Métodos Quantitativos

Pedro Luiz Ramos

UNIP PÓS-GRADUAÇÃO

Apresentação do professor

Formação acadêmica/titulação

Graduação em Estatística - Universidade Estadual Paulista Júlio de Mesquita Filho, UNESP, Brasil.

Mestrado em Matemática Aplicada e Computacional - Universidade Estadual Paulista Júlio de Mesquita Filho, UNESP, Brasil.

Doutorado em Estatística - Universidade de São Paulo, USP e Universidade Federal de São Carlos, UFSCar, Brasil (sanduíche Universidade de Connecticut, EUA).

UNIP PÓS-GRADUAÇÃO

Conteúdo Programático

Conteúdo Programático

1. Análise Exploratória de Dados
2. Espaço probabilístico.
3. Modelos Probabilísticos
4. Dependência e independência de eventos.
5. Eventos condicionados.

UNIP PÓS-GRADUAÇÃO

Bibliografia

Básica:

- BUSSAB, W. O. e MORETTIN, P. A. Estatística Básica. São Paulo: Editora Saraiva, 5ed, 2006.
- LOUZADA, F. Estatística Básica Usando R, Em desenvolvimento, 2016.

UNIP
PÓS-GRADUAÇÃO

Material Didático

Este material é um resumo obtido a partir do Livro "Estatística Básica Usando R"(2015) desenvolvido pelo Prof Dr. Francisco Louzada Neto, com a minha colaboração em algumas seções. É importante ressaltar que tanto a parte teórica quando os exemplos foram retirado do livro apresentado. Sendo estes de total autoria do autor principal.

UNIP PÓS-GRADUAÇÃO

Introdução

Nas últimas décadas a grande revolução da informática possibilitou o desenvolvimento e a aplicação de métodos quantitativos em diversas áreas do conhecimento, dentre as quais podemos citar desde áreas básicas como física, a química e a biologia.

Segundo Louzada, (2015) algumas aplicações de métodos quantitativos em áreas específicas são:

Demografia: Estudo sobre fenômenos populacionais, sociais ou ambientais;

Ecologia: Estimação de tamanho populacional ou estudo da dinâmica de populações;

Sumarização de dados

As variáveis são classificadas em:

Qualitativas (ou categórica): São aquelas para as quais uma medição numérica não é possível. Subdivide-se em:

1. *Nominal:* não existe ordem definida. Exemplos: sexo, raça, grupo sanguíneo, cor de flor, sabor etc.
2. *Ordinal:* existe uma ordem definida. Exemplos: gravidade da doença (leve, moderada ou grave), nível sócio-econômico (classes A a E) etc.

UNIP PÓS-GRADUAÇÃO

Sumarização de dados

Quantitativas (ou numéricas): São aquelas para as quais é possível realizar-se uma medição numérica. Subdivide-se em:

1. *Discretas:* próprias de dados de contagem, ou seja, só assumem valores inteiros. Exemplos: número de filhos, número de acidentes de trânsito ocorridos num certo período, número de ovos depositados por um inseto, número de pessoas desempregadas numa família etc.

2. *Contínuas:* são aquelas originárias de medições que, deste modo, podem assumir qualquer valor real entre dois extremos. Exemplos: peso corporal, altura, resistência a ruptura, volume, índice de massa corporal, tempo que um medicamento demora para fazer efeito etc.

Sumarização de dados

A representação gráfica é uma maneira eficiente e simples de apresentar os dados. As variáveis qualitativas podem ser representadas por:

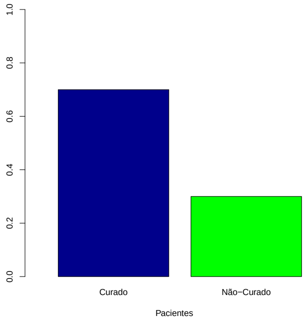
- Gráfico em barras;
- Gráfico em setores (Gráfico de "pizza");
- Gráfico em retângulo.

As variáveis quantitativas, podem ser representadas por:

- Diagrama de pontos;
- Histogramas;
- Polígono de frequências;

Sumarização de dados

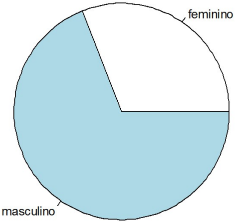
Gráfico em Barra: Considere um exemplo aplicado a pacientes que participaram de um determinado tipo de terapia em que uma das variáveis qualitativas presente é a cura dos pacientes. A Figura a seguir apresenta o gráfico em barras para a variável referente ao numero de pacientes curados 700 (70%).



UNIP PÓS-GRADUAÇÃO

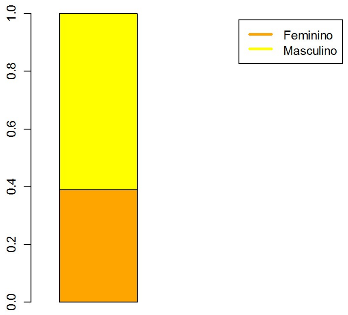
Sumarização de dados

Gráfico em setores: A figura a seguir apresenta o gráfico em setores para a variável sexo dos pacientes analisados no exemplo anterior. Por meio deste gráfico pode-se perceber que a maioria (69,09%) dos indivíduos analisados são do sexo masculino.



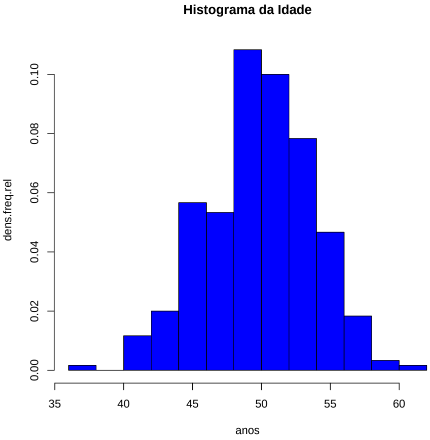
Sumarização de dados

Gráfico em retângulo: A figura a seguir apresenta o gráfico em retângulo para a variável sexo dos pacientes analisados no exemplo anterior. Por meio deste gráfico pode-se perceber que a maioria (69,09%) dos indivíduos analisados são homens.



Sumarização de dados

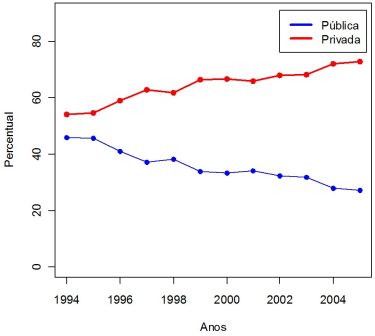
Histograma: A figura a seguir apresenta o histograma para a variável idade do paciente. Notamos que a idade dos pacientes em estudo está compreendida entre 35 e 62 anos, tratando-se de uma distribuição simétrica com alta concentração de clientes de 50 anos.



UNIP PÓS-GRADUAÇÃO

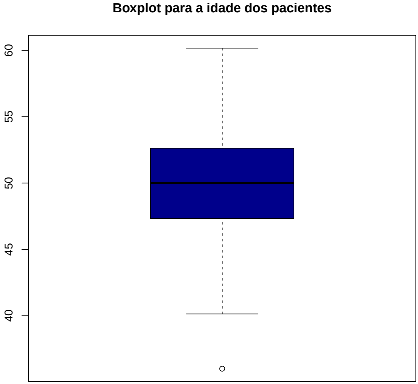
Sumarização de dados

Gráfico temporal: A figura a seguir mostra um gráfico da série temporal para o percentual de alunos ingressantes na UFSCar no período de 1994 a 2005 que fizeram o ensino médio em instituições públicas e privadas.



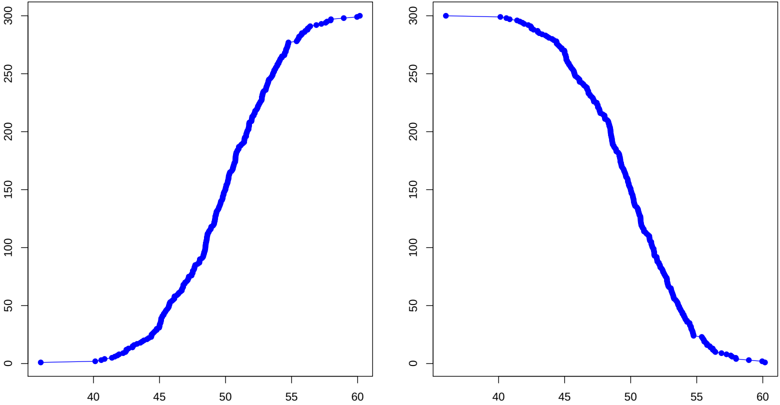
Sumarização de dados

Box plot: O box plot é um tipo de gráfico que mostra simultaneamente as características de centro, dispersão, desvios da simetria e identificação de observações discrepantes de um conjunto de dados. A Figura a seguir representa o box plot para a idade dos pacientes.



Sumarização de dados

Polígono de frequências



Sumarização de dados

Média aritmética: medida de tendência central

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

Exemplo 10: Seja o conjunto de dados abaixo formado pela altura de 10 pacientes em estudo.

$$\begin{aligned}
\overline{X} &= \frac{\sum_{i=1}^{n} X_i}{n} = \frac{\sum_{i=1}^{8} X_i}{8} \\
&= \frac{X_1 + X_2 + X_3 + X_4 + X_5 + X_6 + X_7 + X_8}{8} \\
&= \frac{1,85 + 1,90 + 1,35 + 1,75 + 1,70 + 1,50 + 1,65 + 1,70}{8} \\
&= \frac{13,40}{8} = 1,675kg.
\end{aligned}$$

UNIP PÓS-GRADUAÇÃO

Sumarização de dados

Moda: Seja um conjunto de dados formado por $X_1, X_2, ..., X_n$, então a moda ou o valor modal, denominado de Mo, é dado por:

$$Mo = X_{freq}$$

em que X_{freq} é o valor mais frequente. Dizemos que o conjunto de dados é:

Amodal: Quando não apresenta nenhum valor mais frequente;

Unimodal: Quando apresenta um valor mais frequente;

Bimodal: Quando apresenta dois valores mais frequentes;

Trimodal: Quando apresenta três valores mais frequentes;

Multimodal: Quando o conjunto de dados apresenta

quatro ou mais de quatro valores mais frequentes.

UNIP PÓS-GRADUAÇÃO

Exemplos:

$17, 20, 18, 25, 11, 28, 13, 23$ - Amodal, pois não há valor mais frequente;

$11, 13, 8, 5, 14, 9, 8, 12, 8$ - Unimodal, pois $Mo = 8$;

$53, 48, 50, 48, 49, 51, 51, 55$ - Bimodal, pois $Mo = 48$ e $Mo = 51$;

$81, 85, 81, 74, 82, 83, 82, 86, 83$ - Trimodal, pois $Mo = 81$, $Mo = 82$ e $Mo = 83$.

Sumarização de dados

Mediana: A mediana é uma medida de tendência central que deixa 50% dos dados abaixo e 50% dos dados acima de si mesma, dividindo as observações ordenadas em duas partes iguais.

Caso 1. Quando o número de dados (n) for ímpar, a mediana é dada por:

$$Me = X_{\left(\frac{n+1}{2}\right)}.$$

Caso 2. Quando o número de dados (n) for par, a mediana é dada por:

$$Me = \frac{X_{\left(\frac{n}{2}\right)} + X_{\left(\frac{n+2}{2}\right)}}{2}.$$

Exemplo 11: Seja o conjunto de dados (ordenado) formado por $(34, 37, 40, 41, 41, 41, 44)$. Então, como n é ímpar $(n = 7)$ temos que a mediana é dada por:

$$Me = X_{\left(\frac{n+1}{2}\right)} = X_{(4)} = 41.$$

Seja o conjunto de dados (ordenado) formado por $(7, 8, 8, 9, 10, 10, 12, 14)$. Então, como n é par $(n = 8)$ temos que a mediana é dada por:

$$Me = \frac{X_{\left(\frac{n}{2}\right)} + X_{\left(\frac{n+2}{2}\right)}}{2} = \frac{X_{(4)} + X_{(5)}}{2} = \frac{9 + 10}{2} = \frac{19}{2} = 9,5$$

Sumarização de dados

Abaixo seguem alguns exemplos de conjuntos de dados diferentes com média igual.

$$
\begin{array}{llcl}
\text{Conjunto A:} & (30, 30, 30, 30, 30) & \Rightarrow & \overline{X} = 30; \\
\text{Conjunto B:} & (60, 10, 10, 10, 60) & \Rightarrow & \overline{X} = 30; \\
\text{Conjunto C:} & (20, 40, 20, 40, 30) & \Rightarrow & \overline{X} = 30; \\
\text{Conjunto D:} & (10, 20, 30, 40, 50) & \Rightarrow & \overline{X} = 30; \\
\text{Conjunto E:} & (10, 20, 90, 20, 10) & \Rightarrow & \overline{X} = 30.
\end{array}
$$

Sumarização de dados

Amplitude: É a medida de dispersão mais simples, em trata-se da diferença entre o maior e o menor valor observado. A amplitude (A) é dada por

$$A = X_{\max} - X_{\min}.$$

Considerando os conjuntos anteriores, vamos determinar a amplitude de cada um deles.

Conjunto A: $A = X_{\max} - X_{\min} = 30 - 30 = 0$;

Conjunto B: $A = X_{\max} - X_{\min} = 60 - 10 = 50$;

Conjunto C: $A = X_{\max} - X_{\min} = 40 - 20 = 20$;

Conjunto D: $A = X_{\max} - X_{\min} = 50 - 10 = 40$;

Conjunto E: $A = X_{\max} - X_{\min} = 90 - 10 = 80$.

UNIP PÓS-GRADUAÇÃO

Sumarização de dados

Variância: Esta medida pode ser entendida como se fosse praticamente a "média" da soma de quadrados de desvios em relação à média.

Definição: Seja um conjunto de dados formado por $X_1, X_2, ..., X_n$, e seja μ a média populacional, então a variância populacional é dada por:

$$\sigma^2 = \frac{\sum_{i=1}^{n} (X_i - \mu)^2}{n}.$$

Para o caso da variância amostral temos:

$$s^2 = \frac{\sum_{i=1}^{n} \left(X_i - \overline{X}\right)^2}{n-1}.$$

UNIP PÓS-GRADUAÇÃO

Sumarização de dados

Desvio Padrão: É a medida de dispersão mais utilizada na estatística. Trata-se da raíz quadrada da variância. **Definição:** Considere um conjunto de dados formado por $X_1, X_2, ..., X_n$, e seja μ a média populacional, então o desvio-padrão populacional é dada por:

$$\sigma = \sqrt{\sigma^2} = \sqrt{\frac{\sum_{i=1}^{n} (X_i - \mu)^2}{n}}.$$

Para o caso do desvio-padrão amostral temos:

$$s = \sqrt{s^2} = \sqrt{\frac{\sum_{i=1}^{n} \left(X_i - \overline{X}\right)^2}{n-1}}.$$

UNIP PÓS-GRADUAÇÃO

Sumarização de dados

Exemplos: Considerando os conjuntos anteriores, vamos determinar o desvio-padrão populacional de cada um deles.

Conjunto A:	$(30, 30, 30, 30, 30)$	$\Rightarrow$	$\sigma = 0$	e	$s = 0$;
Conjunto B:	$(60, 10, 10, 10, 60)$	$\Rightarrow$	$\sigma = 24,49$	e	$s = 27,39$;
Conjunto C:	$(20, 40, 20, 40, 30)$	$\Rightarrow$	$\sigma = 8,94$	e	$s = 10$;
Conjunto D:	$(10, 20, 30, 40, 50)$	$\Rightarrow$	$\sigma = 14,14$	e	$s = 15,81$;
Conjunto E:	$(10, 20, 90, 20, 10)$	$\Rightarrow$	$\sigma = 30,33$	e	$s = 33,91$.

Elementos de Probabilidade

Economia: Estudo sobre a evolução/previsão da inflação ou rendimento da bolsa de valores ao longo do tempo;

Indústria: Controle da qualidade ou estudo do tempo de garantia de um produto;

Medicina: Estudo do tempo de vida de pacientes com uma determinada doença, comparação da eficácia de tratamentos ou lançamento de um novo medicamento;

Política: Pesquisa de intenção de votos numa eleição ou pesquisa sobre popularidade de um determinado candidato.

Meteorologia: Previsão de temperaturas ou chuvas.

UNIP PÓS-GRADUAÇÃO

Elementos de Probabilidade

Todos os dias somos confrontados com situações, que nos conduzem a utilizar, intuitivamente, a noção de probabilidade:

Dizemos que existe uma grande probabilidade de não chover num dia de Verão;

Dizemos que existe uma pequena probabilidade de ganhar na loteria;

O político deseja saber qual a sua probabilidade de ganhar as eleições;

O médico deseja saber qual a probabilidade de um doente sobreviver ao ser tratado com um novo medicamento.

Exemplo 1 : Considere inicialmente o lançamento de um dado. Supondo que o composto do material é homogêneo, de tal forma que todas as faces tenham igual probabilidade. Teremos as seguintes possibilidade

Face cima	1	2	3	4	5	6
Probabilidade	$1/6$	$1/6$	$1/6$	$1/6$	$1/6$	$1/6$

O fato de admitir este modelo de probabilidade para o número da face que fica virada para cima ao lançar um dado permite-nos agora construir modelos para experimentos mais elaborados.

Elementos de Probabilidade

Alguns outros experimentos:

1. Lançamento de uma moeda e leitura da figura da face voltada para cima;
2. Lançamento de um dado comum e leitura do número voltado para cima;
3. Nascimento de uma criança;
4. Sorteio de uma carta de baralho;
5. Altura (em cm) de uma pessoa sorteada da população;
6. Peso (em gramas) de um recém-nascido.

PÓS-GRADUAÇÃO

Classificador naive Bayes

$$P(A_k \mid B) = \frac{P(B \mid A_k)\, P(A_k)}{\sum_{i=1}^{n} P(B \mid A_i)\, P(A_i)},\ k = 1, 2, ..., n.$$

Elementos de Probabilidade

Experimentos aleatórios estão sujeitos a lei do acaso, além disto os mesmos estão associados a um espaço amostral (denotado por Ω).

Considere os seguintes experimentos:

No lançamento de uma moeda temos que o espaço amostral é:

$$\Omega = \{c, k\}\text{, em que } c = cara \text{ e } k = coroa.$$

Em lançamentos independentes de uma moeda até ocorrer a primeira cara, temos o seguinte espaço amostral:

$$\Omega = \{c, (k, c), (k, k, c), (k, k, k, c), ..., (k, k, ..., k, c), ...\}.$$

UNIP PÓS-GRADUAÇÃO

Elementos de Probabilidade

Exemplo 2: Considere o lançamento de um dado e observe a face voltada para cima. Temos então que $\Omega = \{1, 2, 3, 4, 5, 6\}$. Agora considere os seguintes eventos:

Evento A: "Ocorre face par" $A = \{2, 4, 6\}$;

Evento B: "Ocorre face menor ou igual a 3" $B = \{1, 2, 3\}$;

Evento C: "Ocorre face ímpar" $C = \{1, 3, 5\}$;

Evento D: "Ocorre face maior que 5" $D = \{6\}$;

Evento E: "Ocorre face maior que 20". $E = \{\varnothing\}$;

Evento F: "Ocorre face maior ou igual a 1 e menor ou igual a 6" $F = \{\Omega\}$;

UNIP PÓS-GRADUAÇÃO

Elementos de Probabilidade

A teoria dos conjuntos é um ramo da matemática extremamente útil no estudo probabilístico de eventos uma vez que estes nada mais são que subconjuntos de um espaço amostral. Consideremos um espaço amostral finito dado por:

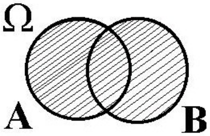
$$\Omega = \{\omega_1, \omega_2, ..., \omega_n\}.$$

Sejam A e B dois eventos de Ω. Temos três operações básicas com eventos aleatórios: união, intersecção e complementação.

UNIP PÓS-GRADUAÇÃO

União: O evento união é formado pelos pontos amostrais ω que pertencem a pelo menos um dos eventos A e B.

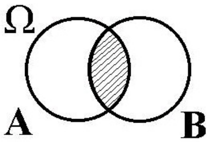
Definição: $A \cup B = \{\omega \in \Omega : \omega \in A \text{ ou } \omega \in B\}$.



Elementos de Probabilidade

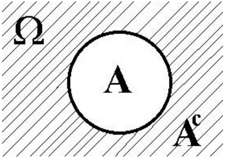
Intersecção: O evento intersecção é formado pelos pontos amostrais ω que pertencem simultaneamente aos eventos A e B.

Definição: $A \cap B = \{\omega \in \Omega : \omega \in A \text{ e } \omega \in B\}$.



Complementação: O evento complementação é formado pelos pontos amostrais ω que não pertencem a ao evento em questão.

Definição: $A^c = \Omega - A = \{\omega \in \Omega : \omega \notin A\}$.



Exemplo 3: Considere novamente o lançamento de um dado e observe a face voltada para cima em que

$$A \cup B \qquad A \cap B \qquad A \cup C \qquad A \cap C \qquad B \cup C$$
$$B \cap C \qquad A \cup E \qquad A \cap E \qquad A \cup D \qquad A \cap D$$

Temos que:

$A \cup B = \{1, 2, 3, 4, 6\}$.
$A \cap B = \{2\}$.
$A \cup C = \{1, 2, 3, 4, 5, 6\} = \Omega$.
$A \cap C = \{\} = \varnothing$.
$B \cup C = \{1, 2, 3, 5\}$.
$B \cap C = \{1, 3\}$.
$A \cup E = \{2, 4, 6\} = A$.
$A \cap E = \{\} = \varnothing$.
$A \cup D = \{2, 4, 6\} = A$.
$A \cap D = \{6\}$.

Elementos de Probabilidade

A Probabilidade é a possibilidade ou a chance de ocorrência de um evento definido sobre um espaço amostral. Note-se que a probabilidade é a proporção ou fração própria cujos valores variam de 0 a 1 inclusives.

Considere a três abordagens sobre o esse tema:

1. Qual é a chance de se retirar uma carta de ouros de um baralho comum?

2. Qual é a chance de que um indivíduo prefira um produto a outro?

3. Qual é a chance de que um novo produto, lançado no mercado, tenha sucesso junto ao consumidor?

UNIP PÓS-GRADUAÇÃO

Exemplo 4: No lançamento de um dado honesto, observando-se a face voltada para cima, determinar a probabilidade de ocorrência dos eventos:

a) Face ímpar.

b) Face maior do que 2.

c) Face ímpar ou maior do que 2.

d) Face maior do que 2 e face ímpar.

Solução: No espaço amostral $\Omega = \{1, 2, 3, 4, 5, 6\}$ tem-se que:

a) $P(\text{face ímpar}) = P(\{1, 3, 5\}) = 3/6 = 0,5$ ou 50%.

b) $P(\text{face maior do que } 2) = P(\{3, 4, 5, 6\}) = 4/6 = 2/3$

$$= 0,667 \text{ ou } 66,7\%.$$

c) $P(\text{face ímpar ou maior do que } 2) = P(\{1, 3, 4, 5, 6\}) =$

$$= 5/6 = 0,833 \text{ ou } 83,3\%.$$

d) $P(\text{face maior do que 2 e ímpar}) = P(\{3, 5\}) = 2/6 = 1/3 =$

$$= 0,333 \text{ ou } 33,3\%.$$

UNIP PÓS-GRADUAÇÃO

Elementos de Probabilidade

Definição: Probabilidade é a função P que associa a cada evento A um número real pertencente ao intervalo $[0, 1]$, satisfazendo os axiomas:

1. $0 \leq P(A) \leq 1$;
2. $P(\Omega) = 1$;
3. Se $A_1, A_2, \ldots$ for uma sequência de eventos mutuamente exclusivos, isto é, $A_i \cap A_j = \emptyset$, $i \neq j$, então temos:

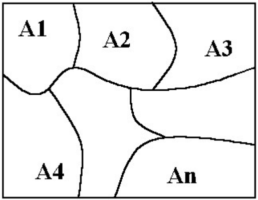
$$P\left(\bigcup_{i=1}^{\infty} A_i\right) = \sum_{i=1}^{\infty} P(A_i).$$

Podemos verificar que se $P(\Omega) = 1$ então $P(\emptyset) = 0$.

UNIP PÓS-GRADUAÇÃO

Partição de um espaço amostral

Seja o espaço amostral Ω conforme o diagrama abaixo:



Definição: Dizemos que os eventos $A_1, A_2, ..., A_n$ formam uma partição do espaço amostral Ω se:

1. $A_i \neq \emptyset,\ i = 1, 2, ..., n;$
2. $A_i \cap A_j = \emptyset,\ \forall i \neq j;$
3. $\bigcup_{i=1}^{n} A_i = \Omega.$

UNIP PÓS-GRADUAÇÃO

Exemplo 5: Considere novamente o lançamento de um dado honesto, observando-se a face voltada para cima, e os seguintes eventos:

A: "Ocorre face par" $A = \{2, 4, 6\}$;

B: "Ocorre face menor ou igual a 3" $B = \{1, 2, 3\}$

Determine a P$(A \cup B)$.

Solução: No espaço amostral $\Omega = \{1, 2, 3, 4, 5, 6\}$ tem-se que:

$$P(A) = P(\{1, 3, 5\}) = 3/6 = 1/2$$

$$P(B) = P(\{1, 2, 3\}) = 3/6 = 1/2$$

$$P(A \cap B) = P(\{1, 3\}) = 1/3$$

Logo

$$P(A \cup B) = P(A) + P(B) - (A \cap B) = 1/2 + 1/2 - 1/3 = 2/3$$

UNIP PÓS-GRADUAÇÃO

Elementos de Probabilidade

Probabilidade condicional e independência

Seja $A \subset \Omega$ e $B \subset \Omega$, então a probabilidade condicional de A dado que B ocorreu é dada por:

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

Lê-se: "Probabilidade de A dado B".

Seja $A \subset \Omega$ e $B \subset \Omega$, então dizemos que A é **independente** de B se sua probabilidade condicional for dada por:

$$P(A \mid B) = P(A).$$

Exemplo 6: Lançam-se três moedas. Deseja-se verificar se são independentes os seguintes eventos:

A : "Saída de cara na primeira moeda".

B : "Saída de coroa na segunda e terceira moeda".

Solução: Temos que o espaço amostral para esse experimento e os eventos propostos são:

$$\Omega = \{(c,c,c), (c,c,k), (c,k,c), (c,k,k), (k,c,c), (k,c,k), (k,k,c), (k,k,k)\}$$

$$A = \{(c,c,c), (c,c,k), (c,k,c), (c,k,k)\}$$

$$B = \{(c,k,k), (k,k,k)\}$$

$$A \cap B = \{(c,k,k)\}$$

Assim,

$P(A \mid B) = \frac{P(A \cap B)}{P(B)} = \frac{1}{2}$ e $P(A) = \frac{4}{8} = \frac{1}{2}$.

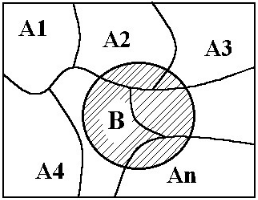
Como $P(A \mid B) = P(A)$ então A e B são dois eventos independentes.

Elementos de Probabilidade

Teorema da Probabilidade Total: Sejam $A_1, A_2, ..., A_n$ eventos que formam uma partição do espaço amostral Ω. Seja B um evento desse espaço. Então temos

$$P(B) = \sum_{i=1}^{n} P(B \mid A_i)\, P(A_i)$$

Esquematicamente temos:



Exemplo 7: Uma urna contém 3 bolas brancas e 2 amarelas. Uma segunda urna contém 4 bolas brancas e 2 amarelas. Escolhe-se ao acaso, uma urna e dela retira-se, também ao acaso, uma bola. Qual a probabilidade de que essa bola retirada seja branca?

Solução: Sejam os eventos:

I : "A urna escolhida é a urna I".

II : "A urna escohida é a urna II".

$B \mid I$: "A bola é branca dado que a urna escolhida foi a I".

$B \mid II$: "A bola é branca dado que a urna escolhida foi a II".

B : "A bola escolhida é branca".

Elementos de Probabilidade

E, as respectivas probabilidades são $P(I) = \frac{1}{2}$, $P(II) = \frac{1}{2}$, $P(B \mid I) = \frac{3}{5}$, $P(B \mid II) = \frac{4}{6}$. Então temos que a probabilidade $P(B)$ é dada por:

$$
\begin{aligned}
P(B) &= \sum_{i=1}^{n} P(B \mid A_i)\, P(A_i) = P(B \mid I)\, P(I) + P(B \mid II)\, P(II) \\
&= \frac{3}{5} \times \frac{1}{2} + \frac{4}{6} \times \frac{1}{2} = \frac{3}{10} + \frac{2}{6} = \frac{38}{60} = \frac{19}{30}.
\end{aligned}
$$

UNIP PÓS-GRADUAÇÃO

Teorema de Bayes

Considere $A_1, A_2, ..., A_n$ eventos que formam uma partição do espaço amostral Ω e que $P(A_i)$ e $P(B \mid A_i)$, $i = 1, 2, ..., n$, sejam conhecidas. Então temos:

$$P(A_k \mid B) = \frac{P(B \mid A_k)\, P(A_k)}{\sum_{i=1}^{n} P(B \mid A_i)\, P(A_i)},\ k = 1, 2, ..., n.$$

Observação: O Teorema de Bayes também é conhecido como Teorema da pro-babilidade *a posteriori*. Ele relaciona uma das parcelas da probabilidade total com a própria probabilidade total.

Variáveis aleatórias

Uma variável é dita aleatória quando o valor da mesma é obtido por meio de observações ou experimentos, e a cada valor estiver associada uma certa probabilidade. Denota-se uma variável por letra maiúscula (por exemplo X, Y, Z) e os valores assumidos por letra minúscula (x, y, z).

Discreta: Uma variável é dita **discreta** quando assume valores em pontos isolados ao longo de uma escala (n^o finito ou infinito enumerável de valores).

Contínua: Uma variável é dita **contínua** quando assume qualquer valor ao longo de um intervalo (n^o infinito não enumerável de valores).

UNIP PÓS-GRADUAÇÃO

Distribuições discretas de probabilidade: Seja X uma variável aleatória e $x_1, x_2, ..., x_n$ um conjunto finito de valores de X. Então a distribuição de probabilidade (ou função de probabilidade) tem que satisfazer:

i) $0 \leq P\left(X = x_k\right) \leq 1,\ k = 1, 2, ..., n;$

ii) $\sum_{k=1}^{n} P\left(X = x_k\right) = 1.$

Exemplo 8: Considere o lançamento de duas moedas e seja X o número de "caras" obtidas. Sabemos que o espaço amostral é dado por $\Omega = \{(c,c), (c,k), (k,c), (k,k)\}$ e, portanto, a quantidade de caras que X pode assumir é $X = 0, 1, 2$, tal que:

$$P\left(X = 0\right) = 1/4, \qquad P\left(X = 1\right) = 2/4,$$
$$P\left(X = 2\right) = 1/4.$$

UNIP PÓS-GRADUAÇÃO

Variáveis aleatórias

Distribuições contínuas de probabilidade: Seja X uma variável aleatória absolutamente contínua associada a uma função $f(x)$. Então sua função densidade de probabilidade tem que satisfazer:

i) $f(x) \geq 0$ para todo $x \in \mathbb{R}$

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

Note que para qualquer $a, b \in \mathbb{R}$, em que $a < b$ temos

$$P(a < X < b) = \int_a^b f(x)dx.$$

Esperança e variância de uma variável aleatória

Exemplo 9: Conside o exemplo em que temos o lançamento de duas moedas e X é o número de "caras"obtidas. Um valor médio ou esperado para X pode ser calculado da seguinte maneira:

$$\begin{aligned} E(X) &= 0 \times P(X=0) + 1 \times P(X=1) + 2 \times P(X=2) \\ &= 0 \times \frac{1}{4} + 1 \times \frac{2}{4} + 2 \times \frac{1}{4} \\ E(X) &= 1. \end{aligned}$$

Variáveis aleatórias

Para determinarmos a variância $VAR(X)$ temos que encontrar primeiramente o segundo momento de X, dado por:

$$
\begin{aligned}
E\left(X^2\right) &= 0^2 \times P(X=0) + 1^2 \times P(X=1) + 2^2 \times P(X=2) \\
&= 0^2 \times \frac{1}{4} + 1^2 \times \frac{2}{4} + 2^2 \times \frac{1}{4} \\
E\left(X^2\right) &= 3/2.
\end{aligned}
$$

E por sua vez, a variância é dada da seguinte forma:

$$
\begin{aligned}
VAR(X) &= E\left(X^2\right) - [E(X)]^2 = \frac{3}{2} - 1^2 \\
VAR(X) &= 1/2.
\end{aligned}
$$

UNIP PÓS-GRADUAÇÃO

Distribuições de Probabilidade

A distribuição Normal desempenha papel preponderante em métodos quantitativos, e os processos de inferência nela baseados têm larga aplicação.

Definição: Dizemos que X tem distribuição Normal se sua função densidade de probabilidade (f.d.p) é dada por:

$$f(x) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left\{-\frac{1}{2\sigma^2}(x-\mu)^2\right\},$$

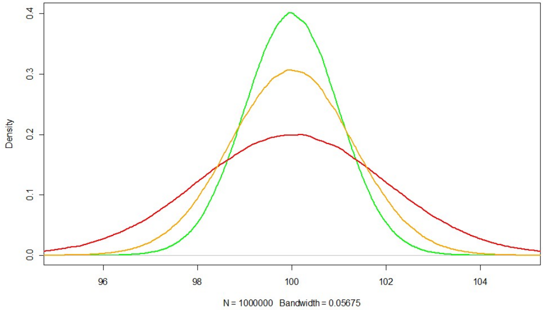
onde $-\infty < x < \infty$, $-\infty < \mu < \infty$ e $\sigma > 0$.

Notação: $X \sim N\left(\mu, \sigma^2\right)$

UNIP PÓS-GRADUAÇÃO

Distribuições de Probabilidade

A Figura a seguir mostra três distribuições Normais com a mesma média, mas com variâncias diferentes.



UNIP PÓS-GRADUAÇÃO

Distribuições de Probabilidade

A curva Normal tem as seguintes características

1. $E(X) = \mu$ e $VAR(X) = \sigma^2$.
2. O ponto máximo de $f(x)$ é μ.
3. A distribuição tem forma de sino e é simétrica em torno de μ.
4. A moda e a mediana são iguais a média, $Me = Mo = \mu$.
5. Os pontos de inflexão da curva são $[\mu - \sigma; \mu + \sigma]$.
6. O intervalo $[\mu - 1\sigma; \mu + 1\sigma]$ compreende pelo menos $68,26\%$ dos dados.
7. O intervalo $[\mu - 2\sigma; \mu + 2\sigma]$ compreende pelo menos $95,44\%$ dos dados.
8. O intervalo $[\mu - 3\sigma; \mu + 3\sigma]$ compreende pelo menos $99,74\%$ dos dados.

UNIP
PÓS-GRADUAÇÃO

Distribuições de Probabilidade

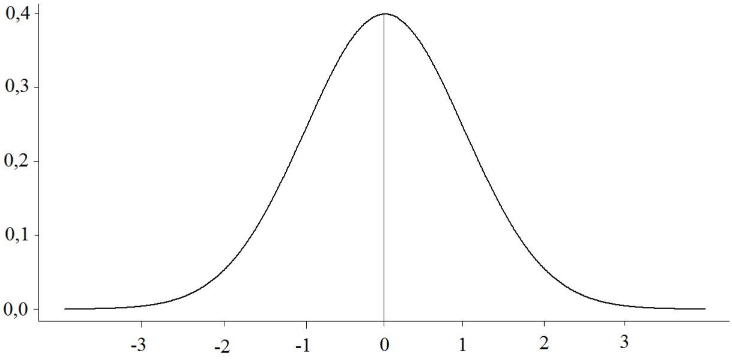
Problema: A determinação das probabilidades são realizadas por meio de aproximações numéricas, sendo difícil de obtê-las analiticamente.

Solução: Esta tarefa é facilitada por meio do uso da distribuição Normal padrão definida a seguir.

Resultado: Se $X \sim N\left(\mu, \sigma^2\right)$ então a v.a.

$$Z = \frac{X - \mu}{\sigma} \sim N\left(0, 1\right).$$

Distribuições de Probabilidade



Cálculo de probabilidades usando a curva padrão

Passos para a determinação das proporções Normais:

Passo 1. Enunciar o problema em termos da variável X observada.

Passo 2. Padronizar X para reformular o problema em termos de uma variável Normal padrão Z. Fazer o gráfico para mostrar a área sob a curva Normal padrão.

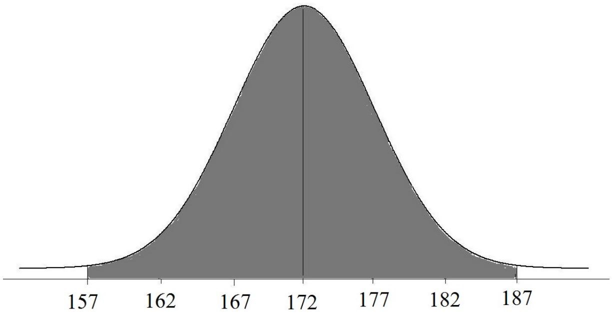
Passo 3. Determinar a área solicitada sob a curva Normal padrão, usando a tabela da curva Normal Z e o fato de que a área total sob a curva é 1.

Exemplo 12: Foi feito um estudo sobre a altura dos alunos de uma faculdade, observando-se que esta é distribuída normalmente com média $1,72m$ e desvio-padrão $5cm$. Qual a porcentagem dos alunos com altura:

a) Entre $1,57m$ e $1,87m$?

b) Acima de $1,90m$?

Solução: a) Seja X a altura dos alunos desta faculdade (em cm), então temos que $X \sim N(172, 25)$. Então a área de interesse é:



Logo, pela padronização temos:

$$
\begin{aligned}
P\,(157 \leq X \leq 187) &= P\left[\left(\frac{157-172}{5}\right) \leq Z \leq \left(\frac{187-172}{5}\right)\right] \\
&= P\,(-3 \leq Z \leq 3) \\
&= P\,(-3 \leq Z \leq 0) + P\,(0 \leq Z \leq 3) \\
&= 0,4987 + 0,4987 = 0,9974.
\end{aligned}
$$

Ou seja, a probabilidade de que um aluno desta faculdade tenha uma altura entre $1,57m$ e $1,87m$ é de $99,74\%$.

Distribuição Amostral

Abordamos agora noções introdutórias das distribuição da média amostral em populações finitas e estendemos esses conceitos para populações infinitas.

Amostra aleatória: Conjunto de n variáveis aleatórias, independentes entre si, com a mesma distribuição de probabilidade $f(\cdot)$, em que cada elemento da população tem a mesma probabilidade de ser incluído na amostra.

Estatísticas: Funções de, e apenas de observações amostrais, ou seja, de variáveis aleatórias (dados) e que, portanto são elas próprias variáveis aleatórias. Dentre elas podemos citar a média amostral a variância e etc...

Estimação: Processo de obtenção de aproximações numéricas para parâmetros associados a $f(\cdot)$.

Estimativa: Uma aproximação numérica particular (ou seja, na dada amostra) para parâmetro(s) associado(s) a $f(\cdot)$.

Estimador: Função ou o processo numérico que permite a geração de estimativas.

Distribuição Amostral

Amostragem: O objetivo da amostragem é estimar os parâmetros da população por meio de amostra(s). A amostragem apresenta muitas vantagens em relação ao censo. Algumas Vantagens para sua utilização são

1 - Custo reduzido;

2- Maior rapidez;

3- Maior amplitude;

4- Maior exatidão.

Distribuição Amostral

Há vários tipos de amostragens, as quais se concentram em dois grupos, probabilísticas e não-probabilísticas, a seguir, apresentamos algumas formas de amostragens em cada um desses grupos.

1. Amostragem aleatória simples;
2. Amostragem aleatória estratificada;
3. Amostragem aleatória sistemática;
4. Amostragem aleatória por conglomerado.

Amostragem não-probabilística:

1. Inacessibilidade a toda a população;
2. Amostragem sem norma;
3. Amostragem intencional.

UNIP PÓS-GRADUAÇÃO

Distribuição Amostral

Ainda no campo de amostragem, nos deparamos com dois tipos de populações, a finita ou infinita enumerável e a infinita.

População finita: é a população em que se pode contar ou enumerar todos os seus elementos. Exemplos:

1. Número de itens produzidos na linha de produção, em um dia;
2. Número de pessoas com certa doença na cidade desejada;

População infinita: é aquela população na qual é impossível contar ou enumerar todos os seus elementos. Exemplos:

1. População de mamíferos selvagens no Pantanal MG;
2. Produção brasileira de certo equipamento eletrônico;
3. Número de acidentes de trânsito em uma determinada rodovia;

Calculo do tamanho Amostral

Considere o caso em que queremos calcular o tamanho amostral n, para populações finitas. Temos que

$$n = \frac{N \times p \times (1-p) \times z^2_{\alpha/2}}{(N-1) \times E^2 + p \times (1-p) \times z^2_{\alpha/2}} \tag{1}$$

- $z_{\alpha/2}$ é o valor da curva normal padrão;
- N é o tamanho da população;
- E é o erro máximo admitido;
- $(1-\alpha)$ é o nível de confiança;
- p é o estimador da proporção

UNIP PÓS-GRADUAÇÃO

Calculo do tamanho Amostral

Para utilizar essa fórmula, é necessário encontrar um estimador para p. É possível encontrar tal estimador baseando em resultados de pesquisas anteriores ou de uma amostra piloto. Quando nenhuma pesquisa é realizada previamente uma forma alternativa é utilizar o fato que $p(1-p) \leq 1/4$. Neste, caso teremos um valor conservativo para n. Desta forma, com um nível de confiança $de 95\%$ teremos os seguintes tamanhos amostrais dependendo do erro admitido com um nível de confiança de 95%:

UNIP PÓS-GRADUAÇÃO

Calculo do tamanho Amostral

Tabela: Tamanho amostral considerando diferentes tipos de erros.

Erro (E)	Tamanho amostral (n)
0.01	3288
0.02	1622
0.03	880
0.04	536
0.05	375
0.06	253
0.07	189
0.08	146
0.09	116
0.10	94
0.11	78
0.12	66
0.13	56

UNIP PÓS-GRADUAÇÃO

Calculo do tamanho Amostral

Observações: Neste caso para o cálculo do tamanho amostral não utilizamos a idéia de probabilidades associadas aos erros tipos I e II popularmente convencionadas como α e β, pois estes valores são utilizados apenas no cálculo comparação de dois grupos.

A interpretação do nível de confiança e do erro máximo admitido é dada a seguir: Quando se afirma que o erro máximo admitido é de cinco pontos $(E = 0.05)$ percentuais, e que o intervalo de confiança é de 95%, está se afirmando que, se na amostra a porcentagem de hipertensos é de 30$\%$, na população a porcentagem deve estar entre 25% e 35% (margem de erro). Além disso, como o intervalo de confiança é de 95%, é possível que uma em cada 20 pesquisas realizadas com a mesma metodologia irá apresentar um resultado fora da margem de erro.

UNIP
PÓS-GRADUAÇÃO

Calculo do tamanho Amostral

Miot, H. A. (2011). Tamanho da amostra em estudos clínicos e experimentais. J Vasc Bras, 10(4), 275-8.

PÓS-GRADUAÇÃO

Testando Correlações

Relação entre as variáveis: Havendo indicativo lógico ou suposição fundamentada de que pode uma variável pode influenciar outra, podemos iniciar a investigação de relação entre elas.

Algumas correlações estranhas

http://www.tylervigen.com/spurious-correlations

UNIP PÓS-GRADUAÇÃO

Testando Correlações

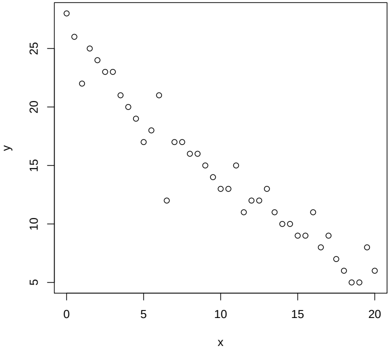
Havendo explicações plausíveis entre a relação de duas ou mais variáveis, podemos estabelecer a relação entre elas. Como por exemplo:

- Peso e altura de uma pessoa;
- Tamanho e idade gestacional de um feto;
- Número de clientes de representante comercial e seu tempo de trabalho no ramo;
- Preço de um produto e quantidade vendida;
- Produção de melancias e a quantidade de chuva no período (irrigação).

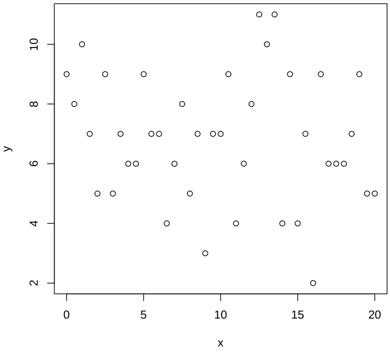
Poderiamos tecer muitas outras possibilidades de variáveis relacionadas, mas vamos abranger outros aspectos da teoria. Primeiramente vamos identificar os tipos de variáveis.

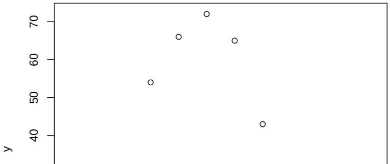
Testando Correlações

Supomos que X e Y são respectivamente variáveis explicativa e resposta.



Testando Correlações





Testando Correlações

O coeficiente de correlação linear de Pearson populacional é geralmente denotado por ρ, mas como trabalhamos com amostras, precisamos de um estimador para ρ que é denominado coeficiente de correlação linear de Pearson amostral, denotado por r e assume valor no intervalo $[-1, 1]$.

$$r = \frac{\sum_{i=1}^{n}(X_i - \overline{X})(Y_i - \overline{Y})}{(n-1)S_x S_y}, \tag{2}$$

em que S_x e S_y são os desvios padrões de X e Y respectivamente; $\overline{X}$ e $\overline{Y}$ são os valores das médias e n é o tamanho da amostra.

Em relação aos valores que r pode assumir, podemos afirmar que:

Se $r < 0$ indica correlação linear negativa;

Se $r > 0$ indicar correlação linear positiva;

Se $r = 0$ indica ausência de correlação linear.

UNIP PÓS-GRADUAÇÃO

Testando Correlações

Exemplo: Considere os seguintes conjuntos de dados:

Tabela: Tamanho amostral considerando diferentes tipos de erros.

X_1	2	4	5	6	8	9	10
Y_1	23	27	30	32	35	37	40
X_2	2	3	4	5	6	7	9
Y_2	12	9	8	5	3	2	2
X_3	2	3	4	5	6	7	8
Y_3	2	15	22	25	21	16	2

Determine o coeficiente de correlação linear de Pearson para os pares de amostras e obtenha os respectivos diagramas de dispersão.

UNIP PÓS-GRADUAÇÃO

Testando Correlações

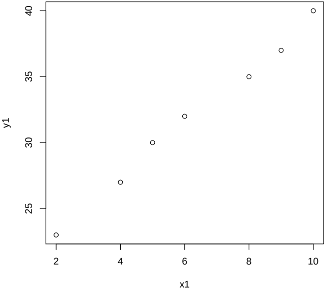
Os coeficientes de correlação linear são:

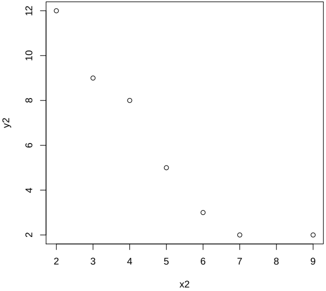
- $r_1 = 0,996$ indicando correlação linear positiva (direta);
- $r_2 = -0,939$ indicando correlação linear negativa (inversa);
- $r_3 = 0,008$ indicando ausência de correlação linear.

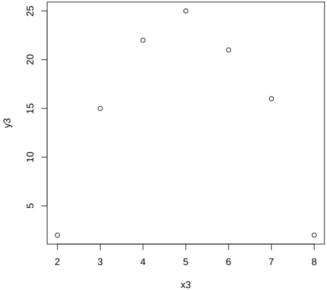
Os diagramas de dispersão são dados a seguir e como ressaltamos anteriormente, mesmo não havendo indicativos de correlação linear, não podemos afirmar que não há associação entre as variáveis X_3 e Y_3. A partir do resultado de $r_3 = 0,008$, podemos somente afirmar que há muito pouca evidência de correlação linear entre elas. E o diagrama de dispersão para os dados, indica que há associação quadrática entre X_3 e Y_3

UNIP PÓS-GRADUAÇÃO

Testando Correlações







Testando Correlações

Mesmo sabendo que quanto mais próximo dos limites do intervalo $[-1, 1]$ for o valor de r mais forte é a correlação, questionamentos em relação a significância da correlação podem surgir. E para valores não muito expressivos de r esses questionamentos só aumentam.

Para responder a estas questões com propriedade, o fazemos considerando o resultado do teste de hipóteses para o coeficiente de correlação. As hipóteses consideradas são:

$$\begin{cases} H_o : \rho = 0; \\ H_a : \rho \neq 0. \end{cases}$$

Considerando o nível de confiança $(1 - \alpha) \times 100\%$, a regra de decisão é: Se p-valor $< \alpha$ rejeitamos H_o.

UNIP
PÓS-GRADUAÇÃO

Testando Correlações

Um valor-p pequeno significa que a probabilidade de obter um valor da estatística de teste como o observado é muito improvável, levando assim à rejeição da hipótese nula.

Contato

pedrolramos@usp.br

UNIP
PÓS-GRADUAÇÃO