc;*
  - *r = 2 para a distribuição normal.*

- Se o teste for de **independência:** ***gl = (l - 1)(c -1)***, (*l* = número de linhas; *c* = número de colunas)

O TESTE DE QUI-QUADRADO (χ^2)

- Teste de Independência

Ho: variável linha **independe** da variável coluna

H1: variável linha **depende** da variável coluna

$$\chi^2_{obs} = \sum_{i=1}^{k} \frac{(fo_i - fe_i)^2}{fe_i}$$

gl = (l-1) (c-1)

l=número de opções variável linha pode assumir

c=número de opções variável coluna pode assumir

$$fe = \frac{(total\ linha).(total\ coluna)}{total\ geral}$$

$RRH_0 = \{\chi^2_{obs} > \chi^2_{\alpha}(gl)\}$

TESTE DE QUI-QUADRADO (χ^2)

- Teste de Aderência

Ho: Dados observados seguem a distribuição esperada

H1: Dados observados não seguem a distribuição esperada

α: Nível de significância (geralmente: 0,05 ou 0,01)

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

v = k - r – 1

$$fe_i = n \, . Probabilidade$$

$RRH_0 = \{\chi^2 > \chi^2_{\alpha}(v)\}$

Determinar o valor de χ^2 , comparar com RRH_0 e concluir.

Regressão Linear Simples

- **Linear:** porque a disposição dos pontos permite interpolar-lhes uma **Reta**;

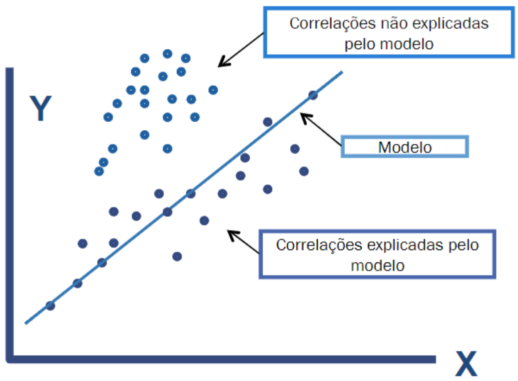
- **Simples:** porque só há **2** variáveis envolvidas no processo.

Regressão Linear Simples

- Equação geral da regressão linear simples

$$\boldsymbol{y_i = a + b x_i}$$

  - *Obter as estimativas "a" e "b" dos parâmetros α e β*
  - *α: coeficiente linear da reta*
  - *β: coeficiente angular da reta*



Regressão Linear Simples

- Equação da Reta

$$\boldsymbol{y_i = a + b.x_i}$$

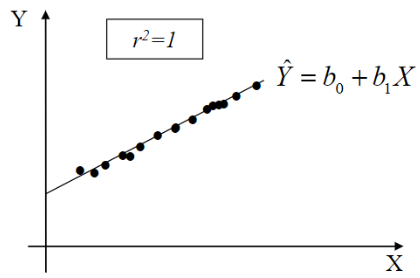
$$a = \bar{y} - b.\bar{x} \qquad b = \frac{\sum XY - \frac{(\sum X)(\sum Y)}{n}}{\sum X^2 - \frac{(\sum X)^2}{n}} = \frac{SPxy}{SQx}$$

Coeficiente de Determinação (r^2)

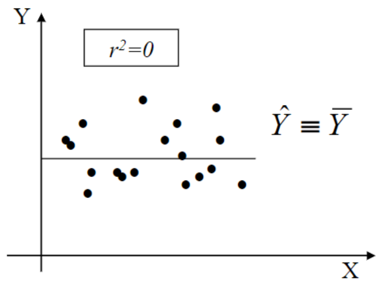
- Representa a proporção da variação total explicada pela regressão de Y para X

$$0 \leq r^2 \leq 1$$

- É o quanto da variabilidade total dos dados é explicada pelo modelo de regressão. Quanto maior o r^2 mais a variação total de Y pode ser deduzida pela introdução da variável preditora X no modelo.



A variável preditora X é responsável por toda a variação nas observações Y_i.



A variável preditora X não ajuda na redução da variação de Y_i com a Reg. Linear

Coeficiente de Determinação (r^2)

- Coeficiente de Correlação Linear

$$r = \frac{\sum xy - \frac{\sum x \, . \sum y}{n}}{\sqrt{\left(\sum x^2 - \frac{(\sum x)^2}{n}\right) . \left(\sum y^2 - \frac{(\sum y)^2}{n}\right)}} .$$

$$r = \frac{1589{,}2 - \frac{69 * 141{,}2}{8}}{\sqrt{\left(767 - \frac{69^2}{8}\right) * (3299{,}42 - \frac{141{,}2^2}{8})}} = 0{,}99$$

- Coeficiente de Determinação (r^2)

$r^2 = 0{,}98$, o que significa que 98% da variação total de y podem ser explicados pela regressão.