Métodos Numéricos e Estatísticos

(Componente de Métodos Estatísticos)

DMat-UA

Este documento, de modo algum, substitui a leitura cuidada da bibliografia recomendada.

(Revisão de Setembro 2020)

Capítulos

Conceitos Básicos

1. Análise Exploratória de Dados

2. Teoria de Probabilidades

3. Inferência Estatística

O que é a estatística?

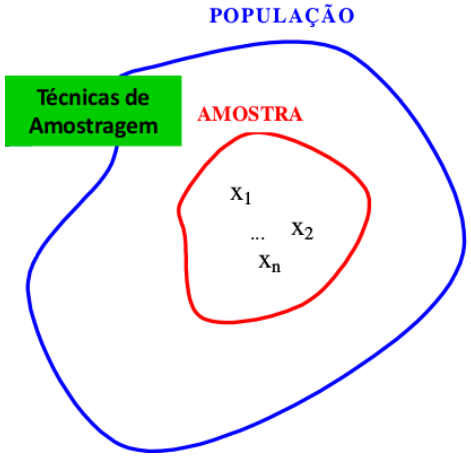
- A Estatística é a ciência que se ocupa da obtenção de informação, seu tratamento inicial, com a finalidade de, através de modelos probabilísticos adequados, inferir de uma amostra para a população e, eventualmente, prever a evolução futura de um fenómeno.

De modo esquemático, vejamos os vários passos envolvidos numa Análise Estatística de uma característica aleatória e as várias áreas do conhecimento envolvidas nessa análise.

Censo e Estudo por Amostragem

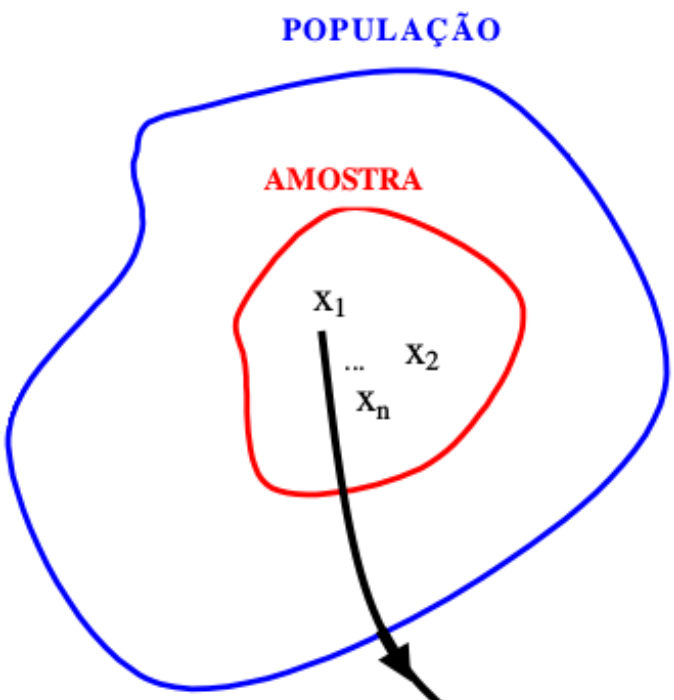
Em termos genéricos, na análise de uma característica "mensurável", duas perspectivas distintas existem:
1. Recolhem-se e analisam-se todos os elementos que constituem a **população** (conjunto constituído por todos os elementos que são objecto de estudo). Neste caso, realiza-se um **Censo**.

2. Recolhe-se e analisa-se uma **amostra** (subconjunto da população).
Neste caso, realiza-se um **Estudo por Amostragem**.



Pretende-se que a amostra seja representativa da população. O processo de recolha e selecção dos elementos que constituem a amostra é objecto de estudo das TÉCNICAS DE AMOSTRAGEM.

Análise Exploratória de Dados



Recolhida a amostra, realiza-se uma descrição numérica e gráfica dos dados amostrais através de **medidas estatísticas** e de **representações gráficas de funções estatísticas** com vista à elaboração de uma

ANÁLISE EXPLORATÓRIA DE DADOS

Nesta fase pretende-se que os dados "falem por si"; espera-se que a informação obtida nesta análise evidencie as características relevantes da população.

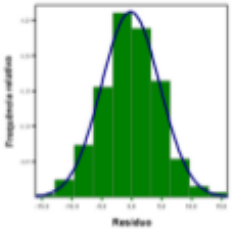
Análise Exploratória de Dados

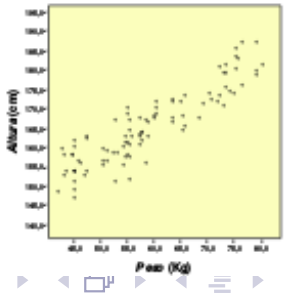
$$\overline{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$$

(média amostral)

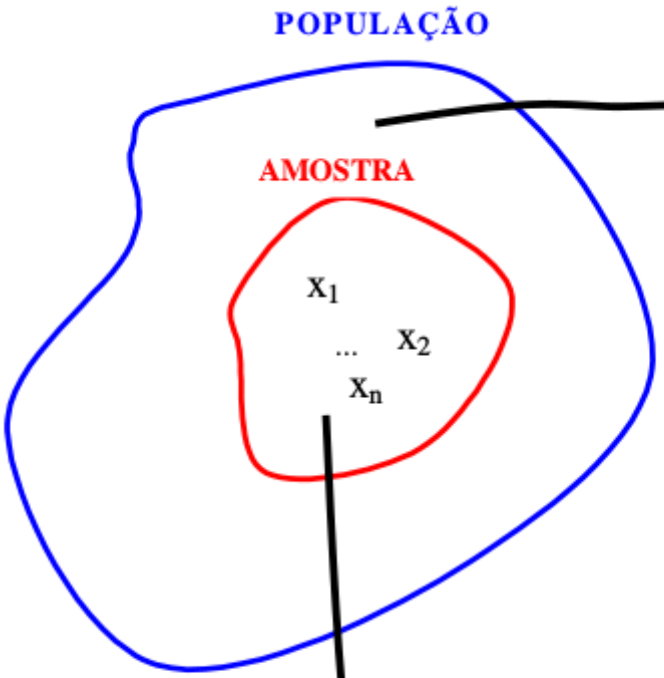
$$s^2$$

(variância amostral)





Inferência Estatística



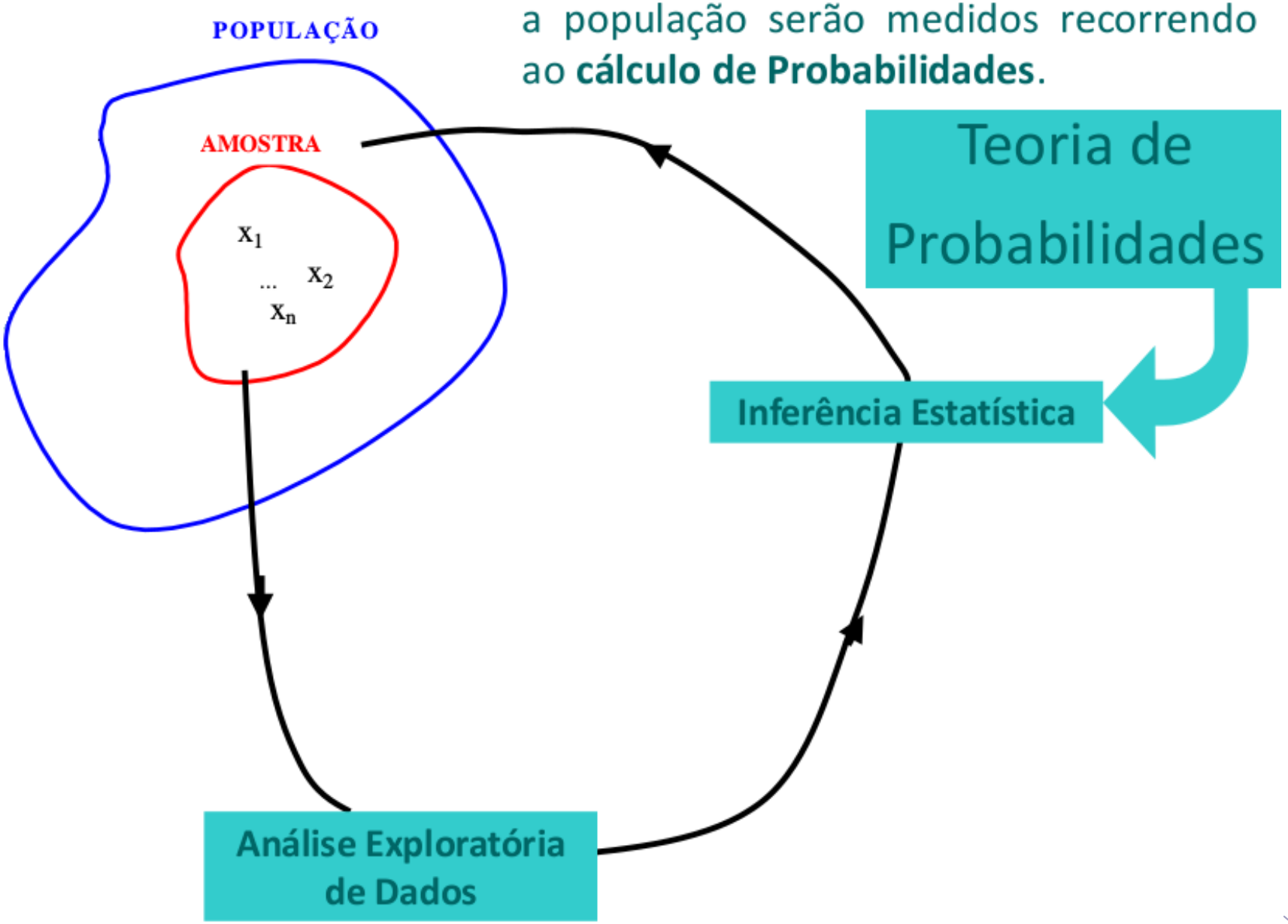
Inferência Estatística

Análise Exploratória de Dados

Se a amostra for representativa da população, os resultados obtidos da "leitura" dos dados amostrais poderão ser estendidos/inferidos para a população, utilizando métodos de **Inferência estatística**.

Teoria das Probabilidades

Os erros que se cometem ao inferir sobre a população serão medidos recorrendo ao **cálculo de Probabilidades**.



Estatística descritiva vs inferencial

- **Estatística Descritiva**: conjunto de métodos estatísticos que visam sumariar e descrever os atributos mais proeminentes dos dados.
- **Estatística Inferencial**: conjunto de métodos estatísticos que visam caracterizar uma população a partir de uma amostra.

População e Amostra

- **População** – conjunto de elementos (indivíduos, empresas, objetos, etc.) que apresentam uma ou mais características em comum e que descrevem um fenómeno que interessa estudar.
  Ex.: os alunos de 1ª matricula na UA, em um determinado ano letivo.

- **Amostra** – subconjunto finito de elementos extraídos da população, segundo uma técnica de amostragem adequada.
  Ex.: conjunto de 20 alunos seleccionados ao acaso do conjunto de todos os alunos de 1ª matricula na UA, em um determinado ano letivo.

Variável

- **Variável** – característica mensurável ou atributo de interesse para estudo na população.
  Ex.: nota de ingresso.

Obs.: Uma amostra pode conter mais de uma característica de interesse para cada uma das unidades observadas. Com duas características temos uma amostra bivariada; com mais do que duas temos uma amostra multivariada.
Ex: na população de alunos de 1ª matrícula da UA, em um determinado ano letivo, pode interessar a análise de várias características em simultâneo como, por exemplo, nota de ingresso e região de proveniência.

Natureza da variável

- **Quantitativa (ou numérica)** – assume valores intrinsecamente numéricos.
  Estas podem ser:
  - **Discretas** – assumem um conjunto finito ou infinito numerável (que se consegue contar) de possibilidades.
    Ex.: número de peças defeituosas por lote.
  - **Contínuas** – assumem valores em intervalos reais.
    Ex.: comprimentos, pesos, tempos.
- **Qualitativa (ou categórica)** – assume valores intrinsecamente não numéricos.
  Ex.: região de proveniência.

Natureza discreta ou contínua?

Natureza discreta ou contínua? A fronteira pode ser pouco clara. Por vezes,

- os dados de uma variável contínua são registados com precisão finita sendo, na prática, tomados como de natureza discreta.
  Ex.: a idade de uma pessoa, em anos.
- os valores possíveis que uma variável discreta pode assumir representam uma gama de valores tão dispersos que, na prática, os dados observados para a variável são tratados como se fossem de natureza contínua, agrupando-se os valores em intervalos (classes).
  Ex.: nº de habitações familiares por freguesia.

Escalas de medidas de uma variável

1. **Escala de razões** – os valores numéricos possuem ordem e diferenças têm significado.
   O valor zero representa ausência da característica e múltiplos de valores possuem significado. São exemplos:
   - medidas de comprimento, pesos, intervalos de tempo;
   - o nº de filhos de um casal.
2. **Escala de intervalos** – os valores numéricos possuem ordem e diferenças têm significado. São exemplos:
   - temperatura medida em graus C ou F;
   - medidas de altitude.

Escalas de medidas de uma variável

3. **Escala ordinal** – os valores (numéricos ou não) possuem uma ordem intrínseca. São exemplos:
   - classificações obtidas no 2º e 3º ciclos do ensino básico (1 a 5);
   - grupos etários (crianças, jovens, adultos e idosos).
4. **Escala nominal** – os valores (numéricos ou não) não possuem uma ordem intrínseca. São exemplos:
   - tipo de sangue de uma pessoa (O, A, B e AB);
   - sexo (0 - Feminino, 1 - Masculino).

Capítulos

Conceitos Básicos

1. Análise Exploratória de Dados
   - Organização dos Dados
   - Descrição Numérica
   - Descrição Gráfica
   - Análise e Interpretação

2. Teoria de Probabilidades

3. Inferência Estatística

Análise Exploratória de Dados

- A Análise Exploratória de Dados é também designada Análise Preliminar de Dados ou Estatística Descritiva, embora alguns autores estabeleçam diferenças subtis entre estas designações.
- Objectivo do Capítulo: pretende-se que o estudante seja capaz de elaborar e interpretar uma análise estatística exploratória simples de dados univariados.
- Apontamentos-base:
  Murteira et al. (2015): Capítulo 1, Sec. 1.1 - 1.7;
  Hall et al. (2011): Capítulo 1.

Fases de uma Análise Exploratória de Dados

Tópicos em estudo:

- **Organização dos dados** através da construção de tabelas de frequências.
- **Descrição numérica** dos dados através do cálculo de medidas amostrais.
- **Descrição gráfica dos dados** de vários tipos.
- **Análise e interpretação** das medidas numéricas e dos gráficos obtidos.

Tabelas de frequências (org. dados)

Tabelas de contagens que resumem a informação contida na amostra, ordenando os seus valores e agrupando-os em categorias (de valores repetidos ou de valores distribuídos por classes/intervalos).

Em geral, são calculadas as

- frequências absolutas (acumuladas ou não acumuladas);
- frequências relativas (acumuladas ou não acumuladas).

Tabelas de frequências (org. dados)

Durante um ano contabilizou-se, diariamente, o número de golfinhos presos nas redes dos pescadores das águas açorianas, tendo-se obtido a seguinte tabela em SPSS:

nº de golfinhos presos num dia

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	0	37	14,4	14,4	14,4
	1	45	17,5	17,5	31,9
	2	84	32,7	32,7	64,6
	3	52	20,2	20,2	84,8
	4	23	8,9	8,9	93,8
	5	11	4,3	4,3	98,1
	6	2	,8	,8	98,8
	8	1	,4	,4	99,2
	9	1	,4	,4	99,6
	13	1	,4	,4	100,0
	Total	257	100,0	100,0	

Qual o significado de

52

84,8 ?

Quantos dias foram analisados?

Tabelas de frequências (org. dados)

Percent e **Valid Percent**

Atributo1

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	0	4	4,0	8,0	8,0
	1	11	11,0	22,0	30,0
	2	15	15,0	30,0	60,0
	3	10	10,0	20,0	80,0
	4	7	7,0	14,0	94,0
	5	2	2,0	4,0	98,0
	6	1	1,0	2,0	100,0
	Total	50	50,0	100,0	
Missing	99	50	50,0		
Total		100	100,0		

Exemplo: 100 observações; em metade não foi possível medir o atributo de interesse (Atributo1); valores omissos classificados com 99.

Medidas amostrais (descrição numérica)

- Uma medida amostral é uma medida numérica que descreve alguma característica de uma amostra.
- É habitualmente representada por letras latinas.
- A maioria das medidas que vamos considerar podem ser obtidas usando a máquina de calcular.
- Exemplos de medidas amostrais:
  $\bar{x}$ (média, *mean, average*)
  s (desvio padrão, *standard deviation*), s^2 (variância,*variance*)
  m_o (moda,*mode*)

Medidas amostrais (descrição numérica)

Nos próximos slides iremos abordar as:

- medidas de localização central: média $\bar{x}$, mediana m_e, moda m_o;
- medidas de localização relativa: mínimo, máximo, quantil x_p cujos casos particulares são: quartil Q_i, decil D_i, percentil P_i;
- medidas de dispersão: amplitude, desvio padrão s, variância s^2, distância interquartil H;
- medidas de forma: coeficiente de assimetria,

podendo, muitas delas, ser obtidas com as funções estatísticas das máquinas de calcular.

Média (medida de localização central)

- Numa amostra de n observações, $x_1, x_2, \ldots, x_n$, a **média** é

$$\bar{x} = \frac{x_1 + x_2 + \cdots + x_n}{n} = \frac{\sum_{i=1}^{n} x_i}{n}.$$

- Se os dados estiverem agrupados em k valores distintos:

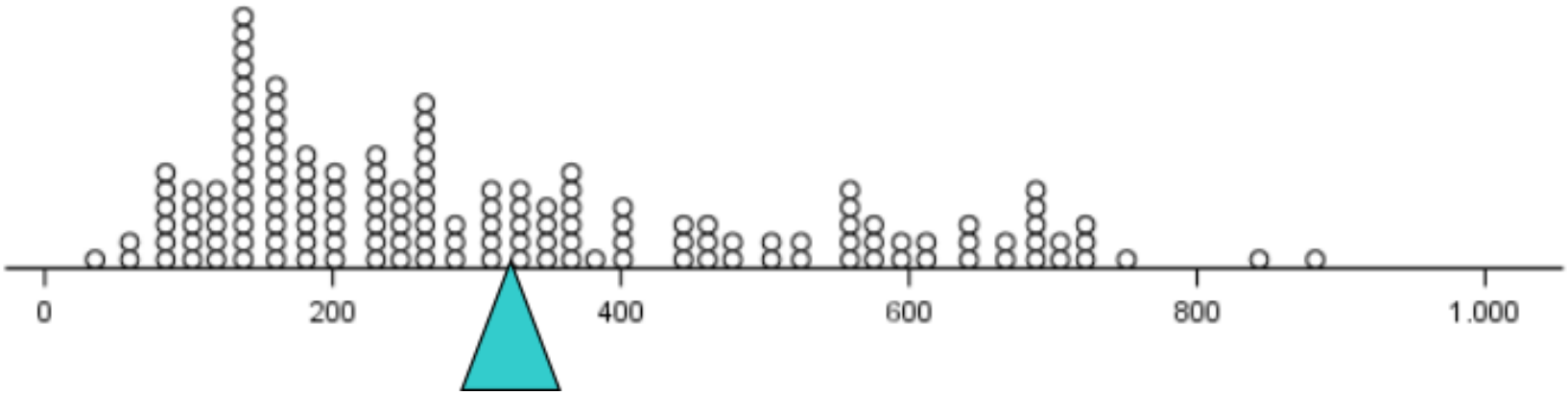
$$\bar{x} = \frac{x_1^* n_1 + x_2^* n_2 + \cdots + x_k^* n_k}{n} = \frac{\sum_{i=1}^{k} x_i^* n_i}{n}$$

onde n_i designa a frequência absoluta de x_i^*.

Pode usar-se, naturalmente, a frequência relativa no cálculo da média. E, no caso de dados agrupados em classes, a média poderá ser aproximada através do ponto médio da classe (x_i^* será substituído pelo ponto médio da classe k).

Média (medida de localização central)

- A média pode ser pensada como o centro de massa dos valores das observações, i.e., o ponto de equilíbrio após dispormos as observações sobre uma régua.



- Observações atípicas ou erros nas observações podem afastar a média do centro das observações.

Mediana (medida de localização central)

- Primeiro, ordena-se a amostra.
- A **mediana** é a observação localizada no centro da **amostra ordenada**.
- Dizer que m_e é a mediana de uma amostra significa que pelo menos 50% dos dados são inferiores ou iguais a m_e e pelo menos 50% dos dados são superiores ou iguais a m_e.
- A mediana é mais robusta do que a média a erros ou a observações afastadas da maioria das observações.

Mediana (medida de localização central)

Assim, se a amostra tiver:

- dimensão ímpar, m_e coincide com a observação central.

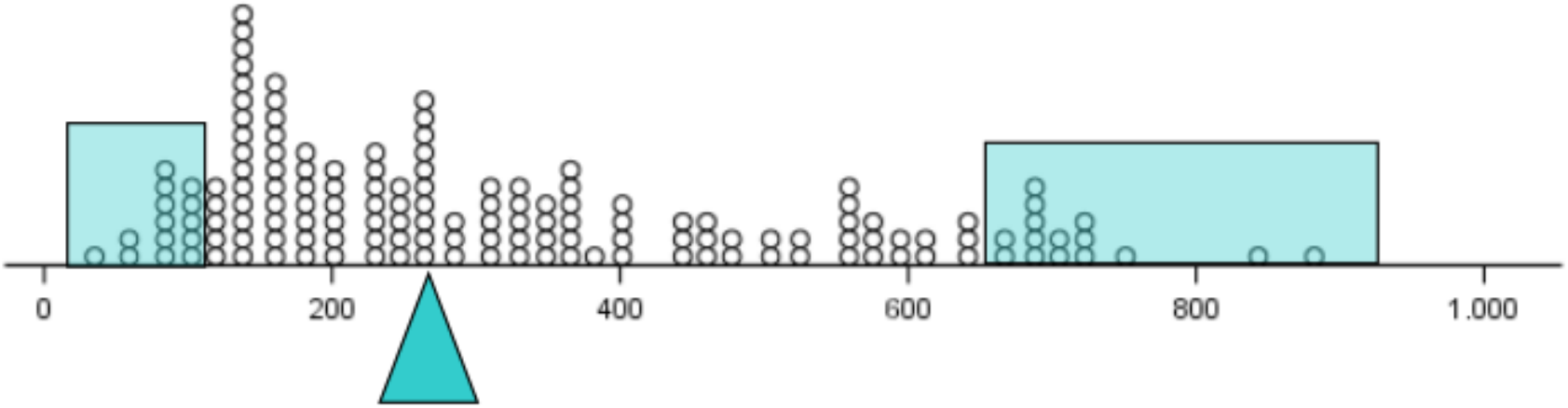
  Exemplo: na amostra 1.2; 2.1; 2.1; 2.2; 2.4, a mediana é 2.1. (Atenção à sua interpretação!)

- dimensão par, a mediana toma o valor da média das duas observações mais centrais.

  Exemplo: na amostra 0.0; 0.7; 0.9; 1.6, a mediana é 0.8.

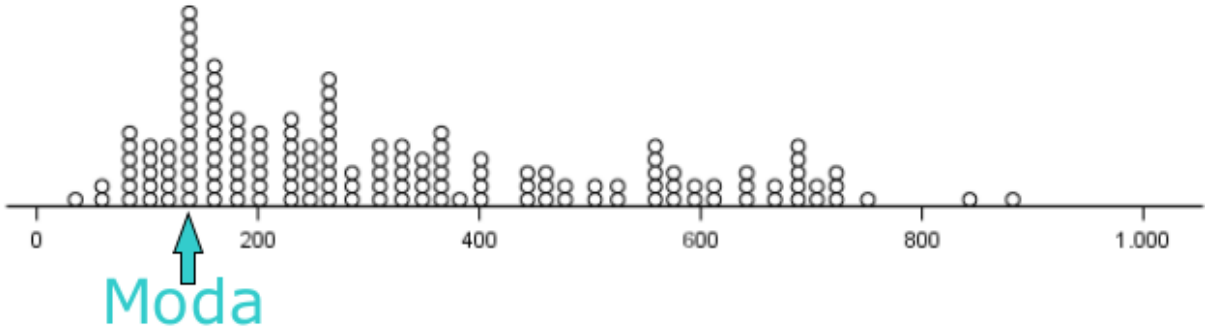
Média aparada (localização central)

Uma **média aparada** a $\alpha\%$ é a média calculada quando se exclui $\alpha\%$ das observações mais extremas, ou seja, $(\alpha/2)\%$ em cada extremo da amostra.



Moda (medida de localização central)

- A **moda** é o valor mais frequente de uma amostra.
- Ao contrário do que acontece com a mediana e a média, uma amostra pode possuir mais do que uma moda.



- A moda é a única medida de localização central que pode ser utilizada para dados numa escala nominal.
- A moda pode não ter significado, especialmente em dados de natureza contínua ou em dados discretos com poucas observações repetidas!
- Quando os dados estão agrupados em classes pode definir-se a classe modal, ou seja, a classe com maior frequência absoluta/relativa.

Mínimo e máximo (localização relativa)

- Mínimo: é o valor mais reduzido da amostra.
- Máximo: é o valor mais elevado da amostra.

Numa amostra ordenada, $x_{(i)}$ representa o elemento na posição i. Por exemplo:

$x_{(1)}$	$x_{(2)}$	$x_{(3)}$	$x_{(4)}$
10	12	13	20

sendo o mínimo=10, $x_{(1)} = 10$, e o máximo=20, $x_{(4)} = 20$.

Quantis (medidas de localização relativa)

- **Quantil de ordem** p $(0 < p < 1)$: é um valor, x_p, que divide a amostra ordenada em duas partes, tal que à esquerda de x_p está uma proporção p dos valores mais reduzidos da amostra e à direita uma proporção $(1 - p)$ dos valores mais elevados da amostra.
- Dizer que x_p é o quantil de ordem p significa que pelo menos $p \times 100\%$ das observações da amostra são menores ou iguais a x_p e que pelo menos $(1 - p) \times 100\%$ das observações são maiores ou iguais a x_p.

Quantis (como calcular)

- Por exemplo, qual o valor do quantil de ordem 0.25, i.e. $x_{0.25}$, na amostra ordenada?

$x_{(1)}$	$x_{(2)}$	$x_{(3)}$	$x_{(4)}$	$x_{(5)}$	$x_{(6)}$	$x_{(7)}$	$x_{(8)}$	$x_{(9)}$	$x_{(10)}$
2	3	5	5	5	5	5	5	5	6

Calcula-se $\boxed{(n+1)p}$[1] com $n = 10$ e $p = 0.25$ obtendo-se a posição $(10+1)0.25 = 2.75$. Este resultado indica que o valor na posição 3 deverá "pesar" mais do que o valor na posição 2. Assim,

$$x_{0.25} = x_{(2)} \times 0.25 + x_{(3)} \times 0.75 = 4.5.$$

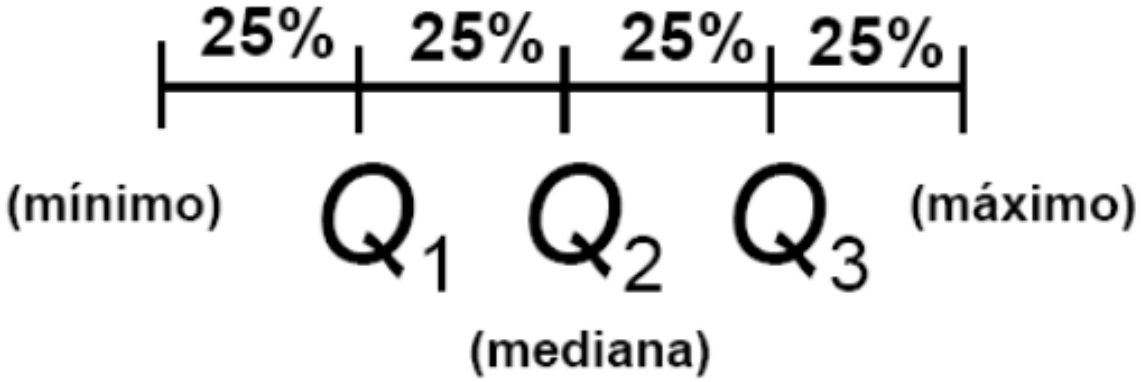
Ou, de forma equivalente,

$$x_{0.25} = x_{(2)} + 0.75 \times (x_{(3)} - x_{(2)}) = 4.5.$$

[1] Alguns autores designam quantil de ordem α (e não ordem p). Naturalmente, nesse caso, considera-se $\boxed{(n+1)\alpha}$.
Note-se que existem outras expressões para identificar a ordem da observação; e.g., Murteira et al. (2015).

Quartis (medidas de localização relativa)

- Consoante o valor de p alguns quantis recebem designações próprias. Existem os
  - quartis: $p \in \{1/4, 2/4, 3/4\}$;
  - decis: $p \in \{1/10, 2/10, \cdots, 9/10\}$;
  - percentis: $p \in \{1/100, 2/100, \cdots, 99/100\}$.
- Os quartis são habitualmente denotados por Q_1, Q_2 e Q_3. Note que Q_2 coincide com a mediana.



Medidas de dispersão (medidas amostrais)

De seguida, iremos abordar as seguintes medidas de dispersão:

- amplitude (*range*);
- distância interquartil (*inter-quartile range*);
- variância (*variance*);
- desvio padrão (*standard deviation*).

Amplitude (medida de dispersão)

- A **amplitude** de uma amostra é a diferença entre o máximo e o mínimo.

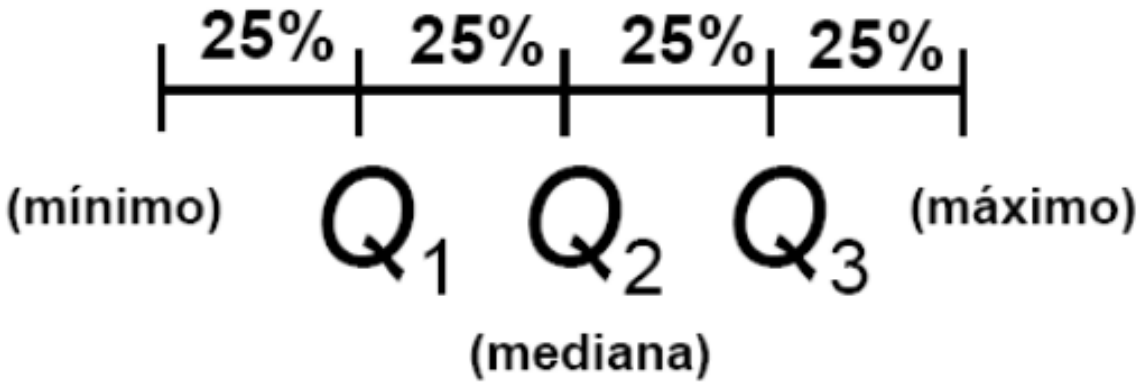
  Exemplo: na amostra

  $$1.2; 1.7; 2.1; 2.2; 2.3$$

  a amplitude é 2.3 - 1.2 $=$ 1.1.

Distância interquartil (dispersão)

- Designa-se por **intervalo interquartil** ao intervalo entre o 1º e o 3º quartis: $[Q_1, Q_3]$.
  Nele encontram-se 50% das observações (as mais centrais).



- A **distância (amplitude) interquartil** é a diferença entre o 3º e o 1º quartis:

$$H = Q_3 - Q_1.$$

Variância (medida de dispersão)

- A **variância** é a média dos quadrados dos desvios das observações em relação à média da amostra:

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

- Habitualmente, também é considerada a **variância corrigida**:

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

Desvio padrão (medida de dispersão)

A variância não é representada na mesma unidade das observações. Todavia, ao considerar-se a raiz quadrada da variância obtém-se o **desvio padrão**, igualmente uma medida de dispersão, que é representado na mesma unidade das observações.

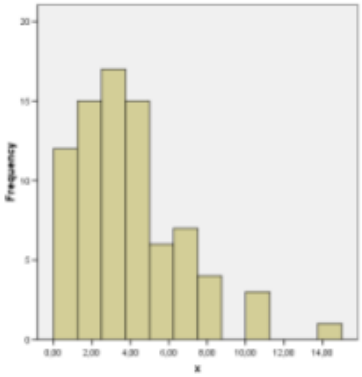
$$s = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2} \qquad s_c = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2}$$

- No SPSS são fornecidas as versões corrigidas da variância e do desvio padrão. Nas máquinas de calcular e no Excel existem as duas.
- O desvio padrão e a variância podem ser fortemente afetados por erros ou observações atípicas.

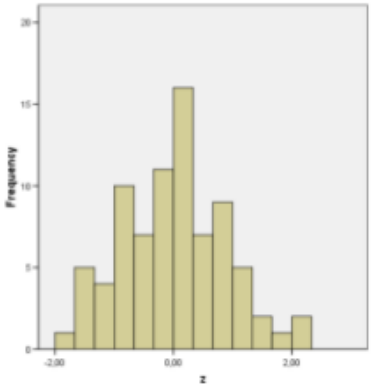
Medidas de forma (medidas amostrais)

- A forma da distribuição de valores da amostra pode ser avaliada pela sua **simetria** ou **assimetria**.
- Diz-se que uma distribuição possui assimetria positiva (negativa) quando existe uma concentração de observações na zona de valores mais reduzidos (elevados) da amostra.

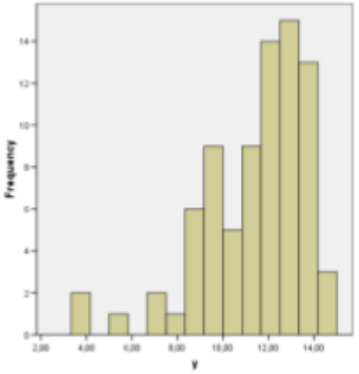
Assimetria positiva



Quase simetria



Assimetria negativa

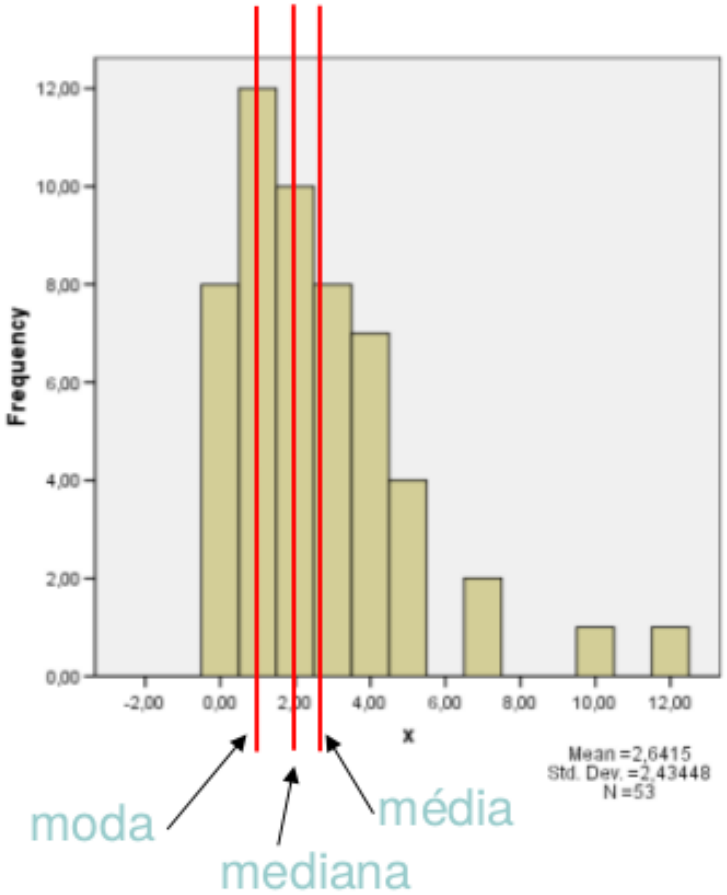


Média, mediana e moda na assimetria

- A assimetria também pode ser avaliada comparando os valores da média, mediana e moda (desde que esta última tenha sentido).

Assimetria positiva:

moda < mediana < média



Média, mediana e moda na assimetria

Assimetria positiva: moda $<$ mediana $<$ média

Assimetria negativa: média $<$ mediana $<$ moda

Simetria pura: média $=$ mediana $=$ moda

Simetria aproximada: média $\approx$ mediana $\approx$ moda

Coeficiente de assimetria (assimetria)

- O *software* SPSS, entre muitos outros, permite o cálculo de uma medida conhecida por coeficiente de assimetria (*skewness*). Trata-se de uma medida amostral que assume o valor zero quando a distribuição de frequências da amostra é completamente simétrica. Quando assume valores positivos: evidência de assimetria positiva. Quando assume valores negativos: evidência de assimetria negativa.
- Note-se que é muito raro observar simetria pura numa amostra, pelo que o coeficiente de assimetria assume quase sempre valores diferentes de zero.

Descrição gráfica dos dados

A construção de gráficos de frequências ou associados a medidas amostrais tem como objectivo dar uma informação visual sobre a forma da distribuição dos dados da amostra e, consequentemente, da população de onde foram extraídos.

Iremos ver os

- Histogramas;
- Gráficos de frequências;
- Caixas de bigodes (*boxplots*).

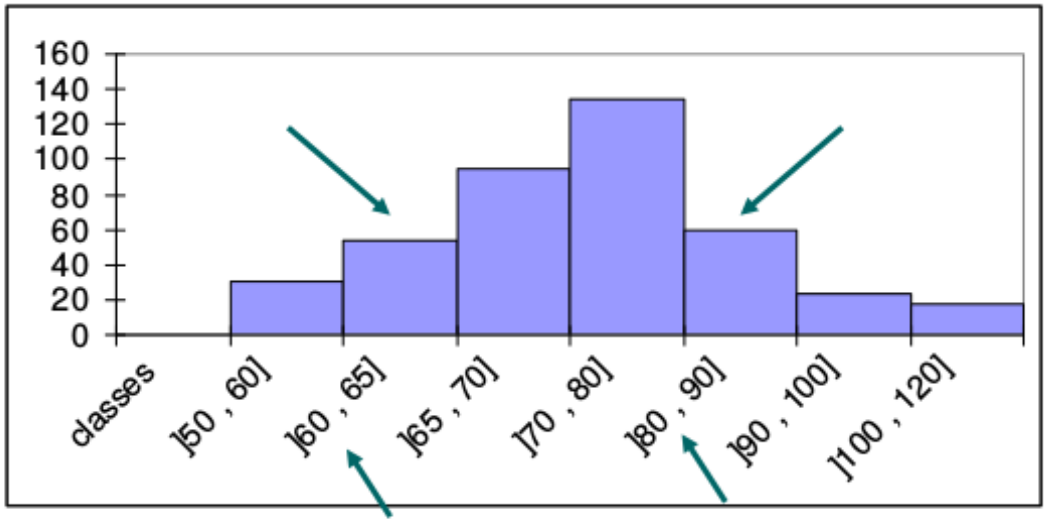
Histogramas (descrição gráfica)

- O histograma é um gráfico de rectângulos justapostos que reflecte a forma da distribuição de frequências da amostra.
- Os histogramas são particularmente úteis para amostras de variáveis contínuas ou amostras com poucas observações repetidas.
- Para construir um histograma é necessário:
  - primeiro repartir os dados por classes (=bases de cada rectângulo);
  - depois, calcular as respectivas frequências ajustadas (=frequência/amplitude da classe).

Classes com diferentes amplitudes?

Quando as classes têm diferentes amplitudes, observa-se que alguns *softwares* disponíveis não produzem histogramas correctos. Por exemplo, o seguinte conjunto de dados relativos ao peso de 414 mulheres com 8 meses de gestação é representado, no Excel, pelo seguinte "histograma"!

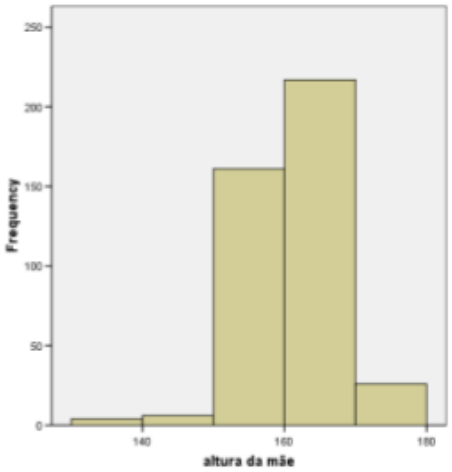
classes	Freq absoluta
]50 , 60]	30
]60 , 65]	54
]65 , 70]	95
]70 , 80]	134
]80 , 90]	60
]90 , 100]	23
]100 , 120]	18
	414

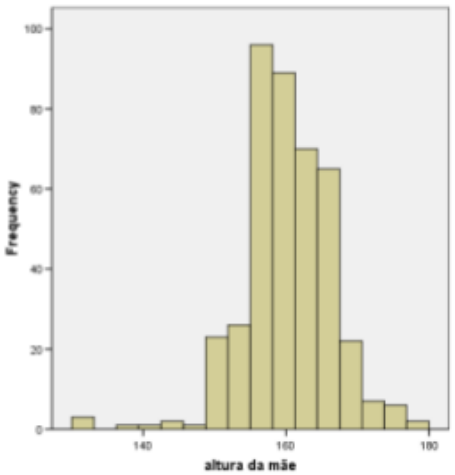


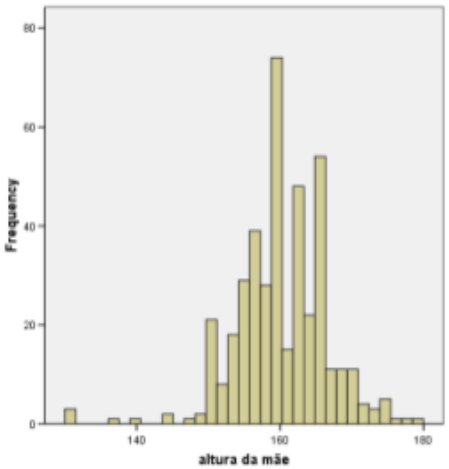
Diferentes amplitudes

Número de classes (histogramas)

Os histogramas dependem do número de classes considerado. Muitas classes podem produzir um histograma com demasiada irregularidade, enquanto que um histograma com poucas classes pode ocultar a forma da distribuição (e perder-se informação).







Os histogramas anteriores são referentes ao mesmo conjunto de observações. Embora a escolha do número "adequado" de classes dependa de vários fatores (forma da distribuição, objetivos do estudo, etc.), existem diversas regras práticas, sendo as mais populares: a de Sturges, a de Doane e a do inteiro mais próximo do valor da raiz quadrada do número de observações.

Gráficos de frequências (descrição gráfica)

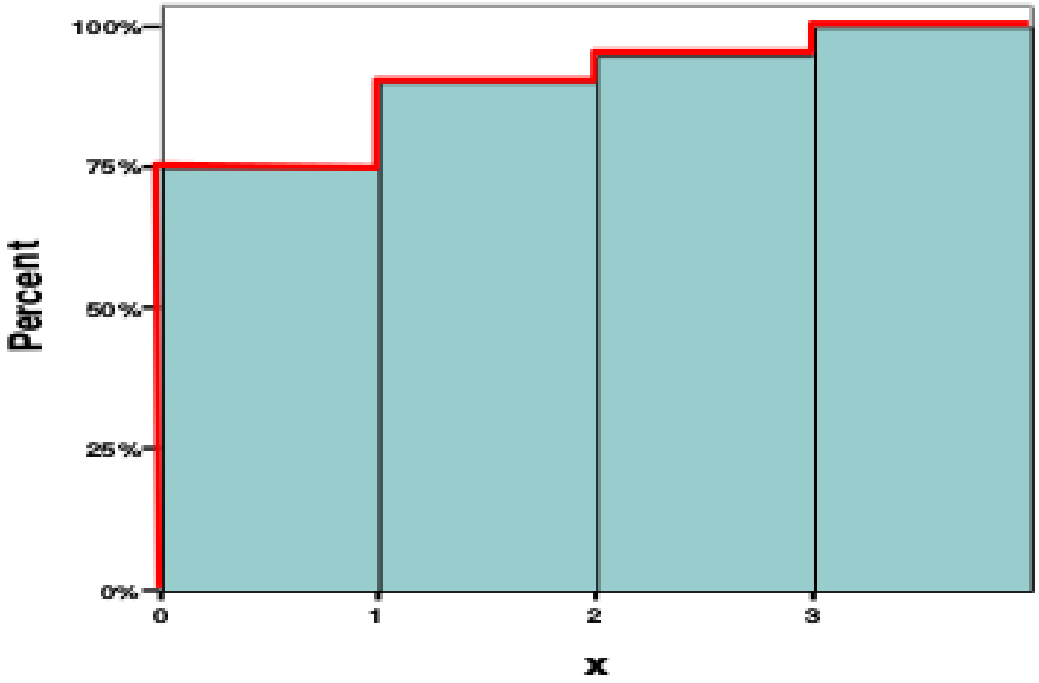
- Gráficos de frequências são gráficos de barras que traduzem graficamente o conteúdo da tabela de frequências. Os mais habituais são os gráficos de frequências absolutas ou relativas, mas também podemos construir gráficos de frequências absolutas ou relativas acumuladas.
- Os gráficos de frequências (não acumuladas) são apropriados para dados qualitativos ou numéricos discretos. Quando as frequências absolutas são reduzidas e a gama de valores da amostra é dispersa os gráficos de frequências tornam-se pouco interessantes (muito irregulares).

Função de distribuição empírica

Chama-se função de distribuição empírica à função definida através dos valores das frequências relativas acumuladas.

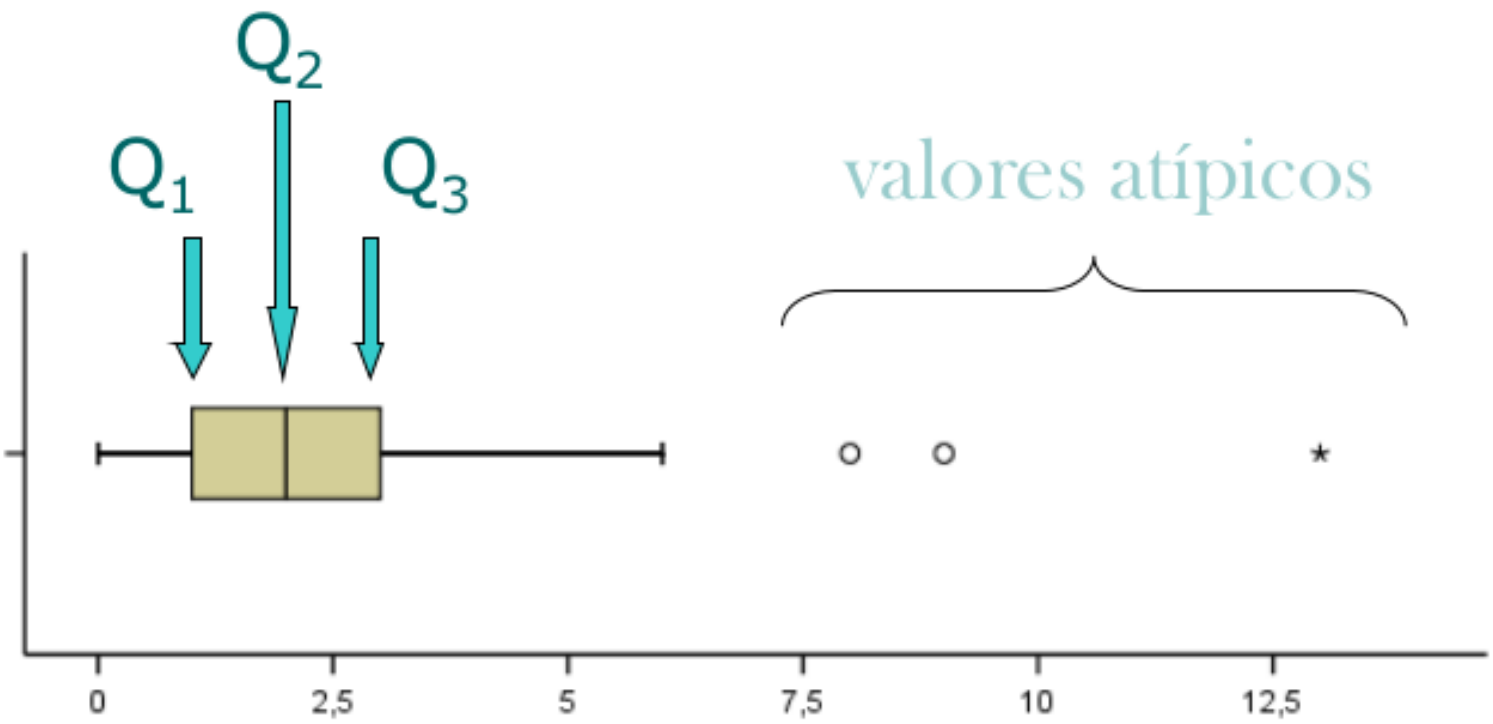
$$F_n(x) = \frac{\text{número de observações} \leq x}{n}$$

Exemplo:



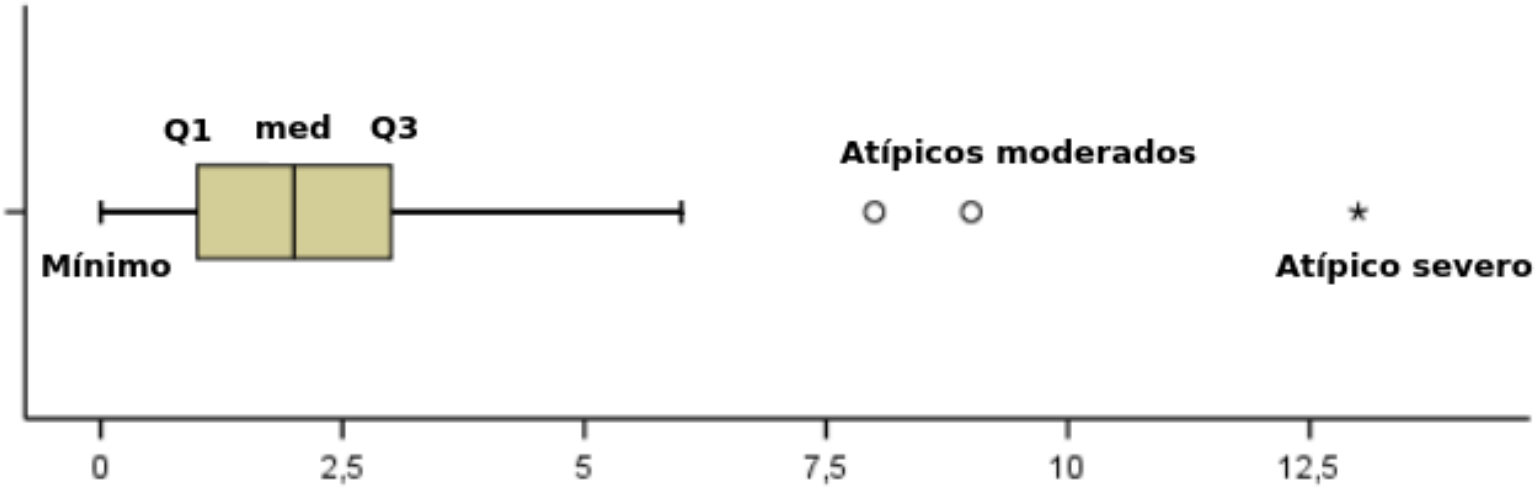
Caixas de bigodes (descrição gráfica)

- A caixa de bigodes (*boxplot*) corresponde a uma representação gráfica de algumas medidas de localização.
- As máquinas gráficas fazem *boxplots*, mas saber como os construir é essencial.



Caixas de bigodes (construção)

1. A "caixa" é marcada com os quartis Q_1, Med, Q_3;
2. Bigodes e valores atípicos (*outliers*):
   - a distância interquartil é $H = Q_3 - Q_1$;
   - as barreiras[1] de valores atípicos moderados (ou primeiras barreiras de *outliers*) são $Q_1 - 1.5H$ e $Q_3 + 1.5H$;
   - as barreiras[1] de valores atípicos severos (ou segundas barreiras de *outliers*) são $Q_1 - 3H$ e $Q_3 + 3H$.



[1] Estes limites, embora sejam os habitualmente usados, poderão ser modificados em situações específicas.

Caixas de bigodes (construção)

Note que os "bigodes" não podem ultrapassar as primeiras barreiras de *outliers*, pelo que, quando existem observações para além destas barreiras, os "bigodes" são desenhados somente até à última observação que não a ultrapassa (barreira superior ou inferior).

As observações entre barreiras são desenhadas com $\circ$ (atípico moderado ou *outlier* moderado) e as observações para além das segundas barreiras com $*$ (atípico severo ou *outlier* severo).

Para concretizar, considere a seguinte informação relativa ao número de pagamentos em atraso a fornecedores, verificado no último ano, num determinado sector empresarial:

Nº de pagamentos em atraso	0	1	2	3	4	6	9
Quantidade de empresas	64	34	32	18	2	1	1

Caixas de bigodes (construção)

Mínimo: 0; Máximo: 9. Para calcular os quartis:

1º quartil: 0.
$(152 + 1) \times 0.25 = 38.25$ (posição 38: 0; posição 39: 0);

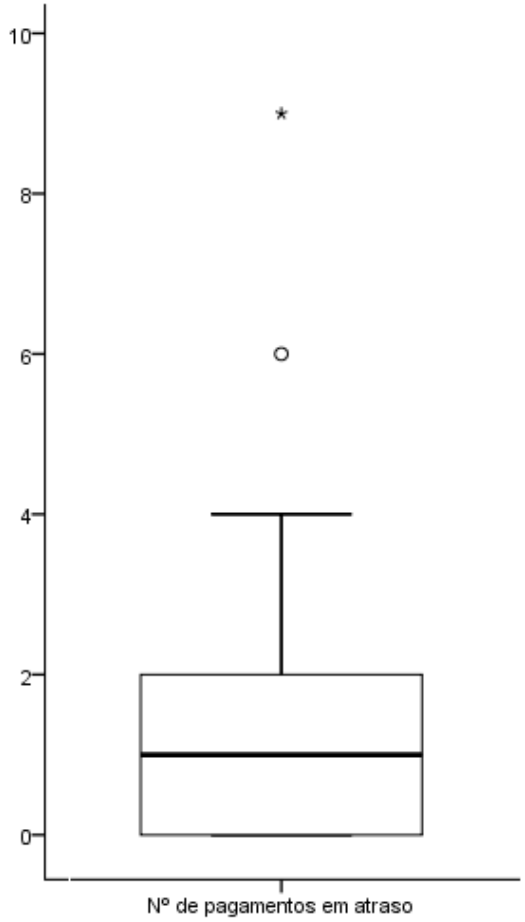
2º quartil (mediana): 1.
$(152 + 1) \times 0.50 = 76.50$ (posição 76: 1; posição 77: 1);

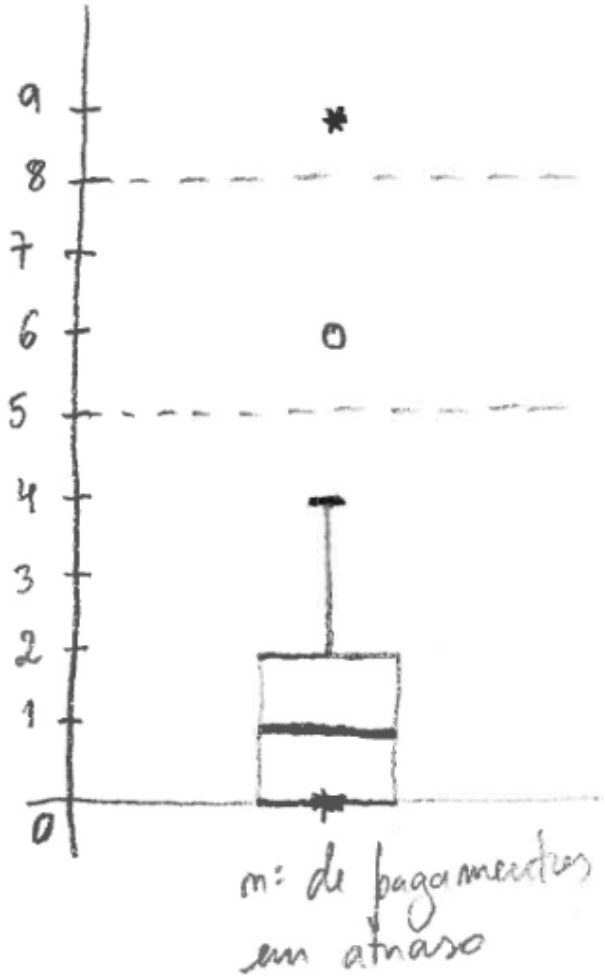
3º quartil: 2.
$(152 + 1) \times 0.75 = 114.75$ (posição 114: 2; posição 115: 2).

Naturalmente, neste exercício, os *outliers*, a existirem, serão somente para valores elevados da amostra. Assim,

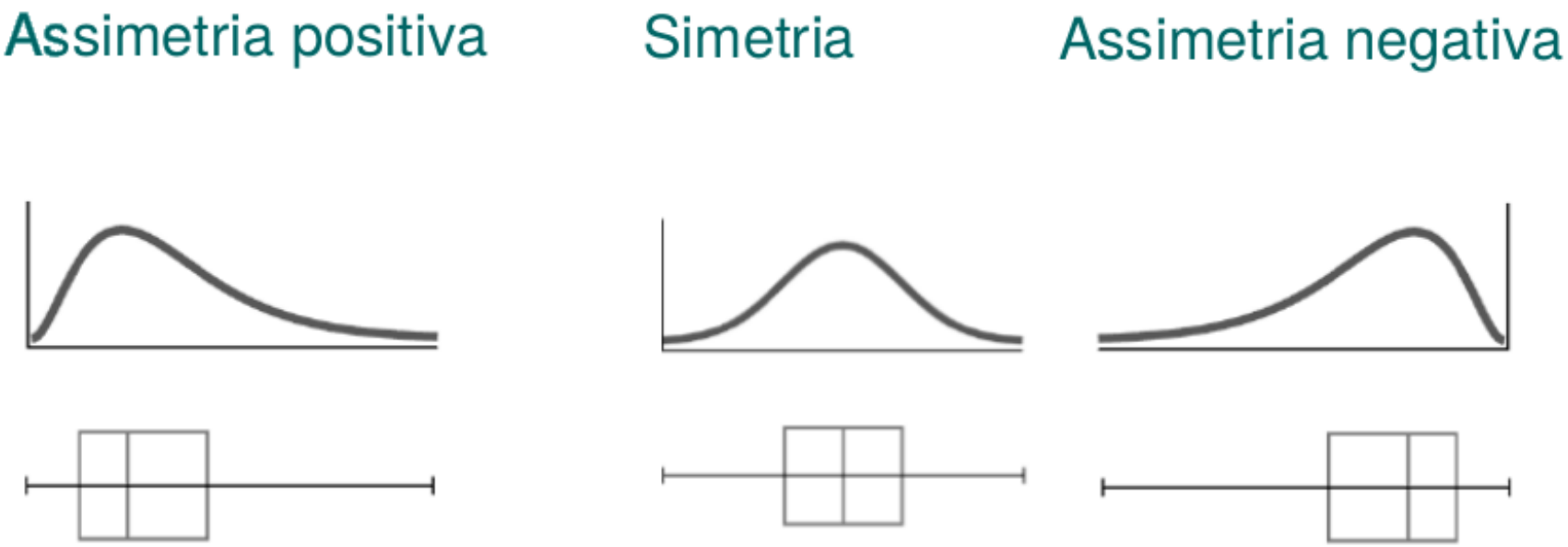
1ª barreira superior: $2 + 1.5 \times (2 - 0) = 5$;
2ª barreira superior: $2 + 3.0 \times (2 - 0) = 8$.

Caixas de bigodes (construção)





Caixas de bigodes (assimetria)



Nota: esta ilustração não implementa a identificação de valores atípicos (com $\circ$ e/ou com $*$) no *boxplot*.

Caixas de bigodes (resumo)

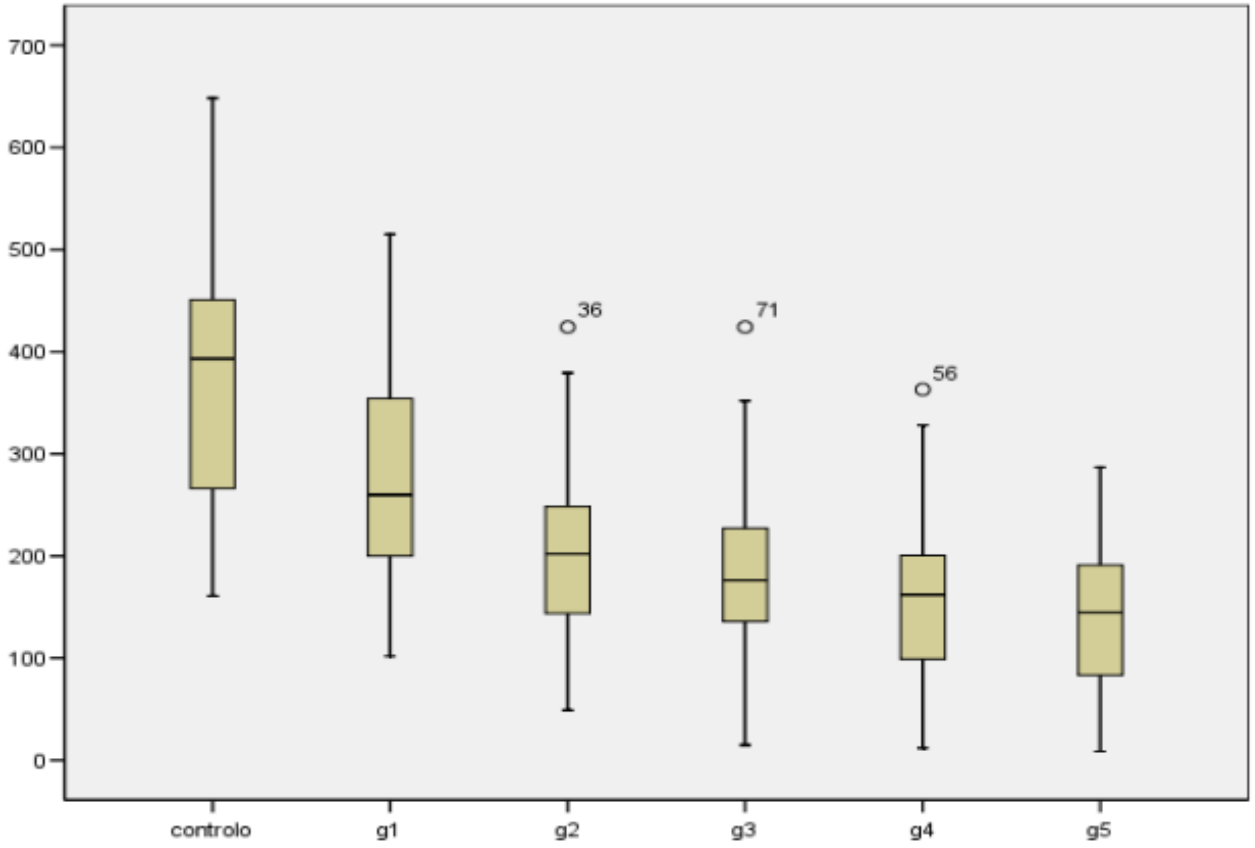
As caixas de bigodes fornecem informação sobre

- Localização central: mediana.
- Localizações relativas: 1º e 3º quartis, mínimo e máximo.
- Dispersão: amplitude e distância interquartil.
- Assimetria: pela posição relativa da mediana na caixa, pelo comprimento dos bigodes e pela eventual existência de *outliers*.

Caixa de bigodes comparativas

As caixas de bigodes também são úteis para comparar várias amostras num mesmo gráfico: caixas de bigodes comparativas.

Tempos de vida de porquinhos da Índia infectados com várias doses de bacilos da tuberculose



O que representam as *labels* no gráfico?

Diagrama de caule-e-folhas

Dados:
1.5, 1.6, 1.6, 1.6, 1.7, 1.7, 1.8, 1.9, 2.1, 2.5, 2.7, 2.7, 3.8, 4.1.

```
Exemplo Stem-and-Leaf Plot

 Frequency    Stem &  Leaf

     8,00        1 .  56667789
     4,00        2 .  1577
     1,00        3 .  8
     1,00        4 .  1

 Stem width:       1,0
 Each leaf:        1 case(s)
```

Diagrama de caule-e-folhas

Dados:
15, 16, 16, 16, 17, 17, 18, 19, 21, 25, 27, 27, 38, 41.

```
Exemplo Stem-and-Leaf Plot

 Frequency    Stem &  Leaf

     8,00        1 .  56667789
     4,00        2 .  1577
     1,00        3 .  8
     1,00        4 .  1

 Stem width:        10
 Each leaf:        1 case(s)
```

Diagrama de dispersão

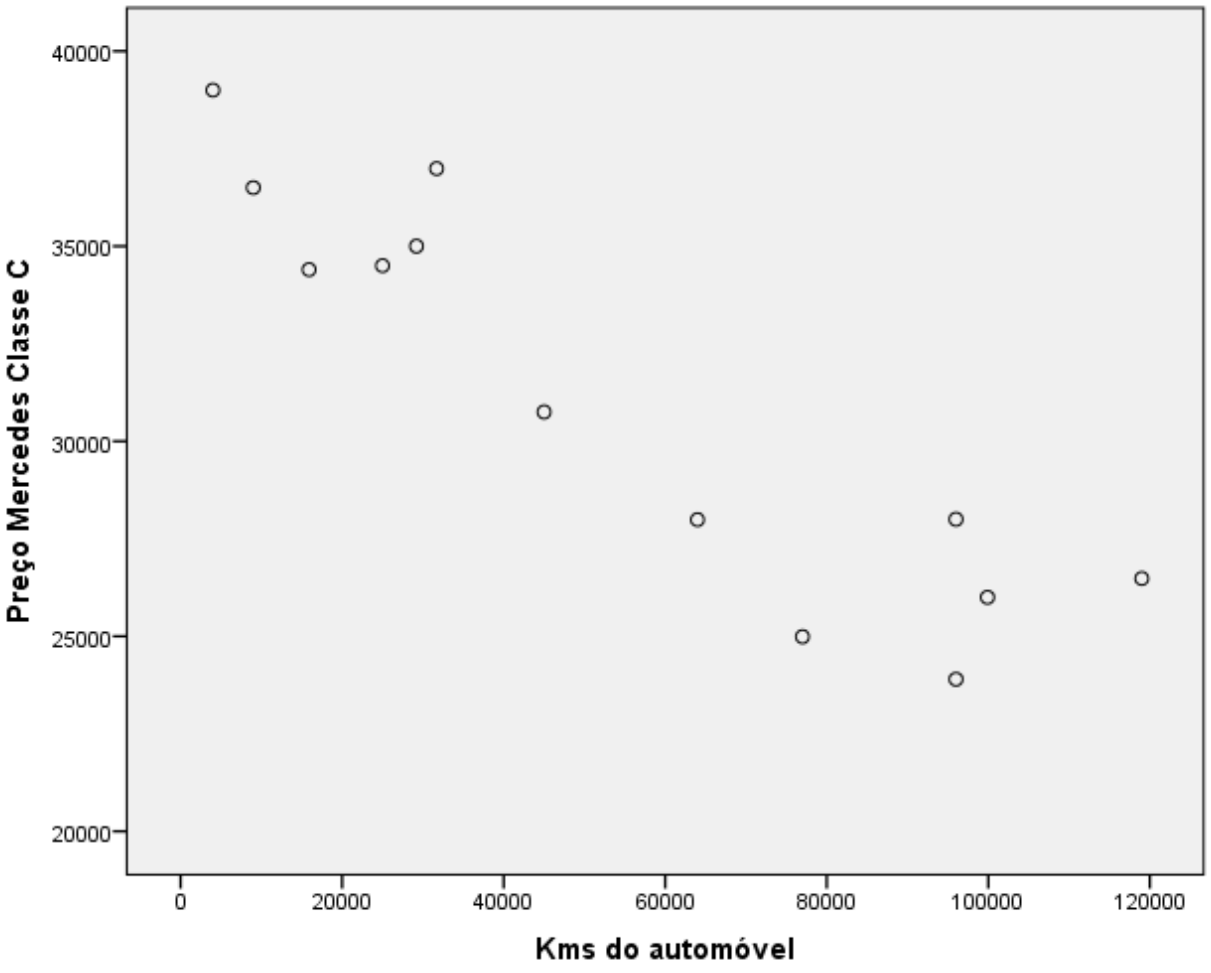
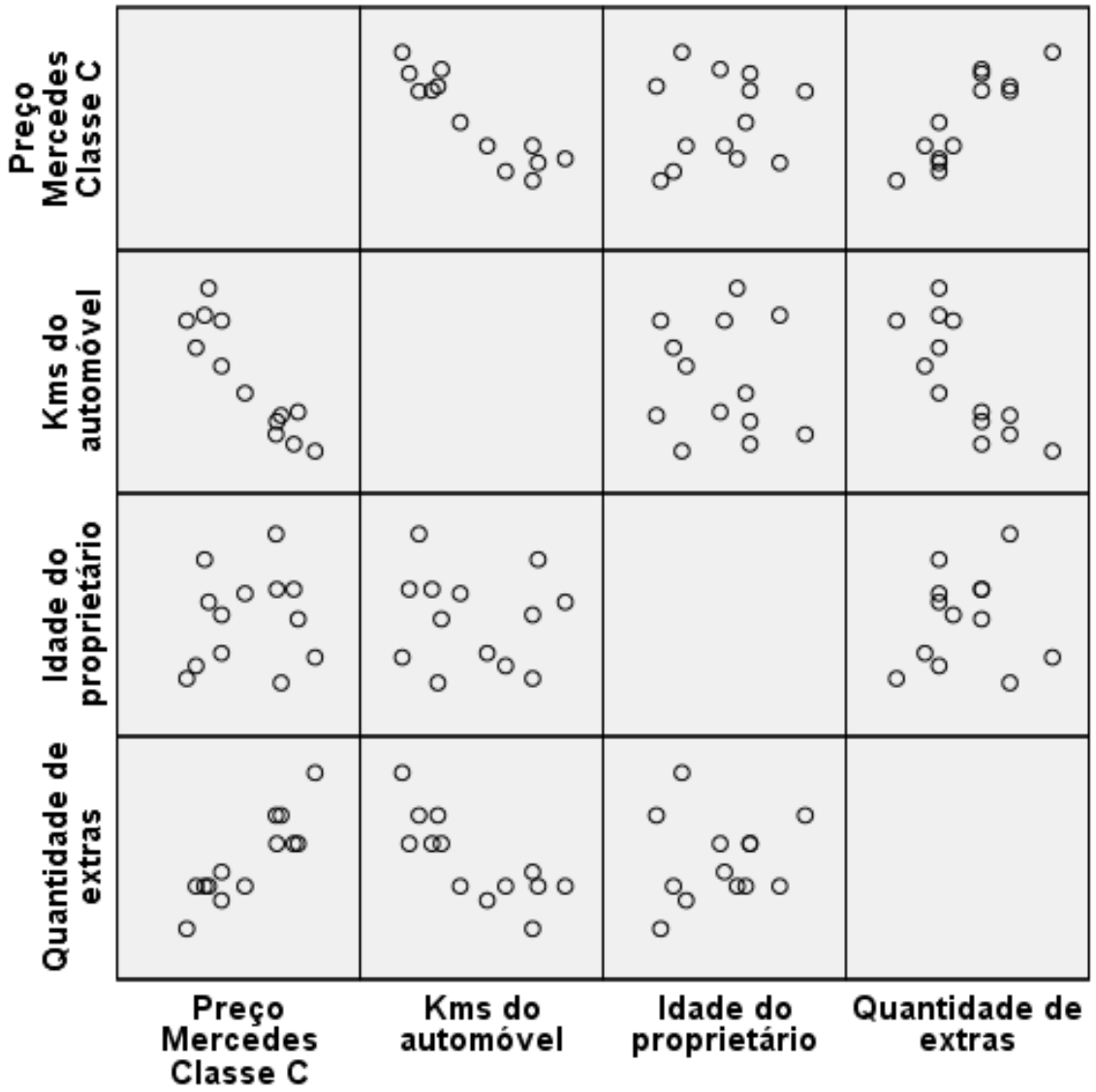


Diagrama de dispersão



Análise e interpretação

- Com base em medidas amostrais calculadas e em gráficos construídos, deve ser elaborado um relatório que descreva diferentes medidas amostrais informativas (medidas de localização, de dispersão e de forma) e ilustre a forma da distribuição dos dados através de gráficos adequados.

- Nesta fase exploratória de dados é essencial

"deixar os dados falarem por si"!

Capítulos

Conceitos Básicos

1. Análise Exploratória de Dados

2. Teoria de Probabilidades
   Conceitos Gerais de Probabilidade
   Noção de Variável Aleatória
   Variáveis Aleatórias Discretas
   Variáveis Aleatórias Contínuas
   Teorema do Limite Central

3. Inferência Estatística

Revisões de conceitos gerais

Experiência aleatória: qualquer experiência cujo resultado pertence a um conjunto previamente fixado e que não pode ser previsto com exatidão antes da realização da experiência.

Espaço de resultados (Ω): conjunto formado por todos os resultados que, hipoteticamente, são possíveis de obter quando se realiza a experiência aleatória.

Acontecimento: qualquer subconjunto do espaço de resultados Ω.

Acontecimento elementar: qualquer acontecimento constituído por um só elemento.

Revisões de conceitos gerais

Exemplo:

Experiência aleatória: lançamento de duas moedas e observação do lado que fica voltado para cima.

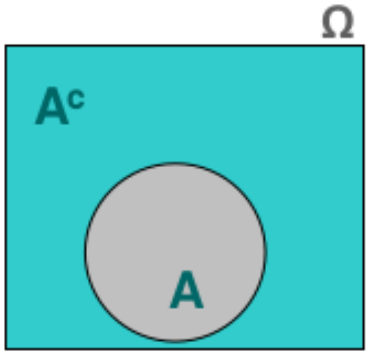
Espaço de resultados: $\Omega = \{KK, KC, CK, CC\}$, onde K representa a "saída de coroa" e C a "saída de cara".

Acontecimento: $\{KC, CK, CC\}$, por exemplo, que denota a ocorrência de "saída de pelo menos uma cara".

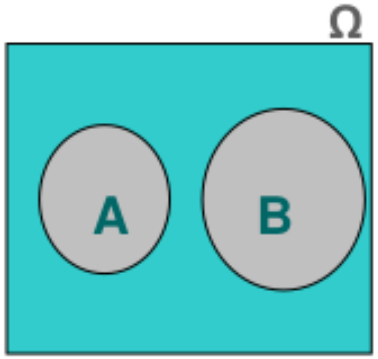
Acontecimento elementar: $\{CC\}$, por exemplo.

Revisões de conceitos gerais

Se A é um acontecimento de Ω, então $A^c = \overline{A} = \Omega \backslash A$ é o **complementar** de A.



Os acontecimentos A e B de Ω dizem-se **disjuntos** (ou **incompatíveis** ou **mutuamente exclusivos**) se não podem ocorrer em simultâneo, ou seja, $A \cap B = \emptyset$. (É impossível a realização simultânea de A e B.)



Revisões de conceitos gerais

Cálculo de probabilidades (conceito[1] clássico): seja Ω composto por n acontecimentos elementares distintos, em que cada um tem a *mesma chance* de ocorrer. Se o acontecimento A pode ocorrer em k desses n acontecimentos elementares, então a probabilidade de A ocorrer é

$$P(A) = \frac{\text{n}^{\underline{\text{o}}} \text{ de casos favoráveis a } A}{\text{n}^{\underline{\text{o}}} \text{ de casos possíveis}} = \frac{k}{n}.$$

Para o cálculo do número de casos favoráveis a A ou para o cálculo do número de casos possíveis, recorre-se, comummente, à análise combinatória (arranjos, permutações ou combinações).

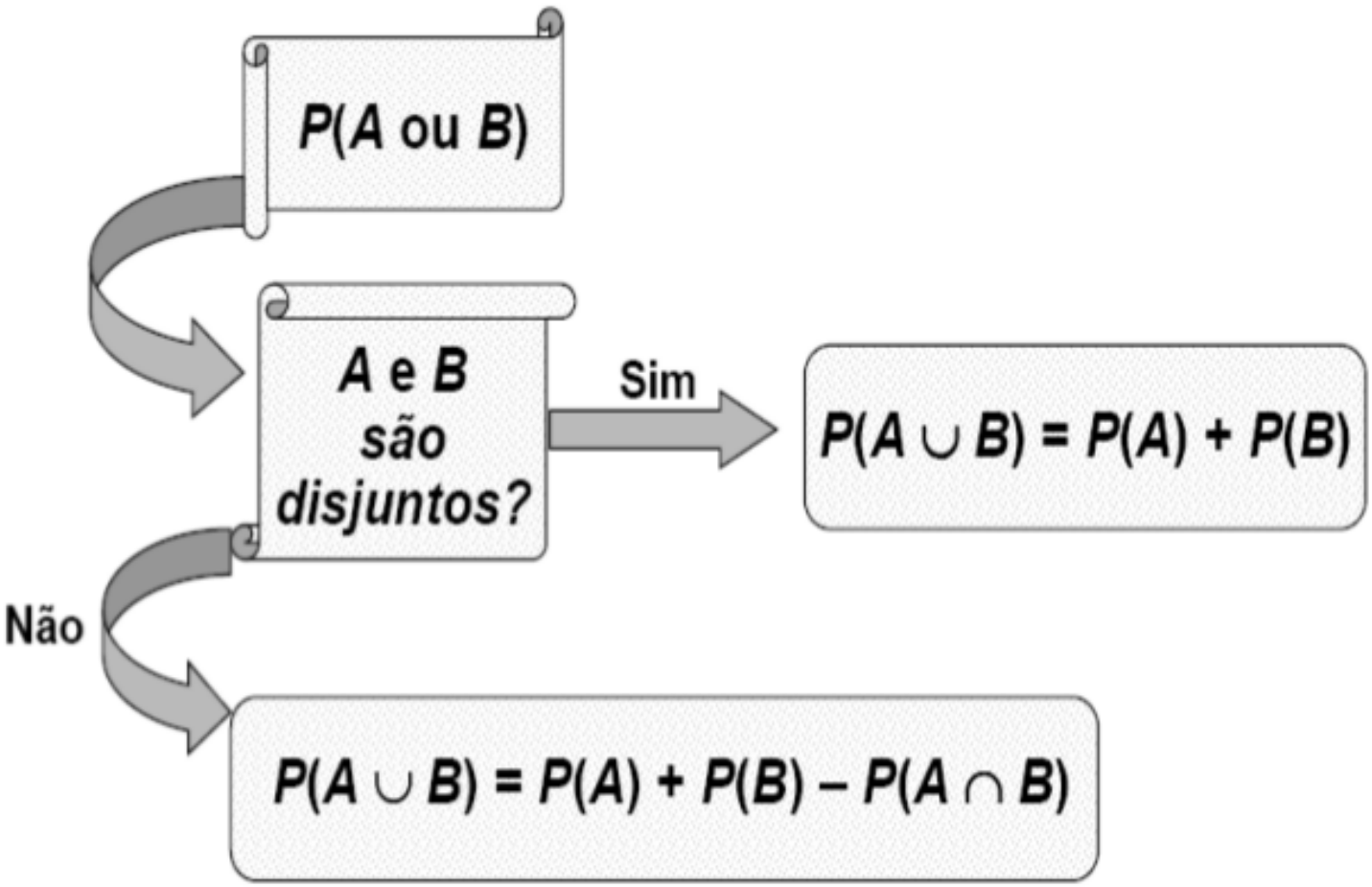
[1] Consulte a literatura para outros conceitos de probabilidade.

Revisões de conceitos gerais

Propriedades da medida de probabilidade

- A probabilidade do acontecimento impossível ($\emptyset$) é zero;
- A probabilidade do acontecimento certo (Ω) é um;
- Para qualquer acontecimento A: $0 \leq P(A) \leq 1$;
- $P(A^c) = P(\overline{A}) = 1 - P(A)$;
- $P(A \text{ e } B) = P(AB) = P(A \cap B) =$
  $= P(A \text{ e } B \text{ ocorrem simultaneamente})$;
- $P(A \text{ ou } B) = P(A \cup B) =$
  $= P(\text{ou ocorre } A \text{ ou ocorre } B \text{ ou ocorrem ambos})$;
- $P(A \cup B) = P(A) + P(B) - P(A \cap B)$.

Cálculo da probabilidade da reunião



Exercício

Em determinada população, 9.8% das pessoas adquirem a revista Caras, 22.9% a revista Lux e 5.1% ambas as revistas. Um individuo é seleccionado ao acaso dessa população. Qual a probabilidade de este

(a) adquirir somente a revista Caras?

(b) adquirir pelo menos uma das revistas?

(c) não adquirir nenhuma das revistas?

Resolução do exercício (em Excel)

Sejam os acontecimentos

A = "Adquirir a revista Caras"
B = "Adquirir a revista Lux"

Dados do Problema

P(**A**)=	0,098
P(**B**)=	0,229
P(**A**∩**B**)=	0,051

Logo,

	B	B^c	
A	0,051	0,047	0,098
A^c	0,178	0,724	0,902
	0,229	0,771	1

(a) $P(A \cap B^c) = P(A) - P(A \cap B) = 0.098 - 0.051 = 0.047$

(b) $P(A \cup B) = P(A) + P(B) - P(A \cap B) = 0.098 + 0.229 - 0.051 = 0.276$

(c) $P(A^C \cap B^c) = 1 - P(A \cup B) = 1 - 0.276 = 0.724$

Probabilidade Condicionada

Se existe informação adicional relevante sobre a ocorrência de um dado acontecimento, essa informação pode ser incorporada no cálculo das probabilidades recorrendo ao conceito de **probabilidade condicionada**.

Exemplo: seleciona-se, ao acaso, um aluno de uma sala de aula.

Qual a probabilidade do aluno ter nascido a 01 de janeiro?

E, caso fosse conhecido que o aluno faz anos em janeiro, qual a probabilidade do aluno ter nascido a 01 de janeiro?

Probabilidade Condicionada

$P(A|B)$, que se lê "probabilidade de A dado B", representa a probabilidade de o acontecimento A ocorrer, tendo conhecimento que (sabendo que ou dado que) o acontecimento B já se realizou e é dada por

$$P(A|B) = \frac{P(A \cap B)}{P(B)}, \text{ com } P(B) > 0.$$

O conhecimento da ocorrência de B implica uma restrição do espaço de resultados da experiência ao espaço definido por B, ou seja, $P(A|B)$ representa a probabilidade de A no espaço de resultados de B.

Pode ainda interpretar-se $P(A|B)$ como uma reavaliação da probabilidade de A quando se tem a informação de que o acontecimento B se realizou.

Exercício (Caderno de Exercícios)

Uma fábrica produz discos de plástico da qual se conhecem as probabilidades de um disco ter alta resistência a choque e a ferimentos, de acordo com a seguinte tabela:

	B	$\overline{B}$
A	0.80	0.09
$\overline{A}$	0.06	0.05

Sejam A o acontecimento "um disco tem alta resistência a choque" e B o acontecimento "um disco tem alta resistência a ferimentos". Nestas condições,

(a) encontre $P(A)$, $P(B)$, $P(A \cup B)$, $P(A \cup \overline{B})$ e $P(A|B)$.

Resolução do exercício

	B	$\overline{B}$	
A	0.80	0.09	0.89
$\overline{A}$	0.06	0.05	0.11
	0.86	0.14	1.00

$P(A) = 0.89;$

$P(B) = 0.86;$

$P(A \cup B) = P(A) + P(B) - P(A \cap B) = 0.89 + 0.86 - 0.80 = 0.95;$

$P(A \cup \overline{B}) = P(A) + P(\overline{B}) - P(A \cap \overline{B}) = 0.89 + 0.14 - 0.09 = 0.94;$

$$P(A|B) = \frac{P(A \cap B)}{P(B)} = \frac{0.80}{0.86} \approx 0.93.$$

Regra da multiplicação das probabilidades

A partir da definição de probabilidade condicionada vem

$$P(A|B) = \frac{P(A \cap B)}{P(B)} \Rightarrow P(A \cap B) = P(A|B)P(B), \text{ com } P(B) > 0,$$

$$P(B|A) = \frac{P(B \cap A)}{P(A)} \Rightarrow P(B \cap A) = P(B|A)P(A), \text{ com } P(A) > 0.$$

Como $P(A \cap B) = P(B \cap A)$, então,

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

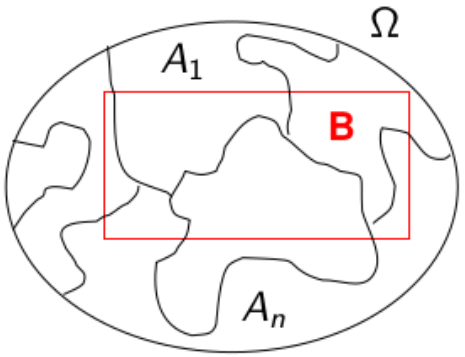
Esta regra pode ser generalizada para três ou mais acontecimentos. Para três acontecimentos vem

$$P(A \cap B \cap C) = P(A)P(B|A)P(C|A \cap B).$$

Lei da Probabilidade Total

A probabilidade condicionada tem especial interesse quando o espaço de resultados Ω admite uma partição[1] com n acontecimentos A_i, $i = 1, 2, \ldots, n$, da forma

- $\Omega = A_1 \cup A_2 \cup \ldots \cup A_n$;
- $A_i \cap A_j = \emptyset \ (i \neq j)$;
- $P(A_i) > 0$.



Assim, para qualquer acontecimento B,

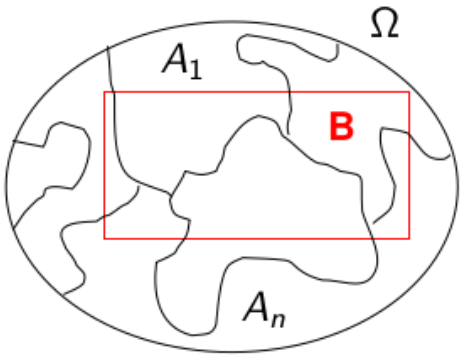
$$\begin{aligned} P(B) &= P(B \cap A_1) + P(B \cap A_2) + \ldots + P(B \cap A_n) \\ &= P(A_1)P(B|A_1) + P(A_2)P(B|A_2) + \ldots + P(A_n)P(B|A_n). \end{aligned}$$

[1] A partição do espaço de resultados também é possível com um número infinito não numerável de acontecimentos.

Teorema de Bayes

Quando o espaço de resultados Ω admite uma partição com n acontecimentos A_i, $i = 1, 2, \ldots, n$, da forma

- $\Omega = A_1 \cup A_2 \cup \ldots \cup A_n$;
- $A_i \cap A_j = \emptyset \ (i \neq j)$;
- $P(A_i) > 0$;



então, para qualquer acontecimento B, com $P(B) > 0$,

$$\begin{aligned} P(A_i|B) &= \frac{P(A_i)P(B|A_i)}{P(B)} \\ &= \frac{P(A_i)P(B|A_i)}{P(A_1)P(B|A_1) + P(A_2)P(B|A_2) + \ldots + P(A_n)P(B|A_n)}. \end{aligned}$$

Exercício

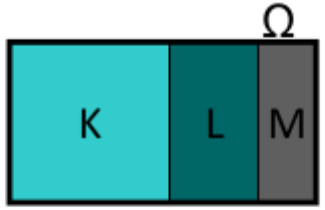
Num hospital dão entrada 50% de indivíduos com a doença K, 30% com a doença L e 20% com a doença M, sendo que só dão entrada naquele hospital indivíduos com a doença K, L ou M e estes só podem sofrer de uma e uma só daquelas doenças. A probabilidade de cura da doença K é 0.7; para as doenças L e M, a respectiva probabilidade é de 0.8 e 0.9.

Calcule:

a) a probabilidade de um individuo ser curado e ser da doença K;

b) a probabilidade de um doente internado ter alta (ie, estar curado).

c) a probabilidade de que doente internado tenha sofrido da doença K sabendo que a esse indivíduo lhe foi dada alta.



Resolução do exercício (em Excel)

Sejam os acontecimentos

K = Ter a doença K

L = Ter a doença L

M = Ter a doença M

C = Estar curado

Do enunciado tem-se:

P(**K**)=0,5 ; P(**L**)= 0,3; P(**M**)=0,2

P(**C** | **K**) = 0,7

P(**C** | **L**)= 0,8

P(**C** | **M**)= 0,9

Resolução via fórmulas:

a) $P(K \cap C) = P(K) * P(C \mid K) = 0,5 * 0,7 = 0,35$

b) $P(C) = P(K \cap C) + P(L \cap C) + P(M \cap C) = P(K)P(C \mid K) + P(L)P(C \mid L) + P(M)P(C \mid M)$

$= 0,5*0,7 + 0,3*0,8 + 0,2*0,9 = 0,77$

c) $P(K \mid C) = P(K \cap C) / P(C) = 0,35 / 0,77 = 0,45$

Resolução via tabela

	C	C^c	
K	0,35	0,15	0,5
L	0,24	0,06	0,3
M	0,18	0,02	0,2
	0,77	0,23	1

Dados do Problema

P(**K**)=0,5 ; P(**L**)= 0,3; P(**M**)=0,2

P(**C** | **K**) = 0,7

P(**C** | **L**)= 0,8

P(**C** | **M**)= 0,9

$P(K \cap C) = P(K) * P(C \mid K)$

$P(L \cap C) = P(L) * P(C \mid L)$

$P(M \cap C) = P(M) * P(C \mid M)$

Com os dados do problema, completar a tabela. Da tabela conclui-se que:

a) $P(K \cap C) = 0,35$; **b)** $P(C) = 0,77$; **c)** $P(K \mid C) = 0,35/0,77 = 0,45$

Exercício

Numa experiência de aprendizagem, um indivíduo realiza duas vezes seguidas uma determinada tarefa, podendo falhar ou ser bem-sucedido. A probabilidade de não ser bem-sucedido na primeira é 0.53. Se falhar na primeira, a probabilidade de ser bem-sucedido na segunda é de 0.64. Se for bem-sucedido na primeira, a probabilidade de ser sucedido na segunda é 0.91. Calcule a probabilidade de falhar na segunda tentativa.

Resolução

Sejam os acontecimentos: $T_i -$ Tarefa i bem sucedida $(i = 1, 2)$.
Sabe-se que $P\left(\overline{T}_1\right) = 0.53$, pelo que $P(T_1) = 0.47$. Sabe-se, ainda, que $P\left(T_2|\overline{T}_1\right) = 0.64$ e $P(T_2|T_1) = 0.91$. Pretende-se $P\left(\overline{T}_2\right)$. Assim, pela Lei da Probabilidade Total,

$$\begin{aligned} P\left(\overline{T}_2\right) &= P\left(T_1 \cap \overline{T}_2\right) + P\left(\overline{T}_1 \cap \overline{T}_2\right) \\ &= P\left(\overline{T}_2|T_1\right) P(T_1) + P\left(\overline{T}_2|\overline{T}_1\right) P\left(\overline{T}_1\right) \\ &= 0.09 \times 0.47 + 0.36 \times 0.53 \\ &= 0.2331. \end{aligned}$$

Exercício (Caderno de Exercícios)

Suponha um teste de escolha múltipla com cinco respostas possíveis para cada questão. Um aluno sabe a resposta com probabilidade $\frac{1}{3}$. Sabendo a resposta, responde corretamente com probabilidade 0.99; se não sabe, responde corretamente com probabilidade $\frac{1}{5}$. Determine a probabilidade de o aluno saber a resposta à pergunta dado que respondeu corretamente.

Resolução

Sejam os acontecimentos:
$S -$ o aluno sabe a resposta e $R -$ o aluno responde corretamente.
Sabe-se que $P(S) = \frac{1}{3}$, $P(R|S) = 0.99$ e $P\left(R|\overline{S}\right) = \frac{1}{5}$. Pretende-se $P(S|R)$. Assim, pelo Teorema de Bayes,

$$\begin{aligned} P(S|R) &= \frac{P(S \cap R)}{P(R)} = \frac{P(S)P(R|S)}{P(S)P(R|S) + P\left(\overline{S}\right) P\left(R|\overline{S}\right)} \\ &= \frac{\frac{1}{3} \times 0.99}{\frac{1}{3} \times 0.99 + \frac{2}{3} \times \frac{1}{5}} \\ &\approx 0.71. \end{aligned}$$

Acontecimentos independentes

Os **acontecimentos** A e B, do mesmo espaço de resultados, são **independentes** sse

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

com $P(A) \geq 0$ e $P(B) \geq 0$. Como

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

se A e B são independentes, então

$$P(A|B) = P(A), \text{ com } P(B) > 0,$$
$$P(B|A) = P(B), \text{ com } P(A) > 0.$$

Note que se A e B são independentes, então os acontecimentos $\overline{A}$ e B, A e $\overline{B}$ ou $\overline{A}$ e $\overline{B}$ também são independentes.

Independência e Disjunção

Note que **independência** e **disjunção** não são o mesmo conceito!

A independência é uma propriedade da medida de probabilidade,

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

enquanto que a disjunção (ou incompatibilidade) é uma propriedade dos acontecimentos/conjuntos,

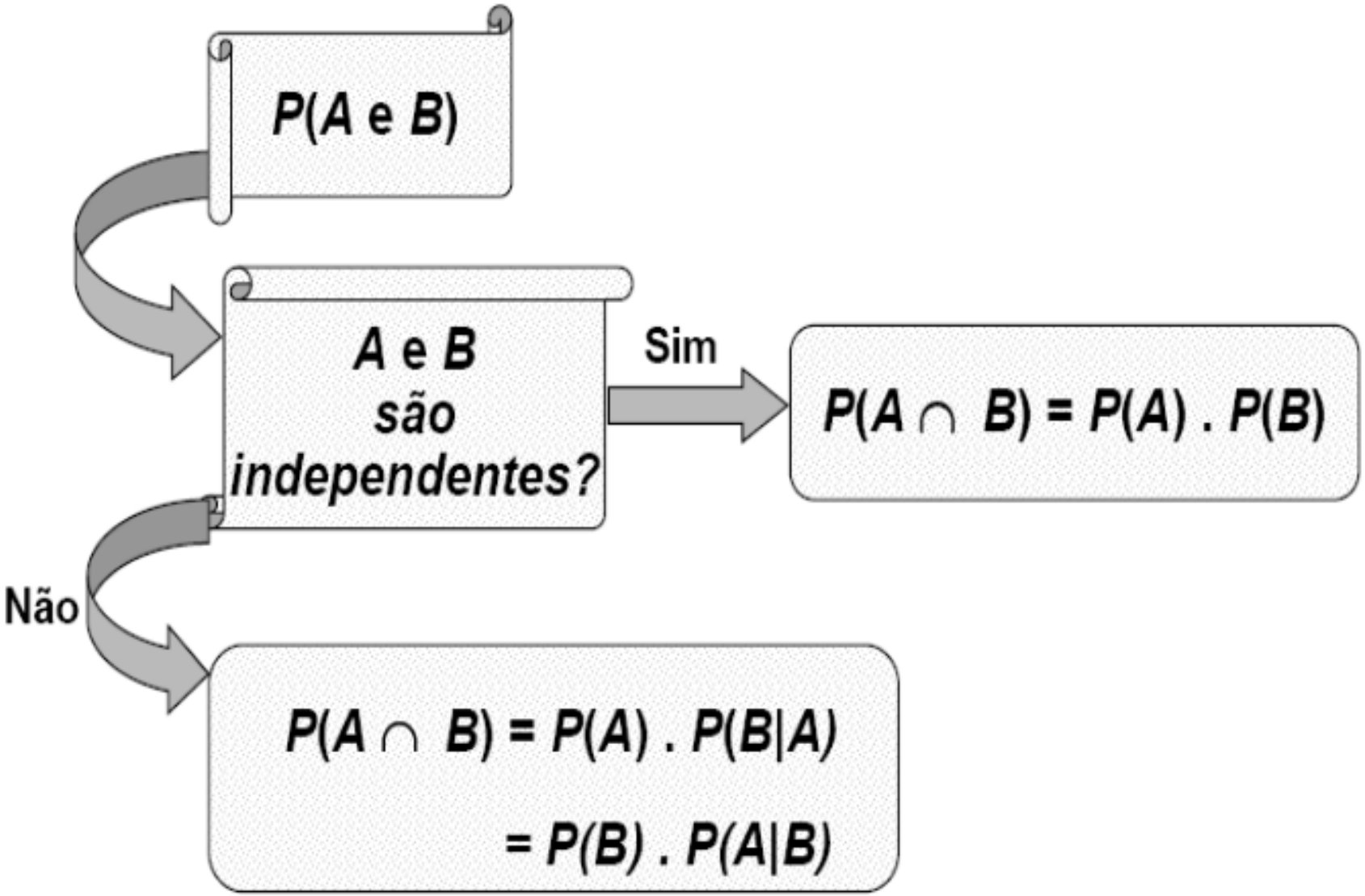
$$A \text{ e } B \text{ são disjuntos sse } A \cap B = \emptyset.$$

Por exemplo, admita que A e B são acontecimentos independentes e disjuntos (ou incompatíveis ou mutuamente exclusivos). Então,

$$P(A \cap B) = P(A)P(B) = P(\emptyset) = 0,$$

pelo que, $P(A) = 0$ ou $P(B) = 0$. Logo, com $P(A) > 0$ e $P(B) > 0$, se os acontecimentos são disjuntos, não podem ser independentes!

Intersecção de acontecimentos



Exemplo

Considere a experiência aleatória que consiste em tirar ao acaso uma carta de uma baralho de 52 cartas. Seja A o acontecimento "saída de copa" e B o acontecimento "saída de figura".

Os acontecimentos A e B são independentes, pois

$$P(A \cap B) = P(B|A)P(A) = \frac{3}{13} \times \frac{13}{52} = \frac{3}{52}$$

e

$$P(A)P(B) = \frac{13}{52} \times \frac{12}{52} = \frac{3}{52}$$

pelo que

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

Nota: facilmente se constata que $P(B|A) = P(B)$.

Exercício (Caderno de Exercícios)

Uma fábrica produz discos de plástico da qual se conhecem as probabilidades de um disco ter alta resistência a choque e a ferimentos, de acordo com a seguinte tabela:

	B	$\overline{B}$
A	0.80	0.09
$\overline{A}$	0.06	0.05

Sejam A o acontecimento "um disco tem alta resistência a choque" e B o acontecimento "um disco tem alta resistência a ferimentos". Nestas condições,

(a) encontre $P(A)$, $P(B)$, $P(A \cup B)$, $P(A \cup \overline{B})$ e $P(A|B)$.
(b) os acontecimentos A e B são independentes?
(c) os acontecimentos A e B são mutuamente exclusivos?

Resolução do exercício

	B	$\overline{B}$	
A	0.80	0.09	0.89
$\overline{A}$	0.06	0.05	0.11
	0.86	0.14	1.00

(a) Resolvida anteriormente. Obtém-se $P(A \cup B) = 0.95$, $P(A \cap B) = 0.80$, $P(A) = 0.89$, $P(B) = 0.86$ e $P(A|B) \approx 0.93$.

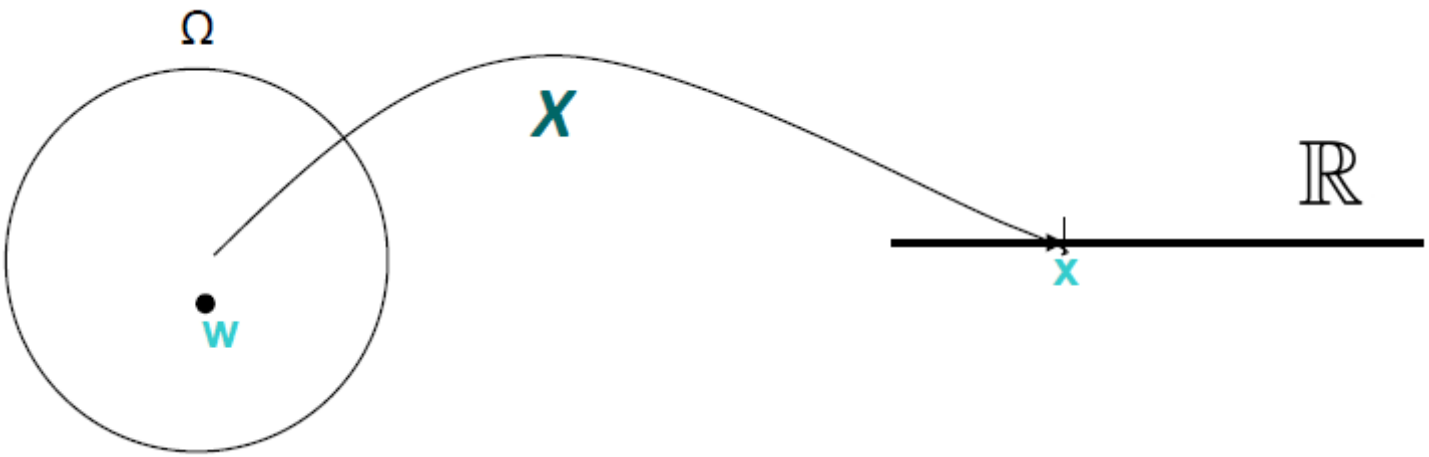
(b) Não são acontecimentos independentes, pois $P(A \cap B) \neq P(A)P(B)$, ou, analogamente, $P(A|B) \neq P(A)$.

(c) Os acontecimentos não são mutuamente exclusivos, pois $P(A \cap B) = 0.80 > 0$, pelo que $A \cap B \neq \emptyset$. Ou, analogamente, $P(A \cup B) \neq P(A) + P(B)$, com $P(A) + P(B) > 1$ (absurdo).

Noção de Variável Aleatória

Uma variável aleatória (v.a.) X é uma função com domínio Ω e com contradomínio em $\mathbb{R}$, a qual associa um número real x a cada acontecimento elementar w de Ω, de tal forma que a probabilidade de um intervalo real é igual à probabilidade dos acontecimentos que lhe deram origem em Ω.

$$X : w \in \Omega \to X(w) = x \in \mathbb{R}$$



Variáveis Aleatórias

Uma variável aleatória diz-se

- **discreta** se só assume um número finito ou infinito numerável de valores.
  Exemplos:
  - nº de erros ortográficos por página;
  - nº de automóveis que entram numa estrada por dia;
  - nº de avarias de uma máquina.
- **contínua** se assume um número infinito não numerável de valores (intervalos reais).
  Exemplos:
  - duração de uma chamada no telemóvel;
  - peso de um saco de sementes;
  - tempo de duração de um equipamento electrónico.

Variáveis Aleatórias Discretas

A distribuição de probabilidade de uma v.a. discreta fica completamente caracterizada por qualquer uma das seguintes funções:

- **Função (massa) de probabilidade (f.m.p.):**

$$f(x) = P(X = x),$$

  para todo o x possível, isto é, para todos os valores possíveis que a v.a. X possa assumir.
  (Note-se a diferença entre X e x.)

- **Função de distribuição (f.d.):**

$$F(x) = P(X \leq x) = \sum_{x_i \leq x} f(x_i).$$

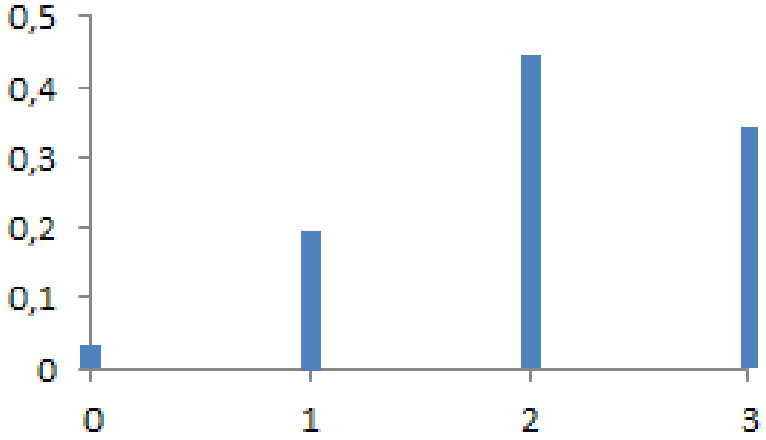
  (Representa a probabilidade acumulada até x.)

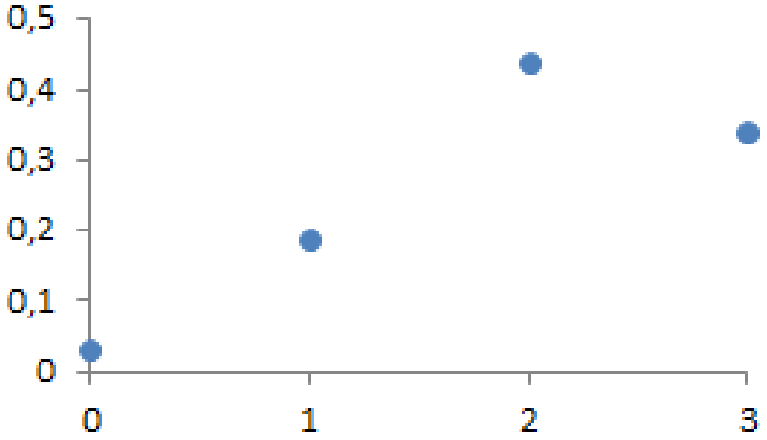
Variáveis Aleatórias Discretas: propriedades da f.m.p.

- $P(X = x) \geq 0$;
- $\sum_x P(X = x) = 1$.

A f.m.p. é comummente representada em tabelas, diagramas de barras ou, simplesmente, através de pontos no plano.

x	0	1	2	3
$f(x)=P(X=x)$	0,03	0,19	0,44	0,34

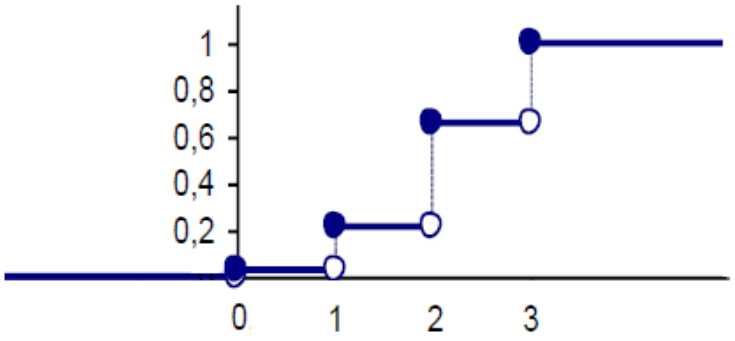




Variáveis Aleatórias Discretas: propriedades da f.d.

- $0 \leq F(X) \leq 1$, tal que $\lim_{x \to -\infty} F(x) = 0$ e $\lim_{x \to +\infty} F(x) = 1$;
- F é uma função contínua à direita e não decrescente;
- $P(X = x) = F(x^+) - F(x^-)$ e $P(a < X \leq b) = F(b) - F(a)$.

$$F(X) = \begin{cases} 0 & x < 0 \\ 0{,}03 & 0 \leq x < 1 \\ 0{,}22 & 1 \leq x < 2 \\ 0{,}66 & 2 \leq x < 3 \\ 1 & x \geq 3 \end{cases}$$



Variáveis Aleatórias Discretas: parâmetros

- Esperança, valor esperado ou média:

$$\mu = E[X] = \sum_x x\, P(X = x).$$

- Variância:

$$\sigma^2 = Var[X] = \sum_x (x - \mu)^2\, P(X = x).$$

- Desvio Padrão:

$$\sigma = \sqrt{Var[X]}.$$

Variáveis Aleatórias Discretas: parâmetros

- Quantil de ordem p $(0 < p < 1)$: é o valor x_p tal que

$$P(X \leq x_p) \geq p \;\text{ e }\; P(X \geq x_p) \geq 1 - p.$$

O quantil de ordem p é calculado determinando o valor x_p que satisfaz a dupla desigualdade

$$p \leq F(x_p) \leq p + P(X = x_p).$$

Exemplos de quantis:

- mediana (quantil de ordem 0.5);
- quartis (quantis de ordem 0.25, 0.50 e 0.75);
- decis;
- percentis.

Variáveis Aleatórias Discretas: propriedades da esperança

Quaisquer que sejam as v.a.'s X e Y, e as constantes reais a e b,

- $E[a] = a$;
- $E[aX + b] = aE[X] + b$;
- $E[X + Y] = E[X] + E[Y]$. Generalizando,

$$E\left[\sum_i X_i\right] = \sum_i E[X_i], \text{ quaisquer que sejam as v.a.'s } X_i.$$

Variáveis Aleatórias Discretas: propriedades da variância

Quaisquer que sejam as v.a.'s X e Y, e as constantes reais a e b,

- $Var[a] = 0$;
- $Var[aX + b] = a^2 Var[X]$;
- $Var[X + Y] = Var[X] + Var[Y]$, desde que X e Y sejam **independentes**. Generalizando,

  $$Var\left[\sum_i X_i\right] = \sum_i Var[X_i], \text{ para v.a.'s } X_i \textbf{ independentes};$$
- $Var[X] = E\left[X^2\right] - (E[X])^2$.

Exemplo

Considere a função de distribuição

$$F(x) = \begin{cases} 0, & x < -1 \\ 0.2, & -1 \leq x < 0 \\ a, & 0 \leq x < 1 \\ 0.6, & 1 \leq x < 3 \\ b, & x \geq 3 \end{cases}$$

Sabendo que $\mathrm{E}[X] = 1.3$, calcule $\mathrm{Var}[3 + 2X]$.

Exemplo - Resolução

Função massa de probabilidade:[1]

	-1	0	1	3
$f(x) = P(X = x)$	0.2	$f(0)$	$f(1)$	0.4

Como

$$\mathrm{E}[X] = -1 \times 0.2 + 0 \times f(0) + 1 \times f(1) + 3 \times 0.4 = 1.3,$$

então

$$f(1) = 1.3 + 0.2 - 1.2 = 0.3$$

e, consequentemente,

$$f(0) = 1 - (0.2 + 0.3 + 0.4) = 0.1.$$

Logo, uma vez que

$$\mathrm{Var}[X] = (-1 - 1.3)^2 \times 0.2 + (0 - 1.3)^2 \times 0.1 + (1 - 1.3)^2 \times 0.3 + \\ + (3 - 1.3)^2 \times 0.4 = 2.41,$$

então

$$\mathrm{Var}[3 + 2X] = 4 \times \mathrm{Var}[X] = 4 \times 2.41 = 9.64.$$

[1] Note que $b = 1$, $a \in [0.2, 0.6]$, $f(0) = a - 0.2$ e $f(1) = 0.6 - a$.

Distribuições Discretas: Binomial

Se

$X =$ "nº de *sucessos* em n provas[1] independentes"

diz-se que X segue uma distribuição Binomial de parâmetros n e p, e escreve-se $X \sim B(n, p)$, sendo p a probabilidade de sucesso em cada prova.

Nestas condições,

$$f(x) = P(X = x) = C_x^n \, p^x \, (1-p)^{n-x}, \; x = 0, 1, \ldots, n,$$

$$E[X] = np,$$

$$Var[X] = np(1-p).$$

[1] Estas são denominadas **provas de Bernoulli**. Consultar Murteira et al. (2015) para detalhes adicionais.

Distribuições Discretas: Binomial

Alguns exemplos:

- $X =$ "nº de caras em 10 lançamentos de uma moeda"
  $X \sim B(10, 0.5)$;
- $X =$ "nº de defeitos em 35 peças de roupa avaliadas"
  $X \sim B(35, p)$, sendo p a probabilidade de uma peça de roupa ser defeituosa;
- $X =$ "nº de praticantes de desporto no conjunto de 43 alunos presentes na sala de aula"
  $X \sim B(43, p)$, sendo p a probabilidade de um estudante praticar desporto;
- $X =$ "nº de erros de medição superiores a 0.01mg em 100 pesagens realizadas"
  $X \sim B(100, p)$, sendo p a probabilidade de se cometer um erro de medição superior a 0.01mg numa pesagem.

Distribuições Discretas: propriedade da Binomial

Sejam X e Y duas v.a.'s independentes tais que

$$X \sim B(n_1, p) \text{ e } Y \sim B(n_2, p).$$

Assim, nestas condições,

$$X + Y \sim B(n_1 + n_2, p).$$

Exemplo: admita que a probabilidade de uma peça ser defeituosa é 0.01. Seja
$X =$ "nº de peças defeituosas num lote A de 100 peças" e
$Y =$ "nº de peças defeituosas num lote B de 500 peças". Assim,
$X \sim B(100, 0.01)$ e $Y \sim B(500, 0.01)$. Logo,
$Z = X + Y =$ "nº de peças defeituosas no conjunto dos dois lotes"
segue uma distribuição $B(600, 0.01)$.

Distribuições Discretas: utilização de *software*

- SPSS (em *Transform* / *Compute Variable*)
  $f(x) = P(X = x) = \text{PDF.BINOM}(x, n, p)$
  $F(x) = P(X \leq x) = \text{CDF.BINOM}(x, n, p)$

- Excel[1] (em *Insert Function*)
  $f(x) = P(X = x) = \text{Binomdist}(x; n; p; \textit{false})$
  $F(x) = P(X \leq x) = \text{Binomdist}(x; n; p; \textit{true})$

- Máquina de calcular[2]
  $f(x) = P(X = x) = \text{pdfbinom}\ldots$
  $F(x) = P(X \leq x) = \text{cdfbinom}\ldots$

[1] Depende da versão instalada no seu computador.
[2] Depende da marca e versão da sua máquina de calcular.

Distribuições Discretas: exemplo com Binomial

Seja $X \sim B(10, 0.2)$. Assim, por exemplo, recorrendo ao SPSS,

- $P(X = 4) = \text{PDF.BINOM}(4, 10, 0.2) \approx 0.0881$;
- $P(X \geq 6) = 1 - P(X < 6) = 1 - P(X \leq 5) = 1 - F(5)$. Logo, $P(X \geq 6) = 1 - \text{CDF.BINOM}(5, 10, 0.2) \approx 0.0064$.

E, ainda, manualmente,

- $E[4X - 1] = 4E[X] - 1 = 4 \times (10 \times 0.2) - 1 = 7$;
- $Var[4X - 1] = 16Var[X] = 16 \times (10 \times 0.2 \times 0.8) = 25.6$.

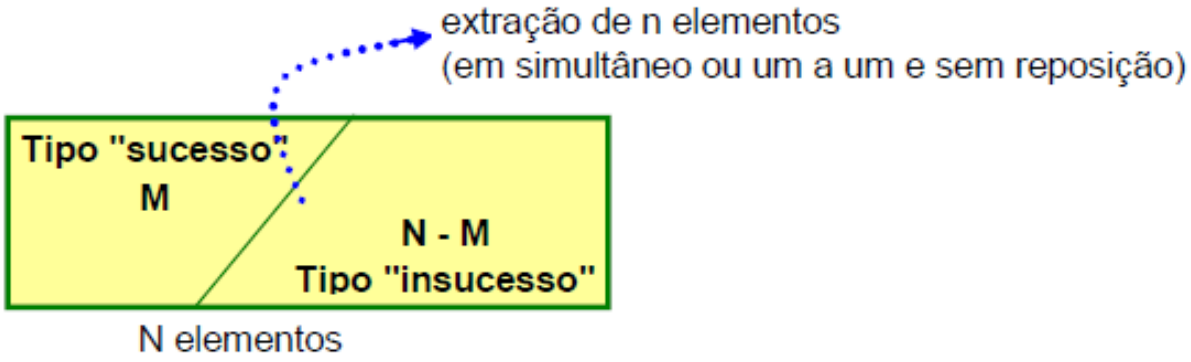
Distribuições Discretas: Hipergeométrica

Se

$X =$ "nº de *sucessos* em n provas dependentes num universo de N elementos com M elementos de tipo *sucesso*"

diz-se que X segue uma distribuição Hipergeométrica de parâmetros N, M e n, e escreve-se $X \sim H(N, M, n)$.

Nota: as provas são dependentes quando as extracções são sem reposição, uma vez que a probabilidade de *sucesso* não se mantém constante de prova para prova.



Distribuições Discretas: Hipergeométrica

Nestas condições,

$$f(x) = P(X = x) = \frac{C_x^M \, C_{n-x}^{N-M}}{C_n^N}, \; x = 0, 1, \ldots, n,$$

$$E[X] = np,$$

$$Var[X] = np(1-p)\frac{N-n}{N-1},$$

onde $p = \frac{M}{N}$.

Nota: quando $n << N$, a $H(N, M, n)$ aproxima-se da $B\left(n, \frac{M}{N}\right)$. Quando o número N de elementos do universo é muito grande, comparado com o número n de elementos que dele se extraem, extrair sem reposição (provas dependentes) ou com reposição (provas independentes) conduz a resultados semelhantes.

Distribuições Discretas: utilização de *software*

- SPSS (em *Transform* / *Compute Variable*)
  $f(x) = P(X = x) = \text{PDF.HYPER}(x, N, n, M)$
  $F(x) = P(X \leq x) = \text{CDF.HYPER}(x, N, n, M)$

- Excel[1] (em *Insert Function*)
  $f(x) = P(X = x) = \text{Hypgeomdist}(x; n; M; N; \textit{false})$
  $F(x) = P(X \leq x) = \text{Hypgeomdist}(x; n; M; N; \textit{true})$

- Máquina de calcular[2]
  $f(x) = P(X = x) = \text{pdfhyper}\ldots$
  $F(x) = P(X \leq x) = \text{cdfhyper}\ldots$

[1] Depende da versão instalada no seu computador.
[2] Depende da marca e versão da sua máquina de calcular. Algumas máquinas não possuem a Hipergeométrica.

Distribuições Discretas: Geométrica

Se

$X =$ "nº de provas independentes até obter o primeiro *sucesso*"

diz-se que X segue uma distribuição Geométrica de parâmetro p, e escreve-se $X \sim Geom(p)$ ou $X \sim G(p)$, sendo p a probabilidade de *sucesso* em cada prova.

Nestas condições,

$f(x) = P(X = x) = p\,(1-p)^{x-1},\ x = 1, 2, \ldots,$

$E[X] = \frac{1}{p},$

$Var[X] = \frac{1-p}{p^2}.$

Distribuições Discretas: Geométrica

Alguns exemplos:

- $X =$ "nº de lançamentos de uma moeda até obter a primeira cara"
  $X \sim Geom(0.5)$;

- $X =$ "nº de verificações até encontrar a primeira peça de roupa defeituosa"
  $X \sim Geom(p)$, sendo p a probabilidade de uma peça de roupa ser defeituosa;

- $X =$ "nº de estudantes a inquirir até obter o primeiro a afirmar ser fumador"
  $X \sim Geom(p)$, sendo p a probabilidade de um estudante ser fumador.

Distribuições Discretas: utilização de *software*

- SPSS (em *Transform* / *Compute Variable*)
  $f(x) = P(X = x) = \text{PDF.GEOM}(x, p)$
  $F(x) = P(X \leq x) = \text{CDF.GEOM}(x, p)$

- Excel[1] (em *Insert Function*)
  $f(x) = P(X = x) = \text{Negbinomdist}(x + 1; 1; p; \textit{false})$
  $F(x) = P(X \leq x) = \text{Negbinomdist}(x + 1; 1; p; \textit{true})$

- Máquina de calcular[2]
  $f(x) = P(X = x) = \text{pdfgeom}\ldots$
  $F(x) = P(X \leq x) = \text{cdfgeom}\ldots$

[1] Depende da versão instalada no seu computador. Note-se que é aqui utilizada a distribuição Binomial Negativa. A distribuição Geométrica é um caso particular ($r = 1$) da Binomial Negativa. Consultar Murteira et al. (2015) para detalhes adicionais.
[2] Depende da marca e versão da sua máquina de calcular.

Distribuições Discretas: Poisson

Se, em geral,

$X =$ "contagem do nº de *eventos* em intervalos de tempo ou regiões do espaço"

diz-se que X segue uma distribuição de Poisson[1] de parâmetro λ e escreve-se $X \sim P(\lambda)$.

Nestas condições,

$$f(x) = P(X = x) = e^{-\lambda} \frac{\lambda^x}{x!}, \; x = 0, 1, 2, \ldots, \; \lambda > 0,$$

$E[X] = \lambda$ e $Var[X] = \lambda$.

[1] Consultar Murteira et al. (2015) para detalhes adicionais, nomeadamente os pressupostos de um **processo de Poisson aproximado**.

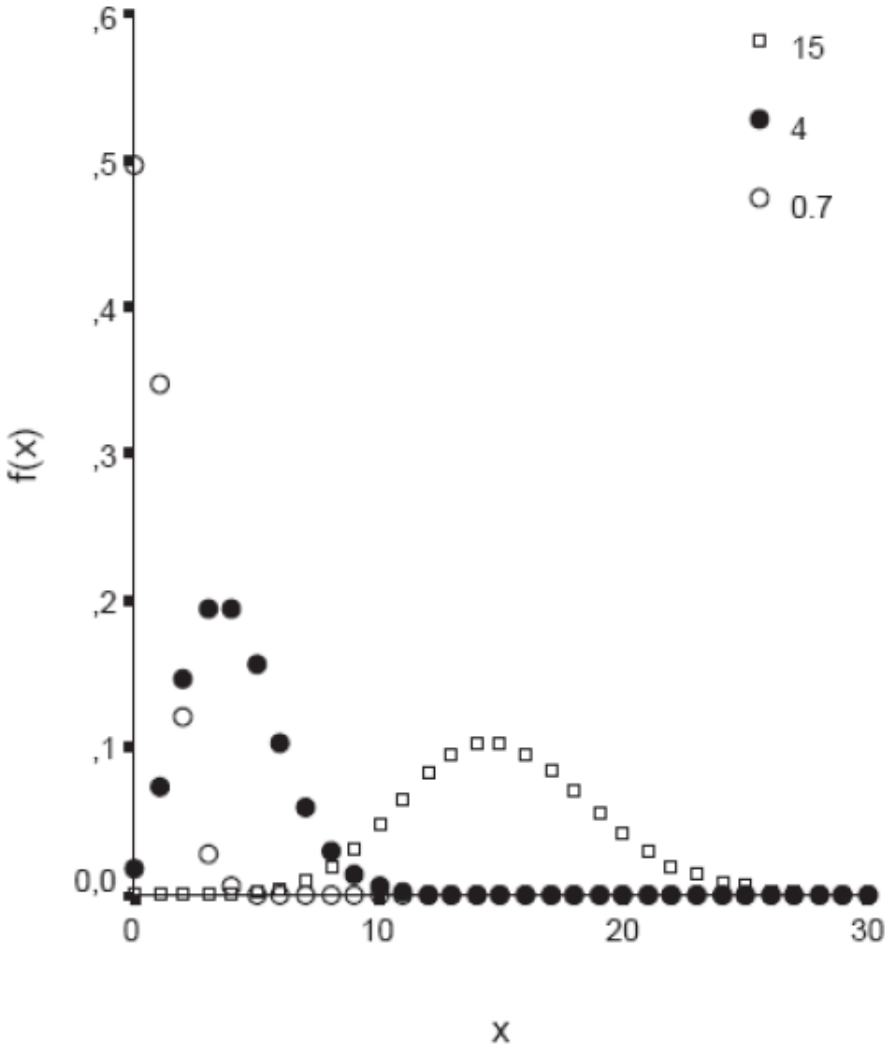
Distribuições Discretas: Poisson

Alguns exemplos:

- $X =$ "nº de carros que abastecem o depósito, numa determinada bomba de gasolina, por dia";
- $X =$ "nº de chamadas telefónicas realizadas diariamente numa determinada empresa";
- $X =$ "nº de bactérias existentes no teclado de um determinado computador";
- $X =$ "nº de avarias mensais de uma determinada máquina".

Distribuições Discretas: Poisson

Gráfico da f.m.p. para diferentes valores de λ:



Distribuições Discretas: propriedade da Poisson

A soma de duas v.a.'s de Poisson independentes é ainda uma v.a. de Poisson com parâmetro igual à soma dos respectivos parâmetros. Generalizando, se $X_1, X_2, \ldots, X_n$ são v.a.'s independentes com $X_i \sim P(\lambda_i), i = 1, 2, \ldots, n$, então

$$\sum_{i=1}^{n} X_i \sim P\left(\sum_{i=1}^{n} \lambda_i\right).$$

Exemplo: seja
$X =$ "nº de viaturas que abastecem numa determinada estação de serviço durante a manhã de um certo dia", com $X \sim P(5.1)$, e
$Y =$ "nº de viaturas que abastecem nessa mesma estação de serviço durante a tarde do mesmo dia", com $Y \sim P(8.4)$. Assim,
$Z = X + Y =$ "nº de viaturas que, no dia inteiro, abastecem nesta estação de serviço" segue uma distribuição $P(13.5)$.

Distribuições Discretas: utilização de *software*

- SPSS (em *Transform* / *Compute Variable*)
  $f(x) = P(X = x) = \text{PDF.POISSON}(x, \lambda)$
  $F(x) = P(X \leq x) = \text{CDF.POISSON}(x, \lambda)$

- Excel[1] (em *Insert Function*)
  $f(x) = P(X = x) = \text{Poisson}(x; \lambda; \textit{false})$
  $F(x) = P(X \leq x) = \text{Poisson}(x; \lambda; \textit{true})$

- Máquina de calcular[2]
  $f(x) = P(X = x) = \text{pdfpoisson}\ldots$
  $F(x) = P(X \leq x) = \text{cdfpoisson}\ldots$

[1] Depende da versão instalada no seu computador.
[2] Depende da marca e versão da sua máquina de calcular.

Variáveis Aleatórias Contínuas

A distribuição de probabilidade de uma v.a. contínua fica completamente caracterizada por qualquer uma das seguintes funções:

- **Função densidade de probabilidade (f.d.p.):**
  $f(x)$ definida para todo o x real.

  (Note-se que a f.d.p. não representa $P(X = x)$ como definido no caso discreto.)

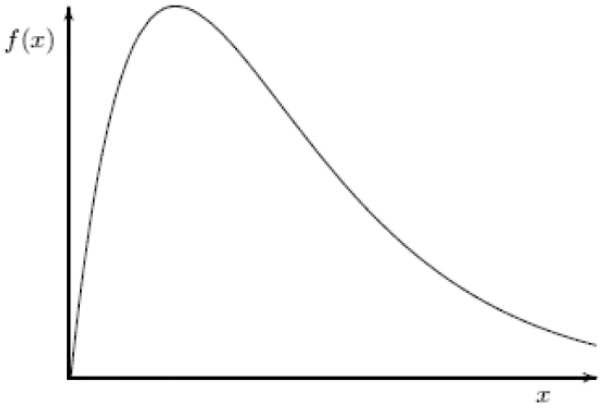
- **Função de distribuição (f.d.):**
  $F(x) = P(X \leq x) = \int_{-\infty}^{x} f(t)\,dt$, para todo o x real.

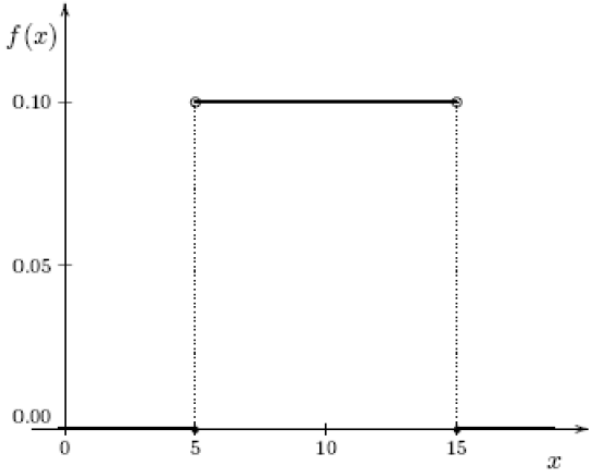
  (Note-se que, à semelhança do caso discreto, $F(x)$ representa a probabilidade acumulada até x.)

Variáveis Aleatórias Contínuas: propriedades da f.d.p.

- $f(x) \geq 0$;
- $\displaystyle\int_{-\infty}^{+\infty} f(x)\,dx = 1$,
  ou seja, a área total sob a curva da f.d.p é unitária.

O gráfico da f.d.p., ou curva de densidade, traduz a distribuição de probabilidade de uma v.a. contínua. Exemplos:

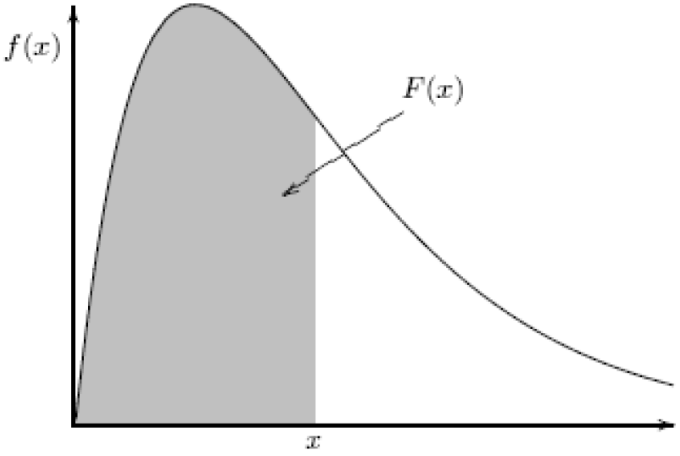




Variáveis Aleatórias Contínuas: propriedades da f.d.

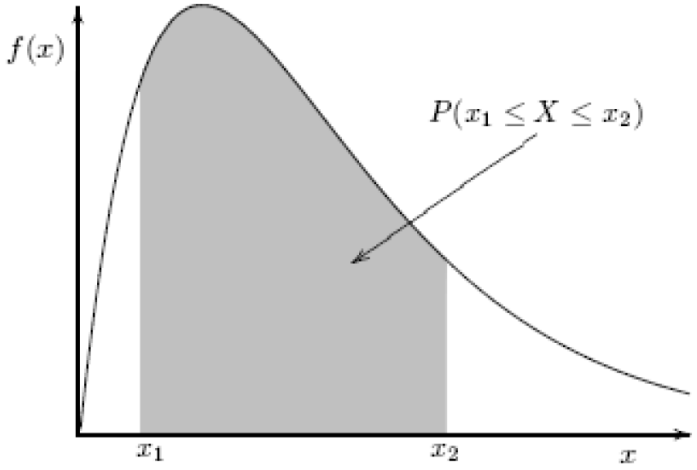
- $0 \leq F(X) \leq 1$, tal que $\lim_{x \to -\infty} F(x) = 0$ e $\lim_{x \to +\infty} F(x) = 1$;
- F é uma função **contínua** e não decrescente;
- $P(a < X \leq b) = F(b) - F(a)$;
- $F'(x) = f(x)$.

Geometricamente, um exemplo:



VA Contínuas: cálculo de probabilidades

As probabilidades obtêm-se a partir de áreas sob partes da curva de densidade. Por exemplo,



$$P(x_1 \leq X \leq x_2) = \int_{x_1}^{x_2} f(x)\,dx = F(x_2) - F(x_1).$$

Naturalmente, $P(X = x) = 0$, para todo o x real. Consequentemente, em distribuições contínuas,

- $P(X \leq a) = P(X < a)$;
- $P(a \leq X \leq b) = P(a < X \leq b) = P(a \leq X < b) = P(a < X < b)$;
- $P(X > a) = P(X \geq a)$.

Variáveis Aleatórias Contínuas: parâmetros

- Esperança, valor esperado ou média:

$$\mu = E[X] = \int_{-\infty}^{+\infty} x\, f(x)\, dx.$$

- Variância:

$$\sigma^2 = Var[X] = \int_{-\infty}^{+\infty} (x-\mu)^2\, f(x)\, dx.$$

- Desvio Padrão:

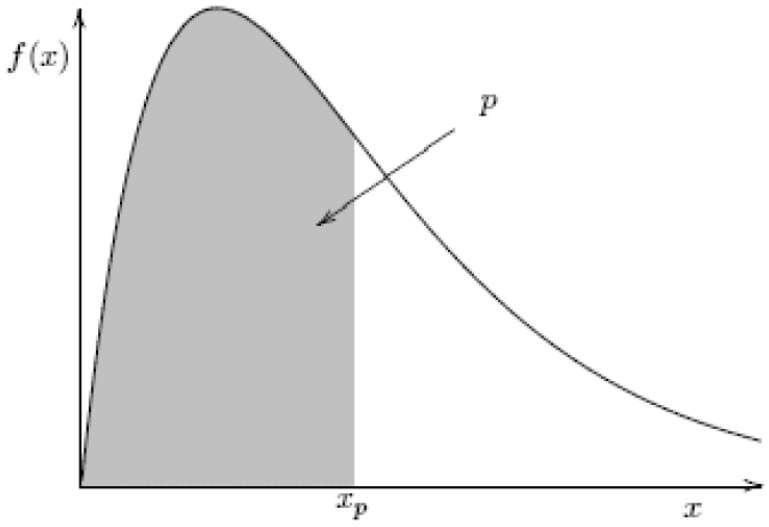
$$\sigma = \sqrt{Var[X]}.$$

Variáveis Aleatórias contínuas: parâmetros

- Quantil de ordem p $(0 < p < 1)$: é o valor x_p tal que

$$P(X \leq x_p) \geq p \;\text{ e }\; P(X \geq x_p) \geq 1 - p.$$

O quantil de ordem p de uma v.a. contínua é calculado determinando o valor x_p que satisfaz a igualdade $F(x_p) = p$.



Variáveis Aleatórias Contínuas: propriedades da esperança e da variância

Quaisquer que sejam as v.a.'s X e Y, e as constantes reais a, b e c, tal como no caso discreto, resumidamente,

- $E[aX + bY + c] = aE[X] + bE[Y] + c$;
- $Var[aX + b] = a^2 Var[X]$;
- $Var[X + Y] = Var[X] + Var[Y]$, desde que X e Y sejam **independentes**. Generalizando,

  $$Var\left[\sum_i X_i\right] = \sum_i Var[X_i], \text{ para v.a.'s } X_i \textbf{ independentes};$$
- $Var[X] = E\left[X^2\right] - (E[X])^2$.

Variáveis Aleatórias Contínuas: exemplo

Estudos recentes na área da Biotecnologia conduziram à produção de uma classe de milho mais tolerante a herbicidas através da modificação de um gene. Considere que, em condições climáticas propícias e estáveis, o tempo (em dias) necessário para o amadurecimento dessa nova classe de milho geneticamente modificado pode ser modelado por uma v.a. X contínua, através de uma distribuição Exponencial, cuja f.d. é dada por

$$F(x) = 1 - e^{-\frac{x}{90}}, \; x > 0.$$

(a) Sabendo que o milho foi semeado há mais de 80 dias, qual a probabilidade de que este não amadureça nos próximos 15 dias (admitindo condições climáticas propícias e estáveis)?

(b) Calcule o 1º quartil da distribuição de X e interprete o seu significado no contexto do problema.

(c) Indique, justificando, o valor de $P(X = 90 \mid X \geq 80)$.

(d) Encontre a função densidade de probabilidade de X.

Variáveis Aleatórias Contínuas: exemplo

(a) Seja,
$X =$ "tempo (em dias) necessário ao amadurecimento do milho geneticamente modificado." Logo,

$$\begin{aligned}
P(X > 80 + 15 \mid X > 80) &= \frac{P(X > 95 \cap X > 80)}{P(X > 80)} \\
&= \frac{P(X > 95)}{P(X > 80)} \\
&= \frac{1 - P(X \leq 95)}{1 - P(X \leq 80)} = \frac{1 - F(95)}{1 - F(80)} \\
&= \frac{1 - \left(1 - e^{-\frac{95}{90}}\right)}{1 - \left(1 - e^{-\frac{80}{90}}\right)} \\
&= e^{-\frac{15}{90}} \approx 0.85.
\end{aligned}$$

Nota: esta alínea pode ser resolvida através de uma propriedade da Exponencial denominada **perda de memória**. Consultar Murteira et al. (2015) ou Hall et al. (2011) para detalhes adicionais.

Variáveis Aleatórias Contínuas: exemplo

(b) Basta resolver $F(x_{0.25}) = 0.25$. Assim,

$$\begin{aligned} F(x_{0.25}) = 0.25 \quad &\Leftrightarrow \quad 1 - e^{-\frac{x_{0.25}}{90}} = 0.25 \\ &\Leftrightarrow \quad e^{-\frac{x_{0.25}}{90}} = 0.75 \\ &\Leftrightarrow \quad \ln\left(e^{-\frac{x_{0.25}}{90}}\right) = \ln(0.75) \\ &\Leftrightarrow \quad -\frac{x_{0.25}}{90} = \ln(0.75) \\ &\Leftrightarrow \quad x_{0.25} = -90\,\ln(0.75) \approx 25.89\,\mathrm{dias}. \end{aligned}$$

Interpretação:

- pelo menos 25% do milho amadurece em menos de 26 dias; ou,
- como $P(X \leq 25.89) \approx 0.25$, pode dizer-se que a probabilidade do tempo (em dias) necessário...

Nota: no SPSS basta executar IDF.EXP(0.25, 1/90).

Variáveis Aleatórias Contínuas: exemplo

(c) $P(X = 90 \mid X \geq 80) = 0$, pois X é uma v.a. contínua e, devido à continuidade da f.d., $P(X = 90) = 0$.

(d) Sabe-se que $F'(x) = f(x)$. Logo,

$$f(x) = \left(1 - e^{-\frac{x}{90}}\right)' = \frac{1}{90} e^{-\frac{x}{90}},\ x > 0.$$

Distribuições Contínuas: Uniforme

Diz-se que X segue uma distribuição Uniforme no intervalo $[a, b]$, e escreve-se $X \sim U[a, b]$, se a f.d.p. de X é constante nesse intervalo.[1]

Nestas condições,

$$f(x) = \begin{cases} \dfrac{1}{b-a}, & x \in [a, b] \\ 0, & x \notin [a, b] \end{cases},$$

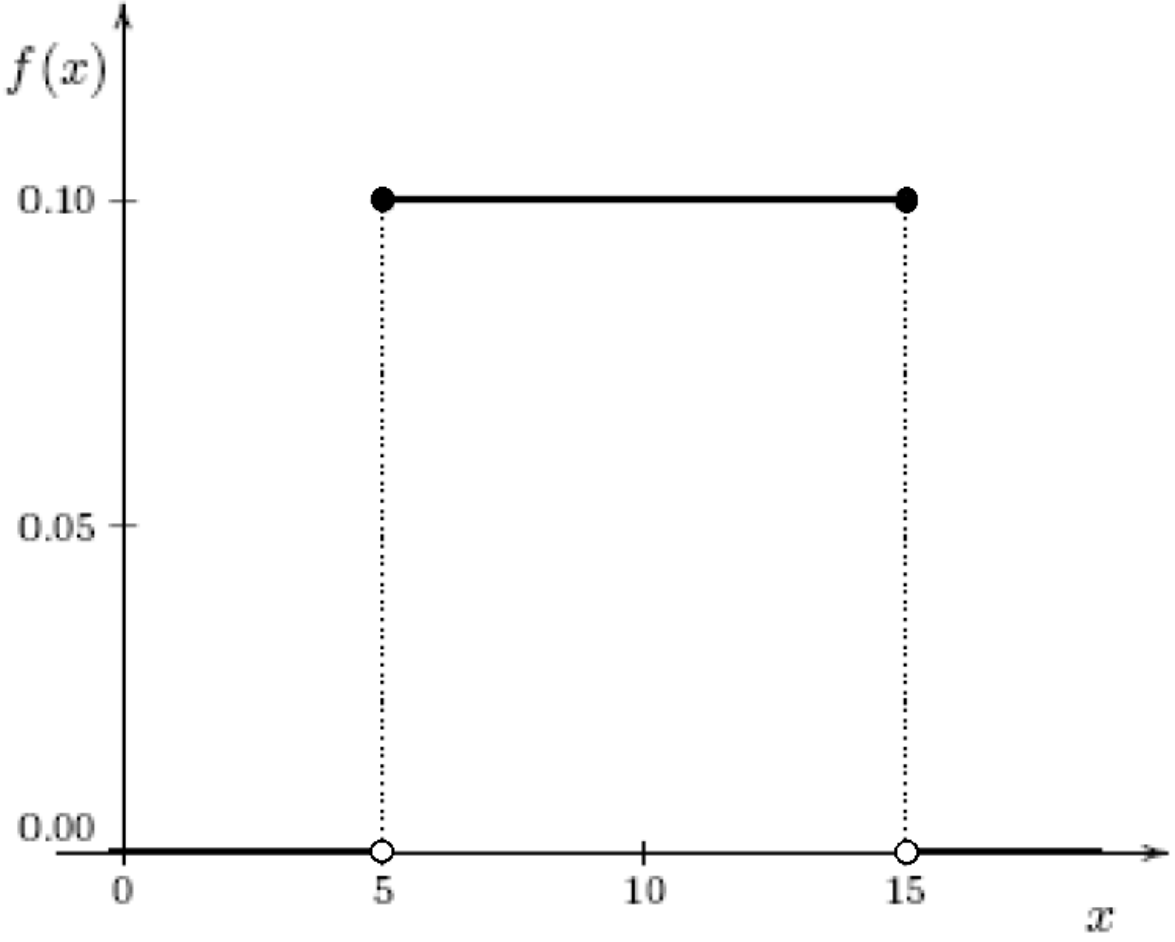
$$E[X] = \frac{a+b}{2},$$

$$Var[X] = \frac{(b-a)^2}{12}.$$

[1] Naturalmente, pode definir-se X num intervalo (a, b) e escrever-se $X \sim U(a, b)$.

Distribuições Contínuas: Uniforme

Gráfico da f.d.p. de uma $U[5, 15]$:



Distribuições Contínuas: Exponencial

Diz-se que X segue uma distribuição Exponencial, e escreve-se $X \sim Exp(\lambda)$, se a f.d.p. de X é definida por

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

Nestas condições,
$E[X] = \dfrac{1}{\lambda}$ e $Var[X] = \dfrac{1}{\lambda^2}$.

Esta distribuição é comummente usada para modelar v.a.'s que representam tempos de duração (e.g., tempos de vida de equipamentos electrónicos), tempos entre ocorrências consecutivas de um processo de Poisson, etc.[1]

[1] Consultar Murteira et al. (2015) ou Hall et al. (2011) para detalhes adicionais e exemplos de aplicação.

Distribuições Contínuas: Normal ou Gaussiana

Diz-se que X segue uma distribuição Normal de média μ e variância σ^2, e escreve-se[1] $X \sim N(\mu, \sigma^2)$, se a f.d.p. de X é definida por

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}}\, e^{-\dfrac{(x-\mu)^2}{2\sigma^2}}, \; x \in \mathbb{R}, \mu \in \mathbb{R}, \sigma > 0.$$

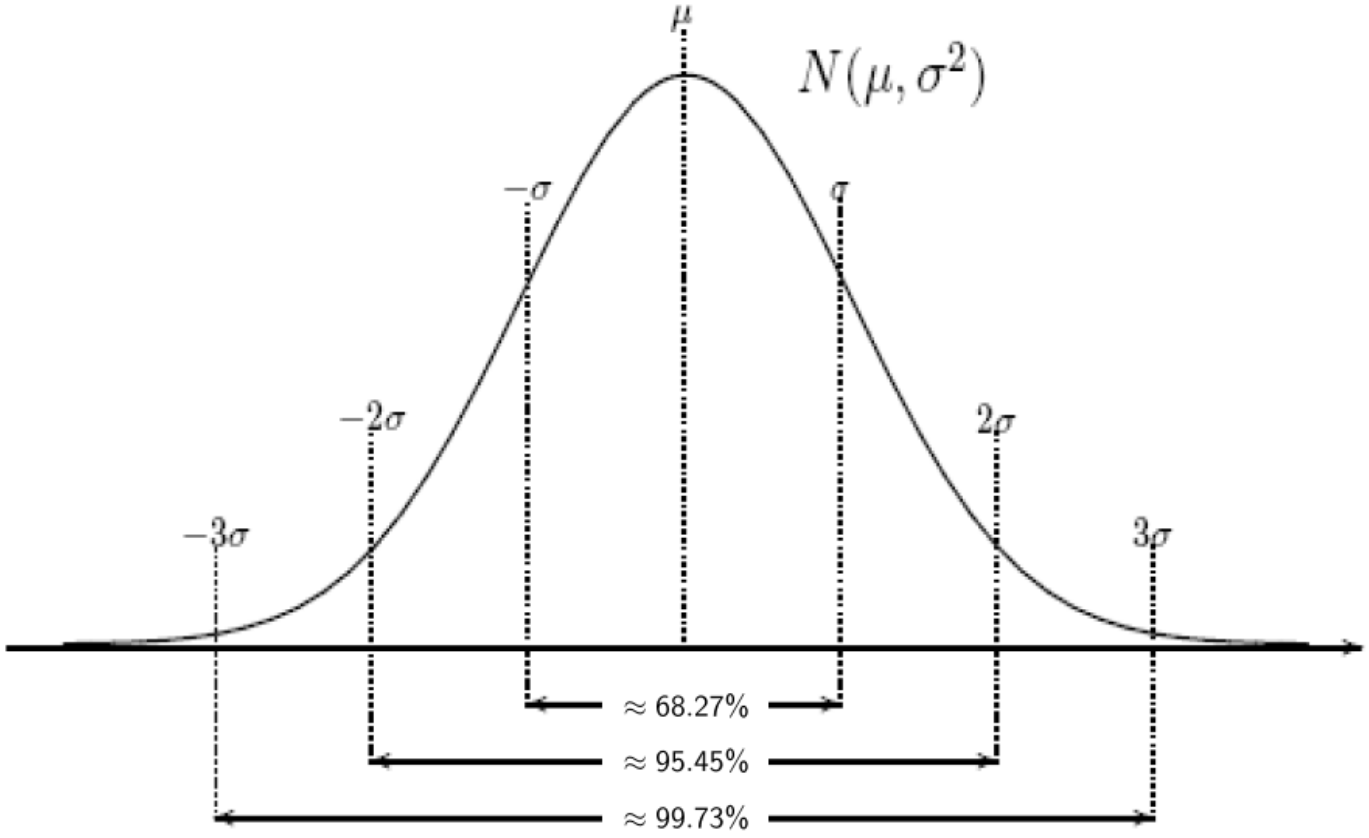
Nestas condições, $E[X] = \mu$ e $Var[X] = \sigma^2$.

A distribuição Normal é importantíssima em estatística, quer do ponto de vista teórico, quer do ponto de vista prático, como doravante se observará.

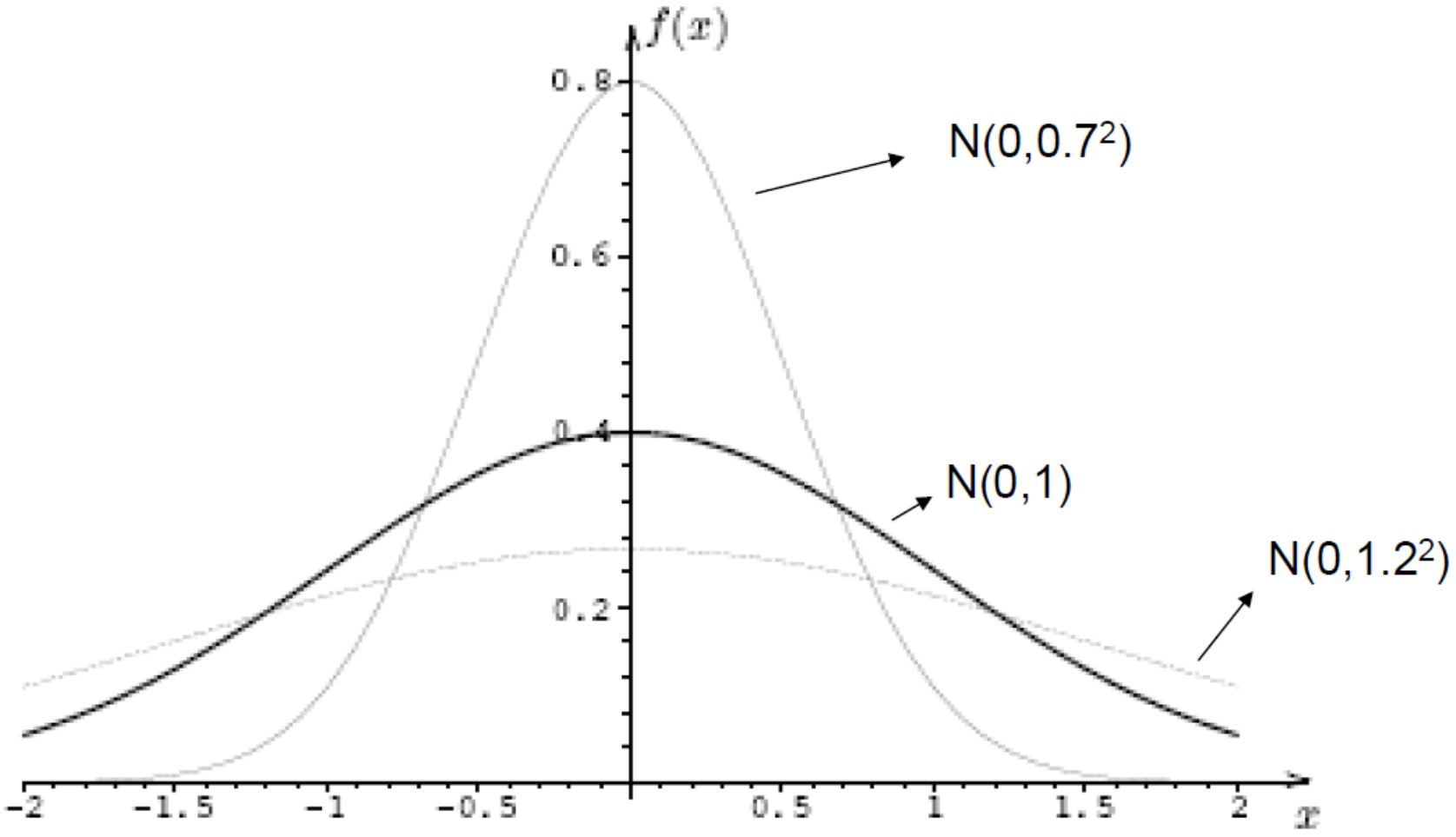
[1] Alguns autores utilizam a notação $N(\mu, \sigma)$, assinalando o valor do desvio padrão, σ, em detrimento do valor da variância, σ^2. A maioria do *software* solicita o desvio padrão.

Distribuição Normal: gráfico da f.d.p.

Chama-se curva de Gauss à curva da f.d.p. da $N(\mu, \sigma^2)$. Trata-se de uma curva simétrica em torno da média μ e em forma de sino.



Distribuição Normal: significado do parâmetro σ^2



Distribuição Normal: propriedades

Se $X \sim N(\mu, \sigma^2)$, então

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

  A distribuição da v.a. Z é conhecida como
  - normal *standard*;
  - normal padrão;
  - normal "centrada e reduzida";
  - normal "zero um".

  A sua f.d. é habitualmente representada por

$$\Phi(z) = F(z) = P(Z \leq z),\ z \in \mathbb{R}.$$

- $aX + b \sim N(a\mu + b, a^2\sigma^2)$, com a e b constantes reais.

Distribuição Normal: propriedades

Se $X_i \sim N(\mu_i, \sigma_i^2)$, $i = 1, 2, \ldots, n$, são v.a.'s **independentes**,

$$X_1 + X_2 + \ldots + X_n \sim N(\underbrace{\mu_1 + \mu_2 + \ldots + \mu_n}_{E[X_1 + X_2 + \ldots + X_n]}, \underbrace{\sigma_1^2 + \sigma_2^2 + \ldots + \sigma_n^2}_{Var[X_1 + X_2 + \ldots + X_n]}),$$

ou seja,

$$\sum_{i=1}^{n} X_i \sim N\left(\sum_{i=1}^{n} \mu_i, \sum_{i=1}^{n} \sigma_i^2\right).$$

Em particular, se $\mu_i = \mu$ e $\sigma_i^2 = \sigma^2$, então

$$X_1 + X_2 + \ldots + X_n = \sum_{i=1}^{n} X_i \sim N(n\mu, n\sigma^2)$$

e

$$\overline{X} = \frac{X_1 + X_2 + \ldots + X_n}{n} = \frac{1}{n}\sum_{i=1}^{n} X_i \sim N\left(\mu, \frac{\sigma^2}{n}\right).$$

Distribuição Normal: algumas demonstrações

$$E\left[\sum_{i=1}^{n} X_i\right] = \sum_{i=1}^{n} E[X_i] = \sum_{i=1}^{n} \mu = n\mu$$

$$Var\left[\sum_{i=1}^{n} X_i\right] = \sum_{i=1}^{n} Var[X_i] = \sum_{i=1}^{n} \sigma^2 = n\sigma^2$$

(v.a.'s independentes)

$$E[\overline{X}] = E\left[\frac{1}{n}\sum_{i=1}^{n} X_i\right] = \frac{1}{n} E\left[\sum_{i=1}^{n} X_i\right] = \frac{1}{n} n\mu = \mu$$

$$Var[\overline{X}] = Var\left[\frac{1}{n}\sum_{i=1}^{n} X_i\right] = \frac{1}{n^2} Var\left[\sum_{i=1}^{n} X_i\right] = \frac{1}{n^2} n\sigma^2 = \frac{\sigma^2}{n}$$

Distribuição Normal: cálculo de probabilidades e quantis

Para além da consulta de tabelas da $N(0,1)$, o cálculo de probabilidades ou quantis é facilitado com a utilização de *software*:

- SPSS (em *Transform* / *Compute Variable*)
  $F(x) = P(X \leq x) = \text{CDF.NORMAL}(x, \mu, \sigma)$
  Quantil de ordem p, valor x_p: $\text{IDF.NORMAL}(p, \mu, \sigma)$

- Excel[1] (em *Insert Function*)
  $F(x) = P(X \leq x) = \text{Normdist}(x; \mu; \sigma; \textit{true})$
  Quantil de ordem p, valor x_p: $\text{Norminv}(p, \mu, \sigma)$

- Máquina de calcular[2]
  $F(x) = P(X \leq x) = \text{cdfnormal}\ldots$
  Quantil de ordem p, valor x_p: invnormal...

[1] Depende da versão instalada no seu computador.
[2] Depende da marca e versão da sua máquina de calcular.

Distribuição Normal: cálculo de probabilidades e quantis

Algumas máquinas de calcular exigem um valor mínimo (m) e um valor máximo (M) no cálculo de probabilidades, na forma

$$\mathrm{cdfnormal}(m, M, \mu, \sigma).$$

Assim, por exemplo, para calcular

- $F(b) = P(X \leq b)$, considera-se $m = -\infty$ e $M = b$;
- $P(X > a) = P(X \geq a)$, considera-se $m = a$ e $M = +\infty$.

Para as máquinas que não permitem a utilização de $-\infty$ e $+\infty$, devem utilizar-se valores negativos/positivos, suficientemente elevados, em sua substituição.

Distribuição Normal: cálculo de probabilidades e quantis

Exemplo 1: sabendo que $X \sim N(5, 4)$, calcular $P(X \leq 7)$.

- SPSS (em *Transform* / *Compute Variable*)
  $P(X \leq 7) = F(7) = \text{CDF.NORMAL}(7, 5, 2)$

- Excel[1] (em *Insert Function*)
  $P(X \leq 7) = F(7) = \text{Normdist}(7; 5; 2; \textit{true})$

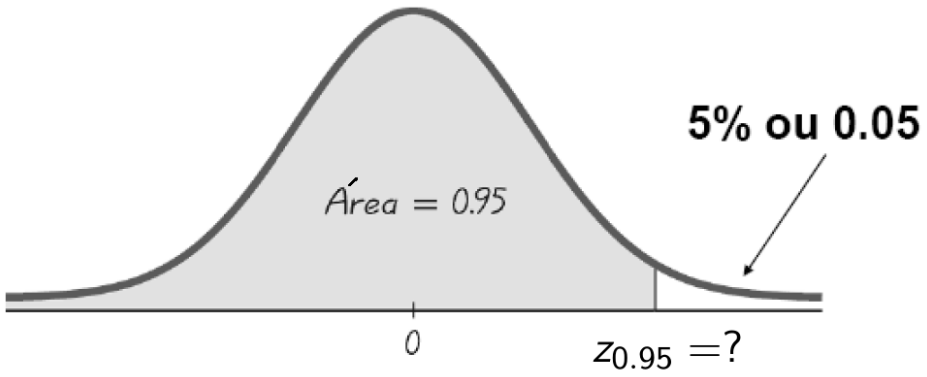
- Máquina de calcular[2]
  $P(X \leq 7) = F(7) = \text{cdfnormal}(-\infty, 7, 5, 2)$
  ou
  $P(X \leq 7) = F(7) \approx \text{cdfnormal}(-10^{10}, 7, 5, 2)$

Note-se que, nos três casos, **é requerido o desvio padrão**.

[1] Depende da versão instalada no seu computador.
[2] Depende da marca e versão da sua máquina de calcular.

Distribuição Normal: cálculo de probabilidades e quantis

Exemplo 2: determinar o quantil de ordem 0.95 da $N(0,1)$.



Habitualmente, os quantis de ordem p de uma $N(0,1)$ são denotados por z_p.

- SPSS (em *Transform* / *Compute Variable*)
  $z_{0.95} = \text{IDF.NORMAL}(0.95, 0, 1)$

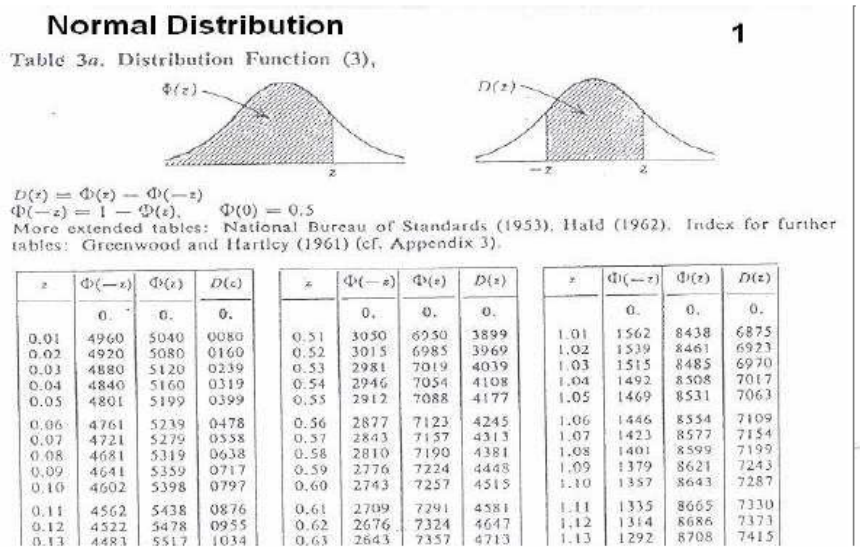
- Excel[1] (em *Insert Function*)
  $z_{0.95} = \text{Norminv}(0.95; 0; 1)$

[1] Depende da versão instalada no seu computador.

Distribuição Normal: cálculo de probabilidades e quantis

- Máquina de calcular[1]
  $z_{0.95} = \text{invnormal}(0.95, 0, 1)$

Obtém-se $z_{0.95} = 1.6448536\ldots \approx 1.645$.
Nas tabelas:



Normal Distribution 1

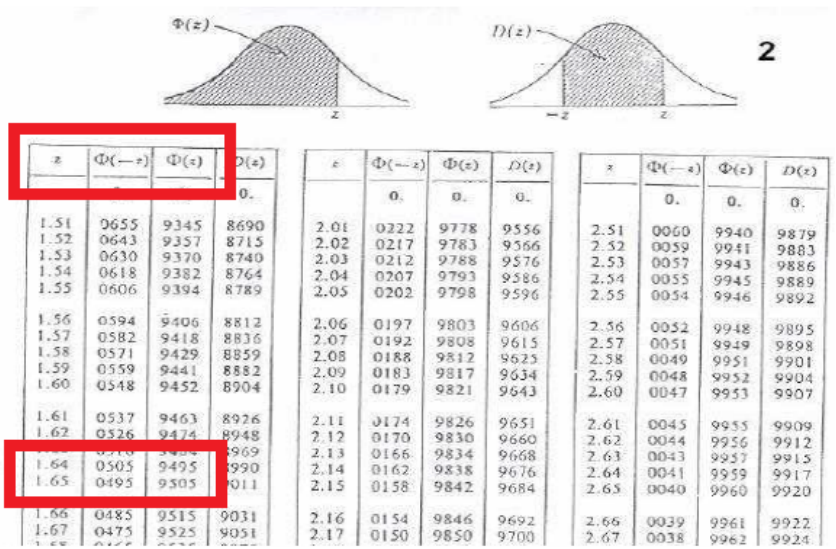
Table 3a. Distribution Function (3),

$D(z) = \Phi(z) - \Phi(-z)$
$\Phi(-z) = 1 - \Phi(z)$, $\Phi(0) = 0.5$
More extended tables: National Bureau of Standards (1953), Hald (1962). Index for further tables: Greenwood and Hartley (1961) (cf. Appendix 3).

z	$\Phi(-z)$	$\Phi(z)$	$D(z)$
	0.	0.	0.
0.01	4960	5040	0080
0.02	4920	5080	0160
0.03	4880	5120	0239
0.04	4840	5160	0319
0.05	4801	5199	0399
0.06	4761	5239	0478
0.07	4721	5279	0558
0.08	4681	5319	0638
0.09	4641	5359	0717
0.10	4602	5398	0797
0.11	4562	5438	0876
0.12	4522	5478	0955
0.13	4483	5517	1034

z	$\Phi(-z)$	$\Phi(z)$	$D(z)$
	0.	0.	0.
0.51	3050	6950	3899
0.52	3015	6985	3969
0.53	2981	7019	4039
0.54	2946	7054	4108
0.55	2912	7088	4177
0.56	2877	7123	4245
0.57	2843	7157	4313
0.58	2810	7190	4381
0.59	2776	7224	4448
0.60	2743	7257	4515
0.61	2709	7291	4581
0.62	2676	7324	4647
0.63	2643	7357	4713

z	$\Phi(-z)$	$\Phi(z)$	$D(z)$
	0.	0.	0.
1.01	1562	8438	6875
1.02	1539	8461	6923
1.03	1515	8485	6970
1.04	1492	8508	7017
1.05	1469	8531	7063
1.06	1446	8554	7109
1.07	1423	8577	7154
1.08	1401	8599	7199
1.09	1379	8621	7243
1.10	1357	8643	7287
1.11	1335	8665	7330
1.12	1314	8686	7373
1.13	1292	8708	7415



2

z	$\Phi(-z)$	$\Phi(z)$	$D(z)$
	0.	0.	0.
1.51	0655	9345	8690
1.52	0643	9357	8715
1.53	0630	9370	8740
1.54	0618	9382	8764
1.55	0606	9394	8789
1.56	0594	9406	8812
1.57	0582	9418	8836
1.58	0571	9429	8859
1.59	0559	9441	8882
1.60	0548	9452	8904
1.61	0537	9463	8926
1.62	0526	9474	8948
1.64	0505	9495	8990
1.65	0495	9505	9011
1.66	0485	9515	9031
1.67	0475	9525	9051

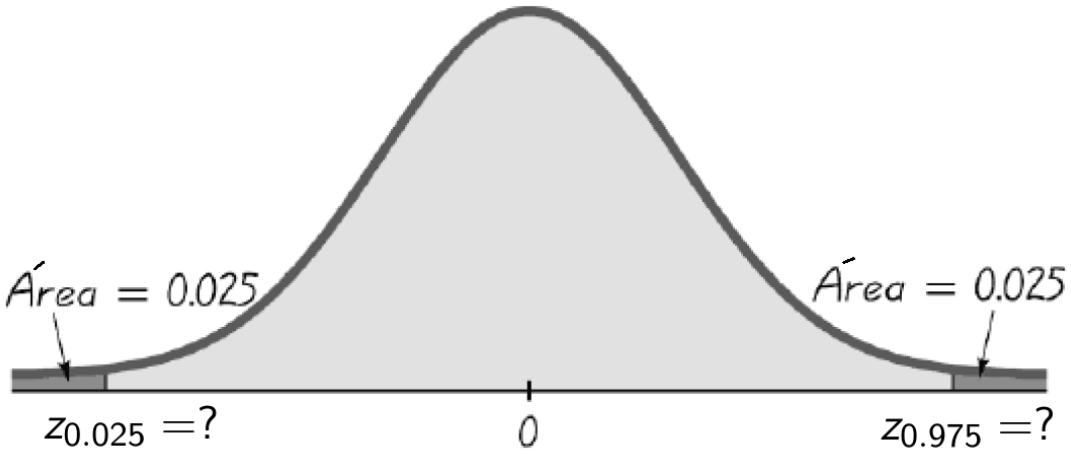
z	$\Phi(-z)$	$\Phi(z)$	$D(z)$
	0.	0.	0.
2.01	0222	9778	9556
2.02	0217	9783	9566
2.03	0212	9788	9576
2.04	0207	9793	9586
2.05	0202	9798	9596
2.06	0197	9803	9606
2.07	0192	9808	9615
2.08	0188	9812	9625
2.09	0183	9817	9634
2.10	0179	9821	9643
2.11	0174	9826	9651
2.12	0170	9830	9660
2.13	0166	9834	9668
2.14	0162	9838	9676
2.15	0158	9842	9684
2.16	0154	9846	9692
2.17	0150	9850	9700

z	$\Phi(-z)$	$\Phi(z)$	$D(z)$
	0.	0.	0.
2.51	0060	9940	9879
2.52	0059	9941	9883
2.53	0057	9943	9886
2.54	0055	9945	9889
2.55	0054	9946	9892
2.56	0052	9948	9895
2.57	0051	9949	9898
2.58	0049	9951	9901
2.59	0048	9952	9904
2.60	0047	9953	9907
2.61	0045	9955	9909
2.62	0044	9956	9912
2.63	0043	9957	9915
2.64	0041	9959	9917
2.65	0040	9960	9920
2.66	0039	9961	9922
2.67	0038	9962	9924

[1] Depende da marca e versão da sua máquina de calcular.

Distribuição Normal: cálculo de probabilidades e quantis

Exemplo 3: determinar os quantis $z_{0.025}$ e $z_{0.975}$ da $N(0,1)$.



- SPSS (em *Transform* / *Compute Variable*)
  $z_{0.975} = \text{IDF.NORMAL}(0.975, 0, 1)$

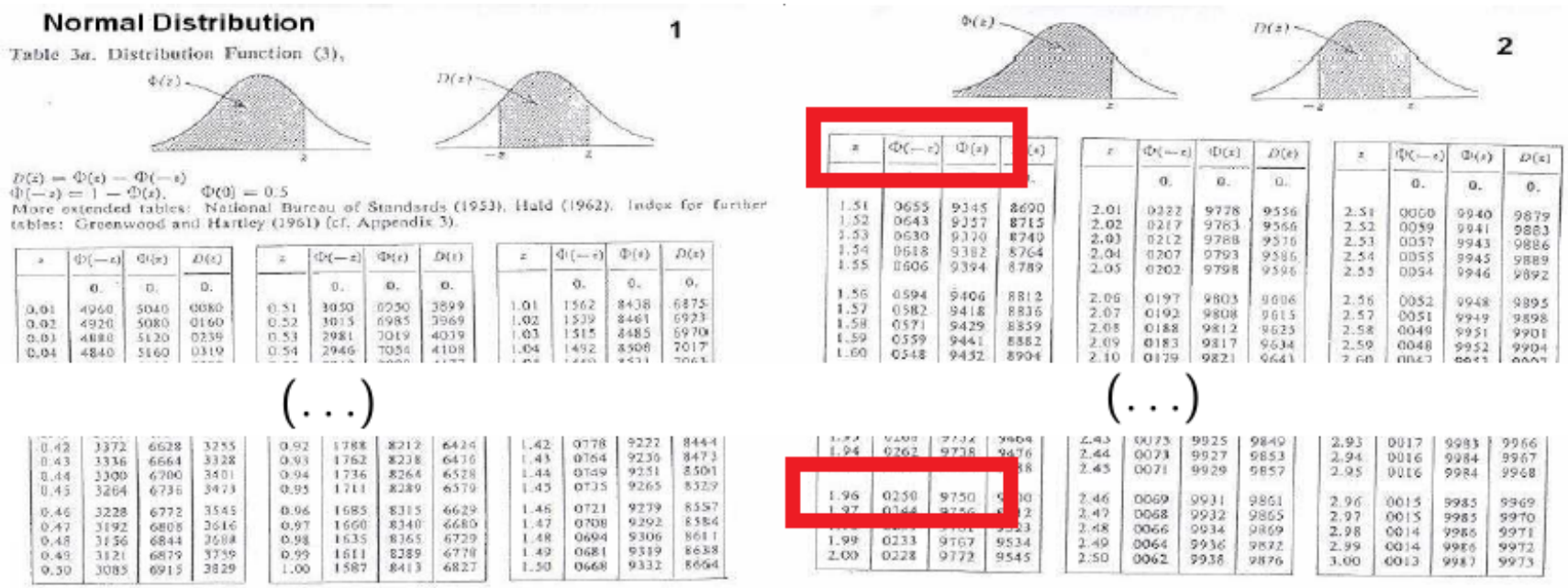
- Excel[1] (em *Insert Function*)
  $z_{0.975} = \text{Norminv}(0.975; 0; 1)$

[1] Depende da versão instalada no seu computador.

Distribuição Normal: cálculo de probabilidades e quantis

- Máquina de calcular[1]
  $z_{0.975} = \text{invnormal}(0.975, 0, 1)$

Obtém-se $z_{0.975} = 1.9599639\ldots \approx 1.96$. Logo, atendendo à simetria da distribuição, $z_{0.025} \approx -1.96$.
Nas tabelas:



Normal Distribution 1

Table 3a. Distribution Function (3),

$D(z) = \Phi(z) - \Phi(-z)$
$\Phi(-z) = 1 - \Phi(z)$, $\Phi(0) = 0.5$
More extended tables: National Bureau of Standards (1953), Hald (1962). Index for further tables: Greenwood and Hartley (1961) (cf. Appendix 3).

z	Φ(−z)	Φ(z)	D(z)
	0.	0.	0.
0.01	4960	5040	0080
0.02	4920	5080	0160
0.03	4880	5120	0239
0.04	4840	5160	0319

z	Φ(−z)	Φ(z)	D(z)
	0.	0.	0.
0.51	3050	6950	3899
0.52	3015	6985	3969
0.53	2981	7019	4039
0.54	2946	7054	4108

z	Φ(−z)	Φ(z)	D(z)
	0.	0.	0.
1.01	1562	8438	6875
1.02	1539	8461	6923
1.03	1515	8485	6970
1.04	1492	8508	7017

(...)

0.42	3372	6628	3255
0.43	3336	6664	3328
0.44	3300	6700	3401
0.45	3264	6736	3473
0.46	3228	6772	3545
0.47	3192	6808	3616
0.48	3156	6844	3688
0.49	3121	6879	3759
0.50	3085	6915	3829

0.92	1788	8212	6424
0.93	1762	8238	6476
0.94	1736	8264	6528
0.95	1711	8289	6579
0.96	1685	8315	6629
0.97	1660	8340	6680
0.98	1635	8365	6729
0.99	1611	8389	6778
1.00	1587	8413	6827

1.42	0778	9222	8444
1.43	0764	9236	8473
1.44	0749	9251	8501
1.45	0735	9265	8529
1.46	0721	9279	8557
1.47	0708	9292	8584
1.48	0694	9306	8611
1.49	0681	9319	8638
1.50	0668	9332	8664

2

z	Φ(−z)	Φ(z)	D(z)
	0.	0.	0.
1.51	0655	9345	8690
1.52	0643	9357	8715
1.53	0630	9370	8740
1.54	0618	9382	8764
1.55	0606	9394	8789
1.56	0594	9406	8812
1.57	0582	9418	8836
1.58	0571	9429	8859
1.59	0559	9441	8882
1.60	0548	9452	8904

z	Φ(−z)	Φ(z)	D(z)
	0.	0.	0.
2.01	0222	9778	9556
2.02	0217	9783	9566
2.03	0212	9788	9576
2.04	0207	9793	9586
2.05	0202	9798	9596
2.06	0197	9803	9606
2.07	0192	9808	9615
2.08	0188	9812	9625
2.09	0183	9817	9634
2.10	0179	9821	9643

z	Φ(−z)	Φ(z)	D(z)
	0.	0.	0.
2.51	0060	9940	9879
2.52	0059	9941	9883
2.53	0057	9943	9886
2.54	0055	9945	9889
2.55	0054	9946	9892
2.56	0052	9948	9895
2.57	0051	9949	9898
2.58	0049	9951	9901
2.59	0048	9952	9904

(...)

1.94	0262	9738	9476
1.96	0250	9750	[illegible]
1.97	[illegible]	[illegible]	[illegible]
1.99	0233	9767	9534
2.00	0228	9772	9545

2.43	0073	9925	9849
2.44	0073	9927	9853
2.45	0071	9929	9857
2.46	0069	9931	9861
2.47	0068	9932	9865
2.48	0066	9934	9869
2.49	0064	9936	9872
2.50	0062	9938	9876

2.93	0017	9983	9966
2.94	0016	9984	9967
2.95	0016	9984	9968
2.96	0015	9985	9969
2.97	0015	9985	9970
2.98	0014	9986	9971
2.99	0014	9986	9972
3.00	0013	9987	9973

[1] Depende da marca e versão da sua máquina de calcular.

Teorema do Limite Central (TLC)

Sucintamente, a soma de n v.a.'s independentes e identicamente distribuídas (i.i.d.) tem, para *n suficientemente* elevado, uma distribuição aproximada Normal.

Seja $X_1, X_2, \ldots, X_n, \ldots$ uma sucessão de v.a.'s i.i.d. com média μ e variância finita σ^2, i.e., $E[X_i] = \mu$ e $Var[X_i] = \sigma^2 > 0$ finita, $i = 1, 2, \ldots$. Então, para *n suficientemente* elevado,

$$X_1 + X_2 + \ldots + X_n = \sum_{i=1}^{n} X_i \underset{aprox.}{\sim} N(n\mu, n\sigma^2)$$

ou, depois de "centrar e reduzir", i.e., estandardizar ou padronizar,

$$\frac{\sum_{i=1}^{n} X_i - n\mu}{\sqrt{n}\,\sigma} \underset{aprox.}{\sim} N(0,1).$$

Teorema do Limite Central (TLC)

Naturalmente, o TLC pode ainda ser formulado por

$$\lim_{n\to+\infty} P\left(\frac{\sum_{i=1}^{n} X_i - n\mu}{\sqrt{n}\,\sigma} \leq x\right) = \Phi(x),\ x \in \mathbb{R}$$

ou, dito de outro modo, para *n suficientemente* elevado,

$$P\left(\frac{\sum_{i=1}^{n} X_i - n\mu}{\sqrt{n}\,\sigma} \leq x\right) \approx \Phi(x),\ x \in \mathbb{R}.$$

Recorde-se que Φ representa a f.d. da $N(0,1)$.

Teorema do Limite Central (TLC)

Nas condições do TLC, é possível ainda estabelecer que

$$\overline{X} = \frac{X_1 + X_2 + \ldots + X_n}{n} = \frac{1}{n}\sum_{i=1}^{n} X_i \underset{aprox.}{\sim} N\left(\mu, \frac{\sigma^2}{n}\right)$$

ou, depois de estandardizar,

$$\frac{\overline{X} - \mu}{\frac{\sigma}{\sqrt{n}}} \underset{aprox.}{\sim} N(0,1).$$

Sugestão: consulte o sítio

`http://www.vias.org/simulations/simusoft_cenlimit.html`

e explore os efeitos do TLC!

TLC e a Correção de Continuidade

O TLC permite efectuar cálculos de probabilidades aproximadas de acontecimentos definidos à custa de somas de um número "elevado" de parcelas i.i.d., não sendo necessário conhecer a distribuição de cada parcela, mas apenas o seu valor esperado e variância.

Porém, quando aquelas parcelas X_i são de **natureza discreta**, ou seja, para uma sucessão $X_1, X_2, \ldots, X_n, \ldots$ de v.a.'s discretas i.i.d. nas condições do TLC, pela aplicação do TLC resultará sempre

$$P(X_1 + X_2 + \ldots + X_n = x) \approx 0,\ x \in \mathbb{R}.$$

Note que, nas condições do TLC, a aproximação é realizada à distribuição Normal, na qual as probabilidades são obtidas a partir de áreas sob partes da curva de densidade.

TLC e a Correção de Continuidade

Assim, para obter, em geral, uma aproximação de boa qualidade **quando aplicamos o TLC a v.a.'s discretas**, aconselha-se efectuar a chamada **correção de continuidade.**

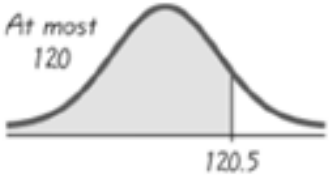
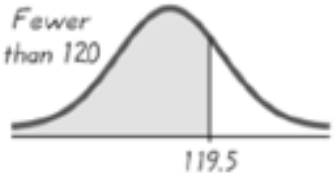
Esta correção consiste em "acrescentar" meia unidade ao valor discreto x em causa, de modo a garantir um intervalo que contenha uma vizinhança à direita e à esquerda de x e que represente, de modo equivalente, o acontecimento de interesse.

Vejam-se os seguintes exemplos:[1]

[1] Consultar Murteira et al. (2015) para detalhes adicionais.

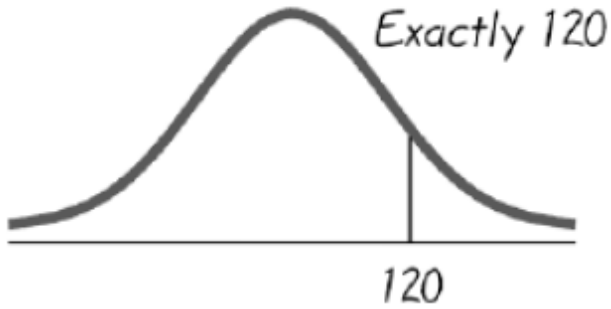
TLC e a Correção de Continuidade

pelo TLC

$P(X \geq 120)$	X = **pelo menos** 120 = 120, 121, 122, . . .		$P(X \geq 119.5)$
$P(X > 120)$	X = **mais do que** 120 = 121, 122, 123, . . .		$P(X \geq 120.5)$
$P(X \leq 120)$	X = **no máximo** 120 = 0, 1, . . . 118, 119, 120		$P(X \leq 120.5)$
$P(X < 120)$	X = **menos do que** 120 = 0, 1, . . . 118, 119		$P(X \leq 119.5)$

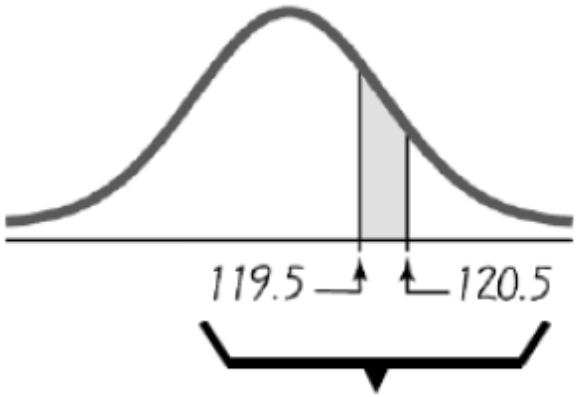
TLC e a Correção de Continuidade

$X = 120$



P(X=120)

pelo TLC



P(119.5≤X≤120.5)

Intervalo que representa o valor discreto 120

Capítulos

Conceitos Básicos

1. Análise Exploratória de Dados

2. Teoria de Probabilidades

3. Inferência Estatística
   - Estimação Pontual
   - Intervalos de Confiança
   - Testes de Hipóteses

Inferência Estatística

O objectivo da **Estatística Inferencial** ou **Inferência Estatística** é caracterizar e, eventualmente, definir regras de decisão sobre uma **população** conhecendo apenas parte dela.

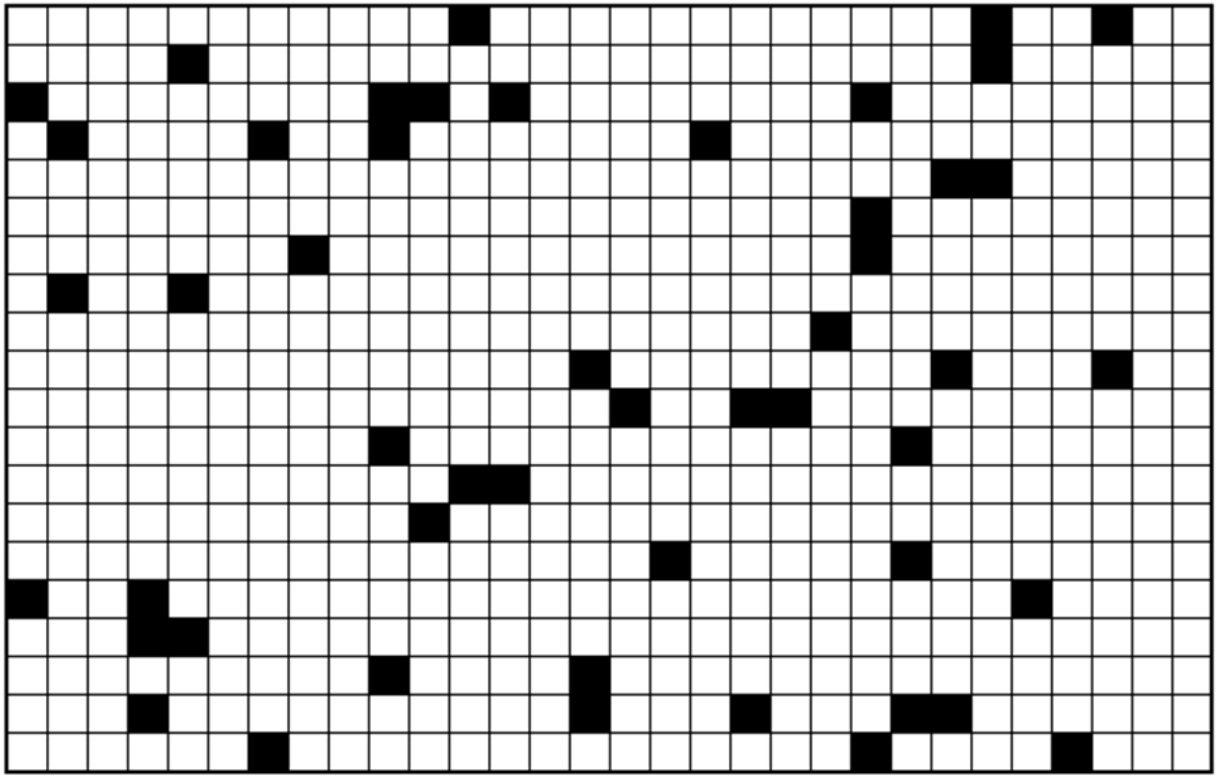
A parte da população que se conhece chama-se **amostra** e o processo pelo qual a amostra é obtida chama-se **amostragem**.[1]
É fundamental assegurar que a amostra é representativa da população da qual foi extraída, isto é, da população em estudo.

Dois objectivos usuais da recolha de amostras:
estimar características desconhecidas e testar hipóteses sobre parâmetros de uma população.

[1] Os quatro *slides* seguintes ilustram, sucintamente, algumas das técnicas usuais de amostragem aleatória. Sobre este tópico, sugerem-se:
Thompson, S. K. (2012). *Sampling*. Wiley;
Coelho, P., Pereira, L., Pinheiro, J. e Xufre, P. (2017). *As Sondagens – Princípios, Metodologias e Aplicações*. Escolar Editora.

Amostragem aleatória simples

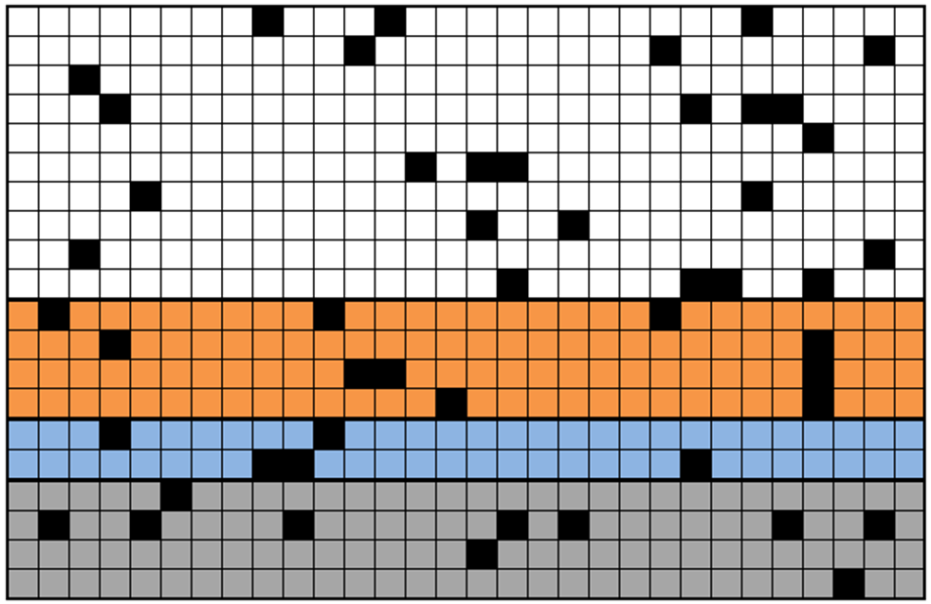
População de 600 elementos. Amostra aleatória simples de 50 elementos.



Amostragem estratificada

A população é dividida em estratos (heterogéneos entre si) e é extraída uma amostra (segundo um determinado plano de amostragem) em cada estrato (onde os elementos são o mais homogéneos possível).

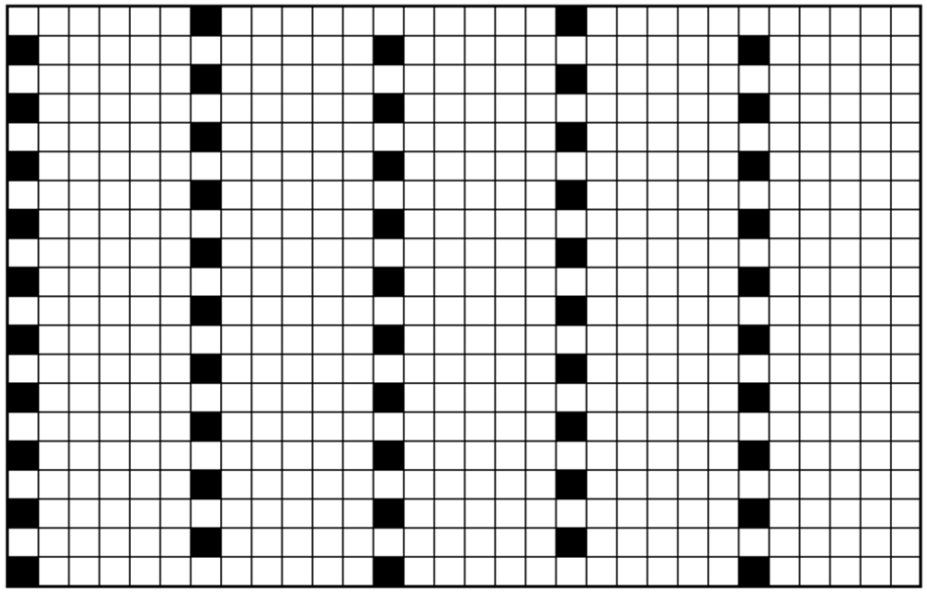
População de 600 elementos. 4 estratos. Amostra aleatória estratificada de 50 elementos (afetação proporcional: 25+10+5+10).



Amostragem sistemática

Escolhe-se aleatoriamente um elemento entre os primeiros k elementos da população, sendo este o primeiro elemento da amostra. Os restantes elementos são obtidos em intervalos sucessivos de amplitude k.

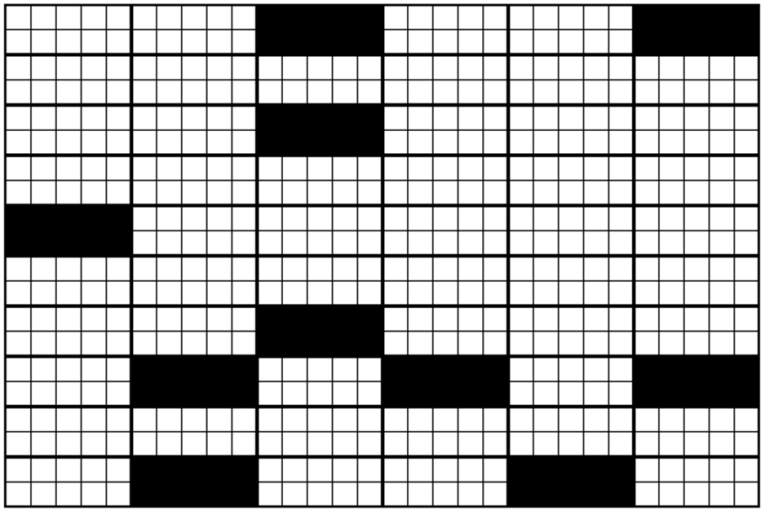
População de 600 elementos. Amostra de 50 elementos ($k = 600/50 = 12$). Elementos da amostra: 7, 19, 31, 43, ..., 583, 595.



Amostragem por grupos (*clusters*)

Considera-se a população dividida em grupos (*clusters*), com uma variabilidade idêntica à da população, preferencialmente. Seleciona-se uma amostra aleatória de grupos e inclui-se na amostra final todos os elementos de cada um dos grupos selecionados.

População de 600 elementos. Amostra aleatória de 10 unidades primárias (grupos), cada uma com 10 unidades secundárias (elementos).



Estimação Pontual

O objectivo usual da **estimação pontual** é inferir valores para os parâmetros desconhecidos que definem uma distribuição F_X. Atribuindo valores aos parâmetros está a definir-se a forma. Por exemplo, a distribuição Normal é caracterizada pela sua média μ e pela sua variância σ^2.

Estatística : função da amostra que não depende de parâmetros desconhecidos. Exemplo: a média amostral $\overline{X}$.

Estimativa pontual : valor obtido a partir da amostra através de uma **estatística** e que se destina a fornecer um valor para um parâmetro desconhecido.

Estimador : nome que se dá à **estatística** que fornece estimativas pontuais.

Estimador

Habitualmente representa-se um estimador de um parâmetro colocando um acento circunflexo sobre a letra que o representa. Por exemplo,

$$\hat{\mu}, \quad \hat{\sigma}, \quad \hat{p}.$$

Um *estimador é uma variável aleatória* e, como tal, tem uma distribuição que o caracteriza designada de **distribuição por amostragem**.

A distribuição *Normal*, *t-Student* ou a *Qui-Quadrado* são famílias de distribuições por amostragem como iremos ver.

Propriedades dos estimadores

No que se segue $\hat{\theta}$ é uma variável aleatória que cumpre a função de estimar um valor para o parâmetro θ.

Enviesamento : Um estimador $\hat{\theta}$ é centrado ou não enviesado se

$$\mathrm{E}[\hat{\theta}] = \theta.$$

Nota: à diferença $\mathrm{E}[\hat{\theta}] - \theta$ chama-se viés.

Eficiência : Entre os estimadores centrados, um estimador $\hat{\theta}_1$ é mais eficiente que outro $\hat{\theta}_2$ se

$$\mathrm{Var}[\hat{\theta}_1] < \mathrm{Var}[\hat{\theta}_2].$$

O Erro Quadrático Médio de um estimador $\hat{\theta}$ é dado por

$$\mathrm{EQM}[\hat{\theta}] = \mathrm{E}[(\hat{\theta} - \theta)^2] = \mathrm{Var}[\hat{\theta}] + (\underbrace{\mathrm{E}[\hat{\theta}] - \theta}_{\text{viés}})^2.$$

Propriedades dos estimadores – Exercício

Sejam $\widehat{\theta}_1$ e $\widehat{\theta}_2$ dois estimadores centrados para o parâmetro populacional $\theta \in \Theta$. Seja $\text{EQM}[\widehat{\theta}_2]$ o erro quadrático médio de $\widehat{\theta}_2$. Assim,

- ☐ $\text{E}[\widehat{\theta}_1] - \theta \neq 0$.
- ☐ se $\text{Var}[\widehat{\theta}_1] < \text{Var}[\widehat{\theta}_2]$, então $\widehat{\theta}_1$ é menos eficiente que $\widehat{\theta}_2$.
- ☐ $\text{EQM}[\widehat{\theta}_2] = \text{Var}[\widehat{\theta}_2]$.
- ☐ Nenhuma das afirmações anteriores é correta.

Estimadores

Entre dois estimadores $\hat{\theta}_1$ e $\hat{\theta}_2$ de um mesmo parâmetro θ devemos escolher o que tem menor EQM, i.e., variância reduzida e viés reduzido (de preferência ser centrado).

Os desafios são:

1. Como encontrar estimadores?
   - Método dos momentos.
   - Método da máxima verosimilhança.
2. Quais as propriedades distribucionais dos estimadores encontrados?

Método dos Momentos

Sucintamente, o método dos momentos traduz-se no seguinte procedimento:

1. Igualam-se os **momentos** populacionais e amostrais,

$$\mathrm{E}[X^k] = \frac{1}{n}\sum_{i=1}^{n} X_i^k,$$

   para os $k \in \mathbb{N}$ parâmetros desconhecidos a estimar.
   Exemplo: se $k = 1$ a equação fica $\mathrm{E}[X] = \overline{X}$.

2. Resolve-se a equação ou equações obtida(s) em ordem ao(s) parâmetro(s) de interesse.
   Exemplo: $\mathrm{E}[X] = \overline{X} \Leftrightarrow \mu = \overline{X}$, pelo que o estimador para μ é dado por $\hat{\mu} = \overline{X}$.

Método dos Momentos – Exemplo

Determine um estimador pelo método dos momentos para o parâmetro b da distribuição Uniforme $U(0, b)$.

Como existe somente um parâmetro a estimar, procede-se a

$$\mathrm{E}[X] = \overline{X} \Leftrightarrow \frac{0+b}{2} = \overline{X} \Leftrightarrow b = 2\overline{X},$$

pelo que se obtém um estimador para b dado por $\hat{b} = 2\overline{X}$.

Método dos Momentos para uma população Normal

Seja $X \sim N(\mu, \sigma^2)$ em que os problemas de estimação mais comuns são:

1. estimar σ^2 com a média μ conhecida (1 incógnita).
2. estimar μ e σ^2 (2 incógnitas).

Método dos Momentos para uma população Normal

Problema 1. Com μ conhecido pretende-se estimar σ^2.
Como há um só parâmetro a estimar, então uma só equação basta. Partindo da relação $\boxed{\sigma^2 = \mathrm{E}[X^2] - \mu^2}$ desenvolve-se a equação do método dos momentos para $k = 2$:

$$\mathrm{E}[X^2] = \frac{1}{n}\sum_{i=1}^{n} X_i^2 \Leftrightarrow \sigma^2 + \mu^2 = \frac{1}{n}\sum_{i=1}^{n} X_i^2 \Leftrightarrow \sigma^2 = \frac{1}{n}\sum_{i=1}^{n} X_i^2 - \mu^2,$$

obtendo-se o estimador $\hat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^{n} X_i^2 - \mu^2$.

M.M. para uma população Normal

Problema 2. Com μ e σ^2 desconhecidos.
Como há dois parâmetros desconhecidos, então necessitamos de um sistema com duas equações ($k = 1$ e $k = 2$ no método dos momentos). Resolve-se o sistema

$$\begin{cases} \mathrm{E}[X] & = \frac{1}{n}\sum_{i=1}^{n} X_i \\ \mathrm{E}[X^2] & = \frac{1}{n}\sum_{i=1}^{n} X_i^2 \end{cases} \Leftrightarrow$$

$$\Leftrightarrow \begin{cases} \mu & = \overline{X} \\ \sigma^2 + \mu^2 & = \frac{1}{n}\sum_{i=1}^{n} X_i^2 \end{cases} \Leftrightarrow$$

$$\Leftrightarrow \begin{cases} \mu & = \overline{X} \\ \sigma^2 & = \frac{1}{n}\sum_{i=1}^{n} X_i^2 - \overline{X}^2 \end{cases} .$$

A solução é o par de estimadores:

$$\begin{cases} \hat{\mu} & = \overline{X} \\ \hat{\sigma}^2 & = S^2 \end{cases} .$$

Método da Máxima Verosimilhança

Para definir o conceito de função de verosimilhança, considera-se aqui, por simplicidade, uma definição sucinta somente para o caso discreto.

Seja $(X_1, X_2, \ldots, X_n)$ uma amostra aleatória de uma população com função massa de probabilidade $f(x|\theta)$. Fixada uma amostra observada, seja $(x_1, x_2, \ldots, x_n)$, fazendo variar θ no seu espaço-parâmetro, Θ, a função de verosimilhança é definida por

$$L(\theta) = L(\theta|x_1, x_2, \ldots, x_n) = \prod_{i=1}^{n} f(x_i|\theta),$$

a qual, para cada $\theta \in \Theta$, quantifica a credibilidade/verosimilhança/ plausibilidade associada a esse valor, dada a amostra concreta observada.

No caso contínuo, salvaguardadas as devidas diferenças (e.g., uso de f.d.p.), a função de verosimilhança define-se como anteriormente. Note-se que a cada amostra observada corresponde uma função de verosimilhança e que a verosimilhança não é uma probabilidade.

Método da Máxima Verosimilhança

Assim, dada uma amostra observada, seja $(x_1, x_2, \ldots, x_n)$, o método da máxima verosimilhança consiste em encontrar, se existir, uma função $\widehat{\theta} = \widehat{\theta}(x_1, x_2, \ldots, x_n)$, tal que

$$L(\widehat{\theta}|x_1, x_2, \ldots, x_n) \geq L(\theta|x_1, x_2, \ldots, x_n), \ \forall \theta \in \Theta.$$

Assim, nestas circunstâncias, diz-se que $\widehat{\theta} = \widehat{\theta}(X_1, X_2, \ldots, X_n)$ é o estimador de máxima verosimilhança de θ.

Com este método, para estimativa de θ, deve escolher-se a função dos dados que torne mais credível/verosímil/plausível a amostra concreta observada, ou seja, deve procurar-se o maximizante de $L(\theta)$.

(Nota: o Método de Newton-Raphson, estudado na Componente de Métodos Numéricos, é comummente usado na obtenção do maximizante da função de verosimilhança.)

Procedimento do método da Máxima Verosimilhança

1. Escreve-se a função de verosimilhança,

$$L(\theta) = \prod_{i=1}^{n} f(x_i|\theta).$$

2. Recorre-se ao logaritmo de $L(\theta)$,

$$\log L(\theta) = \sum_{i=1}^{n} \log f(x_i|\theta),$$

   que permite passar de um produtório a um somatório, sendo esta função mais fácil de maximizar, habitualmente. Note-se que, sendo a função logaritmo uma função monótona crescente, o maximizante de $\log L(\theta)$ coincide com o pretendido maximizante de $L(\theta)$.

3. Maximiza-se $\log L(\theta)$. Normalmente, pois nem sempre tal é possível, toma-se a derivada, iguala-se a zero, e verifica-se, posteriomente, se o ponto estacionário encontrado é, de facto, um máximo.

Máxima Verosimilhança – Exemplo 1

Seja $(X_1, X_2, \ldots, X_n)$ uma amostra aleatória cujas variáveis têm densidade $f(x;\theta) = \theta x^{\theta-1}$, onde $\theta > 0$ e $0 < x < 1$.
Calcule um estimador para θ utilizando o método da máxima verosimilhança.

Máxima Verosimilhança – Resolução

1. Verosimilhança da amostra $(x_1, x_2, \ldots, x_n)$:

$$L(\theta) = \prod_{i=1}^{n} f(x_i) = \prod_{i=1}^{n} \theta x_i^{\theta-1} = \theta^n \prod_{i=1}^{n} x_i^{\theta-1}.$$

2. Logaritmo de $L(\theta)$:

$$\log L(\theta) = n \log(\theta) + (\theta - 1) \sum_{i=1}^{n} \log(x_i).$$

3. Maximizar $\log L(\theta)$:

$$\frac{d}{d\theta} \log L(\theta) = \frac{n}{\theta} + \sum_{i=1}^{n} \log(x_i) = 0 \Leftrightarrow \theta = -\frac{n}{\sum_{i=1}^{n} log(x_i)}.$$

O estimador de máxima verosimilhança de θ é

$$\hat{\theta} = -\frac{n}{\sum_{i=1}^{n} \log(X_i)}.$$

Máxima Verosimilhança – Resolução

Note-se que, como

$$\frac{d^2}{d\theta^2} \log L(\theta)\Big|_{\theta=\hat{\theta}} = -\frac{n}{\theta^2}\Big|_{\theta=\hat{\theta}} < 0,$$

o ponto estacionário da função de verosimilhança,

$$\theta = -\frac{n}{\sum_{i=1}^{n} log(x_i)},$$

é um máximo.

Máxima Verosimilhança – Exemplo 2

Seja $(X_1, X_2, \ldots, X_n)$ uma amostra aleatória de uma população de Bernoulli de parâmetro p (com $0 < p < 1$),

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

Calcule um estimador para p utilizando o método da máxima verosimilhança.

Máxima Verosimilhança – Resolução

1. Verosimilhança da amostra $(x_1, x_2, \ldots, x_n)$:

$$L(p) = \prod_{i=1}^{n} f(x_i) = \prod_{i=1}^{n} p^{x_i}(1-p)^{1-x_i},$$

ou seja, pela lei dos expoentes,

$$L(p) = p^{\sum_{i=1}^{n} x_i}(1-p)^{n-\sum_{i=1}^{n} x_i}.$$

2. Logaritmo de $L(p)$:

$$\log L(p) = \left(\sum_{i=1}^{n} x_i\right)\log(p) + \left(n - \sum_{i=1}^{n} x_i\right)\log(1-p).$$

Máxima Verosimilhança – Resolução

3. Maximizar $\log L(p)$:
(determinar o ponto crítico com $\frac{d}{dp}\log L(p) = 0$)

$$\frac{d}{dp}\log L(p) = \left(\sum_{i=1}^{n} x_i\right)\frac{1}{p} - \left(n - \sum_{i=1}^{n} x_i\right)\frac{1}{1-p} = 0$$

obtendo-se $p = \sum_{i=1}^{n} x_i/n$.

Como facilmente se verifica que

$$\frac{d^2}{dp^2}\log L(p) < 0,$$

o estimador de máxima verosimilhança para p é dado por

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

Máxima Verosimilhança – Outras distribuições

É deixado como exercício verificar que o estimador da máxima verosimilhança para

- $X \sim \mathrm{Poisson}(\lambda)$ é $\hat{\lambda} = \overline{X}$,
- $X \sim \mathrm{Normal}(\mu, \sigma^2)$ são $\hat{\mu} = \overline{X}$ e $\hat{\sigma}^2 = S^2$,
- $X \sim \mathrm{Exp}(\lambda)$ é $\hat{\lambda} = 1/\overline{X}$.

Máxima Verosimilhança – Exemplo 3

Seja $(X_1, X_2, \ldots, X_n)$ uma amostra aleatória genérica de uma população X de Poisson, com função de probabilidade $f(x|\theta) = e^{-\theta}\frac{\theta^x}{x!}$, com $\theta > 0$.

(a) Sabendo que a concretização $(x_1, x_2, \ldots, x_{10})$ de uma amostra aleatória devolve $\sum_{i=1}^{10} x_i = 30$, calcule a estimativa de máxima verosimilhança da variância de X.

(b) Verifique que $\widehat{\theta} = \dfrac{1}{n}\displaystyle\sum_{i=1}^{n} X_i^2$ é um estimador enviesado para θ.

Máxima Verosimilhança – Resolução

(a) Sabe-se que $\widehat{\theta} = \overline{X}$. Como $\overline{x} = \frac{30}{10} = 3$ e $\sigma^2 = \mathrm{Var}[X] = \theta$, a estimativa pretendida é $\widehat{\sigma}^2 = 3$.

(b) O estimador será enviesado se $\mathrm{E}[\widehat{\theta}] \neq \theta$. Neste caso,

$$\mathrm{E}[\widehat{\theta}] = \frac{1}{n}\sum_{i=1}^{n}\mathrm{E}[X_i^2] = \frac{1}{n}\,n\,\mathrm{E}[X^2] = \mathrm{E}[X^2] = \mathrm{Var}[X]+(\mathrm{E}[X])^2 = \theta+\theta^2 \neq \theta.$$

Recorde-se que $\mathrm{Var}[X] = \mathrm{E}[X^2] - (\mathrm{E}[X])^2$ e que, sendo $(X_1, X_2, \ldots, X_n)$ uma amostra aleatória de X, então $\mathrm{E}[X_i] = \mathrm{E}[X], \forall i = 1, 2, ..., n$.

Intervalos de Confiança (IC)

Uma estimativa pontual de um parâmetro não contém informação sobre a precisão do valor obtido.

A variância e o EQM de um estimador fornecem alguma informação, mas uma forma mais completa de abordar a questão consiste em construir um intervalo fixando, previamente, a probabilidade desse intervalo conter o verdadeiro valor do parâmetro.

Intervalos de Confiança (IC)

Um intervalo de confiança para um parâmetro θ a um nível de confiança $1-\alpha$ é um intervalo aleatório $(\hat{\theta}_1, \hat{\theta}_2)$ tal que

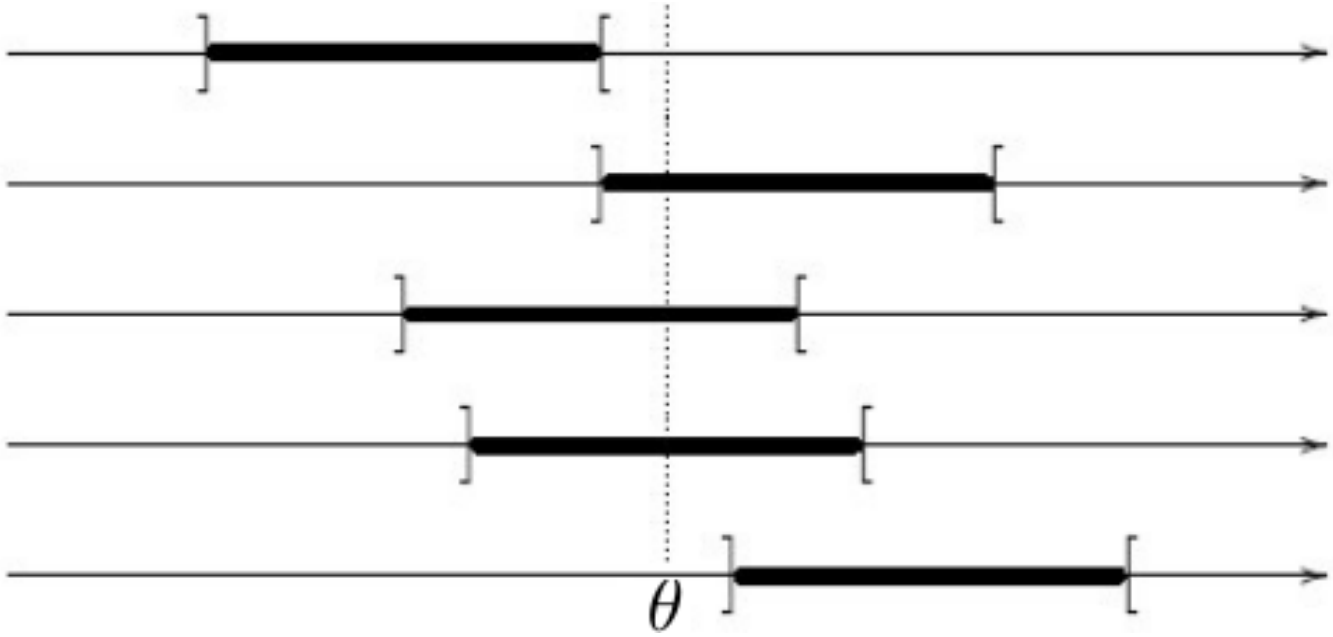
$$P(\hat{\theta}_1 < \theta < \hat{\theta}_2) = 1-\alpha, \quad \alpha \in (0,1), \quad \alpha \text{ pequeno.}$$

- $\hat{\theta}_1$ e $\hat{\theta}_2$ são estimadores do lado esquerdo e direito do intervalo e dependem duma amostra em particular.
- O valor α é reduzido por forma a que se tenham confianças $1-\alpha$ típicas de 0.90, 0.95, 0.99 ou ainda valores superiores.
  Que valores assume α nestes casos?

Intervalos de Confiança (IC)

$$P(\hat{\theta}_1 < \theta < \hat{\theta}_2) = 1 - \alpha, \quad \alpha \in (0, 1).$$

- Para cada amostra que se observa obtém-se, em geral, um intervalo de confiança diferente para o mesmo parâmetro θ.
- Se observarmos muitas amostras distintas, o verdadeiro valor do parâmetro está dentro dos IC calculados $(1 - \alpha) \times 100\%$ das vezes.



Intervalos de Confiança (IC)

Para reforçar esta noção de IC, considere-se esta síntese do exemplo disponível em Murteira et al. (2015), pp. 427-428.

Recolhe-se uma amostra de 20 valores do peso de determinados cigarros. Assume-se que a v.a. que representa o peso tem distribuição $N(\mu, 22.5^2)$. Considerando $\overline{X}$ a v.a. que representa a média da amostra genérica da referida população, obtém-se

$$P\left(\overline{X} - 9.86 < \mu < \overline{X} + 9.86\right) = 0.95.$$

Note-se que $\overline{X}$ é uma v.a. e $(\overline{X} - 9.86, \overline{X} + 9.86)$ é um intervalo aleatório. A probabilidade de este intervalo aleatório conter o valor desconhecido da média da referida população, μ, é igual a 0.95. Assim, considerando a observação de um grande número de amostras com 20 valores de pesos, correspondendo a cada uma delas um particular intervalo $(\overline{x} - 9.86, \overline{x} + 9.86)$, a proporção destes intervalos a que pertence μ é, aproximadamente, 95%.

Intervalos de Confiança (IC)

Resumo dos IC apresentados:

- para a média μ de uma população Normal com σ^2 conhecida e desconhecida;
- para média μ de uma população genérica (em condições TLC);
- para uma proporção p;
- para a diferença de médias $\mu_X - \mu_Y$ de populações Normais, variâncias conhecidas e desconhecidas, amostras independentes;
- para a diferença de médias $\mu_X - \mu_Y$ de populações Normais, amostras emparelhadas;
- para a variância σ^2 de uma população Normal com média μ conhecida e desconhecida.

IC para a média μ com σ^2 conhecida

Intervalo de confiança para a média μ de uma população Normal com variância σ^2 conhecida.

Como construir?

Assumindo que a população é bem modelada por

$$X \sim \mathrm{Normal}(\mu, \sigma^2)$$

temos

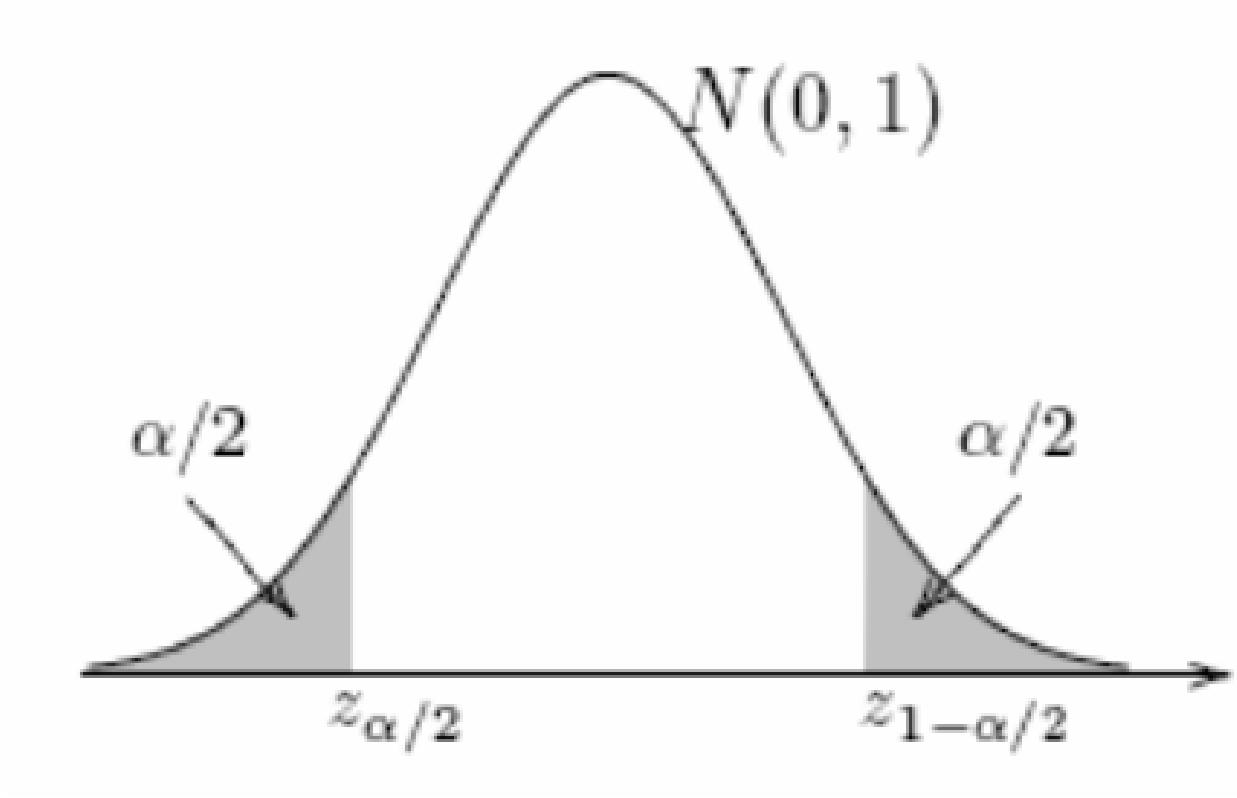
$$\overline{X} \sim N(\mu, \sigma^2/n)$$

ou ainda, "centrando e reduzindo",

$$Z = \frac{\overline{X} - \mu}{\sigma/\sqrt{n}} \sim N(0, 1).$$

IC para a média μ com σ^2 conhecida

Seja $z_{1-\alpha/2}$ o quantil de ordem $1-\alpha/2$ de uma N(0,1), isto é, $\mathrm{P}[Z \leq z_{1-\alpha/2}] = 1-\alpha/2$:



IC para a média μ com σ^2 conhecida

Continuação da construção do intervalo para a média μ.

Recordando que $Z = \frac{\overline{X}-\mu}{\sigma/\sqrt{n}} \sim N(0,1)$, então,

$$P\left(-z_{1-\alpha/2} < Z < z_{1-\alpha/2}\right) = 1-\alpha$$

$$\Leftrightarrow P\left(-z_{1-\alpha/2} < \frac{\overline{X}-\mu}{\sigma/\sqrt{n}} < z_{1-\alpha/2}\right) = 1-\alpha$$

$$\Leftrightarrow P\left(-z_{1-\alpha/2}\frac{\sigma}{\sqrt{n}} < \overline{X}-\mu < z_{1-\alpha/2}\frac{\sigma}{\sqrt{n}}\right) = 1-\alpha$$

$$\Leftrightarrow P\left(\overline{X}-z_{1-\alpha/2}\frac{\sigma}{\sqrt{n}} < \mu < \overline{X}+z_{1-\alpha/2}\frac{\sigma}{\sqrt{n}}\right) = 1-\alpha.$$

IC para a média μ com σ^2 conhecida

Um intervalo de confiança para a média μ de uma população Normal com variância σ^2 conhecida, a um nível de confiança $1-\alpha$, é dado por

$$\left(\overline{X} - z_{1-\alpha/2}\frac{\sigma}{\sqrt{n}}, \overline{X} + z_{1-\alpha/2}\frac{\sigma}{\sqrt{n}}\right)$$

em que $\overline{X}$ é a média da amostra.

Nota: para a construção de Intervalos de Confiança recorrer-se-á ao método da variável fulcral (e.g., Murteira et al., 2015, pp. 428-431). Note que, por exemplo, na construção do IC para a média μ, com σ^2 conhecida, apesar da distribuição de X depender de μ, $X \sim N(\mu, \sigma^2)$, a variável Z depende da amostra e de μ, mas a sua distribuição não depende de μ, $Z \sim N(0,1)$.

IC para a média μ com σ^2 conhecida

Propriedades do $\mathrm{IC}_{1-\alpha}(\mu)$:

$$\left(\overline{X} - z_{1-\alpha/2}\frac{\sigma}{\sqrt{n}}, \overline{X} + z_{1-\alpha/2}\frac{\sigma}{\sqrt{n}}\right)$$

- Quanto maior o nível de confiança, maior a amplitude do intervalo.
- Quando maior a variância, maior a amplitude do intervalo.
- Quanto maior a amostra, menor a amplitude do intervalo.

IC para a média μ com σ^2 conhecida – Exemplo

Adaptado de Hall et al. (2011), p. 132.
(Sugere-se a consulta do livro para detalhes adicionais.)

Pretende-se o intervalo de confiança a 95% para a taxa média de álcool à saída de um bar. Com base numa amostra aleatória de 125 clientes, obteve-se $\overline{x} = 0.6$mg/l. De acordo com a GNR, a quantidade de álcool no sangue é modelada por uma $N(\mu, 0.04)$.

É necessário determinar $z_{1-\frac{\alpha}{2}} = z_{1-\frac{0.05}{2}} = z_{1-0.025} = z_{0.975}$:

Tabelas: para $z = 1.96$, obtém-se $\Phi(1.96) = 0.9750$;

SPSS: IDF.NORMAL(0.975,0,1) $\approx$ 1.96 ou PROBIT(0.975)$\approx$ 1.96;

EXCEL (Office 2010 PT): INV.NORMAL(0,975;0;1) $\approx$ 1.96.

IC para a média μ com σ^2 conhecida – Exemplo

Assim, o IC pretendido é dado por

$$IC_{95\%}(\mu) = \left(0.6 - 1.96\frac{0.2}{\sqrt{125}}, 0.6 + 1.96\frac{0.2}{\sqrt{125}}\right)$$
$$\approx (0.5649, 0.6351),$$

onde μ é a média (populacional) da taxa de álcool à saída do bar.

IC para a média μ com σ^2 conhecida – Exercício

Considerando uma amostra aleatória $X_1, X_2, \ldots, X_n$ proveniente de uma população Normal, $N(\mu, 9)$, foi determinado o seguinte intervalo de confiança para μ:

$$\left(\overline{X} - \frac{6}{\sqrt{n}}, \overline{X} + \frac{6}{\sqrt{n}}\right).$$

Determine o grau de confiança do intervalo.

IC para a média μ com σ^2 conhecida – Resolução

Sendo $X \sim N(\mu, 9)$, então $\sigma = 3$. Do intervalo

$$\left(\overline{X} - z_{1-\alpha/2}\frac{\sigma}{\sqrt{n}}, \overline{X} + z_{1-\alpha/2}\frac{\sigma}{\sqrt{n}}\right)$$

verifica-se que

$$z_{1-\alpha/2}\frac{\sigma}{\sqrt{n}} = \frac{6}{\sqrt{n}},$$

ou seja,

$$z_{1-\alpha/2}\frac{3}{\sqrt{n}} = \frac{6}{\sqrt{n}} \Leftrightarrow z_{1-\alpha/2} = 2.$$

Assim,

$$\mathrm{P}[Z \leq z_{1-\alpha/2}] = 1 - \alpha/2 \Leftrightarrow \mathrm{P}[Z \leq 2] \approx 0.9772 = 1 - \alpha/2,$$

ou seja, $\alpha = 0.0456$, pelo que, $1 - \alpha \approx 0.9544\,(95.44\%)$.

(Tabelas: para $z = 2$, $\Phi(2) = 0.9772$. SPSS: CDF.NORMAL(2,0,1) $\approx$ 0.9772 ou CDFNORM(2) $\approx$ 0.9772.)

IC para μ com σ^2 desconhecida

Intervalo de confiança para a média μ de uma população Normal com variância desconhecida.

Recordando, o intervalo de confiança para μ quando a variância é conhecida foi determinado devido ao facto de

$$Z = \frac{\overline{X} - \mu}{\sigma/\sqrt{n}} \sim N(0,1).$$

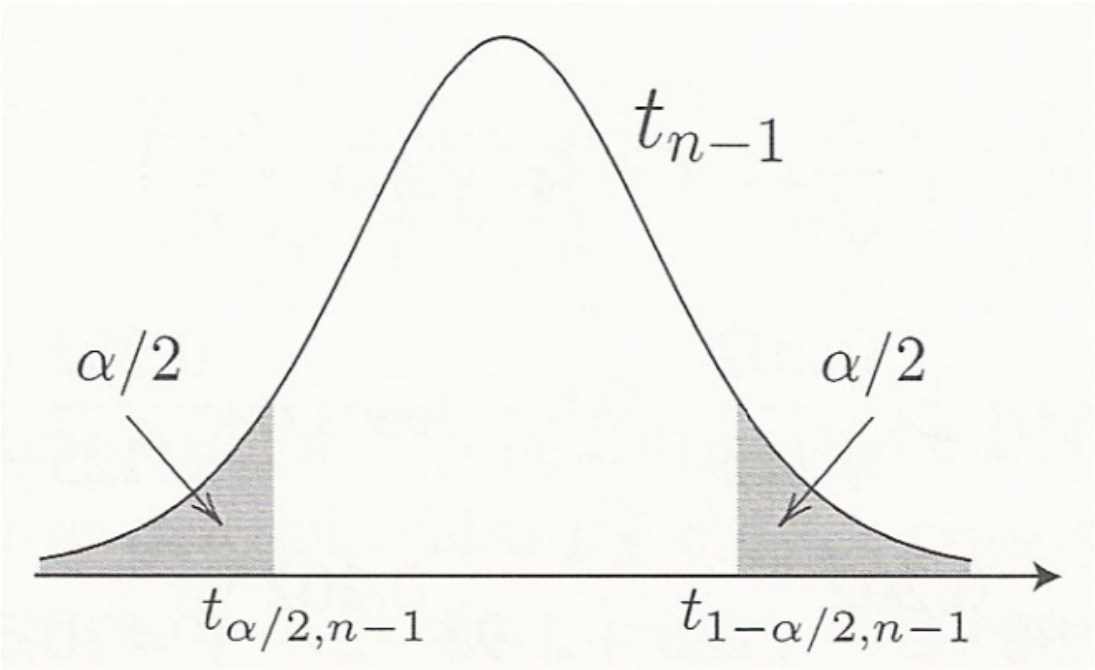
Não dispondo de σ e usando em seu lugar S_c, obtemos

$$T = \frac{\overline{X} - \mu}{S_c/\sqrt{n}} \sim t_{n-1}$$

O que é t_{n-1}?

Distribuição t-Student

t_{n-1} designa a distribuição t-Student com $n-1$ graus de liberdade. Trata-se de uma distribuição contínua e simétrica em torno de zero. À medida que n aumenta, esta distribuição, embora com variância superior, aproxima-se da Normal(0,1).
O gráfico mostra que $\mathrm{P}[T \leq t_{1-\alpha/2,n-1}] = 1 - \alpha/2$:



IC para μ com σ^2 desconhecida

Construção do intervalo:

$$P\left(-t_{1-\alpha/2,n-1} < T < t_{1-\alpha/2,n-1}\right) = 1-\alpha$$

$$\Leftrightarrow P\left(-t_{1-\alpha/2,n-1} < \frac{\overline{X}-\mu}{S_c/\sqrt{n}} < t_{1-\alpha/2,n-1}\right) = 1-\alpha$$

$$\Leftrightarrow P\left(\overline{X} - t_{1-\alpha/2,n-1}\frac{S_c}{\sqrt{n}} < \mu < \overline{X} + t_{1-\alpha/2,n-1}\frac{S_c}{\sqrt{n}}\right) = 1-\alpha.$$

IC para μ com σ^2 desconhecida

Um intervalo de confiança para a média μ de uma população Normal com variância desconhecida, a um nível de confiança $1-\alpha$, é dado por

$$\left(\overline{X} - t_{1-\alpha/2,n-1}\frac{S_c}{\sqrt{n}}, \overline{X} + t_{1-\alpha/2,n-1}\frac{S_c}{\sqrt{n}}\right)$$

em que $\overline{X}$ é média amostral e S_c o desvio padrão amostral corrigido.

IC para μ com σ^2 desconhecida - Exemplo

Adaptado de Hall et al. (2011), pp. 133–134.
(Sugere-se a consulta do livro para detalhes adicionais.)

Pretende-se o intervalo de confiança a 95% para a taxa média de álcool à saída de um bar. Com base numa amostra aleatória de 125 clientes, obteve-se $\overline{x} = 0.6$mg/l e $s_c = 0.202$mg/l. (Assume-se a normalidade da população, mas o conhecimento de σ^2 é irrealista.)

É necessário determinar $t_{1-\frac{\alpha}{2},\, n-1} = t_{1-\frac{0.05}{2},\, 124} = t_{0.975,\, 124}$:

Tabelas: $t_{0.975,100} = 1.98$ e $t_{0.975,200} = 1.97$;

SPSS: IDF.T(0.975,124) $= 1.97928...$

IC para μ com σ^2 desconhecida - Exemplo

Assim, o IC pretendido é dado por

$$IC_{95\%}(\mu) = \left(0.6 - 1.98\frac{0.202}{\sqrt{125}}, 0.6 + 1.98\frac{0.202}{\sqrt{125}}\right)$$
$$\approx (0.5642, 0.6358),$$

onde μ é a média (populacional) da taxa de álcool à saída do bar.

Nota: o IC para a média assumindo variância, σ^2, conhecida é

$IC_{95\%}(\mu) \approx (0.5649, 0.6351)$. Compare a amplitude dos IC.

IC para a média μ de uma população genérica

Intervalo de confiança para a média μ de uma população genérica com variância σ^2 conhecida ou desconhecida.

De uma forma geral, conhecendo a variância duma distribuição e considerando válidas as condições do Teorema do Limite Central (n elevado), sabe-se que

$$\overline{X} \underset{\text{aprox}}{\sim} N(\mu, \sigma^2/n), n \to \infty,$$

ou, de forma equivalente,

$$\frac{\overline{X} - \mu}{\sigma/\sqrt{n}} \underset{\text{aprox}}{\sim} N(0,1), n \to \infty,$$

pelo que, à semelhança do primeiro IC apresentado anteriormente, podemos obter um intervalo de confiança aproximado para μ:

IC para a média μ de uma população genérica

Um intervalo de confiança **aproximado** para a média μ de uma população genérica com variância σ^2 conhecida, a um nível de confiança $1-\alpha$, é dado por

$$\left(\overline{X} - z_{1-\alpha/2}\frac{\sigma}{\sqrt{n}}, \overline{X} + z_{1-\alpha/2}\frac{\sigma}{\sqrt{n}}\right)$$

em que $\overline{X}$ é a média amostral.
Se variância σ^2 for desconhecida, mas a amostra for de elevada dimensão, então σ pode ser substituído por S_c. Note-se que estes IC aproximados serão tanto melhores quanto maior for a dimensão da amostra.

IC para uma proporção p

Intervalo de confiança para a proporção p de sucessos numa população de Bernoulli.

O número de sucessos X numa amostra de dimensão n com experiências independentes de Bernoulli (sucessso/falha) é modelado pela distribuição Binomial

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

em que p é a probabilidade de sucesso.

IC para uma proporção p

Podemos utilizar o Teorema do Limite Central (TLC) para obter intervalos de confiança aproximados para a proporção p.
Estando em condições de aplicar o TLC, vem que

$$X \sim \text{Binomial}(n, p) \simeq \text{Normal}(np, np(1-p))$$

ou ainda

$$\hat{p} = \frac{X}{n} \simeq N(p, p(1-p)/n),$$

em que $\hat{p}$ é um estimador para p.

Nota: sugere-se a análise do Teorema de De Moivre-Laplace.

IC para uma proporção p

Centrando e reduzindo tem-se

$$Z = \frac{\hat{p} - p}{\sqrt{p(1-p)/n}} = \frac{\frac{X}{n} - p}{\sqrt{p(1-p)/n}} \underset{\text{aprox}}{\sim} N(0,1)$$

em que p é desconhecido.

Habitualmente, por simplicidade e porque se trabalha com uma amostra de grande dimensão (recorda-se que se invoca o TLC), usa-se a aproximação

$$\sqrt{p(1-p)/n} \approx \sqrt{\hat{p}(1-\hat{p})/n}.$$

IC para uma proporção p

Seja p a proporção de indivíduos com uma certa característica de interesse ("sucesso") numa população.

Um intervalo de confiança **aproximado** para p, a um nível de confiança $1-\alpha$, é dado por

$$\left(\hat{p} - z_{1-\alpha/2}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}, \hat{p} + z_{1-\alpha/2}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}\right)$$

em que $\hat{p} = X/n$ é a proporção de indivíduos com a característica de interesse ("sucesso") na amostra de dimensão n.

IC para uma proporção p – Exemplo

Adaptado de Hall et al. (2011), pp. 143–144.
(Sugere-se a consulta do livro para detalhes adicionais.)

Pretende-se um intervalo de confiança a 95% para a proporção de clientes que chegam antes da meia-noite a um bar. Com base numa amostra aleatória de 564 clientes verificou-se que 21 deles tinham chegado antes da meia-noite.

Uma vez que $z_{1-\frac{\alpha}{2}} = z_{1-\frac{0.05}{2}} = z_{1-0.025} = z_{0.975} \approx 1.96$ e $\widehat{p} = \frac{21}{564} \approx 0.037$, o IC aproximado pretendido é dado por

$$\left(0.037 - 1.96\sqrt{\frac{0.037(1-0.037)}{564}}, 0.037 + 1.96\sqrt{\frac{0.037(1-0.037)}{564}}\right),$$

ou seja,

$$IC_{95\%}(p) \approx (0.0214, 0.0526).$$

IC para $\mu_X - \mu_Y$ (pop. Normais; var. conhecidas; amostras independentes)

Intervalo de confiança para a diferença de médias $\mu_X - \mu_Y$ de duas populações Normais com variâncias σ_X^2 e σ_Y^2 conhecidas.

Dadas duas a.a. independentes

$$X_1, X_2, \ldots, X_n \quad \text{e} \quad Y_1, Y_2, \ldots, Y_m$$

provenientes de populações Normais $N(\mu_X, \sigma_X^2)$ e $N(\mu_Y, \sigma_Y^2)$, respectivamente, podemos considerar a estatística

$$\overline{X} - \overline{Y} \sim N\left(\mu_X - \mu_Y, \frac{\sigma_X^2}{n} + \frac{\sigma_Y^2}{m}\right)$$

que, após normalização, conduz à variável

$$Z = \frac{\overline{X} - \overline{Y} - (\mu_X - \mu_Y)}{\sqrt{\frac{\sigma_X^2}{n} + \frac{\sigma_Y^2}{m}}} \sim N(0,1).$$

IC para $\mu_X - \mu_Y$ (pop. Normais; var. conhecidas; amostras independentes)

Podemos assim construir um intervalo de confiança para $\mu_X - \mu_Y$:

$$P\left(-z_{1-\alpha/2} < Z < z_{1-\alpha/2}\right) = 1 - \alpha$$

$$\Leftrightarrow P\left(-z_{1-\alpha/2} < \frac{\overline{X} - \overline{Y} - (\mu_X - \mu_Y)}{\sqrt{\frac{\sigma_X^2}{n} + \frac{\sigma_Y^2}{m}}} < z_{1-\alpha/2}\right) = 1 - \alpha$$

$$\Leftrightarrow P\left(\overline{X} - \overline{Y} - z_{1-\alpha/2}\sqrt{\frac{\sigma_X^2}{n} + \frac{\sigma_Y^2}{m}} < \mu_X - \mu_Y <\right.$$

$$\left. < \overline{X} - \overline{Y} + z_{1-\alpha/2}\sqrt{\frac{\sigma_X^2}{n} + \frac{\sigma_Y^2}{m}}\right) = 1 - \alpha.$$

IC para $\mu_X - \mu_Y$ (pop. Normais; var. conhecidas; amostras independentes)

Um intervalo de confiança para a diferença de médias $\mu_X - \mu_Y$ de duas populações Normais com variâncias conhecidas σ_X^2 e σ_Y^2, obtido a partir de duas amostras independentes, a um nível de confiança $1-\alpha$, é dado por

$$\left(\overline{X} - \overline{Y} - z_{1-\alpha/2}\sqrt{\frac{\sigma_X^2}{n} + \frac{\sigma_Y^2}{m}}, \overline{X} - \overline{Y} + z_{1-\alpha/2}\sqrt{\frac{\sigma_X^2}{n} + \frac{\sigma_Y^2}{m}}\right).$$

IC para $\mu_X - \mu_Y$ (pop. Normais; var. desconhecidas; amostras indep.)

Intervalo de confiança para a diferença de médias $\mu_X - \mu_Y$ de duas populações Normais com variâncias desconhecidas.

Como as variâncias populacionais, σ_X^2 e σ_Y^2, são desconhecidas, existem duas situações a considerar:

1. assumir que as variâncias populacionais são iguais.[1]
   A variável fulcral tem distribuição *t de Student* com $n + m - 2$ graus de liberdade. **A exposição seguinte considerará apenas este caso.**
2. assumir que as variâncias populacionais não são iguais.[1]
   A variável fulcral tem uma distribuição assintótica *t de Student*, cujos graus de liberdade são determinados de forma aproximada. Sugere-se a consulta de Murteira et al. (2015), p. 437, para detalhes adicionais, incluindo a fórmula de cálculo dos graus de liberdade.

[1] Para averiguar a igualdade de variâncias populacionais, embora existam outros testes na literatura, recorrer-se-á ao Teste de Levene, que será introduzido no âmbito dos Testes de Hipóteses. Este teste está implementado no SPSS, na tabela **Independent Samples Test**, com a designação *Levene's Test for Equality of Variances*. Esta tabela será discutida no acetato 277.

IC para $\mu_X - \mu_Y$ (pop. Normais; var. desconhecidas; amostras indep.)

Assim, dadas duas a.a. independentes

$$X_1, X_2, \ldots, X_n \quad \text{e} \quad Y_1, Y_2, \ldots, Y_m$$

provenientes de populações Normais $N(\mu_X, \sigma^2)$ e $N(\mu_Y, \sigma^2)$ com a mesma variância, consideramos a estatística

$$\overline{X} - \overline{Y} \sim N\left(\mu_X - \mu_Y, \sigma^2\left(\frac{1}{n} + \frac{1}{m}\right)\right)$$

pelo que, centrando e reduzindo, obtemos a variável

$$Z = \frac{\overline{X} - \overline{Y} - (\mu_X - \mu_Y)}{\sigma\sqrt{\frac{1}{n} + \frac{1}{m}}} \sim N(0,1).$$

Como estimar σ tendo duas populações?

IC para $\mu_X - \mu_Y$ (pop. Normais; var. desconhecidas; amostras indep.)

Vamos estimá-lo sabendo que

$$S_X^2 = \frac{1}{n}\sum_{i=1}^{n}(X_i - \overline{X})^2 \text{ estima } \sigma^2$$

e que

$$S_Y^2 = \frac{1}{m}\sum_{i=1}^{m}(Y_i - \overline{Y})^2 \text{ estima } \sigma^2.$$

Intuitivamente, podia considerar-se um estimador para σ^2 dado pela média ponderada das variâncias anteriores. Todavia, utiliza-se o estimador

$$S_D^2 = \frac{(n-1)S_{c_X}^2 + (m-1)S_{c_Y}^2}{n+m-2},$$

pois é possível mostrar que este é um estimador centrado para σ^2.

IC para $\mu_X - \mu_Y$ (pop. Normais; var. desconhecidas; amostras indep.)

Recorde-se que

$$Z = \frac{\overline{X} - \overline{Y} - (\mu_X - \mu_Y)}{\sigma\sqrt{\frac{1}{n} + \frac{1}{m}}} \sim N(0,1)$$

pelo que, usando S_D para estimar σ e tendo em atenção a definição da t-Student, vem

$$T = \frac{\overline{X} - \overline{Y} - (\mu_X - \mu_Y)}{\sqrt{\frac{(n-1)S^2_{c_X} + (m-1)S^2_{c_Y}}{n+m-2}}\sqrt{\frac{1}{n} + \frac{1}{m}}} \sim t_{n+m-2}.$$

IC para $\mu_X - \mu_Y$ (pop. Normais; var. desconhecidas; amostras indep.)

Com base em T podemos construir o IC desejado:

$$P\left(-t_{1-\alpha/2,n+m-2} < T < t_{1-\alpha/2,n+m-2}\right) = 1-\alpha,$$

ou seja,

$$P\left(\overline{X} - \overline{Y} - t_d < \mu_X - \mu_Y < \overline{X} - \overline{Y} + t_d\right) = 1-\alpha,$$

com

$$t_d = \underbrace{t_{1-\alpha/2,n+m-2}}_{\text{consultar tabela}^*} \sqrt{\frac{1}{n} + \frac{1}{m}} \sqrt{\frac{(n-1)S_{c_X}^2 + (m-1)S_{c_Y}^2}{n+m-2}}.$$

* Ou, alternativamente, máquinas de calcular, folhas de cálculo e *software* estatístico.

IC para $\mu_X - \mu_Y$ (pop. Normais; var. desconhecidas; amostras indep.)

Um intervalo de confiança para a diferença de médias $\mu_X - \mu_Y$ de duas populações Normais com variâncias desconhecidas (consideradas iguais), obtido a partir de duas amostras independentes, a um nível de confiança $1-\alpha$, é dado por

$$\left(\overline{X} - \overline{Y} \mp \underbrace{t_{1-\alpha/2,n+m-2}}_{\text{tabelado}} \sqrt{\frac{1}{n} + \frac{1}{m}} \sqrt{\frac{(n-1)S_{c_X}^2 + (m-1)S_{c_Y}^2}{n+m-2}}\right).$$

IC para $\mu_X - \mu_Y$ (pop. Normais; var. desconhecidas; amostras indep.)

Pretende-se investigar o nível de remuneração de certa categoria profissional. Entre os resultados obtidos destacam-se: amostra de 250 homens ($\overline{x} = 338, s^2_{c_x} = 17.1$) e amostra de 150 mulheres ($\overline{y} = 310, s^2_{c_y} = 30.9$). Admitindo que os dados são provenientes de populações Normais e assumindo que as variâncias populacionais são iguais, construa um intervalo de confiança a 95% para a diferença de remuneração média entre sexos.

Sejam,
$X-$ "Remuneração de homens"; $Y-$ "Remuneração de mulheres".
Assim, $\mathrm{IC}_{95\%}(\mu_X - \mu_Y) =$
$= \left(338 - 310 \mp 1.96\sqrt{\frac{1}{250} + \frac{1}{150}}\sqrt{\frac{249\times 17.1 + 149\times 30.9}{250+150-2}}\right)$
$\approx (27.04, 28.96)$,

pois $t_{1-\alpha/2,n+m-2} = t_{0.975,398} \approx 1.96$.

(Tabelas: $t_{0.975,398} = 1.96$ (coluna ∞); SPSS: IDF.T(0.975,398) = 1.96594...)

IC para $\mu_X - \mu_Y$ (pop. Normais; var. desc.; amostras emparelhadas)

Intervalo de confiança para a diferença de médias $\mu_X - \mu_Y$ de duas populações Normais com variâncias desconhecidas – amostras emparelhadas.

Quando queremos comparar a localização central de duas populações com base em amostras dependentes, não sabemos especificar, em geral, qual a distribuição da diferença das médias amostrais.

Iremos considerar apenas a situação em que as amostras são dependentes na medida em que têm a mesma dimensão e cada observação X_i depende da observação Y_i, mas os pares (X_i, Y_i) e (X_j, Y_j) são independentes $(i \neq j)$. Este tipo de amostras denominam-se de amostras emparelhadas.

IC para $\mu_X - \mu_Y$ (pop. Normais; var. desc.; amostras emparelhadas)

O procedimento é o seguinte:
Dadas duas amostras aleatórias emparelhadas $(X_1, X_2, \ldots, X_n)$ e $(Y_1, Y_2, \ldots, Y_n)$ provenientes de populações Normais, consideram-se as diferenças

$$D_i = X_i - Y_i \sim N(\mu_D, \sigma_D^2),$$

onde μ_D é a diferença das médias das duas populações $(\mu_D = \mu_X - \mu_Y)$ e σ_D^2 representa a variância das diferenças D_i.
Assim, usamos a variável

$$T = \frac{\overline{D} - \mu_D}{S_{c_D}/\sqrt{n}} \sim t_{n-1},$$

onde S_{c_D} representa o desvio padrão amostral corrigido das diferenças.

IC para $\mu_X - \mu_Y$ (pop. Normais; var. desc.; amostras emparelhadas)

Construção do intervalo de confiança:

$$P\left(-t_{1-\alpha/2,n-1} < T < t_{1-\alpha/2,n-1}\right) = 1-\alpha$$

$$\Leftrightarrow P\left(-t_{1-\alpha/2,n-1} < \frac{\overline{D}-\mu_D}{S_{c_D}/\sqrt{n}} < t_{1-\alpha/2,n-1}\right) = 1-\alpha$$

$$\Leftrightarrow P\left(\overline{D} - t_{1-\alpha/2,n-1}\frac{S_{c_D}}{\sqrt{n}} < \mu_D < \overline{D} + t_{1-\alpha/2,n-1}\frac{S_{c_D}}{\sqrt{n}}\right) = 1-\alpha.$$

(Alguma semelhança com um IC construído anteriormente?)

IC para $\mu_X - \mu_Y$ (pop. Normais; var. desc.; amostras emparelhadas)

Um intervalo de confiança para a diferença de médias $\mu_D = \mu_X - \mu_Y$ de duas populações Normais com variâncias desconhecidas, obtido a partir de duas amostras emparelhadas, a um nível de confiança $1 - \alpha$, é dado por

$$\left(\overline{D} - t_{1-\alpha/2,n-1}\frac{S_{c_D}}{\sqrt{n}}, \overline{D} + t_{1-\alpha/2,n-1}\frac{S_{c_D}}{\sqrt{n}}\right).$$

IC para $\mu_X - \mu_Y$ (pop. Normais; var. desc.; amostras emparelhadas)

Pretende-se investigar os resultados de uma determinada dieta. Para tal, 10 indivíduos obesos foram submetidos a um regime dietético durante um período de 35 dias, tendo-se registado os pesos no início e no fim do período:

Início	95	110	98	104	80	86	92	91	92	84
Fim	92	109	94	100	81	80	84	88	90	79

Admitindo que os dados são provenientes de populações Normais e que os participantes tiveram estilos de vida semelhantes durante este período, construa um intervalo de confiança a 99% para a diferença de médias populacionais.

Considere-se,
$D-$ "Peso no início" $-$ "Peso no fim" do estudo.

IC para $\mu_X - \mu_Y$ (pop. Normais; var. desc.; amostras emparelhadas)

Assim,

Início	95	110	98	104	80	86	92	91	92	84
Fim	92	109	94	100	81	80	84	88	90	79
D	3	1	4	4	-1	6	8	3	2	5

e, com base na amostra das diferenças, calculam-se

$$\overline{d} = \frac{3+1+4+\ldots+2+5}{10} = 3.5$$

e

$$s_{c_D} = \sqrt{\frac{(3-3.5)^2+(1-3.5)^2+\ldots+(5-3.5)^2}{9}} \approx 2.55.$$

IC para $\mu_X - \mu_Y$ (pop. Normais; var. desc.; amostras emparelhadas)

Sabendo que
$t_{1-\frac{\alpha}{2},n-1} = t_{1-\frac{0.01}{2},9} = t_{0.995,9} \approx 3.25$, obtém-se

$$\mathrm{IC}_{99\%}(\mu_D) = \left(3.5 - 3.25\frac{2.55}{\sqrt{10}}, 3.5 + 3.25\frac{2.55}{\sqrt{10}}\right) \approx (0.88, 6.12).$$

Por curiosidade,

$$\mathrm{IC}_{90\%}(\mu_D) = \left(3.5 - 1.83\frac{2.55}{\sqrt{10}}, 3.5 + 1.83\frac{2.55}{\sqrt{10}}\right) \approx (2.02, 4.98);$$

$$\mathrm{IC}_{95\%}(\mu_D) = \left(3.5 - 2.26\frac{2.55}{\sqrt{10}}, 3.5 + 2.26\frac{2.55}{\sqrt{10}}\right) \approx (1.68, 5.32).$$

Compare as amplitudes dos três IC e comente.

IC para $\mu_X - \mu_Y$ (pop. Normais; var. desc.; amostras emparelhadas)

No SPSS, o $\mathrm{IC}_{99\%}(\mu_D)$ pode ser obtido de duas formas distintas:

Paired Samples Test

		Paired Differences					t	df	Sig. (2-tailed)
		Mean	Std. Deviation	Std. Error Mean	99% Confidence Interval of the Difference				
					Lower	Upper			
Pair 1	Início - Fim	3,50000	2,54951	,80623	,87990	6,12010	4,341	9	,002

One-Sample Test

	Test Value = 0					
	t	df	Sig. (2-tailed)	Mean Difference	99% Confidence Interval of the Difference	
					Lower	Upper
D	4,341	9	,002	3,50000	,8799	6,1201

(com a amostra das diferenças)

IC para $\mu_X - \mu_Y$ (pop. Normais; var. desc.; amostras emparelhadas)

No EXCEL (Office 2010 PT), entre outras possibilidades, o $\mathrm{IC}_{99\%}(\mu_D)$ pode ser obtido desta forma:

	A	B	C	D	E	F	G	H	I	J	K	L	M
1													
2		Início	95	110	98	104	80	86	92	91	92	84	
3		Fim	92	109	94	100	81	80	84	88	90	79	
4		D	3	1	4	4	-1	6	8	3	2	5	
5													
6						Média amostral:	3,50		=MÉDIA(C4:L4)				
7						Desvio padrão amostral corrigido:	2,55		=DESVPADA(C4:L4)				
8													
9						$IC_{99\%}(\mu_D)$=							
10							0,88		=G6-INT.CONFIANÇA.T(0,01;G7;10)				
11							6,12		=G6+INT.CONFIANÇA.T(0,01;G7;10)				
12													

IC para σ^2 (população Normal, μ conhecida)

Intervalo de confiança para a variância σ^2 de uma população Normal com média μ conhecida.

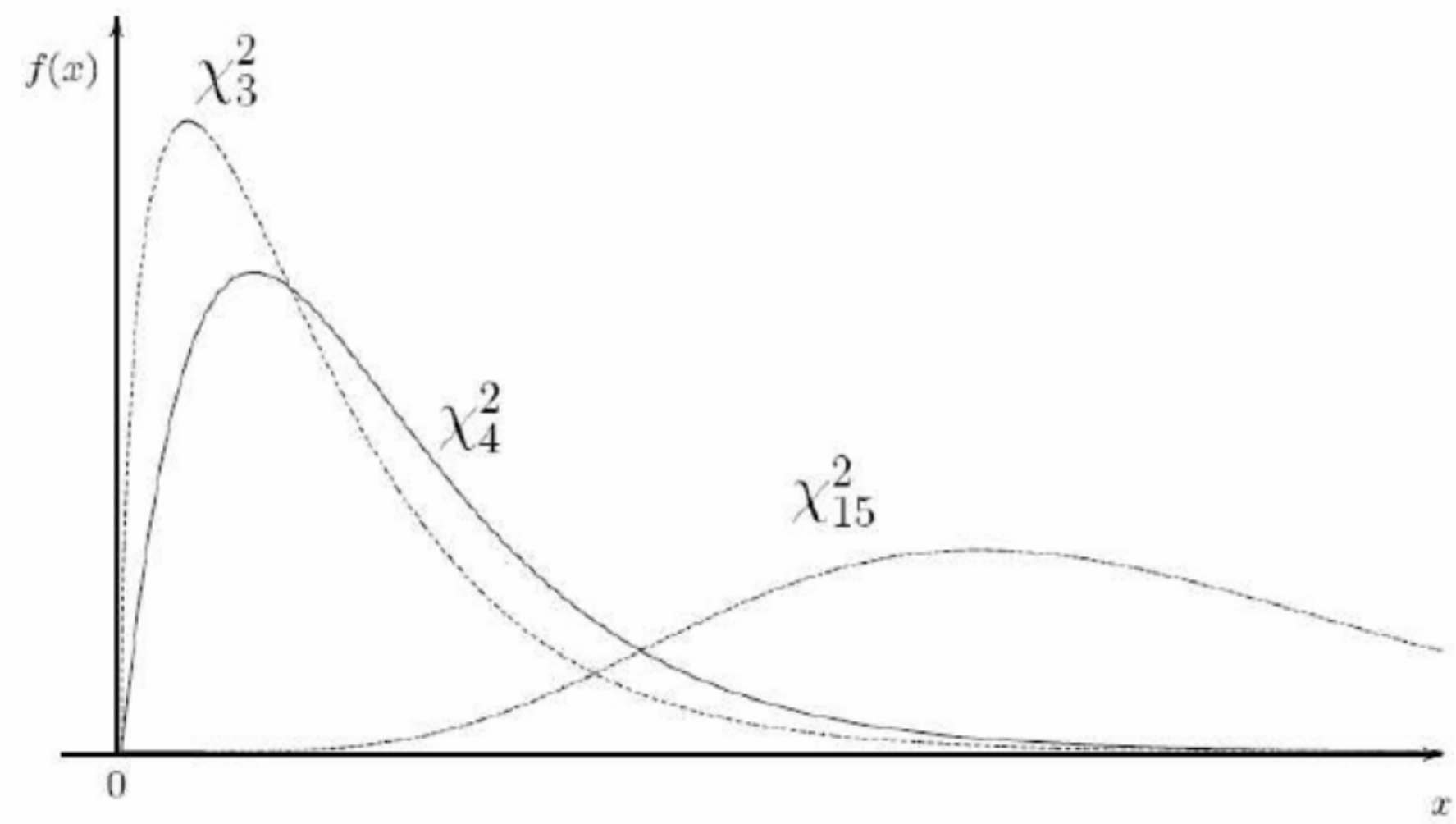
Prova-se que a seguinte v.a. V tem distribuição Qui-Quadrado com n graus de liberdade (χ^2_n):

$$V = \sum_{i=1}^{n} \left(\frac{X_i - \mu}{\sigma} \right)^2 = \frac{\sum_{i=1}^{n} (X_i - \mu)^2}{\sigma^2} \sim \chi^2_n.$$

Qual o aspeto da família de distribuições por amostragem χ^2_n?

IC para σ^2 (população Normal, μ conhecida)

Família de distribuições por amostragem χ^2_n:



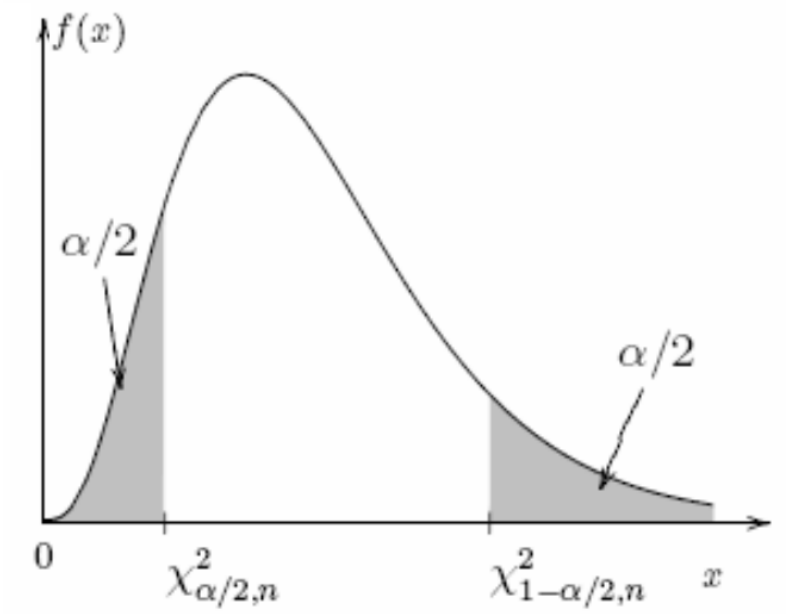
IC para σ^2 (população Normal, μ conhecida)

Retomando,

$$V = \sum_{i=1}^{n} \left(\frac{X_i - \mu}{\sigma} \right)^2 = \frac{\sum_{i=1}^{n} (X_i - \mu)^2}{\sigma^2} \sim \chi^2_n.$$

A distribuição está definida para $x > 0$ e não é simétrica, pelo que escolhemos dois quantis não simétricos e calculamos:

$$P(\chi^2_{\alpha/2,n} < V < \chi^2_{1-\alpha/2,n}) = 1 - \alpha.$$



IC para σ^2 (população Normal, μ conhecida)

E a construção do intervalo segue como habitualmente:

$$\mathrm{P}\left(\chi^2_{\alpha/2,n} < V < \chi^2_{1-\alpha/2,n}\right) = 1-\alpha$$

$$\Leftrightarrow \mathrm{P}\left(\chi^2_{\alpha/2,n} < \frac{\sum_{i=1}^n (X_i-\mu)^2}{\sigma^2} < \chi^2_{1-\alpha/2,n}\right) = 1-\alpha$$

$$\Leftrightarrow \mathrm{P}\left(\frac{\sum_{i=1}^n (X_i-\mu)^2}{\chi^2_{1-\alpha/2,n}} < \sigma^2 < \frac{\sum_{i=1}^n (X_i-\mu)^2}{\chi^2_{\alpha/2,n}}\right) = 1-\alpha.$$

IC para σ^2 (população Normal, μ conhecida)

Um intervalo de confiança para a variância σ^2 de uma população Normal com média μ conhecida, a um nível de confiança $1-\alpha$, é dado por

$$\left(\frac{\sum_{i=1}^{n}(X_i-\mu)^2}{\chi^2_{1-\alpha/2,n}}, \frac{\sum_{i=1}^{n}(X_i-\mu)^2}{\chi^2_{\alpha/2,n}} \right).$$

IC para σ (população Normal, μ conhecida)

Facilmente podemos obter um IC para o desvio padrão:

Um intervalo de confiança para o desvio padrão σ de uma população Normal com média μ conhecida, a um nível de confiança $1-\alpha$, é dado por

$$\left(\sqrt{\frac{\sum_{i=1}^{n}(X_i-\mu)^2}{\chi^2_{1-\alpha/2,n}}}, \sqrt{\frac{\sum_{i=1}^{n}(X_i-\mu)^2}{\chi^2_{\alpha/2,n}}}\right).$$

IC para σ^2 (população Normal, μ desconhecida)

Prova-se que

$$V = \frac{(n-1)S_c^2}{\sigma^2} \sim \chi^2_{n-1}.$$

Assim, procedendo como anteriormente, temos

$$P\left(\chi^2_{\alpha/2,n-1} < V < \chi^2_{1-\alpha/2,n-1}\right) = 1-\alpha$$

$$\Leftrightarrow P\left(\chi^2_{\alpha/2,n-1} < \frac{(n-1)S_c^2}{\sigma^2} < \chi^2_{1-\alpha/2,n-1}\right) = 1-\alpha$$

$$\Leftrightarrow P\left(\frac{(n-1)S_c^2}{\chi^2_{1-\alpha/2,n-1}} < \sigma^2 < \frac{(n-1)S_c^2}{\chi^2_{\alpha/2,n-1}}\right) = 1-\alpha.$$

IC para σ^2 e σ (população Normal, μ desconhecida)

Para a variância σ^2:

Um intervalo de confiança para a variância σ^2 de uma população Normal com média desconhecida, a um nível de confiança $1-\alpha$, é dado por

$$\left(\frac{(n-1)S_c^2}{\chi^2_{1-\alpha/2,n-1}}, \frac{(n-1)S_c^2}{\chi^2_{\alpha/2,n-1}}\right).$$

Para o desvio padrão σ:

Um intervalo de confiança para o desvio padrão σ de uma população Normal com média desconhecida, a um nível de confiança $1-\alpha$, é dado por

$$\left(\sqrt{\frac{(n-1)S_c^2}{\chi^2_{1-\alpha/2,n-1}}}, \sqrt{\frac{(n-1)S_c^2}{\chi^2_{\alpha/2,n-1}}}\right).$$

IC para o desvio padrão σ – Exercício

Suponha-se que o tempo de vida (em horas) de determinados componentes eletrónicos tem distribuição $\mathrm{N}(\mu, \sigma^2)$. Obtida uma amostra casual de $n = 20$ componentes da produção em determinado dia, calculou-se a média amostral, $\overline{x} = 1832$ horas, e o desvio padrão amostral corrigido dos tempos de vida, $s_c = 497$ horas. Determine um intervalo de confiança a 90% para o desvio padrão, σ, do tempo de vida dos componentes eletrónicos produzidos.

IC para o desvio padrão σ – Resolução

Sendo $1-\alpha = 0.90$, então $\alpha = 0.10$ e $\alpha/2 = 0.05$. Por outro lado, como $n-1 = 19$, consultando os quantis da χ^2_{19} na tabela, temos

$$\chi^2_{0.05,19} = 10.12 \quad \text{e} \quad \chi^2_{1-0.05,19} = 30.14,$$

pelo que,

$$IC_{90\%}(\sigma) = \left(\sqrt{\frac{19 \times 497^2}{30.14}}, \sqrt{\frac{19 \times 497^2}{10.12}}\right) \approx (394.6, 681.1).$$

SPSS: IDF.CHISQ(0.05,19) $\approx$ 10.12 e IDF.CHISQ(0.95,19) $\approx$ 30.14;
EXCEL (Office 2010 PT): INV.CHIQ(0,05;19) e INV.CHIQ(0,95;19).

Testes de Hipóteses (TH)

Uma hipótese estatística é uma conjectura sobre uma característica da população.

Um teste de hipóteses é um procedimento estatístico que averigua se os dados sustentam uma determinada hipótese.

Num teste de hipóteses há sempre duas hipóteses:

$$\text{Hipótese Nula} - H_0 \text{ vs. Hipótese alternativa} - H_1$$

Testes de Hipóteses (TH)

Exemplos:

1. $H_0 : \mu = 0$ vs. $H_1 : \mu \neq 0$; População Normal.
2. $H_0 : \mu = 3$ vs. $H_1 : \mu < 3$; População Normal.
3. $H_0 : \sigma^2 = 1$ vs. $H_1 : \sigma^2 > 1$; População Normal.
4. $H_0 : \mu = 4$ vs. $H_1 : \mu = 7$; População Normal.
5. $H_0 : \lambda = 5$ vs. $H_1 : \lambda \neq 5$; População Poisson.
6. $H_0 : \mu > 1$ vs. $H_1 : \mu \leq 1$; População Normal.
7. $H_0 : X \sim \textit{Poisson}$ vs. $H_1 : X \sim \text{outra distribuição}$.
8. $H_0 : X \sim \textit{Binomial}(10, 0.4)$ vs. $H_1 : X \sim \text{outra distribuição}$.

Testes de Hipóteses (TH)

Tipos de hipóteses: as várias hipóteses podem ser simples ou compostas. Uma hipótese simples apenas contempla uma possibilidade (sinal $=$).

Iremos apenas considerar testes em que H_0 é simples.

Tipos de testes: os testes podem ser unilaterais ou bilaterais.

Nos testes unilaterais, a hipótese alternativa apenas contempla possibilidades à direita ou à esquerda da hipótese nula. Nos testes bilaterais, a hipótese alternativa contempla possibilidades à direita e à esquerda da hipótese nula.

Testes de Hipóteses (TH)

Exemplos de testes unilaterais à direita:

$$H_0 : \mu = 1 \text{ vs. } H_1 : \mu > 1$$
$$H_0 : \mu = 1 \text{ vs. } H_1 : \mu = 7$$

Exemplos de testes unilaterais à esquerda:

$$H_0 : \mu = 1 \text{ vs. } H_1 : \mu < 1$$
$$H_0 : \mu = 4 \text{ vs. } H_1 : \mu = 2$$

Exemplo de um teste bilateral:

$$H_0 : \mu = 1 \text{ vs. } H_1 : \mu \neq 1$$

TH – Noções básicas

Estatística de teste: é uma estatística calculada a partir da amostra e que é usada para tomar a decisão acerca de rejeitar ou não a hipótese nula. Costuma representar-se por T, embora possa assumir outras designações, como Z, V ou ET.

Região de rejeição (RR) ou **região crítica** (RC): é o conjunto de valores da estatística de teste que conduzem à rejeição da hipótese nula.

Região de não rejeição: é o complementar da região de rejeição.

Se o valor observado da estatística de teste, t_{obs} (ou z_{obs}, v_{obs}, etc., em concordância com a representação da estatística de teste), pertencer à região crítica, rejeita-se H_0 a favor de H_1, isto é, acredita-se em H_1. Caso contrário, não se rejeita H_0 e, neste caso, não se acredita em H_1.

TH – Noções básicas

Naturalmente, pode acontecer que o valor observado para T pertença à região de rejeição, mas H_0 seja verdadeira. Isto pode acontecer porque T tem, em geral, probabilidade não nula de pertencer à região de rejeição, mesmo sendo válida H_0. Neste caso, rejeita-se indevidamente H_0.
Analogamente, pode acontecer que o valor observado para T esteja na região de não rejeição, mas H_1 é a hipótese verdadeira. Neste caso, não se rejeita H_0 indevidamente.
É impossível evitar completamente estas situações e tudo o que se pode fazer é planear o teste de forma a que estes erros ocorram com uma probabilidade reduzida.

TH – Tipos de Erros

Tipos de erros:

		Hipótese verdadeira	
		H_0	H_1
Decisão do teste	Rejeitar H_0	**Erro de Tipo I**	$\surd$
	Não rejeitar H_0	$\surd$	**Erro de Tipo II**

P(Erro de Tipo I) $= \alpha$

P(Erro de Tipo II) $= \beta$

TH – Nível de significância e Potência do teste

O ideal seria $\alpha = \beta = 0$. Como na prática isto é impossível, procura-se minimizar β após se ter fixado um valor de α reduzido.
Um bom teste deve ter ambos os valores de α e β próximos de zero.
Valores típicos para α são 0.001, 0.01, 0.05 ou 0.1.

Nível de significância (ou Tamanho do teste):

$$\alpha = P(\text{Erro de Tipo I}) = P(\text{rejeitar } H_0 | H_0 \text{ verdadeira}).$$

Potência do teste:

$$1 - \beta = 1 - P(\text{Erro de Tipo II}) = P(\text{rejeitar } H_0 | H_1 \text{ verdadeira}).$$

Nota: sugere-se a leitura dos Exemplos 3.4 e 3.5 em Hall et al. (2011), pp. 147–150, assim como as páginas iniciais do Cap. 8 de Murteira et al. (2015).

Cálculo do $p - value$

O $p - value$ do teste é a probabilidade de se observar um valor da estatística de teste tanto ou mais afastado do que o valor observado na amostra (t_{obs}), assumindo que H_0 é verdadeira.

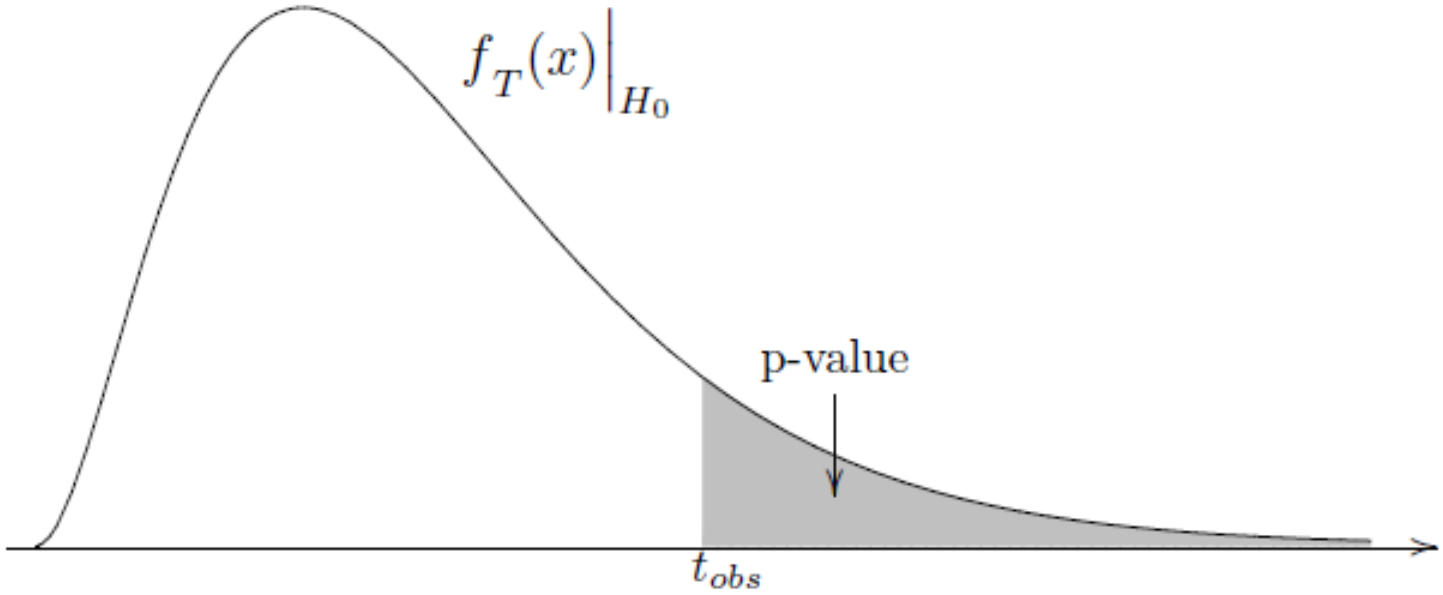
Ou seja, dito de outro modo, é a **probabilidade** de se obter este valor da estatística de teste ou outro mais desfavorável para a hipótese nula, admitindo, naturalmente, que H_0 é verdadeira. Pode dizer-se, assim, que o $p - value$ mede a evidência estatística que a amostra utilizada fornece em favor da hipótese nula.

Três casos podem ser ilustrados:

- a região de rejeição é da forma $T > c$;
- a região de rejeição é da forma $T < c$;
- a região de rejeição é da forma $T < c_1$ ou $T > c_2$.

Cálculo do $p - value$

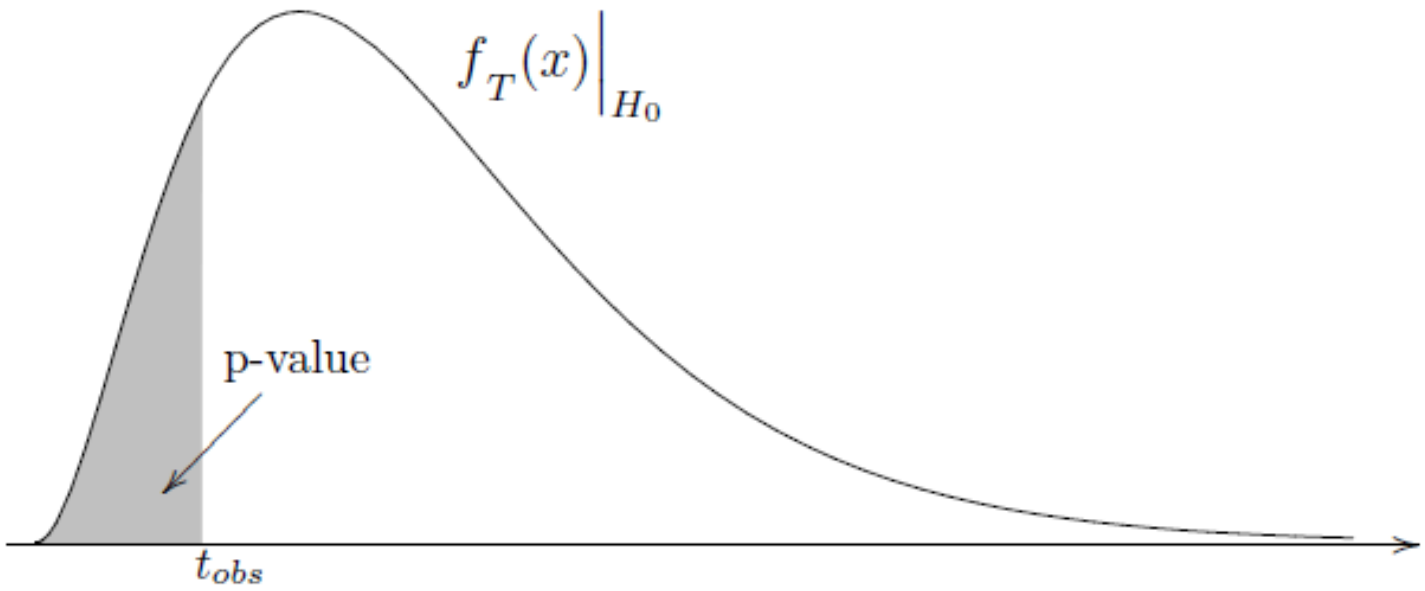
Quando a região de rejeição é da forma $T > c$ (rejeitar para valores elevados da estatística de teste), o $p - value$ é igual a $P(T > t_{\text{obs}}|H_0)$. Graficamente, por exemplo, obtém-se



(Consultar Hall et al. (2011), p. 151.)

Cálculo do $p - value$

Quando a região de rejeição é da forma $T < c$ (rejeitar para valores reduzidos da estatística de teste), o $p - value$ é igual a $P(T < t_{\text{obs}}|H_0)$. Graficamente, por exemplo, obtém-se



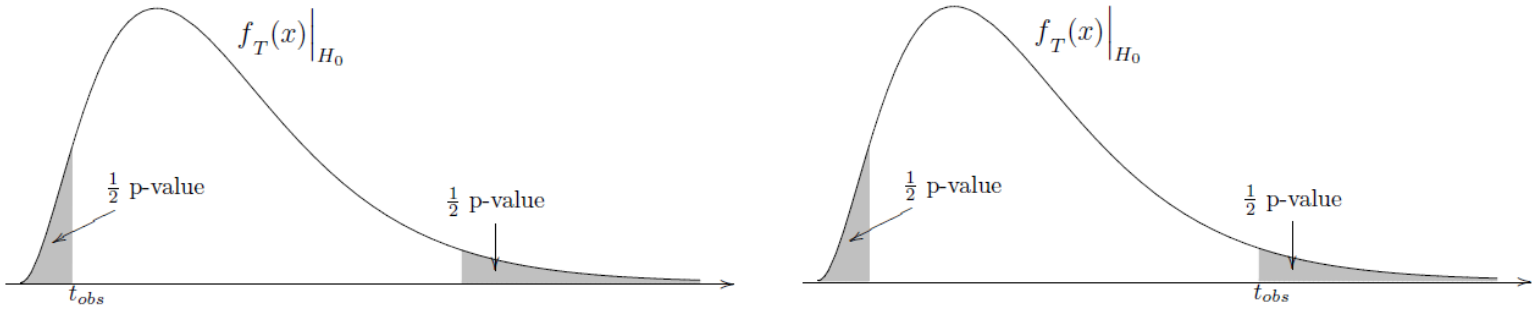
(Consultar Hall et al. (2011), p. 152.)

Cálculo do $p - value$

Quando a região de rejeição é da forma $T < c_1$ ou $T > c_2$ (com igual probabilidade para os dois casos), o $p - value$ é igual a

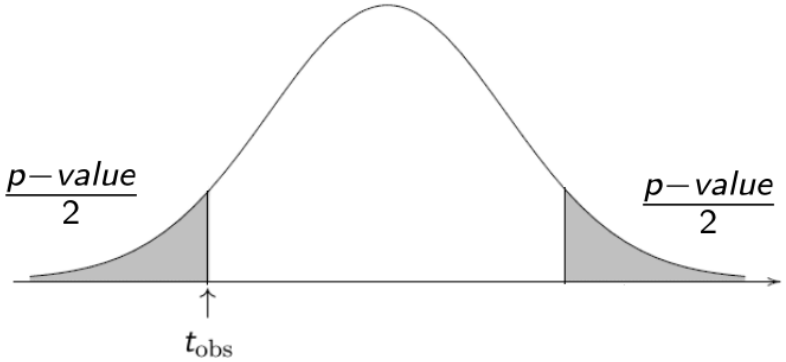
- $2\,P(T < t_{\text{obs}}|H_0)$, se t_{obs} for reduzido;
- $2\,P(T > t_{\text{obs}}|H_0)$, se t_{obs} for elevado.

Dizer que t_{obs} é reduzido (elevado) significa dizer que a estimativa que se obtém para o parâmetro a testar é inferior (superior) ao valor especificado em H_0.
Graficamente, por exemplo, obtém-se

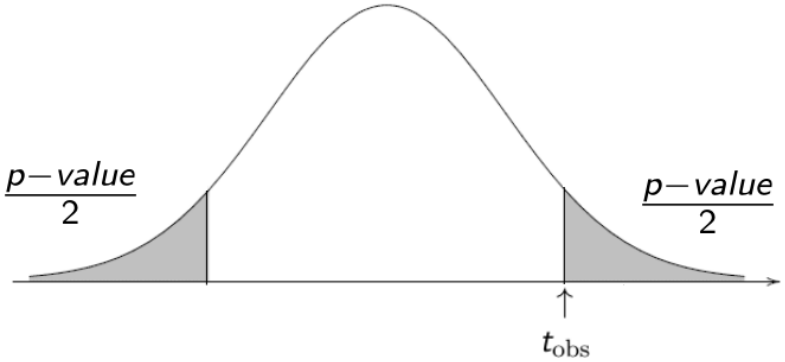


(Consultar Hall et al. (2011), pp. 152–153.)

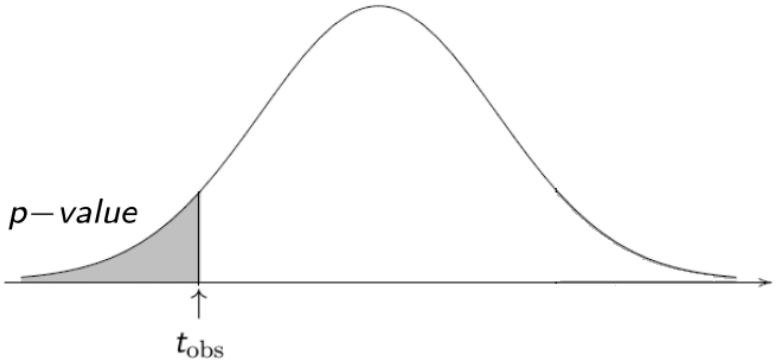
Cálculo do $p - value$ (resumo)



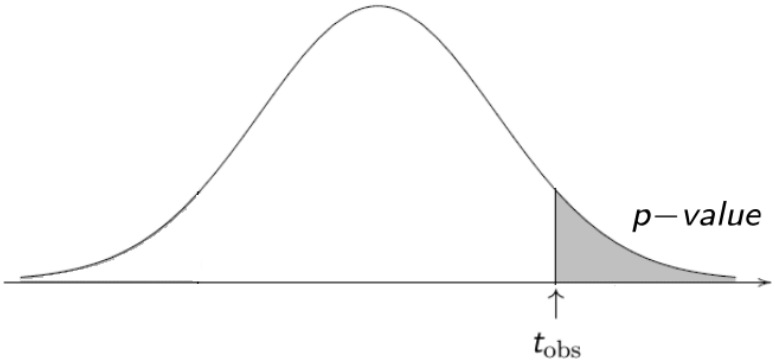
$p\text{-}value = 2\,P\left(T < t_{\text{obs}} | H_0\right)$



$p\text{-}value = 2\,P\left(T > t_{\text{obs}} | H_0\right)$



$p\text{-}value = P\left(T < t_{\text{obs}} | H_0\right)$



$p\text{-}value = P\left(T > t_{\text{obs}} | H_0\right)$

Transformação de *p-values* bilaterais em unilaterais

Frequentemente, o software estatístico fornece apenas o *p-value* bilateral, mas o interesse do investigador reside num teste unilateral. Assim, neste caso, é necessário converter o *p-value* bilateral fornecido pelo software no *p-value* unilateral apropriado.

- Se a(s) amostra(s) aponta(m) no sentido da hipótese alternativa, então
$$\textit{p-value}_{\text{unilateral}} = \frac{\textit{p-value}_{\text{bilateral}}}{2}.$$
- Se a(s) amostra(s) não aponta(m) no sentido da hipótese alternativa, então
$$\textit{p-value}_{\text{unilateral}} = 1 - \frac{\textit{p-value}_{\text{bilateral}}}{2}.$$

Uso do $p - value$

Uso do $p - value$

Procedimentos para a realização de um TH: com base na região de rejeição

Genericamente:

1. Identificar o parâmetro de interesse e especificar as hipóteses H_0 e H_1;

2. Escolher uma estatística de teste, T, com distribuição conhecida (admitindo que H_0 é verdadeira);

3. Determinar a região de rejeição ou região crítica (RC);

4. Calcular t_{obs} (valor que T assume para os dados observados);

5. Tomar uma decisão: rejeitar H_0 se $t_{\text{obs}} \in$ RC;

6. Concluir.

Procedimentos para a realização de um TH: com base no intervalo de confiança

Nota: no âmbito desta disciplina, este procedimento é válido somente para testes bilaterais.

Genericamente:

1. Identificar o parâmetro de interesse e especificar as hipóteses H_0 e H_1;

2. Construir o intervalo de confiança apropriado para o parâmetro;

3. Tomar uma decisão: rejeitar H_0 se o valor do parâmetro especificado em H_0 não pertencer ao intervalo de confiança. Ou seja, o intervalo de confiança fornece uma região de não rejeição do teste;

4. Concluir.

Procedimentos para a realização de um TH: com base no *p-value*

Genericamente:

1. Identificar o parâmetro de interesse e especificar H_0 e H_1;
2. Escolher uma estatística de teste, T, com distribuição conhecida (admitindo que H_0 é verdadeira);
3. Identificar somente a forma da região de rejeição;
4. Calcular t_{obs} (valor que T assume para os dados observados);
5. Calcular o *p-value* do teste;
6. Tomar uma decisão: rejeitar H_0 se *p-value* $\leq \alpha$;
7. Concluir.

Exemplo com os três procedimentos

Adaptado de Hall et al. (2011), pp. 154–155.
(Sugere-se a consulta do livro para detalhes adicionais.)

No controlo de qualidade de um medicamento, pretende-se testar

$$H_0 : \mu = 20 \quad vs. \quad H_1 : \mu \neq 20$$

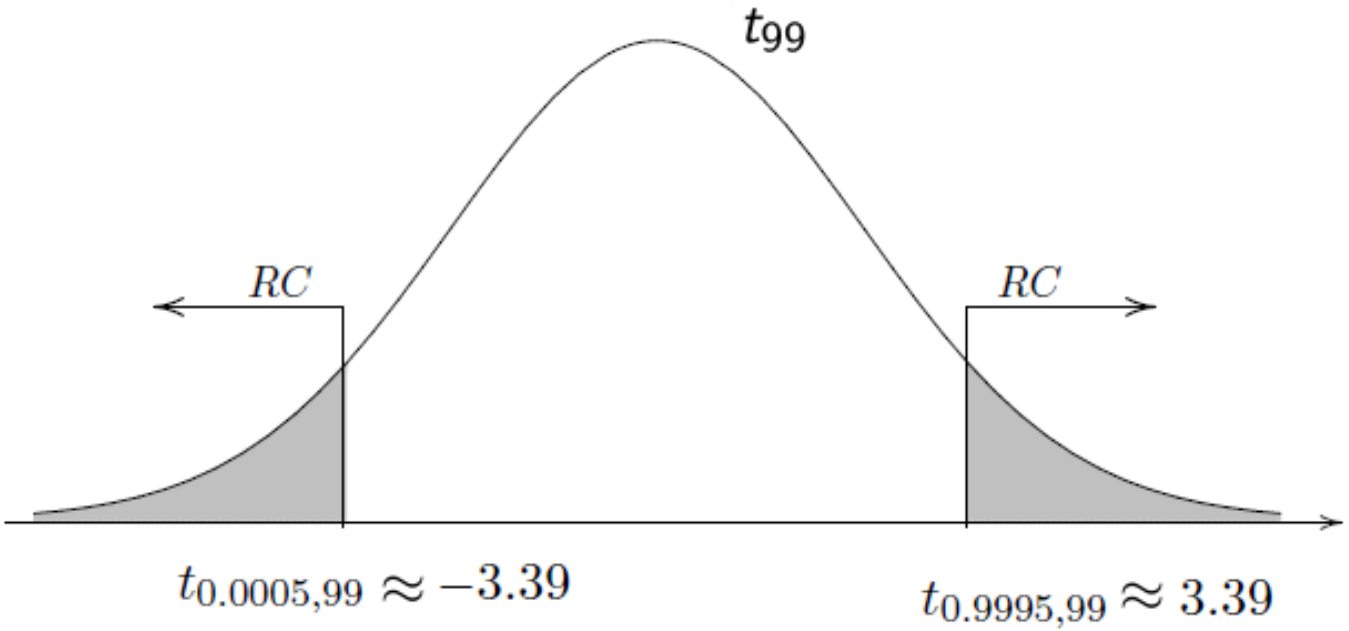
onde μ representa a verdadeira média da quantidade (em mg) de substância activa (fluoxetina) em cada cápsula do medicamento. Com uma certa amostra aleatória de 100 cápsulas, obteve-se $\overline{x} = 20.04$ e $s_c = 0.2$. Considera-se $\alpha = 0.001$. Assume-se uma população Normal.

Estatística de teste:

$$T = \frac{\overline{X} - \mu_0}{\frac{S_c}{\sqrt{n}}} \underset{sob\ H_0}{\sim} t_{n-1}.$$

Exemplo com os três procedimentos

Região de rejeição:



$$t_{1-\frac{\alpha}{2},n-1} = t_{1-\frac{0.001}{2},99} = t_{1-0.0005,99} = t_{0.9995,99}$$

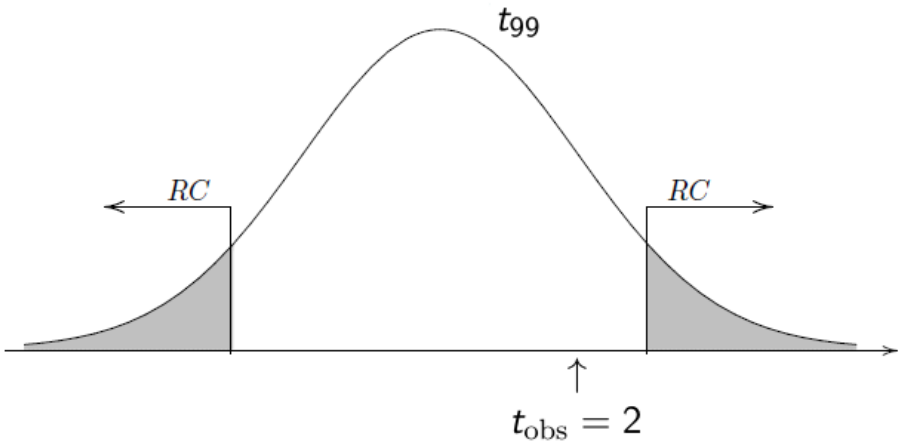
SPSS: IDF.T(0.9995,99) $= 3.3915\ldots$

Exemplo com os três procedimentos

Estatística de teste observada:

$$t_{\text{obs}} = \frac{20.04 - 20}{\frac{0.2}{\sqrt{100}}} = 2.$$

Como t_{obs} não pertence à região de rejeição, não se rejeita H_0.



Com base nesta amostra e ao nível de significância de 0.1%, não há evidência para considerar que a quantidade média de fluoxetina seja significativamente diferente de 20mg por cápsula, pelo que a produção do medicamento pode continuar.

Exemplo com os três procedimentos

Procedimento baseado no *p-value*:
como a estimativa que se obtém para o parâmetro a testar ($\overline{x} = 20.04$) é superior ao valor especificado em H_0 ($\mu = 20$), considera-se t_{obs} elevado, pelo que

$$p\text{-}value = 2\,P(T > t_{\text{obs}}|H_0) = 2\,P(T > 2|H_0),$$

ou seja, recorrendo ao SPSS,

$$p\text{-}value = 2(1 - P(T \leq 2)) = 2 * (1 - \text{CDF.T}(2, 99)) \approx 0.048.$$

Como $p\text{-}value \approx 0.048 > \alpha = 0.001$, não se rejeita H_0.

A conclusão repete-se: com base nesta amostra e ao nível de significância de 0.1%, não há evidência para assumir que a quantidade média de fluoxetina seja significativamente diferente de 20mg por cápsula, pelo que a produção do medicamento pode continuar.

Exemplo com os três procedimentos

Procedimento baseado no IC para a média:

$$\mathrm{IC}_{99.9\%}(\mu) = \left(20.04 - 3.39\frac{0.2}{\sqrt{100}},\ 20.04 + 3.39\frac{0.2}{\sqrt{100}}\right)$$

$$\approx (19.97, 20.11).$$

Como o valor do parâmetro especificado em H_0 pertence ao intervalo de confiança, $20 \in \mathrm{IC}_{99.9\%}(\mu)$, não se rejeita H_0.

A conclusão repete-se: com base nesta amostra e ao nível de significância de 0.1%, não há evidência para assumir que a quantidade média de fluoxetina seja significativamente diferente de 20mg por cápsula, pelo que a produção do medicamento pode continuar.

Normalidade populacional?

Averiguar se determinada amostra provém de uma **população modelada por uma distribuição Normal**.

- Teste de Kolmogorov-Smirnov com correcção de Lilliefors (amostras grandes; o que é uma "amostra grande"?);
- Teste de Shapiro-Wilk (amostras pequenas).

Tests of Normality

	Kolmogorov-Smirnov[a]			Shapiro-Wilk		
	Statistic	df	Sig.	Statistic	df	Sig.
pesos	,075	100	,177	,989	100	,576

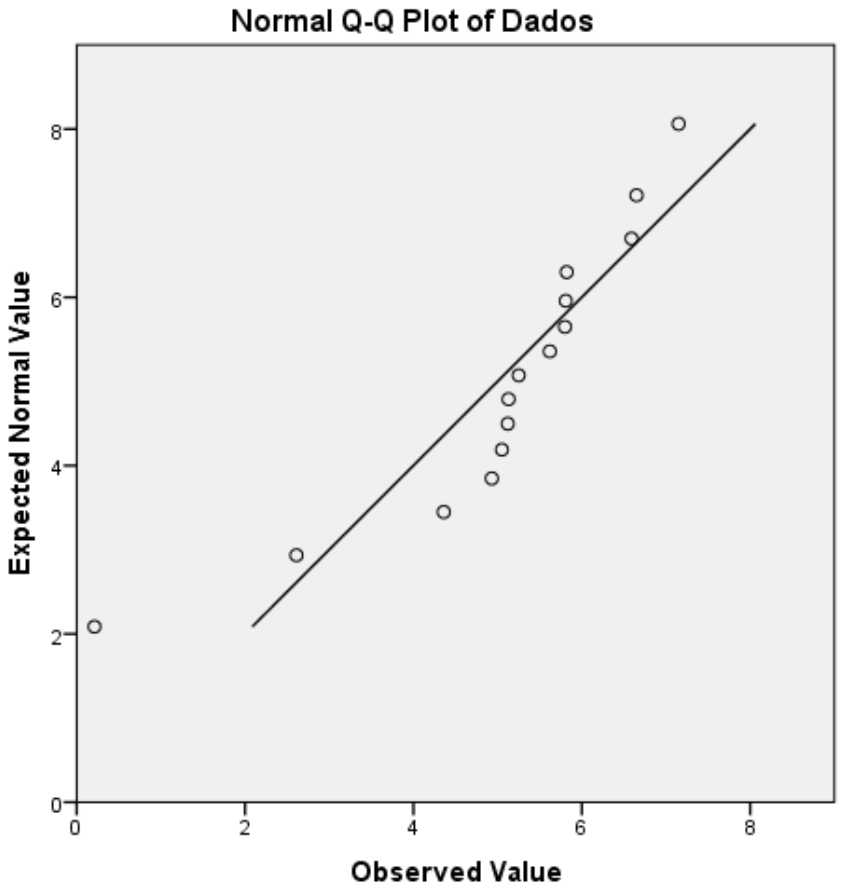
a. Lilliefors Significance Correction

```
Analyze/Descriptive Statistics/Explore/.../Normality plots with tests
```

Nota: atenção às possíveis fragilidades destes testes. Em inúmeras situações, a análise destes testes deverá ser acompanhada de outras análises estatísticas que ajudem a comprovar as suas indicações.

Normalidade populacional?

Q-Q Plot



`Analyze/Descriptive Statistics/Q-Q Plots/...`

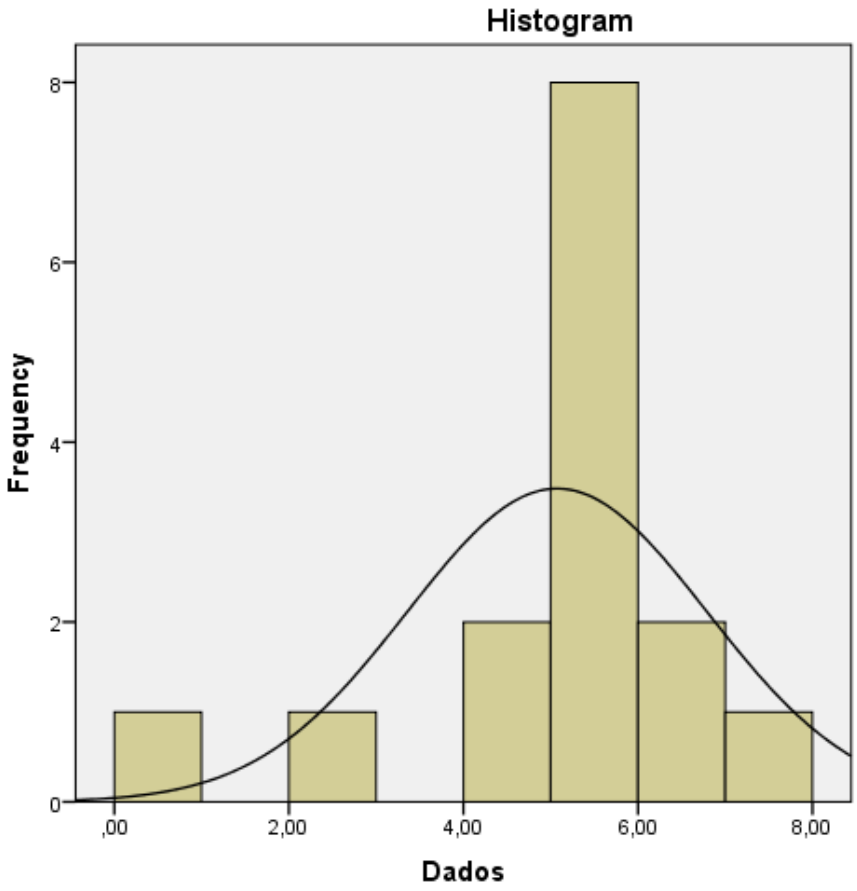
Normalidade populacional?

Confronto de diferente informação

Statistics

Dados

N	Valid	15
	Missing	0
Skewness		-1,831
Std. Error of Skewness		,580
Kurtosis		4,150
Std. Error of Kurtosis		1,121



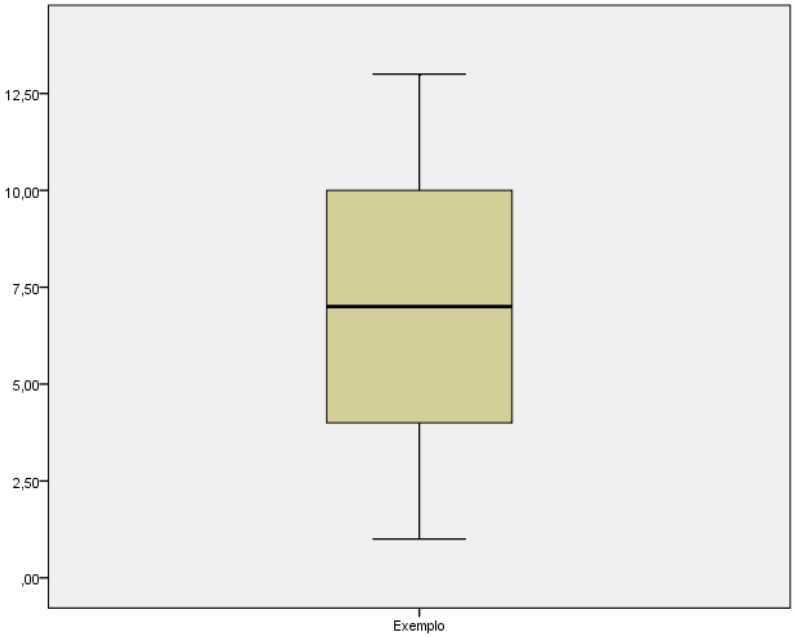
Tests of Normality

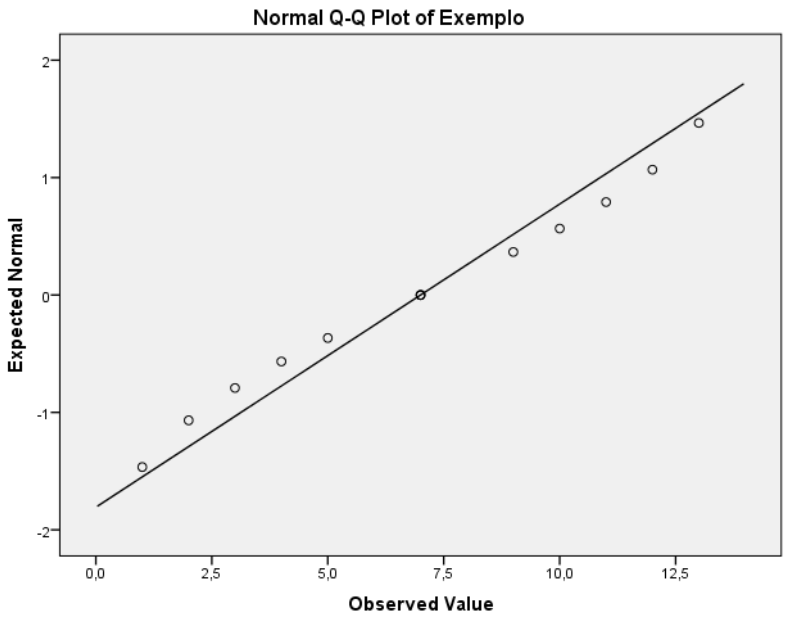
	Kolmogorov-Smirnov[a]			Shapiro-Wilk		
	Statistic	df	Sig.	Statistic	df	Sig.
Dados	,267	15	,005	,819	15	,007

a. Lilliefors Significance Correction

Normalidade populacional?

Confronto de diferente informação





Tests of Normality

	Kolmogorov-Smirnov[a]			Shapiro-Wilk		
	Statistic	df	Sig.	Statistic	df	Sig.
Exemplo	,115	13	,200*	,961	13	,761

*. This is a lower bound of the true significance.

a. Lilliefors Significance Correction

Normalidade populacional?

Em **MATLAB**, por exemplo, consultar a documentação:

Q-Q Plots
`>> doc qqplot`

Alguns testes de ajustamento
`>> doc lillietest`
`>> doc adtest` (Teste de Anderson-Darling)
`>> doc kstest`
`>> doc jbtest` (Teste de Jarque-Bera)

Em **R**, por exemplo, consultar a documentação:

`> help(qqnorm)` e, adicionalmente, `> help(qqline)`
`> help(shapiro.test)`
`> help(ks.test)`

TH em populações Normais

Teste para a média, μ, com variância, σ^2, conhecida.

$$H_0 : \mu = \mu_0 \quad vs. \quad H_1 : \mu \begin{array}{c} < \\ \neq \\ > \end{array} \mu_0$$

Estatística de teste: $T = \dfrac{\overline{X} - \mu_0}{\frac{\sigma}{\sqrt{n}}} \underset{sob\ H_0}{\sim} N(0,1)$.

Região de rejeição:

- se $H_1 : \mu \neq \mu_0$, $RC_\alpha = \{t : |t| > z_{1-\frac{\alpha}{2}}\}$;
- se $H_1 : \mu < \mu_0$, $RC_\alpha = \{t : t < z_\alpha\}$;
- se $H_1 : \mu > \mu_0$, $RC_\alpha = \{t : t > z_{1-\alpha}\}$.

Pressupostos exigidos:

- Observações independentes retiradas da mesma população (amostra aleatória);
- População com distribuição Normal; variância, σ^2, conhecida.

TH em populações Normais

Teste para a média, μ, com variância, σ^2, desconhecida.

$$H_0 : \mu = \mu_0 \quad vs. \quad H_1 : \mu \quad \begin{matrix} < \\ \neq \\ > \end{matrix} \quad \mu_0$$

Estatística de teste: $T = \dfrac{\overline{X} - \mu_0}{\frac{S_c}{\sqrt{n}}} \underset{sob\ H_0}{\sim} t_{n-1}$.

Região de rejeição:

- se $H_1 : \mu \neq \mu_0$, $RC_\alpha = \{t : |t| > t_{1-\frac{\alpha}{2},\, n-1}\}$;
- se $H_1 : \mu < \mu_0$, $RC_\alpha = \{t : t < t_{\alpha,\, n-1}\}$;
- se $H_1 : \mu > \mu_0$, $RC_\alpha = \{t : t > t_{1-\alpha,\, n-1}\}$.

Pressupostos exigidos:

- Observações independentes retiradas da mesma população (amostra aleatória);
- População com dist. Normal; variância, σ^2, desconhecida.

Exemplo

Para equipar as suas instalações, um serviço adquiriu lâmpadas da marca A aceitando a afirmação do fabricante de que a vida média das lâmpadas é de 1650 horas. Numa experiência realizada com 50 lâmpadas escolhidas aleatoriamente, obteve-se $\overline{x} = 1575$ horas e $s_c = 120$ horas. Admitindo que os tempos de vida são modelados por uma distribuição Normal, pode admitir-se correta a afirmação do fabricante? (Considere $\alpha = 5\%$.)

Teste de hipóteses pretendido:

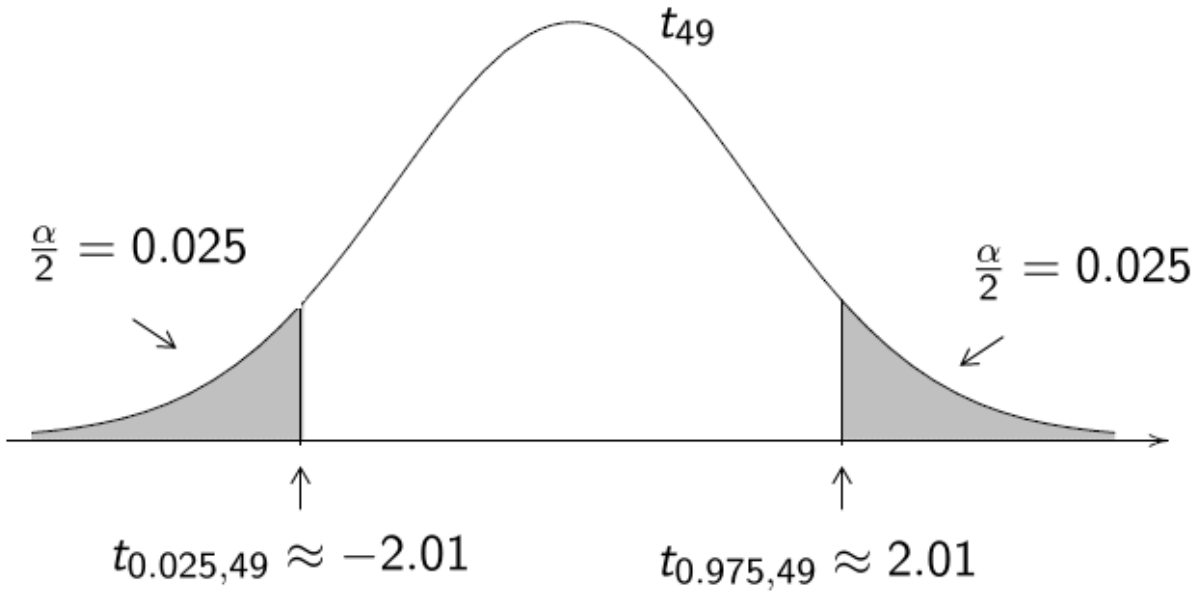
$$H_0 : \mu = 1650 \quad vs. \quad H_1 : \mu \neq 1650$$

Estatística de teste: $T = \dfrac{\overline{X} - \mu_0}{\frac{S_c}{\sqrt{n}}} \underset{sob\ H_0}{\sim} t_{n-1}$.

Região de rejeição:
como $H_1 : \mu \neq 1650$, então $RC_\alpha = \left\{t : |t| > t_{1-\frac{\alpha}{2},\, n-1}\right\}$.

Exemplo

Região crítica ou região de rejeição:



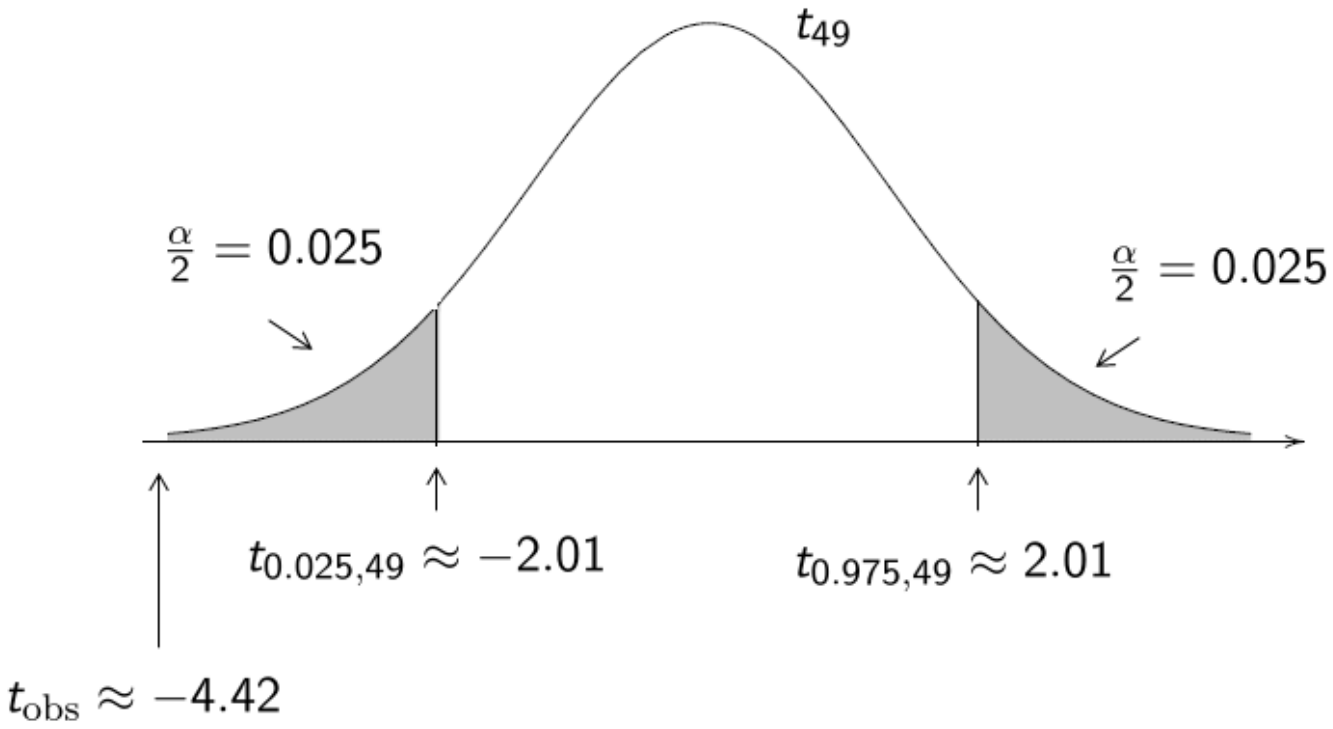
$$t_{1-\frac{\alpha}{2},n-1} = t_{1-\frac{0.05}{2},49} = t_{1-0.025,49} = t_{0.975,49} \approx 2.01$$

SPSS: IDF.T(0.975,49) $= 2.00957\ldots$
Tabelas: $t_{0.975,49} \approx 2.01$ (considerando 50 g.l.)

Exemplo

Estatística de teste observada:

$$t_{\text{obs}} = \frac{1575 - 1650}{\frac{120}{\sqrt{50}}} \approx -4.42.$$



Exemplo

Decisão: como $t_{\text{obs}} \in RC_{5\%}$, rejeita-se H_0.

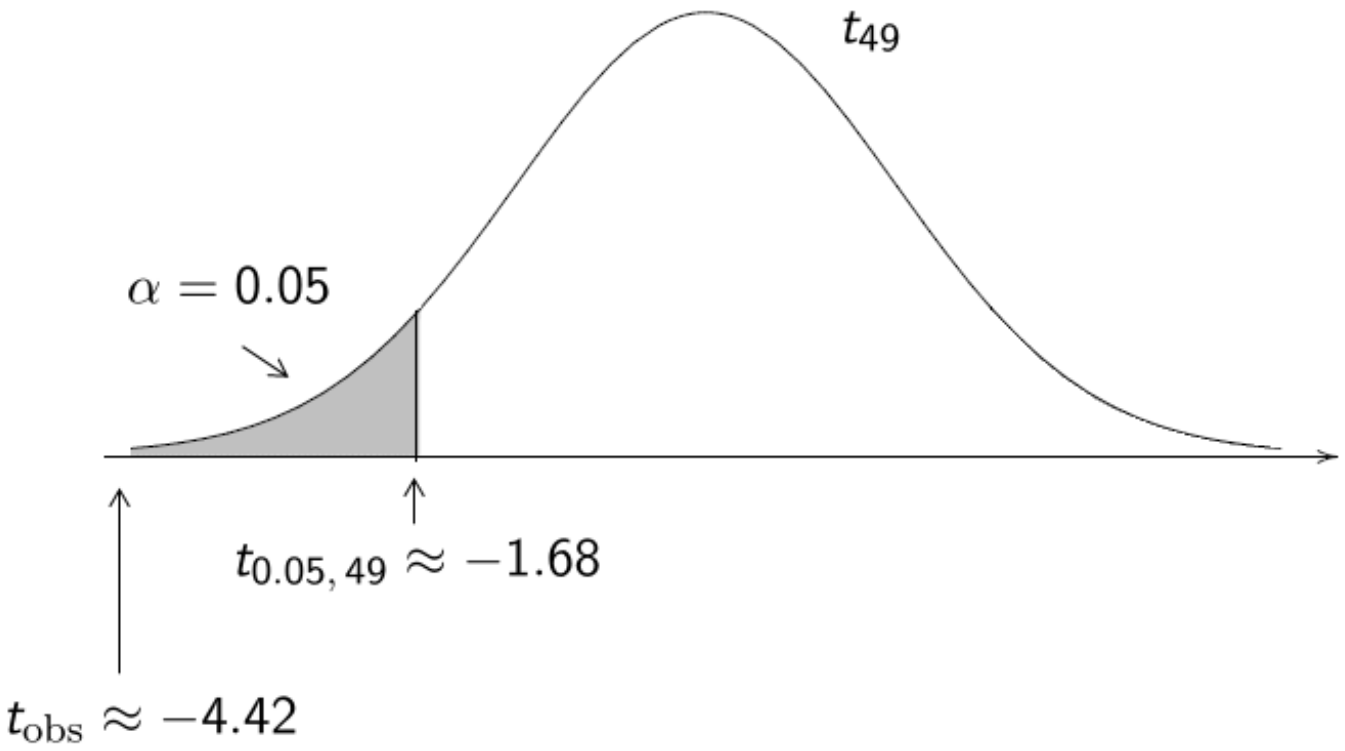
Conclusão: ao nível de significância de 5% e com base na informação desta amostra, há evidência estatística para considerar que o tempo de vida médio das lâmpadas deste fabricante é diferente de 1650 horas, ou seja, que a afirmação do fabricante é falsa.

Uma vez que H_0 foi rejeitada e $\overline{x} = 1575$, é pertinente testar

$$H_0 : \mu = 1650 \quad vs. \quad H_1 : \mu < 1650.$$

A estatística de teste permanece inalterada, assim como a estatística de teste observada. Todavia, a região crítica necessita de ser adaptada:

Exemplo



SPSS: IDF.T(0.95,49) $= 1.67655\ldots$
Tabelas: $t_{0.95,49} \approx 1.68$ (considerando 50 g.l.)

Exemplo

Decisão: como $t_{\text{obs}} \in RC_{5\%}$, rejeita-se H_0.

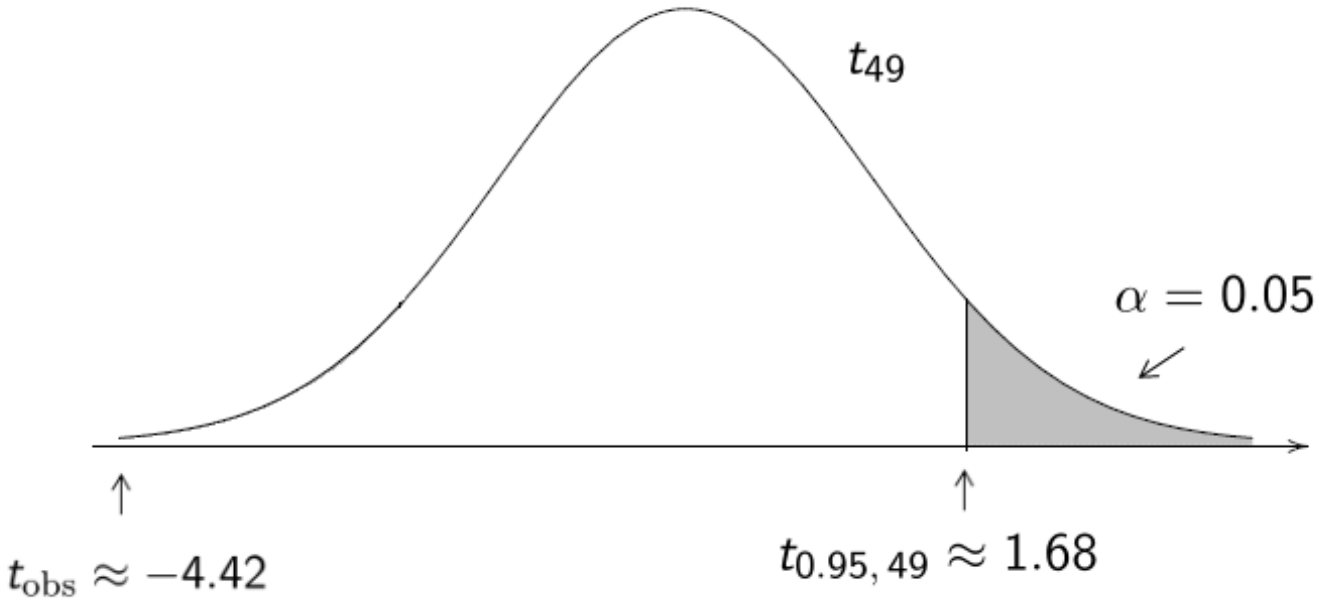
Conclusão: com base nesta amostra e ao nível de significância de 5%, há evidência estatística para suportar a conjectura de que o tempo de vida médio das lâmpadas deste fabricante é inferior a 1650 horas.

Embora desprovido de sentido, admita-se o seguinte teste

$$H_0 : \mu = 1650 \quad vs. \quad H_1 : \mu > 1650.$$

Tal como anteriormente, a estatística de teste permanece inalterada, assim como a estatística de teste observada.
Todavia, a região crítica tem de ser novamente ajustada:

Exemplo



Decisão: como $t_{\text{obs}} \notin RC_{5\%}$, não se rejeita H_0.

Conclusão: com base nesta amostra e ao nível de significância de 5%, não há evidência estatística para considerar que o tempo de vida médio das lâmpadas deste fabricante seja superior a 1650 horas.

TH em populações Normais

Teste para a variância, σ^2, com média, μ, desconhecida.

$$H_0 : \sigma^2 = \sigma_0^2 \quad vs. \quad H_1 : \sigma^2 \begin{array}{c} < \\ \neq \\ > \end{array} \sigma_0^2$$

Estatística de teste: $T = \dfrac{(n-1)S_c^2}{\sigma_0^2} \underset{sob\ H_0}{\sim} \chi^2_{n-1}$.

Região de rejeição:

- se $H_1 : \sigma^2 \neq \sigma_0^2$, $RC_\alpha = \{t : t < \chi^2_{\frac{\alpha}{2},\, n-1} \text{ ou } t > \chi^2_{1-\frac{\alpha}{2},\, n-1}\}$;
- se $H_1 : \sigma^2 < \sigma_0^2$, $RC_\alpha = \{t : t < \chi^2_{\alpha,\, n-1}\}$;
- se $H_1 : \sigma^2 > \sigma_0^2$, $RC_\alpha = \{t : t > \chi^2_{1-\alpha,\, n-1}\}$.

Pressupostos exigidos:

- Observações independentes retiradas da mesma população (amostra aleatória);
- População com dist. Normal; média, μ, desconhecida.

Exemplo

Adaptado de Hall et al. (2011), pp. 163–164.
(Sugere-se a consulta do livro para detalhes adicionais.)

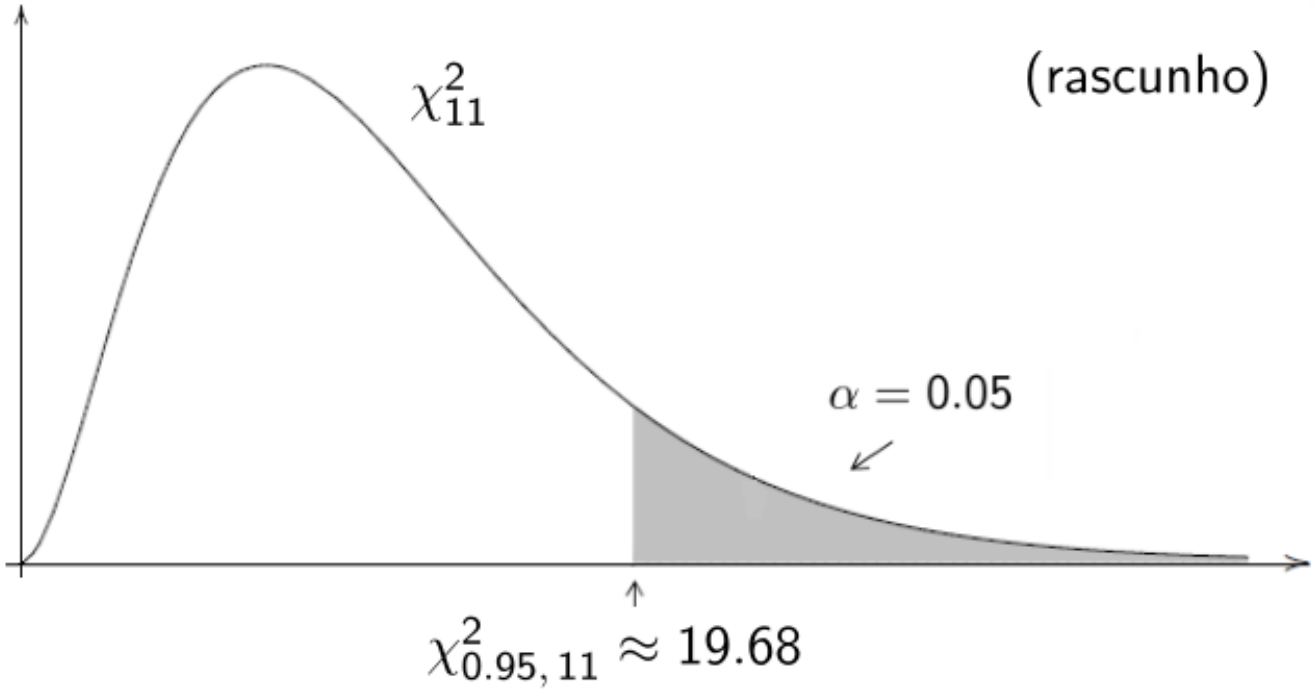
Numa determinada empresa é desejável que a variância da faturação diária não seja superior a 100 Euros. Com base numa amostra aleatória de 12 dias de laboração obteve-se $s_c^2 = 105$ Euros e $\overline{x} = 635$ Euros. Assumindo que as observações são provenientes de uma população Normal, ao nível de significância de 5%, teste

$$H_0 : \sigma^2 = 100 \quad vs. \quad H_1 : \sigma^2 > 100.$$

Estatística de teste: $T = \dfrac{(n-1)S_c^2}{\sigma_0^2} \underset{sob\ H_0}{\sim} \chi^2_{n-1}$.

Região de rejeição:
como $H_1 : \sigma^2 > \sigma_0^2$, então $RC_\alpha = \{t : t > \chi^2_{1-\alpha,\, n-1}\}$.

Exemplo

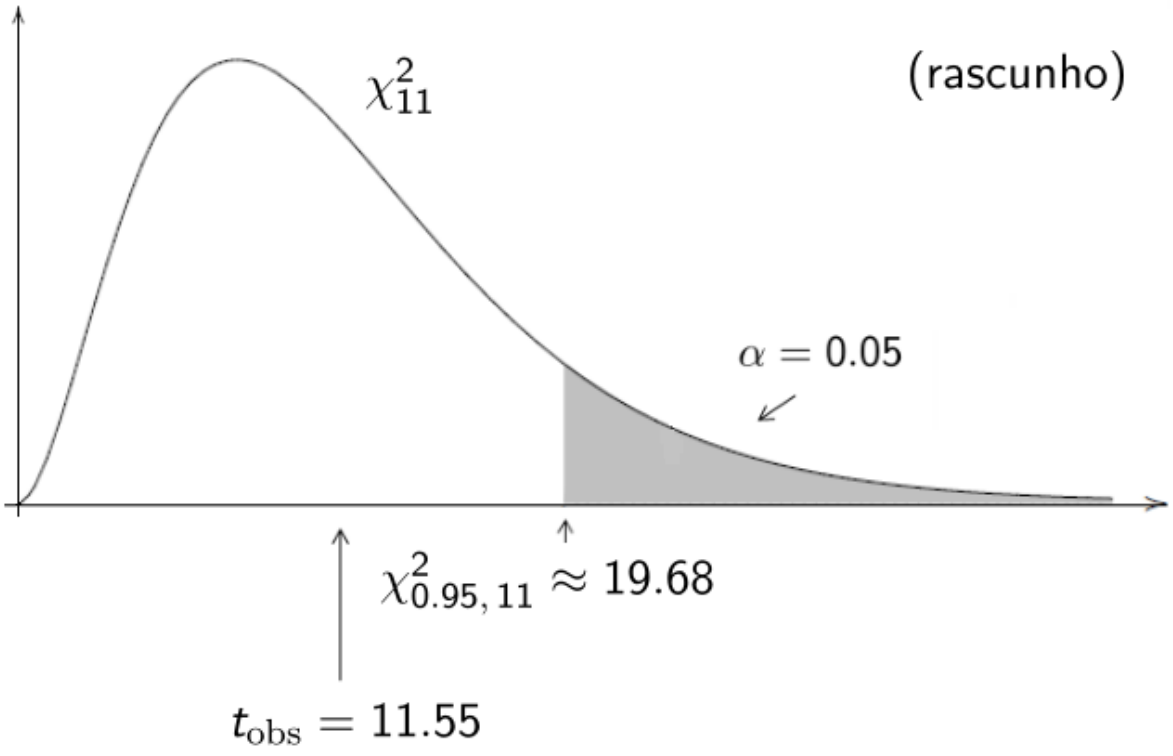


SPSS: IDF.CHISQ(0.95,11) $= 19.6751\ldots$

Estatística de teste observada:

$$t_{\text{obs}} = \frac{11 \times 105}{100} = 11.55.$$

Exemplo



Decisão: como $t_{\text{obs}} \notin RC_{5\%}$, não se rejeita H_0.

Conclusão: com base nesta amostra e ao nível de significância de 5%, não há evidência estatística para assumir que a variância da faturação diária desta empresa seja superior a 100 Euros.

TH em populações Normais

Teste para o desvio padrão, σ.

$$H_0 : \sigma = \sigma_0 \quad vs. \quad H_1 : \sigma \quad \begin{matrix} < \\ \neq \\ > \end{matrix} \quad \sigma_0$$

Realiza-se o teste correspondente para a variância:

$$H_0 : \sigma^2 = \sigma_0^2 \quad vs. \quad H_1 : \sigma^2 \quad \begin{matrix} < \\ \neq \\ > \end{matrix} \quad \sigma_0^2$$

Pressupostos exigidos:

- Observações independentes retiradas da mesma população (amostra aleatória);
- População com distribuição Normal.

TH em populações Normais

Teste para comparação de médias, μ_X e μ_Y, com variâncias, σ_X^2 e σ_Y^2, desconhecidas (mas consideradas iguais); amostras independentes.

$$H_0 : \mu_X - \mu_Y = 0 \quad vs. \quad H_1 : \mu_X - \mu_Y \quad \begin{matrix} < \\ \neq \\ > \end{matrix} \quad 0$$

Estatística de teste:

$$T = \frac{\overline{X} - \overline{Y}}{\sqrt{\frac{1}{n} + \frac{1}{m}} \sqrt{\frac{(n-1)S_{c_X}^2 + (m-1)S_{c_Y}^2}{n+m-2}}} \underset{sob\ H_0}{\sim} t_{n+m-2}.$$

TH em populações Normais

Teste para comparação de médias, μ_X e μ_Y, com variâncias, σ_X^2 e σ_Y^2, desconhecidas (mas consideradas iguais); amostras independentes.

Região de rejeição:

- se $H_1 : \mu_X - \mu_Y \neq 0,\ RC_\alpha = \{t : |t| > t_{1-\frac{\alpha}{2},\, n+m-2}\}$;
- se $H_1 : \mu_X - \mu_Y < 0,\ RC_\alpha = \{t : t < t_{\alpha,\, n+m-2}\}$;
- se $H_1 : \mu_X - \mu_Y > 0,\ RC_\alpha = \{t : t > t_{1-\alpha,\, n+m-2}\}$.

Pressupostos exigidos:

- Amostras independentes: $X_1, X_2, \ldots, X_n$ e $Y_1, Y_2, \ldots, Y_m$;
- Cada amostra constituída por observações independentes retiradas da mesma população (amostras aleatórias);
- Populações com distribuição Normal; variâncias desconhecidas (mas assumidas como iguais).

Exemplo

Procura-se analisar a tensão axial de rutura de dois materiais usados na construção de aviões. Fizeram-se 10 medições aleatórias com cada um dos materiais.

Tradicional	3238	3195	3246	3190	3204	3254	3229	3225	3217	3241
Novo	3261	3187	3209	3212	3258	3248	3215	3226	3240	3234

Considerando assegurados os pressupostos para a realização do teste t para a comparação de médias, será correto assumir que os materiais diferem significativamente?

Resolução com o SPSS através da análise de *p-values*:

Independent Samples Test

		Levene's Test for Equality of Variances		t-test for Equality of Means						
									90% Confidence Interval of the Difference	
		F	Sig.	t	df	Sig. (2-tailed)	Mean Difference	Std. Error Difference	Lower	Upper
Tensão	Equal variances assumed	,060	,810	-,500	18	,623	-5,10000	10,20398	-22,79434	12,59434
	Equal variances not assumed			-,500	17,903	,623	-5,10000	10,20398	-22,79953	12,59953

Exemplo

Sejam,
$X-$ tensão axial de rutura do material tradicional;
$Y-$ tensão axial de rutura do material novo.

Teste de hipóteses pretendido:

$$H_0 : \mu_X - \mu_Y = 0 \quad vs. \quad H_1 : \mu_X - \mu_Y \neq 0.$$

Teste de Levene (teste de homogeneidade de variâncias):

$$H_0 : \sigma_X^2 = \sigma_Y^2 \quad vs. \quad H_1 : \sigma_X^2 \neq \sigma_Y^2.$$

Como *p-value* ≈ 0.810 é muito elevado (maior do que os níveis de significância usuais), não se rejeita H_0. Com base nestas amostras e aos níveis usuais de significância, não há evidência para assumir que as variâncias populacionais sejam significativamente diferentes.

Exemplo

Assim, a análise do teste de igualdade de médias deve ser realizada considerando a informação na linha *Equal variances assumed*.

Como o *p-value* ≈ 0.623 é muito elevado (maior do que os níveis de significância usuais), não se rejeita H_0. Logo, com base nestas amostras e aos níveis usuais de significância, não há evidência para assumir que a tensão axial média de rutura seja significativamente diferente nos dois materiais.

Como base no $IC_{90\%}(\mu_X - \mu_Y)$ da tabela o que pode concluir?

Ao nível de significância de 10%, a conclusão anterior mantém-se, pois $0 \in IC_{90\%}(\mu_X - \mu_Y)$. Adicionalmente, como a amplitude do IC aumenta com o aumento do nível de confiança, a conclusão mantém-se inalterada para níveis de significância inferiores a 10%.

TH em populações Normais

Teste para comparação de médias, μ_X e μ_Y, com variâncias, σ_X^2 e σ_Y^2, desconhecidas; amostras emparelhadas.

$$H_0 : \mu_X - \mu_Y = 0 \quad vs. \quad H_1 : \mu_X - \mu_Y \quad \begin{matrix} < \\ \neq \\ > \end{matrix} \quad 0$$

$$H_0 : \mu_D = 0 \quad vs. \quad H_1 : \mu_D \quad \begin{matrix} < \\ \neq \\ > \end{matrix} \quad 0$$

Estatística de teste: $T = \dfrac{\overline{D}}{\dfrac{S_{c_D}}{\sqrt{n}}} \underset{sob\ H_0}{\sim} t_{n-1}$.

(Amostra das diferenças: $D_i = X_i - Y_i$, com $i = 1, 2, \ldots, n$.)

TH em populações Normais

Teste para comparação de médias, μ_X e μ_Y, com variâncias, σ_X^2 e σ_Y^2, desconhecidas; amostras emparelhadas.

Pressupostos exigidos:

- Amostras emparelhadas: $X_1, X_2, \ldots, X_n$ e $Y_1, Y_2, \ldots, Y_n$;
- Cada amostra constituída por observações independentes retiradas da mesma população (amostras aleatórias);
- Populações com distribuição Normal.

Exemplo

Num estudo sobre o efeito do tabaco na agregação de plaquetas no sangue, mediu-se o grau de agregação a 18 indivíduos antes e depois de fumarem um cigarro. Os valores obtidos foram:

Antes	39	25	25	27	44	30	67	53	53	52	60	28	30	65	52	40	55	60
Depois	38	27	29	37	56	46	82	57	80	61	59	43	43	60	69	62	81	65

Com base nos dados apresentados, poder-se-á concluir, ao nível de significância de 5%, que a alteração do grau de agregação é menosprezável?

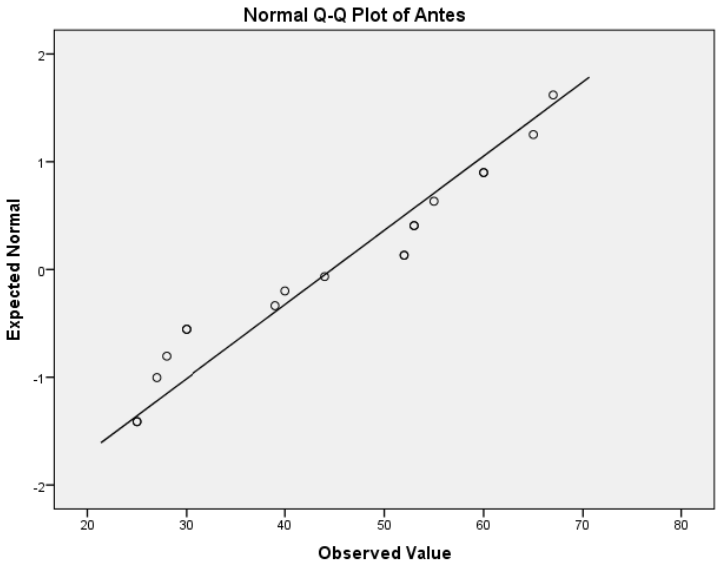
Sejam,
$X-$ grau de agregação antes de fumar o cigarro;
$Y-$ grau de agregação depois de fumar o cigarro.

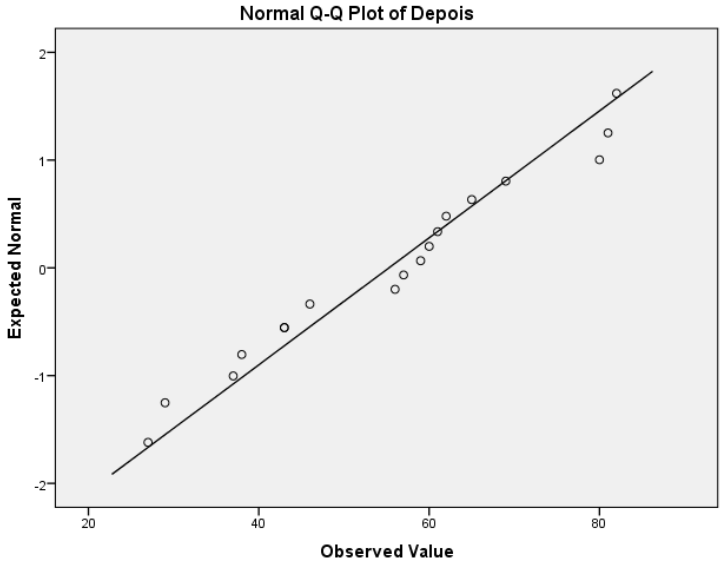
Teste de hipóteses pretendido:

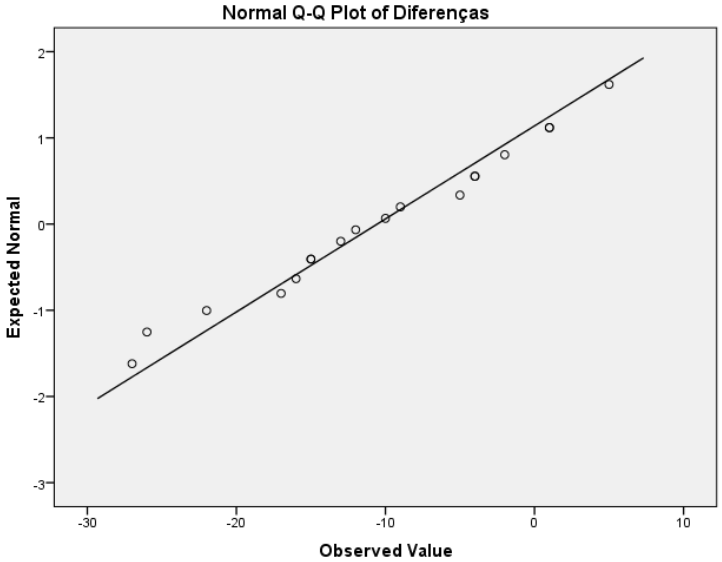
$$H_0 : \mu_D = 0 \quad vs. \quad H_1 : \mu_D \neq 0,$$

com $\mu_D = \mu_X - \mu_Y$.

Exemplo







Tests of Normality

	Kolmogorov-Smirnov[a]			Shapiro-Wilk		
	Statistic	df	Sig.	Statistic	df	Sig.
Antes	,192	18	,078	,909	18	,081
Depois	,128	18	,200*	,952	18	,464
Diferenças	,114	18	,200*	,969	18	,777

*. This is a lower bound of the true significance.

a. Lilliefors Significance Correction

Exemplo

Estatística de teste: $T = \dfrac{\overline{D}}{\dfrac{S_{c_D}}{\sqrt{n}}} \underset{sob\ H_0}{\sim} t_{n-1}$.

Média e desvio padrão corrigido da amostra das diferenças:

$\overline{d} \approx -10.56$ e $s_{c_D} \approx 9.27$.

Estatística de teste observada: $t_{\text{obs}} = \dfrac{-10.56}{\dfrac{9.27}{\sqrt{18}}} \approx -4.83$.

Como a estimativa que se obtém para o parâmetro a testar é inferior ao valor especificado em H_0, então

$$\begin{aligned} p\text{-}value &= 2P(T < -4.83|H0) = 2 * CDF.T(-4.83, 17) \\ &\approx 0.16 \times 10^{-3}. \end{aligned}$$

Exemplo

Como *p-value* $\approx 0.16 \times 10^{-3} < \alpha = 0.05$, rejeita-se H_0.

Assim, com base nestas amostras e ao nível de significância de 5%, pode considerar-se que existem diferenças estatísticas significativas entre as médias dos graus de agregação antes e depois de fumar um cigarro, ou seja, respondendo diretamente à questão inicialmente colocada, a alteração do grau de agregação não é menosprezável.

Sugestão: utilize os procedimentos baseados na região de rejeição e no $IC_{95\%}(\mu_D)$ para confirmar a conclusão obtida anteriormente.

TH em **populações genéricas** (com base no TLC)

Testes para a média, μ.

$$H_0 : \mu = \mu_0 \quad vs. \quad H_1 : \mu \begin{array}{c} < \\ \neq \\ > \end{array} \mu_0$$

Estatísticas de teste:

$$T = \frac{\overline{X} - \mu_0}{\frac{S_c}{\sqrt{n}}} \underset{aprox.}{\sim} N(0,1), \text{ se } \sigma^2 \text{ desconhecida;}$$

$$T = \frac{\overline{X} - \mu_0}{\frac{\sigma}{\sqrt{n}}} \underset{aprox.}{\sim} N(0,1), \text{ se } \sigma^2 \text{ conhecida.}$$

Pressupostos exigidos:

- População nas condições de aplicação do TLC (média e variância finitas);
- Observações independentes retiradas da mesma população (amostra aleatória de grande dimensão[1]).

[1] No âmbito desta disciplina, considera-se uma amostra de "grande" dimensão quando $n \geq 30$.

TH em **populações genéricas** (com base no TLC)

Teste para uma proporção, p. (Teste Binomial.)

Seja p a proporção de indivíduos com uma determinada característica de interesse na população e

$$\widehat{p} = \frac{X}{n}$$

a proporção de indivíduos com essa característica de interesse numa amostra aleatória de dimensão n. Naturalmente, X representa o número de elementos com a característica de interesse na amostra.
(Este teste recorre à aproximação da Binomial pela Normal.)

$$H_0 : p = p_0 \quad vs. \quad H_1 : p \begin{array}{c} < \\ \neq \\ > \end{array} p_0$$

TH em **populações genéricas** (com base no TLC)

Teste para uma proporção, p.

Estatística de teste:

$$T = \frac{\widehat{p} - p_0}{\sqrt{\dfrac{p_0\left(1-p_0\right)}{n}}} = \frac{\frac{X}{n} - p_0}{\sqrt{\dfrac{p_0\left(1-p_0\right)}{n}}} \underset{aprox.}{\sim} N(0,1).$$

Pressupostos exigidos:

- População nas condições de aplicação do TLC (média e variância finitas);
- Observações independentes retiradas da mesma população (amostra aleatória de grande dimensão[1]).

[1] No âmbito desta disciplina, considera-se uma amostra de "grande" dimensão quando $n \geq 30$.

Exemplo

Adaptado de Hall et al. (2011), p. 164.
(Sugere-se a consulta do livro para detalhes adicionais.)

Para avaliar uma possível alteração no horário de funcionamento de um bar, pretende-se testar se a proporção de clientes que chegam antes da meia-noite é inferior a 4%. Com base numa amostra aleatória de 564 clientes, verificou-se que 21 deles tinham chegado antes da meia-noite. Considere um nível de significância de 5%.

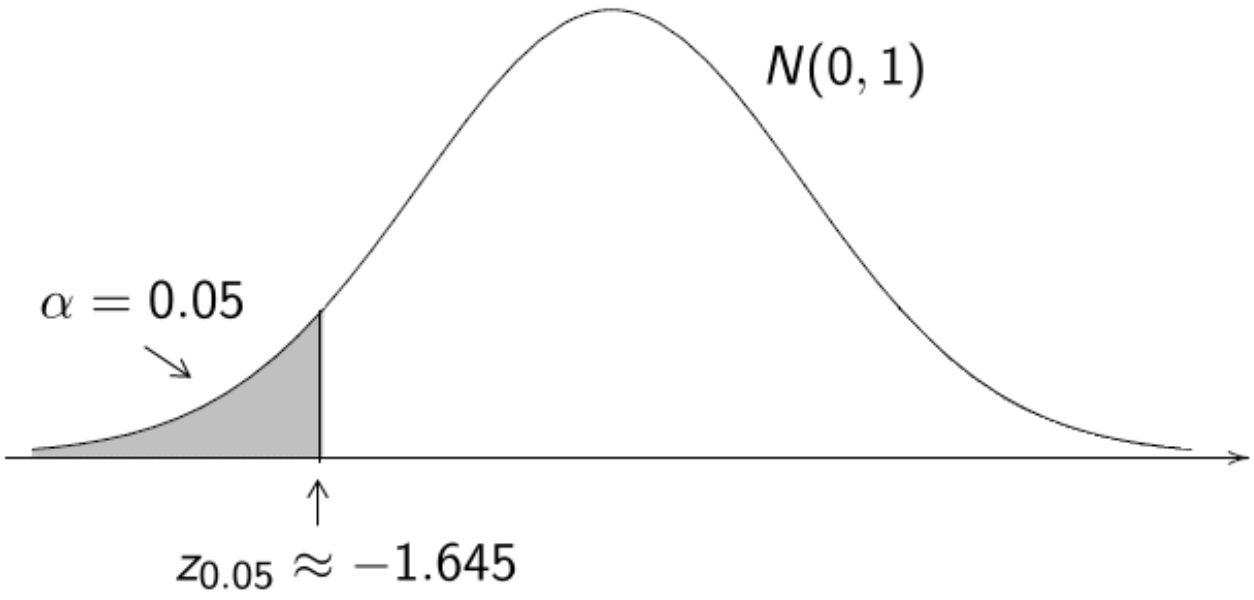
$$H_0 : \ p = 0.04 \quad vs. \quad H_1 : \ p < 0.04$$

(p representa a proporção populacional de clientes que chegam antes da meia-noite.)

Estatística de teste: $T = \dfrac{\frac{X}{n} - p_0}{\sqrt{\dfrac{p_0\,(1-p_0)}{n}}} \underset{aprox.}{\sim} N(0,1).$

Região de rejeição: como $H_1 : p < p_0$, $RC_\alpha = \{t : t < z_\alpha\}$.

Exemplo



Tabelas: para $z = 1.64$, obtém-se $\Phi(1.64) = 0.9495$;
para $z = 1.65$, obtém-se $\Phi(1.65) = 0.9505$;
SPSS: IDF.NORMAL(0.95,0,1) ≈ 1.645 ou PROBIT(0.95)≈ 1.645;
Naturalmente, IDF.NORMAL(0.05,0,1) ≈ -1.645.

Exemplo

Estatística de teste observada:

$$t_{\text{obs}} = \frac{\frac{21}{564} - 0.04}{\sqrt{\frac{0.04 \times (1 - 0.04)}{564}}} \approx -0.34.$$

Decisão: como $t_{\text{obs}} \notin RC_{5\%}$, não se rejeita H_0.

Conclusão: com base nesta amostra e ao nível de significância de 5%, não há evidência estatística para assumir que a proporção (populacional) de clientes que chegam antes da meia-noite seja inferior a 4%.

Note-se que $\widehat{p} = \frac{21}{564} \approx 0.037 < 0.04$ e não se rejeitou H_0. A importância das técnicas de inferência estatística!