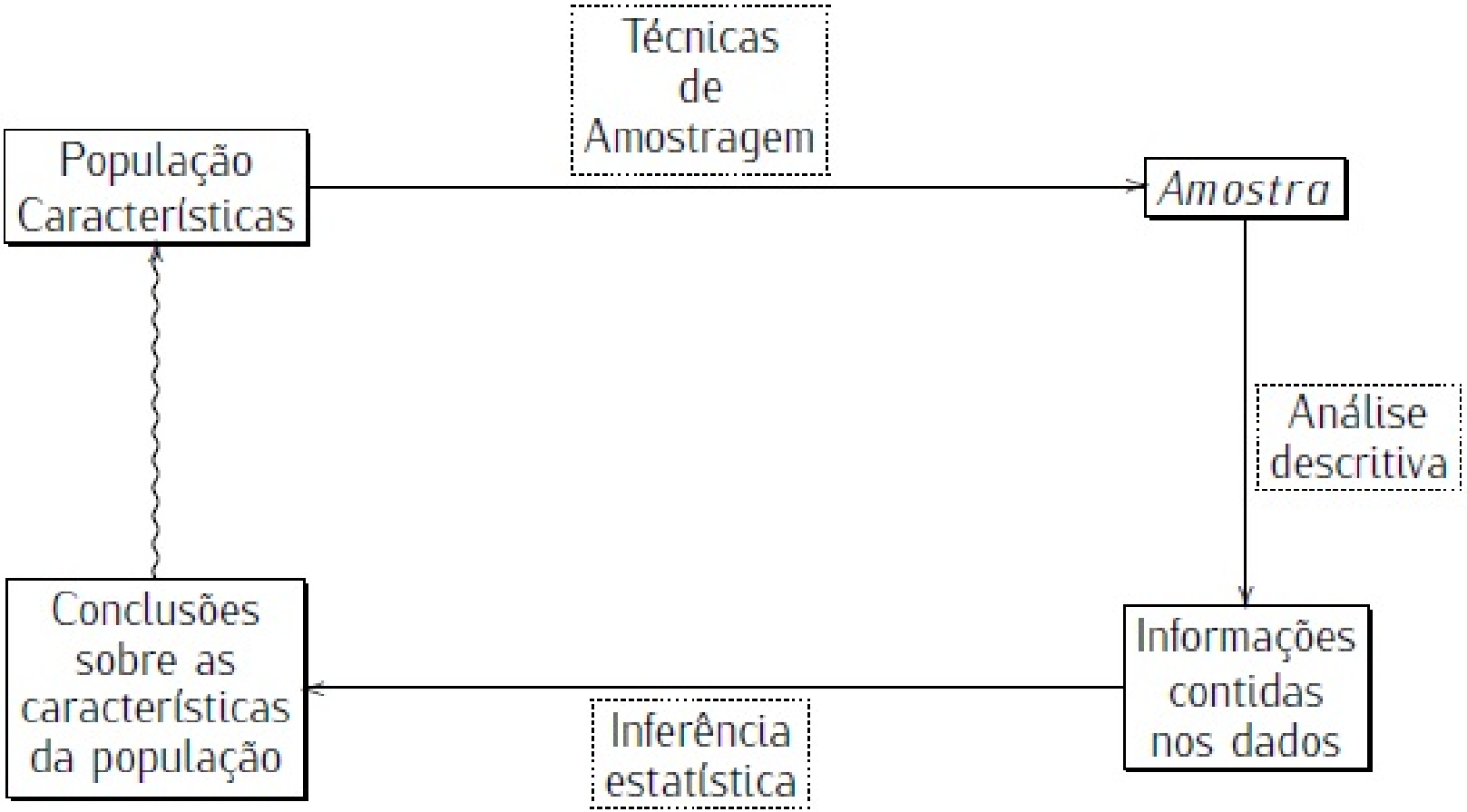
Monitoria de revisão: Uma introdução à análise exploratória de dados e métodos estatísticos

Universidade de São Paulo
Instituto de Matemática e Estatística

29 de Janeiro de 2020

Sumário

Visão geral da metodologia estatística:



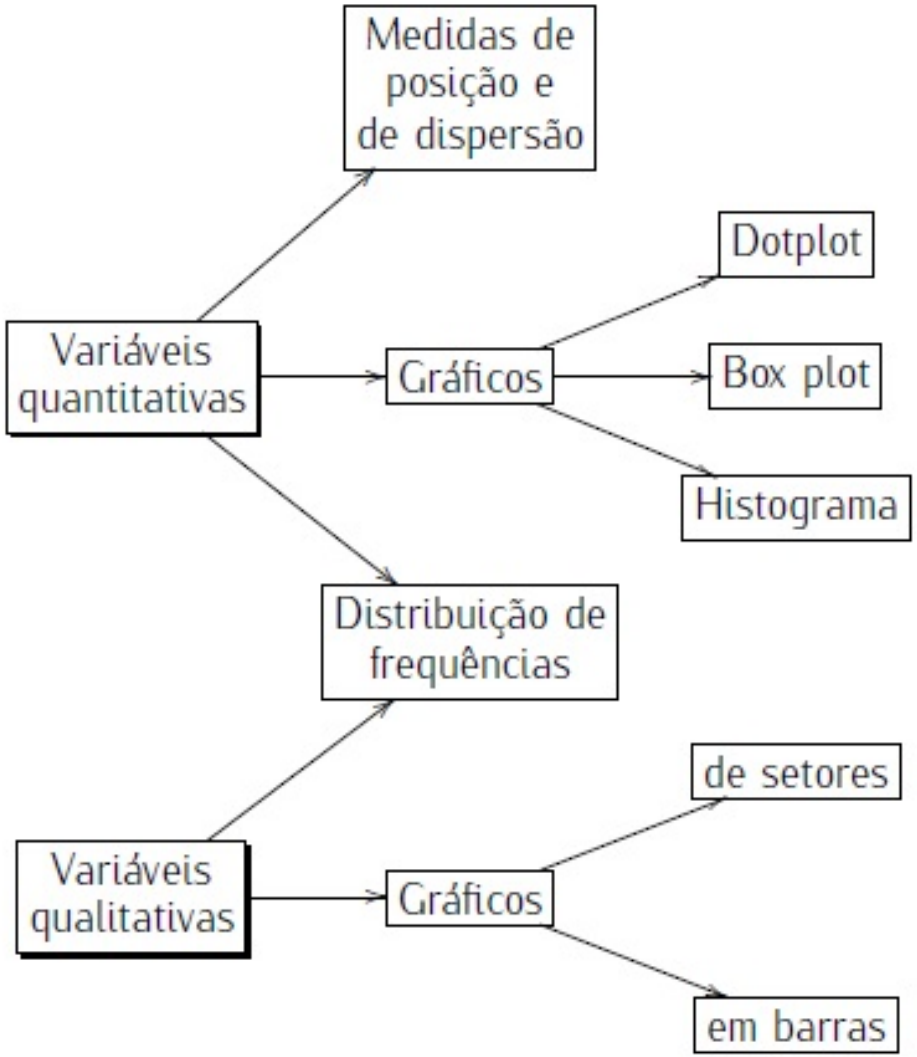


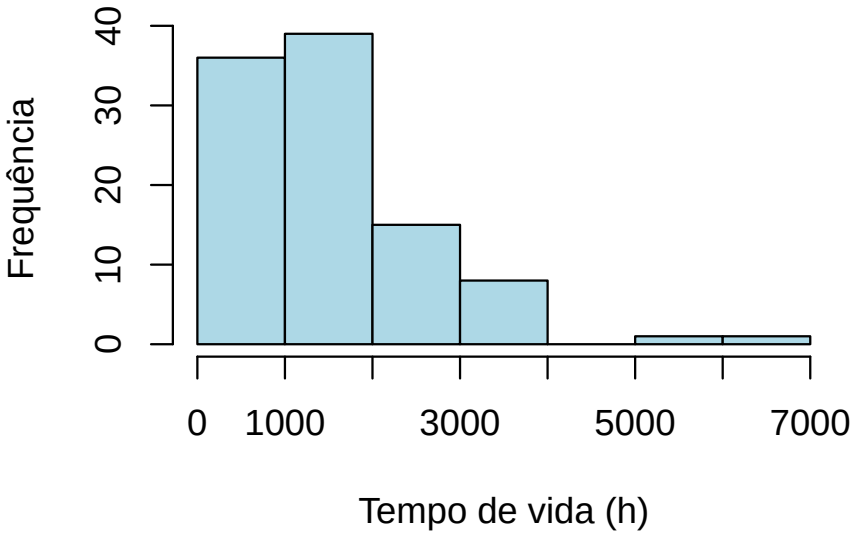
Figura: Análises descritivas unidimensionais de acordo com o tipo de variável.

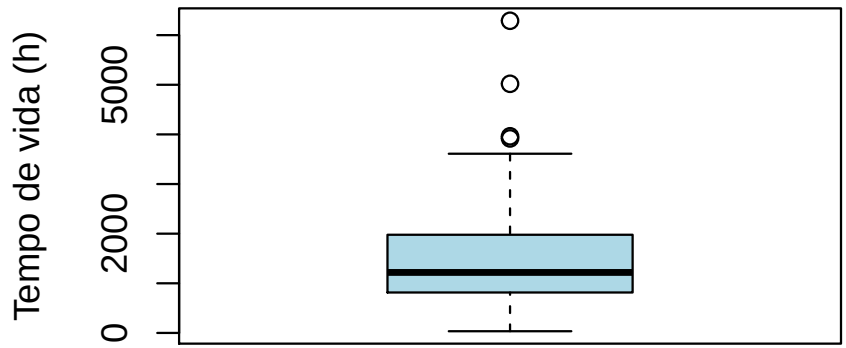
Exemplo

Uma empresa fabricante de lâmpadas incandescentes extraiu uma amostra de tamanho 100 da sua produção a fim de estudar o tempo de vida, mensurado em horas, das lâmpadas produzidas. A seguir são apresentados uma tabela contendo algumas medidas descritivas, e dois gráficos.

Exemplo

$x_{(1)}$	q_1	q_2	$\overline{x}$	q_3	$x_{(n)}$	sd
34,69	820,90	1219,23	1514,29	1964,38	6289,42	1097,018





Exemplo

1. Qual a variável de interesse (resposta)?
2. A distribuição dos dados é simétrica ou assimétrica? Justifique sua resposta com base nos dois gráficos.
3. Em qual intervalo encontra-se a moda da distribuição?
4. Existe algum valor atípico?
5. Mediana ou média: qual a melhor medida de posição para resumir esses dados?
6. A média é maior que a mediana, isso já é esperado? Justifique sua resposta com base no box plot.

Análise bidimensional

Em uma análise bidimensional, o interesse consiste em estudar a associação de duas variáveis. Quando consideramos duas variáveis (ou dois conjuntos de dados), podemos ter as seguintes situações:

(i) **As duas variáveis são quantitativas;**

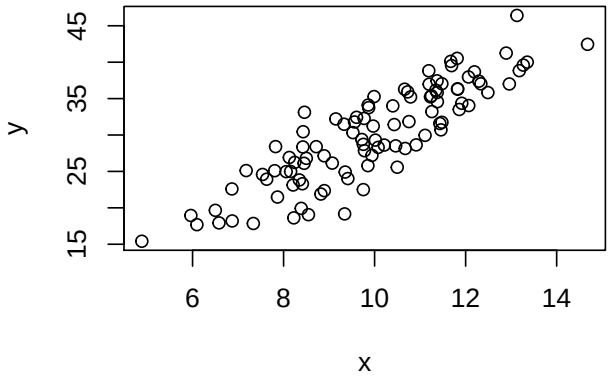
(ii) **As duas variáveis são qualitativas;**

(iii) **Uma variável é qualitativa e outra é quantitativa.**

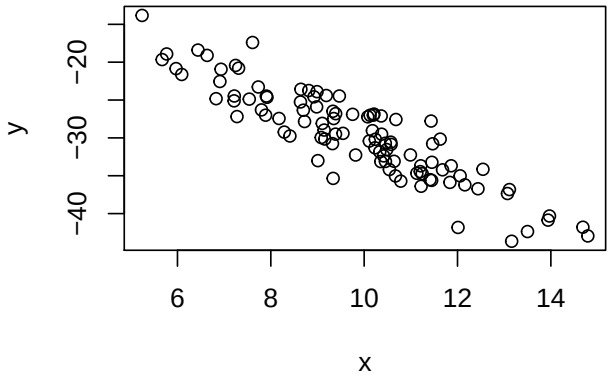
Duas variáveis quantitativas – Gráfico de dispersão

Dadas as observações $(x_1, y_1), \ldots, (x_n, y_n)$ das variáveis X e Y, através do gráfico de dispersão é possível ter uma ideia inicial de como as variáveis estão relacionadas.

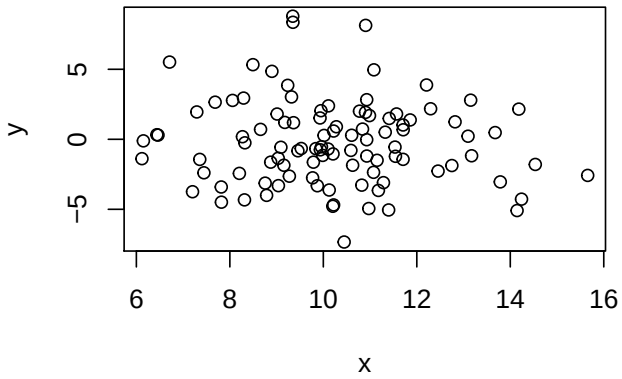
Podemos observar a direção da correlação (isto é, o que ocorre com os valores de Y quando os valores de X aumentam?), e também a "força" da correlação (o quão forte é essa tendência).



(c) Correlação linear positiva

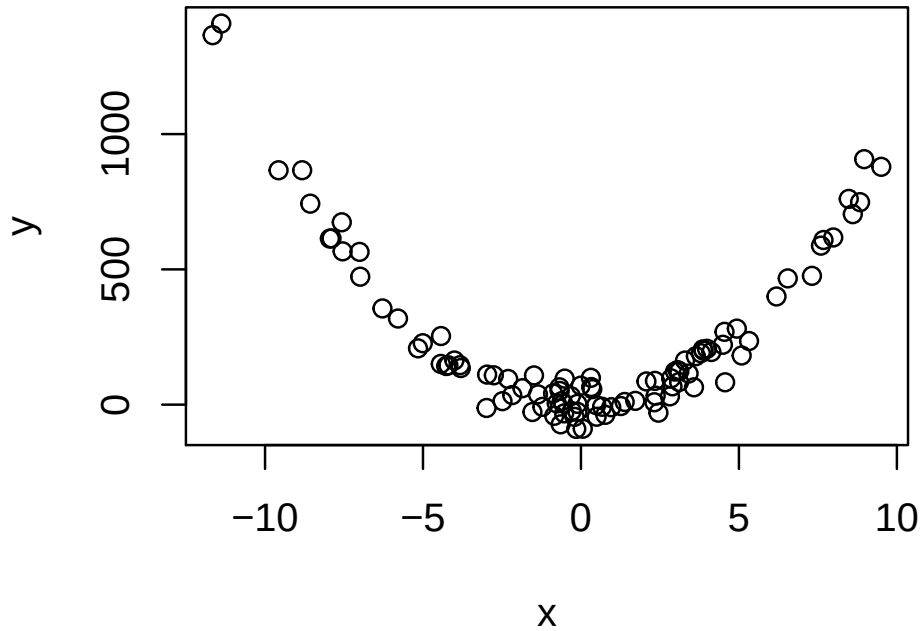


(d) Correlação linear negativa



(e) Ausência de correlação

Com base no gráfico de dispersão abaixo você diria que existe correlação entre as observações das variáveis X e Y?



Duas variáveis quantitativas – Coeficiente de correlação linear

Dados n pares de valores $(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)$, o coeficiente de correlação linear entre as observações das variáveis X e Y é definido por

$$r_{X,Y} = \frac{\sum_{i=1}^{n}(x_i - \overline{x})(y_i - \overline{y})}{(n-1)S_X S_Y},$$

ou de maneira alternativa,

$$r_{X,Y} = \frac{\sum_{i=1}^{n} x_i y_i - n\overline{xy}}{(n-1)S_X S_Y}.$$

Duas variáveis qualitativas – Tabela de contingência

Considere as variáveis X: "Região de procedência" com categorias *capital, interior e outra*; e Y: "Grau de instrução" com níveis *ensino fundamental, médio e superior*. A tabela de contingências entre as duas variáveis é da forma:

Tabela: Distribuição conjunta das variáveis grau de instrução e região de procedência.

	Região de procedência			
Grau de instrução	capital	interior	outra	Total
ensino fundamental	4	3	5	12
ensino médio	5	7	6	18
ensino superior	2	2	2	6
Total	11	12	13	36

Duas variáveis qualitativas – Tabela de contingência

Tabela de contingência com proporções (frequências relativas) em relação ao total geral:

Tabela: Distribuição conjunta das variáveis grau de instrução e região de procedência.

	Região de procedência			
Grau de instrução	capital	interior	outra	Total
ensino fundamental	11%	8%	14%	33%
ensino médio	14%	19%	17%	50%
ensino superior	6%	6%	6%	17%
Total	31%	33%	36%	100%

Duas variáveis qualitativas – Tabela de contingência

Também podemos construir tabelas de contingência em relação ao total de cada linha ou de cada coluna, sendo uma delas mais conveniente de acordo com o objetivo do problema em estudo.

Tabela: Distribuição conjunta das frequências relativas, em relação ao total de cada linha, das variáveis grau de instrução e região de procedência.

	Região de procedência			
Grau de instrução	capital	interior	outra	Total
ensino fundamental	33%	25%	42%	100%
ensino médio	28%	39%	33%	100%
ensino superior	33%	33%	33%	100%
Total	31%	33%	36%	100%

Duas variáveis qualitativas – Estatística χ^2

Uma forma de quantificar a relação entre as frequências observadas e esperadas de uma tabela de contingência é através da quantidade χ^2, definida por

$$\chi^2 = \sum_{i=1}^{k} \frac{(o_i - e_i)^2}{e_i},$$

em que k é o número de caselas na tabela de contingência, o_i e e_i são, respectivamente, as frequências absolutas observadas e esperadas da i-ésima casela da tabela de contingência.

Quanto maior o valor de χ^2, maior é a evidência de associação entre as variáveis.

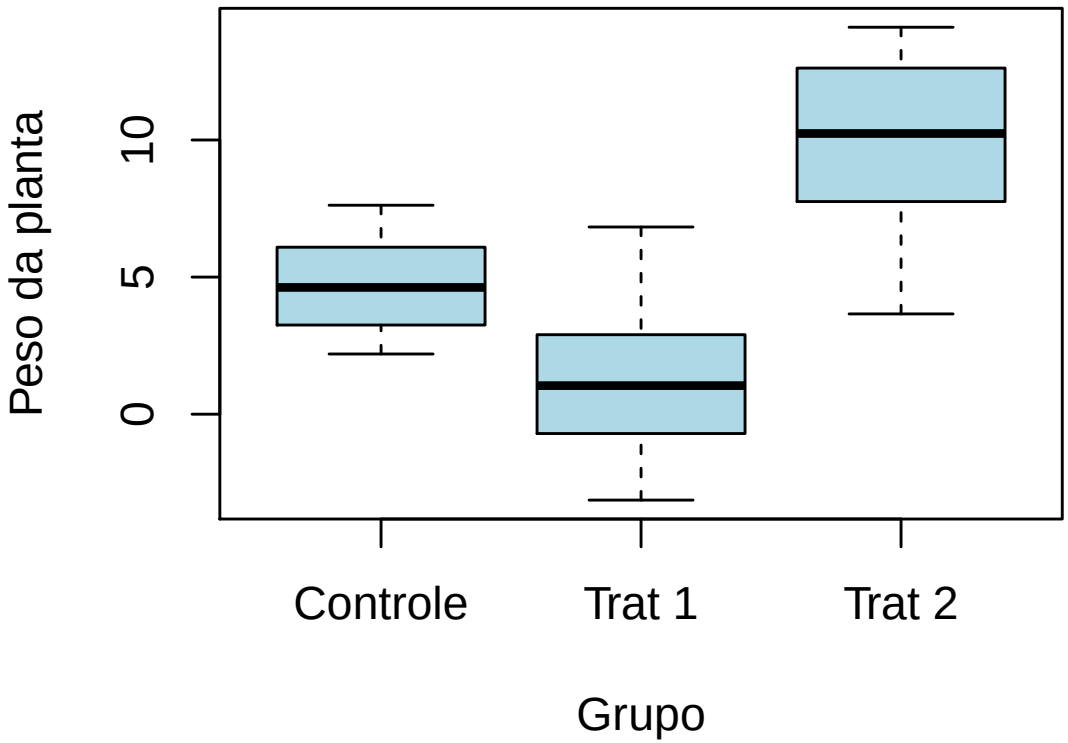
Uma variável é qualitativa e outra é quantitativa

- O boxplot é uma ferramenta importante neste tipo de análise;
- Geralmente, a variável quantitativa é a variável resposta observada, enquanto que a qualitativa é um fator com alguns níveis, e existe o interesse em analisar a distribuição da resposta em cada nível deste fator.

Uma variável é qualitativa e outra é quantitativa

Considere um estudo sobre o peso de uma determinada espécie de planta. O interesse neste estudo consiste em verificar se existem diferenças entre dois tratamentos que serão aplicados.

Para isso, foram selecionadas aleatoriamente 27 plantas que não irão receber nenhum tratamento, 39 para receberem o tratamento 1, e 34 para receberem o segundo tratamento. A figura a seguir apresenta o boxplot do peso das plantas em cada grupo considerado.


Peso da planta
0
5
10
Controle
Trat 1
Trat 2
Grupo

Conceitos importantes

Experimento aleatório: Experimentos aleatórios são fenômenos que, quando repetidos em processos semelhantes, possuem resultados imprevisíveis.

Espaço amostral (Ω): Conjunto de todos os possíveis resultados de um experimento aleatório. *Os elementos do espaço amostral podem ser chamados de pontos amostrais*.

Evento aleatório: Um único ponto amostral, ou uma reunião deles.

Exemplo:
Experimento aleatório: Testes de medição do tempo de vida de um sistema elétrico.
Espaço amostral: Conjunto dos reais positivos, isto é, $\Omega = (0, +\infty)$.
Evento aleatório: Durar mais do que 30h.

Conceitos importantes

Como calcular probabilidades quando os eventos são equiprováveis?

Considere um caso em que o espaço amostral é finito, isto é, podemos escrever $\Omega = \{\omega_1, \ldots, \omega_n\}$, e todos os pontos têm a mesma probabilidade de ocorrência, $1/n$. Se A for um evento contendo m pontos amostrais, então

$$\mathrm{P}(A) = \frac{m}{n}$$

Obs.: Note que não precisamos explicitar completamente Ω e A, mas sim calcularmos o valor m e n, chamados, respectivamente, número de casos favoráveis e número de casos possíveis.

Variáveis aleatórias

- Algumas vezes, é mais interessante associarmos um número a um evento aleatório e calcularmos a probabilidade da ocorrência desse número do que a probabilidade do evento.

- Uma variável aleatória é uma função de Ω em $\mathbb{R}$. Ou seja, uma variável aleatória é uma função que leva cada ponto do espaço amostral em um único ponto de $\mathbb{R}$.

- Uma variável aleatória discreta deve assumir valores em um conjunto finito ou enumerável, enquanto que uma variável aleatória contínua assume valores em um conjunto infinito.

Variáveis aleatórias discretas

Uma variável aleatória discreta pode ser caracterizada através da sua função (ou distribuição) de probabilidade.

Sejam $x_1, x_2, \ldots,$ os possíveis valores assumidos por uma variável aleatória discreta X. A função de probabilidade de X é a função que atribue a cada valor x_i sua probabilidade de ocorrência e pode ser representada por

x	x_1	x_2	$\ldots$
$\mathrm{P}(X = x)$	$\mathrm{P}(X = x_1)$	$\mathrm{P}(X = x_2)$	$\ldots$

Distribuição Bernoulli

Considere um experimento aleatório que possui apenas dois resultados possíveis. Geralmente, estes resultados são conhecidos como sucesso ou fracasso.

Por exemplo,

- Lançamento de uma moeda;
- Uma peça é classificada como boa ou defeituosa;
- Um paciente é submetido a um tratamento, e é verificado se o tratamento é eficaz ou não.

Distribuição Bernoulli

Para cada experimento, podemos definir uma variável aleatória X que assume apenas dois valores: 1, se ocorrer sucesso, e 0, se ocorrer fracasso. Indicaremos por p a probabilidade de sucesso, isto é, P(sucesso) = p, $0 < p < 1$.

A variável aleatória X, que assume apenas os valores 0 ou 1, com função de probabilidade

x	0	1
$\mathrm{P}(X = x)$	$1 - p$	p

possui distribuição Bernoulli, em que denotamos por

$$X \sim \text{Bernoulli}(p).$$

Distribuição Bernoulli

De forma mais resumida, podemos escrever a função de probabilidade da variável aleatória $X \sim \text{Bernoulli}(p)$ como

$$\text{P}(X = x) = p^x(1-p)^{1-x}, \quad \text{para } x = 0, 1.$$

Temos ainda que

$$\text{E}(X) = p \quad \text{e} \quad \text{Var}(X) = p(1-p).$$

Distribuição binomial

Considere agora n repetições independentes de um mesmo experimento aleatório, de modo que cada tentativa admita apenas dois resultados: sucesso ou fracasso; 0 ou 1, e as probabilidades de sucesso e fracasso são as mesmas para cada tentativa.

Dizemos que a variável aleatória X possui distribuição binomial de parâmetros n e p (n é o número de repetições, e p é a probabilidade de sucesso) se sua função de probabilidade é da forma

$$\mathrm{P}(X = k) = \binom{n}{k} p^k (1-p)^{n-k},$$

para $k = 0, 1, 2, \ldots, n$. Neste caso, denotamos $X \sim \mathrm{B}(n, p)$ e temos que

$$\mathrm{E}(X) = np \quad \text{e} \quad \mathrm{Var}(X) = np(1-p).$$

Exemplo

Registros hospitalares mostram que a taxa de mortalidade de uma doença em pacientes de determinada região é de 25%. A fim de estudar esta doença, uma amostra aleatória de 6 pacientes com a enfermidade foi considerada.

1. Qual variável aleatória podemos associar com o problema, e qual a sua função de probabilidade?
2. Qual a probabilidade de que 4 pacientes entre os selecionados consigam se recuperar?
3. Qual a probabilidade de pelo menos um dos pacientes não conseguir se recuperar?

Exemplo

- Item 1:

  Defina: X:"Número de pacientes que conseguem se recuperar da doença entre os 6 pacientes selecionados"

$$X \sim \mathrm{B}(6, 0,75),$$

  com função de probabilidade:

$$\mathrm{P}(X = k) = \begin{pmatrix} 6 \\ k \end{pmatrix} 0,75^k 0,25^{6-k}$$

Exemplo

- Item 2:

$$\begin{aligned} \mathrm{P}(X=4) &= \binom{6}{4} 0,75^4 0,25^{6-4} \\ &\approx 0,2966 = 29,66\%. \end{aligned}$$

- Item 3:

$$\begin{aligned} \mathrm{P}(X \leq 5) &= \mathrm{P}(X=0) + \mathrm{P}(X=1) + \cdots + \mathrm{P}(X=5) \\ &= 1 - \mathrm{P}(X>5) \\ &= 1 - \mathrm{P}(X=6) \\ &= 1 - \binom{6}{6} 0,75^6 0,25^{6-6} \\ &\approx 0,8220 = 82,2\%. \end{aligned}$$

Variáveis aleatórias contínuas

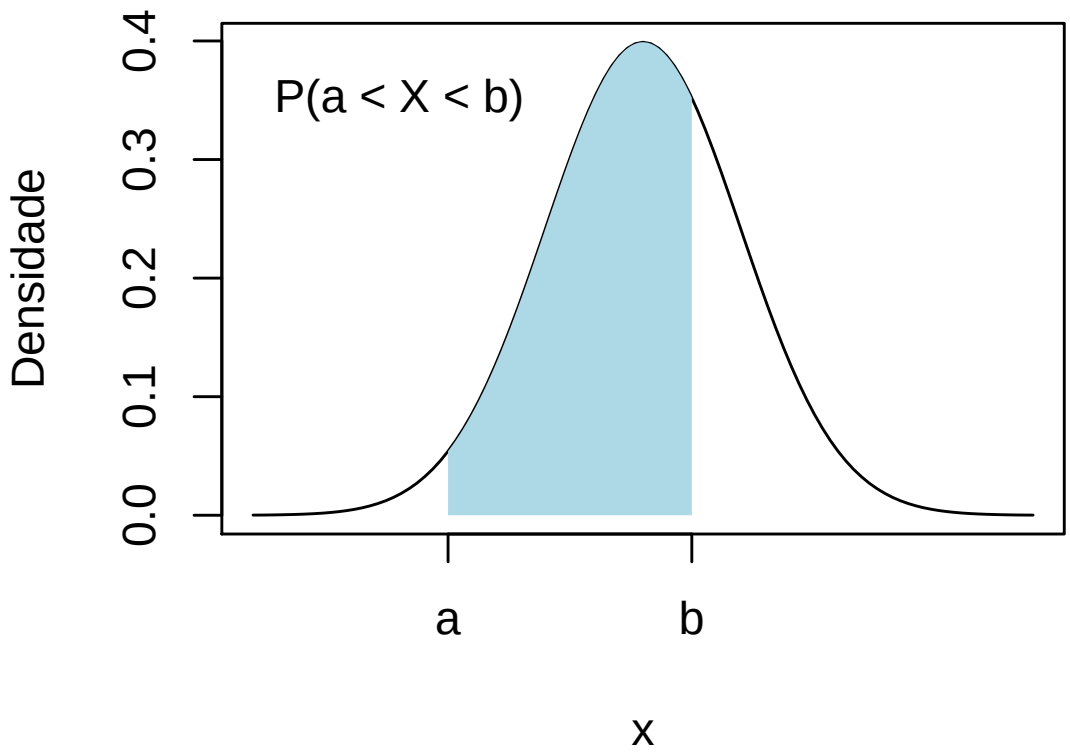
Uma variável aleatória é contínua quando o conjunto de valores possíveis que ela assume for não–enumerável, isto é, ela assume valores num intervalo de números reais.

Neste caso, associamos probabilidades a intervalos de valores da variável.

Variáveis aleatórias contínuas

Uma variável aleatória X contínua é caracterizada por sua função densidade de probabilidade, denotada por $f(x)$ e que possui as seguintes propriedades:

- A área sob a curva da densidade é 1;
- $\mathrm{P}(a \leq X \leq b) =$ área sob a curva da densidade $f(x)$ e acima do eixo x, entre os pontos a e b;
- $f(x) \geq 0$, para todo x;
- $\mathrm{P}(X = x_0) = 0$, para x_0 fixo.


P(a < X < b)
Densidade
0.0
0.1
0.2
0.3
0.4
a
b
x

Distribuição normal

Dizemos que uma variável aleatória X possui distribuição normal com parâmetros μ e σ^2 se sua função densidade de probabilidade é

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

e denotamos por $X \sim \mathrm{N}(\mu, \sigma^2)$.

Temos que:

- A distribuição normal é simétrica;
- μ é o valor esperado (média) de X, com $-\infty < x < +\infty$;
- σ^2 é a variância de X, com $\sigma^2 > 0$.

Cálculo de probabilidades na distribuição normal

- Suponha que estamos interessados em calcular $\mathrm{P}(a < X < b) = \mathrm{P}(a \leq X \leq b)$, em que $X \sim \mathrm{N}(\mu, \sigma^2)$.
- Como vimos, esta probabilidade é dada pela área sob a curva $f(x)$ entre os pontos a e b acima do eixo x.
- Encontrar esta área diretamente não é fácil.

Cálculo de probabilidades na distribuição normal

- Para calculá-la, faremos o uso do seguinte resultado:

$$\text{Se } X \sim \mathrm{N}(\mu, \sigma^2), \text{ então, } Z = \frac{X - \mu}{\sigma} \sim \mathrm{N}(0, 1).$$

- Utilizamos esta relação em razão da existência de tabelas da distribuição normal padrão que fornecem probabilidades acumuladas.

- Assim,

$$\begin{aligned} \mathrm{P}(a < X < b) &= \mathrm{P}\left(\frac{a - \mu}{\sigma} < \frac{X - \mu}{\sigma} < \frac{b - \mu}{\sigma}\right) \\ &= \mathrm{P}\left(\frac{a - \mu}{\sigma} < Z < \frac{b - \mu}{\sigma}\right). \end{aligned}$$

Cálculo de probabilidades na distribuição normal

Exemplo: Verifique na tabela da distribuição normal padrão que

(a) $\mathrm{P}(Z \leq 0,63) = 0,7357$;

(b) $\mathrm{P}(Z \leq -1,25) = 0,1056$;

(c) $\mathrm{P}(-1,25 < Z < 0) = 0,3944$;

(d) $\mathrm{P}(Z \geq 1,5) = 0,0668$.

Cálculo de probabilidades na distribuição normal

Exemplo: Verifique na tabela da normal padrão que

(a) $\mathrm{P}(Z \leq z) = 0,975 \rightarrow z = 1,96$;

(b) $\mathrm{P}(Z \leq z) = 0,3 \rightarrow z \approx -0,52$;

(c) $\mathrm{P}(-z \leq Z \leq z) = 0,80 \rightarrow z \approx 1,28$.

Exemplo

O tempo gasto no exame de vestibular de uma universidade tem distribuição normal, com média 120 min e desvio padrão 15 min.

(a) Defina a variável aleatória.

(b) Sorteando-se um aluno ao acaso, qual é a probabilidade dele terminar o exame antes de 100 minutos?

(c) Sorteando-se um aluno ao acaso, qual é a probabilidade dele não terminar o exame em 2h?

Exemplo

(a) Defina X :"Tempo gasto, em minutos, no exame do vestibular na universidade estudada". De acordo com o enunciado,

$$X \sim \mathrm{N}(120, 15^2)$$

(b) A probabilidade desejada é

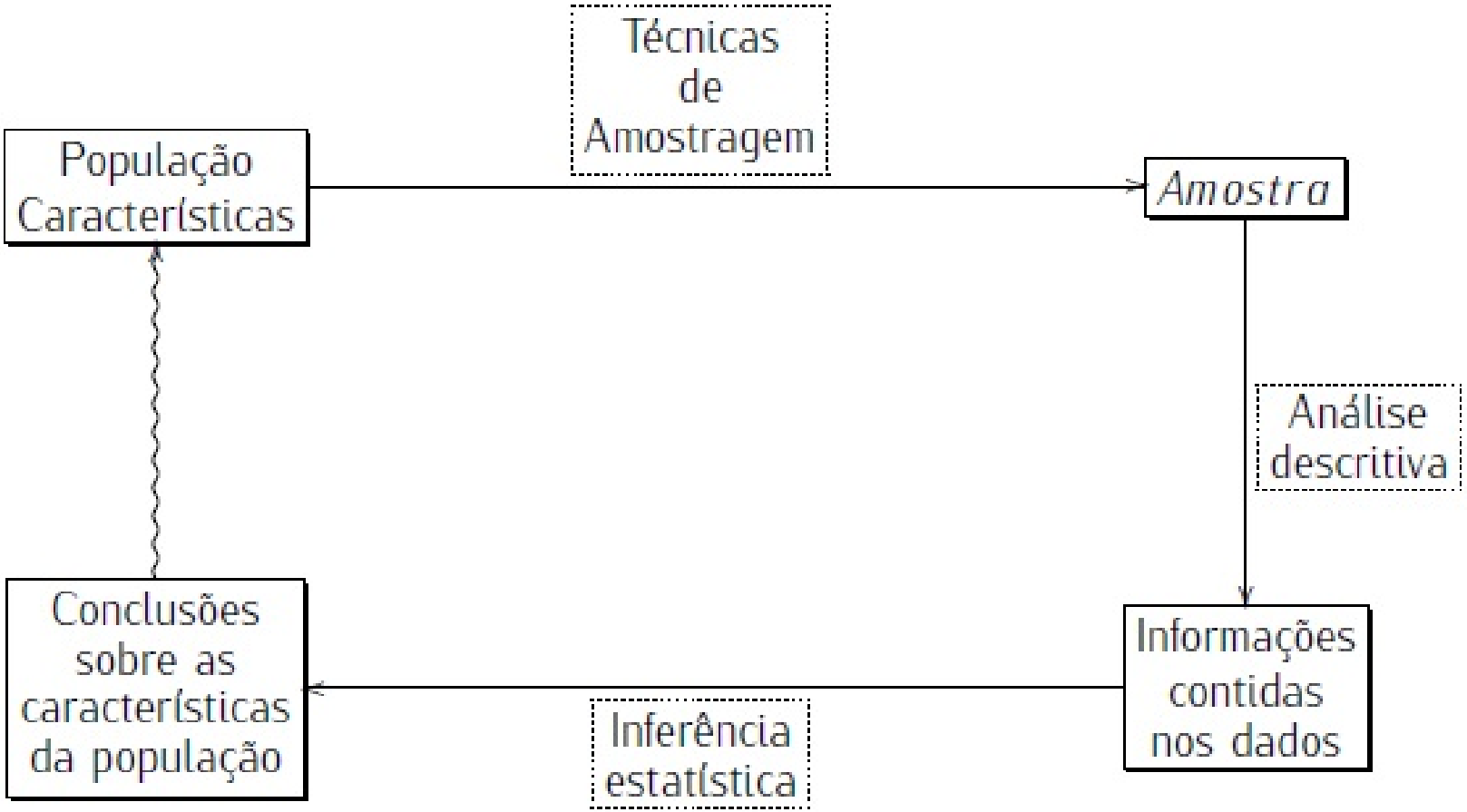
$$\begin{aligned} \mathrm{P}(X < 100) &= \mathrm{P}\left(\frac{X - 120}{15} < \frac{100 - 120}{15}\right) \\ &= \mathrm{P}(Z < -1,33) = 0,0918 \end{aligned}$$

(c) Estamos interessados na seguinte probabilidade

$$\mathrm{P}(X > 120) = 0,5$$

1. Análise exploratória de dados
   - Análise bidimensional
2. Probabilidade e variáveis aleatórias
   - Conceitos importantes
   - Variáveis aleatórias
3. Inferência estatística
   - Conceitos importantes
   - Estimando μ pontualmente
   - Intervalo de confiança para μ
   - Tamanho da amostra
   - Estimando p pontualmente
   - Intervalo de confiança para p
   - Tamanho da amostra

Visão geral da metodologia estatística:



Conceitos importantes

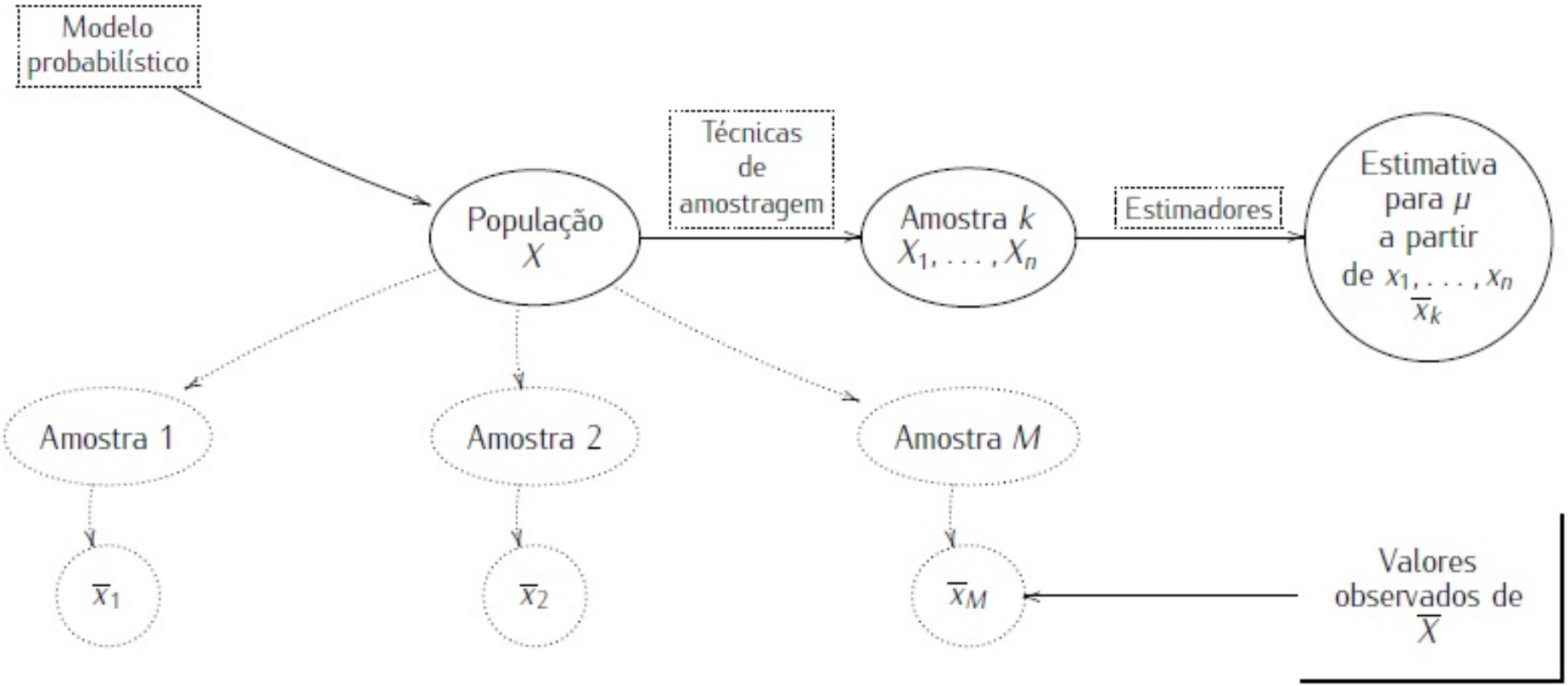
População: Conjunto de todos os elementos ou resultados sob investigação;

Amostra: Qualquer subconjunto da população;

Parâmetro: Quantidade desconhecida que representa uma característica da população, e principal objeto de interesse da inferência estatística;

Estimador: Função dos elementos da amostra, construída com a finalidade de representar, ou estimar, um parâmetro de interesse. O estimador é denotado com letra maiúscula;

Estimativa: É o valor numérico assumido pelo estimador para a amostra selecionada. A estimativa é denotada com letra minúscula.


Modelo probabilístico
População X
Técnicas de amostragem
Amostra k $X_1, \ldots, X_n$
Estimadores
Estimativa para μ a partir de $x_1, \ldots, x_n$ $\overline{x}_k$
Amostra 1
Amostra 2
Amostra M
$\overline{x}_1$
$\overline{x}_2$
$\overline{x}_M$
Valores observados de $\overline{X}$

Estimando μ pontualmente

Um estimador pontual para μ, baseado numa amostra aleatória de tamanho n, é dado pela média amostral,

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

Se observarmos os valores $x_1, \ldots, x_n$ para as variáveis $X_1, \ldots, X_n$ obtemos a estimativa pontual de μ:

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

Intervalo de confiança para μ – Variância conhecida

Sejam $X_1, X_2, \ldots, X_n$ uma amostra aleatória de $X \sim \mathrm{N}(\mu, \sigma^2)$, em que σ é conhecido. Um intervalo de confiança para a média μ, com coeficiente de confiança γ, é dado por

$$\left[\overline{X} - z\frac{\sigma}{\sqrt{n}};\quad \overline{X} + z\frac{\sigma}{\sqrt{n}}\right],$$

em que z é tal que $\mathrm{P}(-z \leq Z \leq z) = \gamma$ com $Z \sim \mathrm{N}(0,1)$.

Intervalo de confiança para μ – Variância conhecida

Podemos resumir o procedimento operacional para a obtenção de uma estimativa por intervalo de μ da seguinte forma

1. Retiramos uma amostra aleatória simples de tamanho n;
2. Calculamos a média amostral $\bar{x}$;
3. Fixamos o coeficiente de confiança γ, e com ele determinamos o valor de z tal que $\mathrm{P}(-\mathrm{z} \leq Z \leq \mathrm{z}) = \gamma$ com $Z \sim \mathrm{N}(0,1)$;
4. Calculamos o erro amostral $\varepsilon = \mathrm{z}\dfrac{\sigma}{\sqrt{n}}$;
5. Escrevemos o intervalo de confiança da forma

$$\left[\bar{x} - \mathrm{z}\frac{\sigma}{\sqrt{n}}; \quad \bar{x} + \mathrm{z}\frac{\sigma}{\sqrt{n}}\right],$$

Exemplo

Deseja-se estimar o tempo médio de estudo (em anos) da população adulta de um município. Sabe-se que o tempo de estudo tem distribuição normal com desvio padrão $\sigma = 2{,}6$ anos. Foram entrevistados $n = 25$ indivíduos, obtendo-se, para essa amostra, um tempo médio de estudo igual a 10,5 anos. Obtenha um intervalo de 90% de confiança para o tempo médio de estudo na população.

Exemplo

Defina

X : "Tempo de estudo da população adulta do município estudado".

De acordo com o problema, temos a informação de que

$$X \sim \mathrm{N}(\mu; 2,6^2).$$

A estimativa pontual de μ foi de $\overline{x} = 10,5$.

Uma vez que $\gamma = 0,9$, obtemos que $\mathrm{z} = 1,65$. Portanto, a estimativa por intervalo com confiança de 90% para μ é dada por

$$\left[\overline{x} - \mathrm{z}\frac{\sigma}{\sqrt{n}}; \quad \overline{x} + \mathrm{z}\frac{\sigma}{\sqrt{n}}\right] = \left[10,5 - 1,65\frac{2,6}{\sqrt{25}}; \quad 10,5 + 1,65\frac{2,6}{\sqrt{25}}\right]$$
$$= [9,64; 11,36].$$

Tamanho da amostra – Variância conhecida

A partir da relação $\varepsilon = z\frac{\sigma}{\sqrt{n}}$, o tamanho da amostra n pode ser determinado por

$$n = \left(\frac{z}{\varepsilon}\right)^2 \sigma^2,$$

conhecendo-se o desvio padrão σ de X, com erro da estimativa ε e com coeficiente de confiança γ tal que $P(-z \leq Z \leq z) = \gamma$ com $Z \sim N(0,1)$.

Exemplo

A renda per-capita domiciliar numa certa região tem distribuição normal com desvio padrão $\sigma = 250$ reais e média μ desconhecida. Se desejamos estimar a renda média μ com erro $\varepsilon = 50$ reais e com uma confiança de $\gamma = 95\%$, quantos domicílios devemos consultar?

Exemplo

Defina

X :"Renda per-capta domiciliar na região de interesse".

De acordo com o enunciado, $X \sim \mathrm{N}(\mu, 250^2)$. Temos que $\varepsilon = 50$ e $\sigma = 250$. Como $\gamma = 0,95$, segue que $\mathrm{z} = 1,96$. Portanto,

$$n = \left(\frac{\mathrm{z}}{\varepsilon}\right)^2 \sigma^2 = \left(\frac{1,96}{50}\right)^2 250^2 = 96,04.$$

Assim, são necessários aproximadamente 97 domicílios.

Intervalo de confiança para μ – Variância desconhecida

Sejam $X_1, X_2, \ldots, X_n$ uma amostra aleatória de $X \sim \mathrm{N}(\mu, \sigma^2)$. Um intervalo de confiança para a média μ, com coeficiente de confiança γ, é dado por

$$\left[\overline{X} - \mathrm{t}_{n-1}^{c}\frac{S}{\sqrt{n}};\quad \overline{X} + \mathrm{t}_{n-1}^{c}\frac{S}{\sqrt{n}}\right],$$

em que S é o estimador do desvio padrão e t_{n-1}^{c} é o ponto crítico da distribuição t-Student com $n-1$ graus de liberdade tal que $\mathrm{P}(-\mathrm{t}_{n-1}^{c} \leq \mathrm{T}_{n-1} \leq \mathrm{t}_{n-1}^{c}) = \gamma$, com $\mathrm{T}_{n-1} \sim \mathrm{t}_{n-1}$.

Estimando p pontualmente

Considere n elementos observados de forma independente e extraídos ao acaso de uma população. Para cada elemento da população, verificamos a presença (sucesso) ou não (fracasso) da característica de interesse.

Note que, neste caso, temos uma amostra aleatória de tamanho n de X, sendo X uma variável aleatória com distribuição Bernoulli, que representamos por

$$X_1, \ldots, X_n,$$

em que X_i é igual a 1 se ocorre sucesso, ou 0 se ocorre fracasso para o i-ésimo elemento da amostra.

Estimando p pontualmente

Um estimador pontual para p, também chamado de proporção amostral, baseado numa amostra aleatória de tamanho n, é dado por

$$\widehat{p} = \frac{X_1 + X_2 + \cdots + X_n}{n} = \frac{1}{n}\sum_{i=1}^{n} X_i.$$

Se observarmos os valores $x_1, x_2, \ldots, x_n$ tal que $x_1 + x_2 + \cdots + x_n = k$, obtemos que $\widehat{p} = k/n$ é uma estimativa pontual para p.

Intervalo de confiança para p

Sejam $X_1, X_2, \ldots, X_n$ uma amostra aleatória de $X \sim \text{Bernoulli}(p)$. Um intervalo de confiança para a proporção p, com coeficiente de confiança γ, é dado por

$$\left[\widehat{p} - z\sqrt{\frac{\widehat{p}(1-\widehat{p})}{n}};\ \ \widehat{p} + z\sqrt{\frac{\widehat{p}(1-\widehat{p})}{n}}\right],$$

em que z é tal que $\mathrm{P}(-z \leq Z \leq z) = \gamma$ com $Z \sim \mathrm{N}(0,1)$.

Tamanho da amostra

De forma análoga ao dimensionamento da amostra no problema de estimar μ, o tamanho da amostra n pode ser determinado por

$$n = \left(\frac{z}{\varepsilon}\right)^2 p(1-p),$$

em que z é tal que $P(-z \leq Z \leq z) = \gamma$ com $Z \sim N(0,1)$.

Note que esta expressão depende de p que é o parâmetro que desejamos estimar. Na prática, substituímos o valor de $p(1-p)$ por o seu valor máximo, obtendo

$$n = \left(\frac{z}{\varepsilon}\right)^2 0,25.$$