De que trata a Estatística?

Objectos de estudo da Estatística - Dados estatísticos

Objectivos da estatística - obter dados (por observação ou recolha planeada), descrevê-los, sumarizá-los, organizá-los, analisá-los e interpretar os resultados da análise.

Unidade experimental - elementos que dão acesso aos dados.

Variável estatística - característica que pode ser diferente nas diversas observações efectuadas.

A Estatística permite-nos descrever e compreender a variabilidade contida na informação (dados) em análise. Desta forma ajuda-nos a tomar decisões em situações de incerteza.

Sub-áreas da Estatística

- Obtenção dos dados: Amostragem e planeamento de experiências.
- Descrição dos dados: Estatística descritiva e análise exploratória de dados.
- Modulação: Teoria da Probabilidade.
- Indução e previsão do futuro: Inferência Estatística

Tipos de variáveis

- qualitativas:
  - nominais ou categóricas;
  - ordinais;
- quantitativas:
  - numa escala intervalar;
  - numa escala de razões;

Em Estatística todas as variáveis são codificadas numericamente pelo que podemos classificar o tipo de escalas numéricas de acordo com o tipo de dados que lhe deram origem:

- escala nominais;
- escala ordinal;
- escala intervalar;
- escala de razões;

Existem metodologias distintas para analisar os vários tipos de variáveis. Após a codificação numérica há que ter o cuidado de seleccionar métodos compatíveis com a natureza dos dados.

Análise Preliminar de dados

Objectivos

- Cálculo numérico de medidas amostrais, tais como valores representativos da tendência central dos dados, e valores representativos da quantidade de variabilidade inerente aos dados.
- Resumo e descrição global dos dados através da construção de tabelas e de gráficos.

Ordenação

Se ordenarmos uma amostra $x_1, x_2, \ldots, x_n$ ascendentemente e representarmos os valores ordenados por $x_{1:n} \leq x_{2:n} \leq \cdots \leq x_{n:n}$ obtemos as chamadas **estatísticas ordinais ascendentes**. Outra forma usual de representar as estatísticas ordinais ascendentes é através da notação $x_{(1)} \leq x_{(2)} \leq \cdots \leq x_{(n)}$.

Tabela de frequências

Sejam $x_1^* \leq x_2^* \leq \cdots \leq x_k^*$, k observações **distintas** na amostra de dimensão n.

Frequência absoluta: $n_i \equiv$número de vezes que se observou o valor x_i^* na amostra;

Frequência relativa: $f_i = \frac{n_i}{n} \equiv$proporção de valores iguais a x_i^* na amostra;

Frequência absoluta acumulada: $N_i = n_1 + n_2 + \cdots + n_i$;

Frequência relativa acumulada: $F_i = f_1 + f_2 + \cdots + f_i$;

Tabela de frequências (em inglês *frequency table*)

x_i^*	n_i	f_i	F_i
x_1^*	n_1	f_1	F_1
x_2^*	n_2	f_2	F_2
$\vdots$	$\vdots$	$\vdots$	$\vdots$
x_k^*	n_k	f_k	1
	$\sum_{i=1}^{k} n_i = n$	$\sum_{i=1}^{k} f_i = 1$	

Função de distribuição empírica

Chama-se função de distribuição empírica à seguinte função real F_n, de variável real:

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

Note-se que $F_n(x_i^*) = F_i$, $i = 1 \ldots, k$ e que é possível recuperar as frequências relativas através das frequências relativas acumuladas mediante a relação $f_i = F_i - F_{i-1}$.

Medidas amostrais

Medidas de localização central:

- **Média** (*sample mean*)

- **Mediana** (*median*)

- **Moda** (*mode*)

Outras medidas de localização:

- **Mínimo** (*minimum*) e **máximo** (*maximum*)

- **Quantis (singular: quantil)** (*quantiles*)

- **Quartis** (*quartiles*)

- **Percentis** (*percentiles*)

Medidas de dispersão:

- **Amplitude da amostra** (*range*)

- **Distância inter-quartil** (*inter-quartile range*)

- **Variância** (*variance*) e **variância corrigida**

- **Desvio-padrão** (*standard deviation*) e **desvio-padrão corrigido**

Medidas de assimetria (*skewness*):

As distribuições de frequências podem ser classificadas em **simétricas** e **assimétricas**.

Tipos de assimetria: **positiva** e **negativa**.

Nas distribuições de frequências simétricas tem-se que $média = mediana = moda$. A assimetria pode ser classificada mediante o estudo da posição relativa destas três medidas de localização, nomeadamente:

- Se $média > mediana > moda$, suspeita-se que haja assimetria positiva por parte da distribuição de frequências;
- Se $média < mediana < moda$, suspeita-se que haja assimetria negativa.

Métodos gráficos

- Histograma

- Gráficos de frequências

- Caixas de bigodes (*boxplot)*

- Gráficos de dispersão (*scatterplot*)

Teoria das Probabilidades

Experiência aleatória é uma experiência que pode ter diferentes resultados, mesmo quando é repetida em circunstâncias análogas.

Resultado elementar é um resultado possível duma experiência aleatória.

Espaço amostral (também designado por espaço de resultados, ou espaço dos possíveis, ou população, ou ainda Universo.) É o conjunto de todos os resultados possíveis numa experiência aleatória. Habitualmente representa-se por Ω.

Acontecimento - subconjunto de Ω contendo um ou mais resultados possíveis (*representa-se com letra maiúscula*).

Intersecção de dois acontecimentos é um acontecimento que ocorre quando ocorrem simultaneamente os dois acontecimentos.

Reunião de dois acontecimentos é um acontecimento que ocorre quando ocorre pelo menos um dos acontecimentos.

Acontecimentos mutuamente exclusivos ou incompatíveis - acontecimentos que não podem ocorrem em simultâneo.

Axiomática da probabilidade

1. $P(\Omega) = 1$
2. $P(A) \geq 0$ para todo o $A \in \Omega$
3. Se A_1 e A_2 forem acontecimentos mutuamente exclusivos

$$P(A_1 \cup A_2) = P(A_1) + P(A_2).$$

Esta relação generaliza-se para qualquer conjunto (numerável) de acontecimentos mutuamente exclusivos.

Independência de acontecimentos

Dois acontecimentos são independentes se e só se

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

Variáveis Aleatórias

Uma variável aleatória (v.a.) associa um número real a cada resultado possível, de tal forma que a probabilidade de um intervalo real é igual à probabilidade dos acontecimentos que lhe deram origem.

As v.a.'s representam-se por letras maiúsculas.

Variáveis Aleatórias Discretas - v.a.'s que só assumem um número finito ou numerável de valores, com probabilidade estritamente positiva.

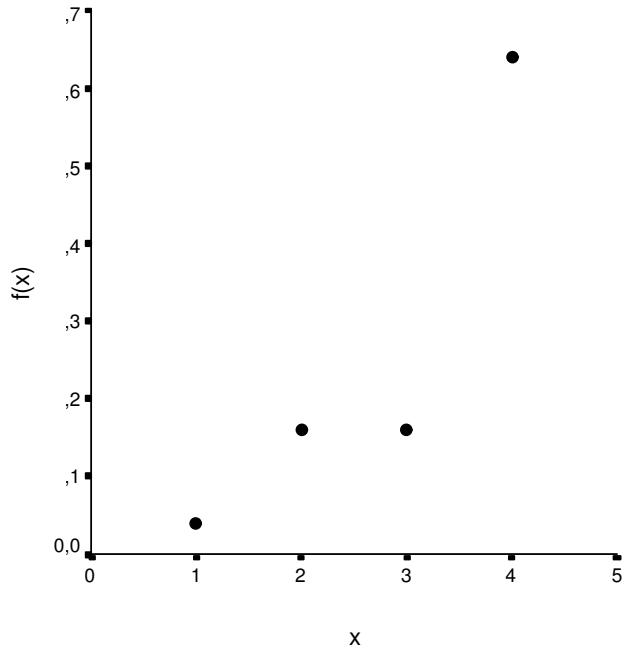
Variáveis Aleatórias Contínuas - v.a.'s que assumem valores em todo o $\mathbb{R}$ ou em intervalos de $\mathbb{R}$.

Caracterização e propriedades duma v.a. discreta

Função Massa de Probabilidade (ou Função de Probabilidade), $f(x)$

$f(x) = P(X = x)$.
$\sum_{x_i} f(x_i) = 1$.

exemplo gráfico



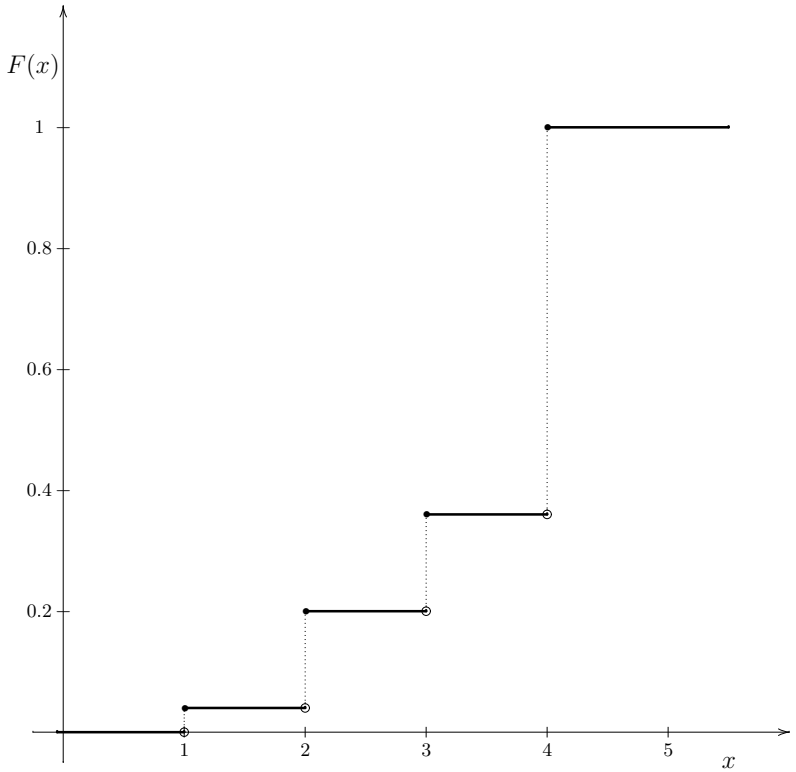
Função de Distribuição, $F(x)$

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

$F(x) = \sum_{x_i \leq x} f(x_i) \quad \Leftrightarrow \quad f(x_i) = F(x_i) - F(x_{i-1})$.

$F(x)$ cresce de 0 a 1 em escada.

Exemplo gráfico:



Caracterização e propriedades de va's contínuas

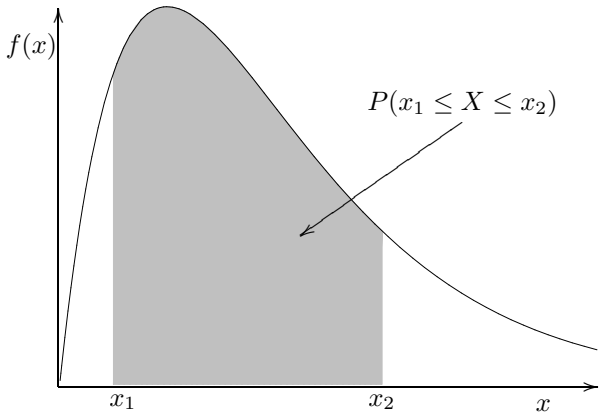
Função Densidade de Probabilidade, $f(x)$

$f(x) \geq 0.$

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

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

Exemplo gráfico



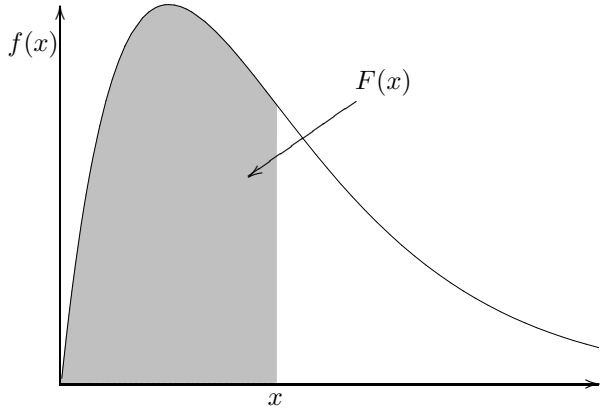
Função de Distribuição

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

$F(x)$ cresce de 0 a 1 de forma contínua.

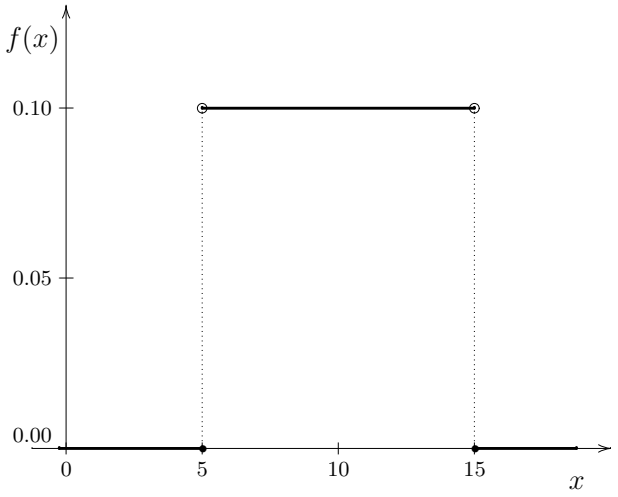
$F(x) = \int_{-\infty}^{x} f(t)dt \quad \Leftrightarrow \quad f(x) = F'(x)$.

Interpretação gráfica:

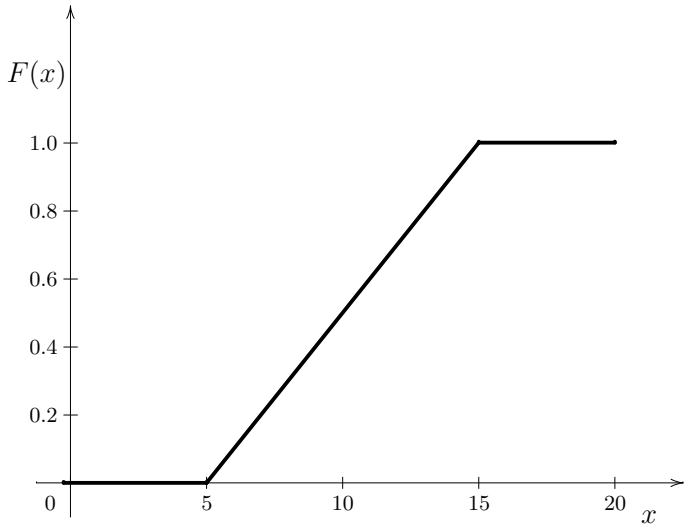


Exemplo gráfico:

$f(x)$ uniforme



$F(x)$ uniforme



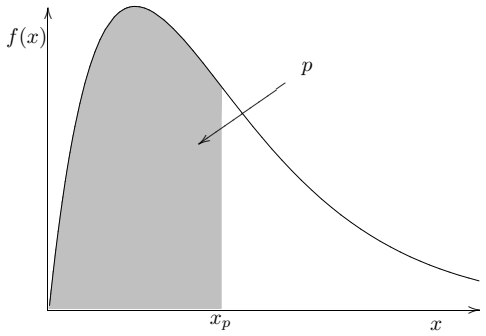
Amostra	**População**
$x_1, \ldots, x_n$	
Medidas amostrais	**Parâmetros da distribuição**
média:	média:
$\bar{x} = \sum_{i=1}^{k} x_i^* f_i$	$\mu = E[X] = \begin{cases} \sum_{x_i} x_i \, f(x_i) & \text{v.a. discreta} \\ \int_{-\infty}^{+\infty} x \, f(x) dx & \text{v.a. contínua} \end{cases}$
variância:	variância:
$s^2 = \sum_{i=1}^{k} (x_i^* - \bar{x})^2 f_i$	$\sigma^2 = Var[X] = \begin{cases} \sum_{x_i} (x_i - \mu)^2 \, f(x_i) & \text{v.a. discreta} \\ \int_{-\infty}^{+\infty} (x - \mu)^2 \, f(x) dx & \text{v.a. contínua} \end{cases}$]
desvio padrão:	desvio padrão:
$s = \sqrt{s^2}$	$\sigma = \sqrt{\sigma^2}$

Moda

A moda de uma distribuição é o número real m tal que $f(m)$ é máximo. (Pode haver mais do que uma moda.)

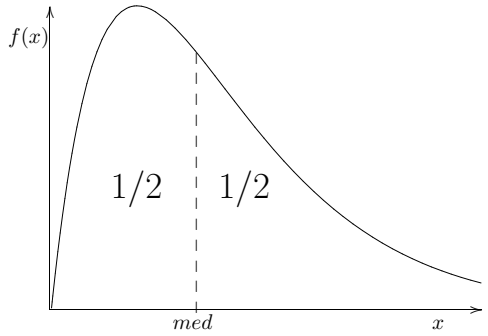
Quantil de ordem p

$$x_p = \begin{cases} x_p : \ F(x_p) = p & \text{v.a.'s contínuas} \\ x_p : \ p \leq F(x_p) \leq p + P(X = x_p) & \text{v.a.'s discretas} \end{cases}$$



Mediana

A mediana duma distribuição é o quantil de ordem 1/2.



Quartil

Quartis são quantis de ordem 0.25, 0.5 e 0.75.

Percentil

Um percentil de ordem x é o quantil de ordem $x/100$.

Amostra	**População**
gráfico de frequências	f.m.p., $f(x)$ (caso discreto)
histograma	f.d.p., $f(x)$ (caso contínuo)
função de distribuição empírica, $F_n(x)$ (frequências relativas acumuladas)	Função de distribuição, $F(x)$

Alguns resultados importantes

Sejam $X_1, X_2, \ldots, X_n$ v.a.'s independentes e identicamente distribuídas (i.i.d.) com $E[X_i] = \mu$ e $Var[X_i] = \sigma^2$. Seja $\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$ a média das variáveis (média amostral).

$$
\begin{aligned}
E[\textstyle\sum_{i=1}^{n} X_i] &= n\mu \\
Var[\textstyle\sum_{i=1}^{n} X_i] &= n\sigma^2 \\
E[\bar{X}] &= \mu \\
Var[\bar{X}] &= \frac{\sigma^2}{n}
\end{aligned}
$$

Distribuições usuais discretas

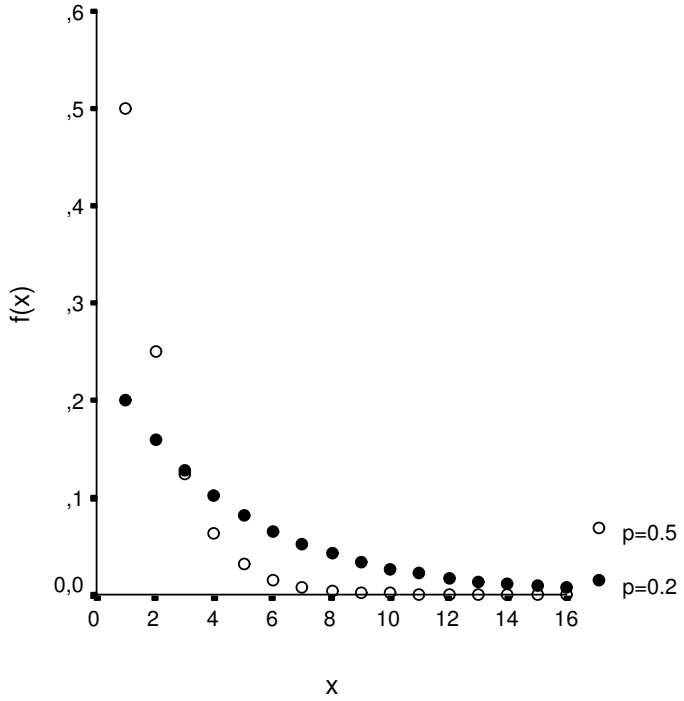
Bernoulli, $X \frown B(p)$.

$$X = \begin{cases} 1, & \text{se } \{\text{sucesso}\} \\ 0, & \text{se } \{\text{insucesso}\}. \end{cases}$$

$$\begin{aligned} f(1) &= P(X = 1) = P\{\text{sucesso}\} = p \in [0, 1], \\ f(0) &= P(X = 0) = P\{\text{insucesso}\} = 1 - p. \end{aligned}$$

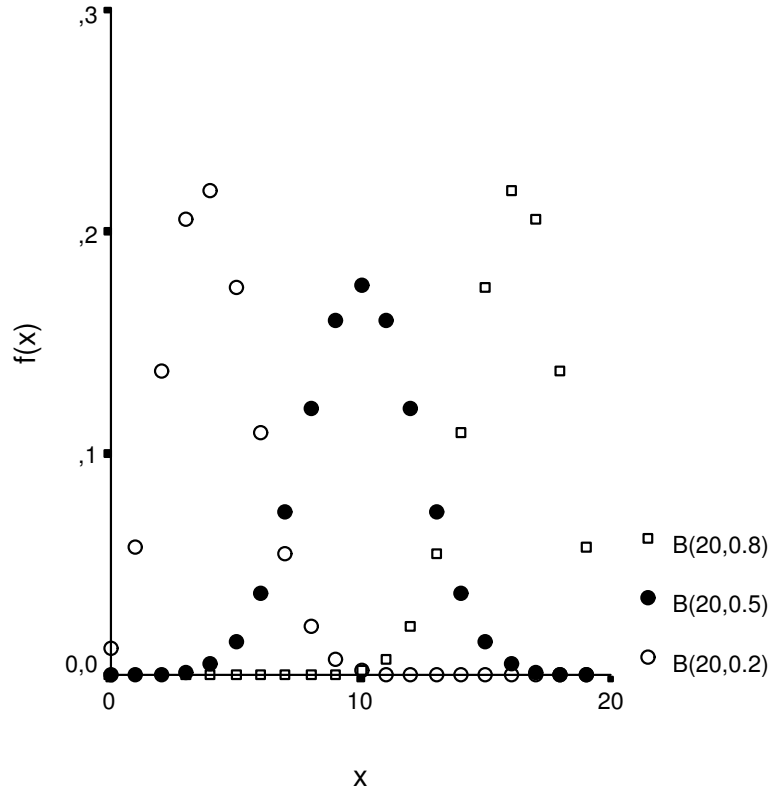
Distribuição Geométrica, $X \frown G(p)$.

X - número de provas (ou insucessos) até ao primeiro sucesso (numa repetição de Bernoullis).



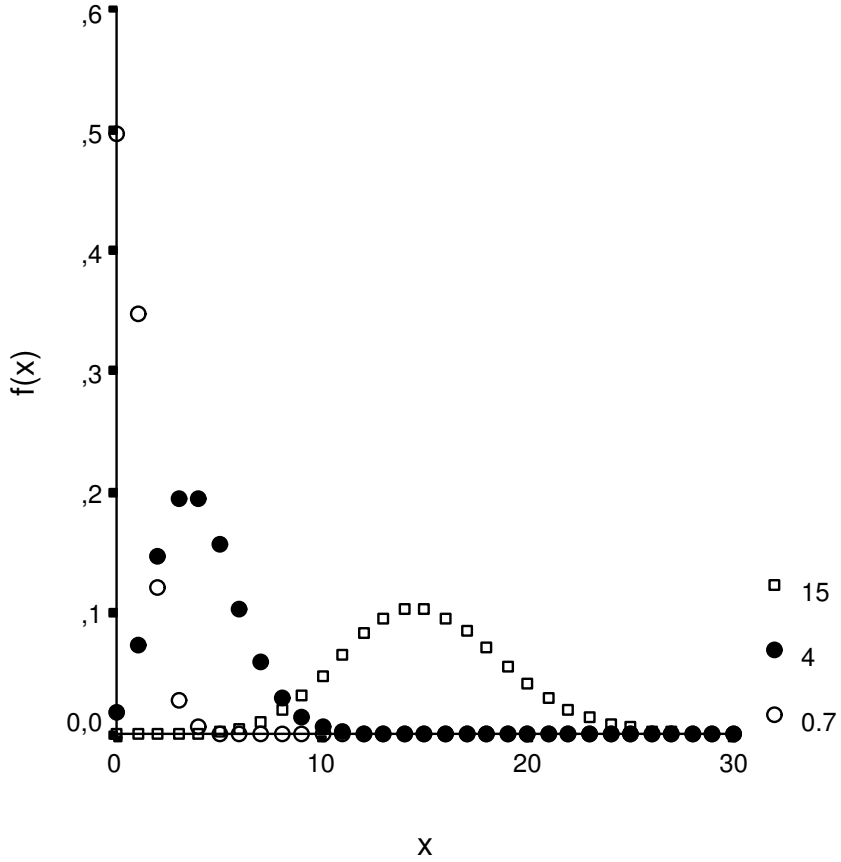
Binomial, $X \frown B(n,p)$

X - número de sucessos num conjunto de n provas de Bernoulli (n fixo).



Poisson, $X \frown P(\lambda)$

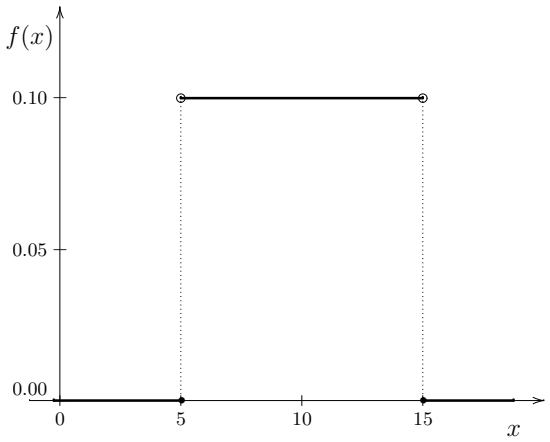
$X =$ número de ocorrências num intervalo de tempo ou numa região do espaço.

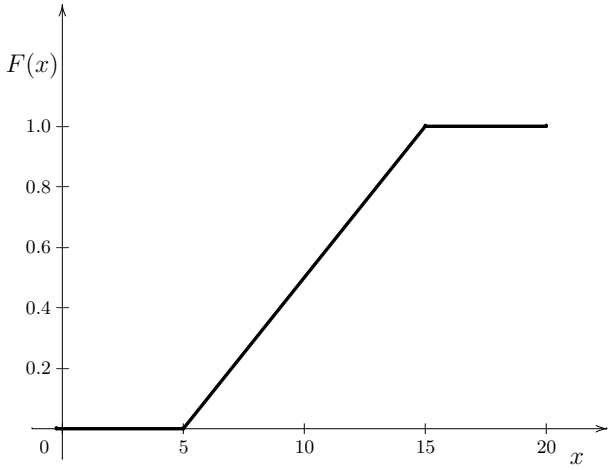


Distribuições usuais contínuas

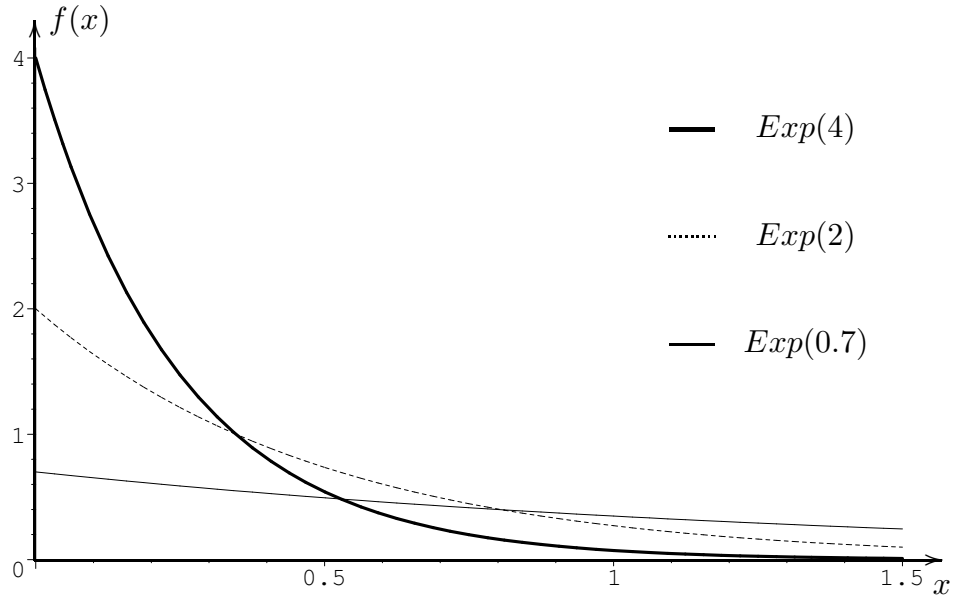
Uniforme, $X \frown U(a, b)$

exemplo para $a = 5$ e $b = 15$:





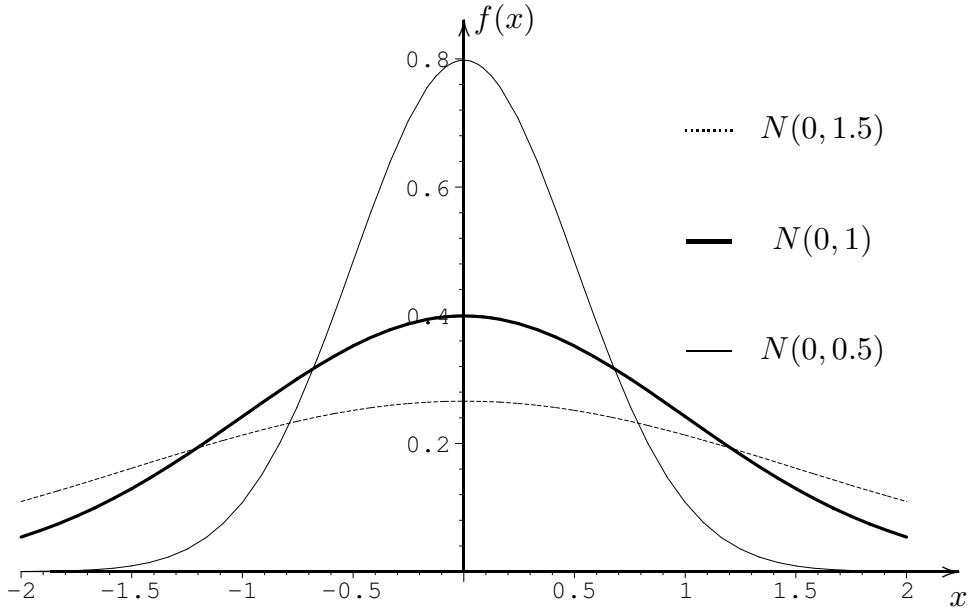
Exponencial, $X \frown E(\lambda)$

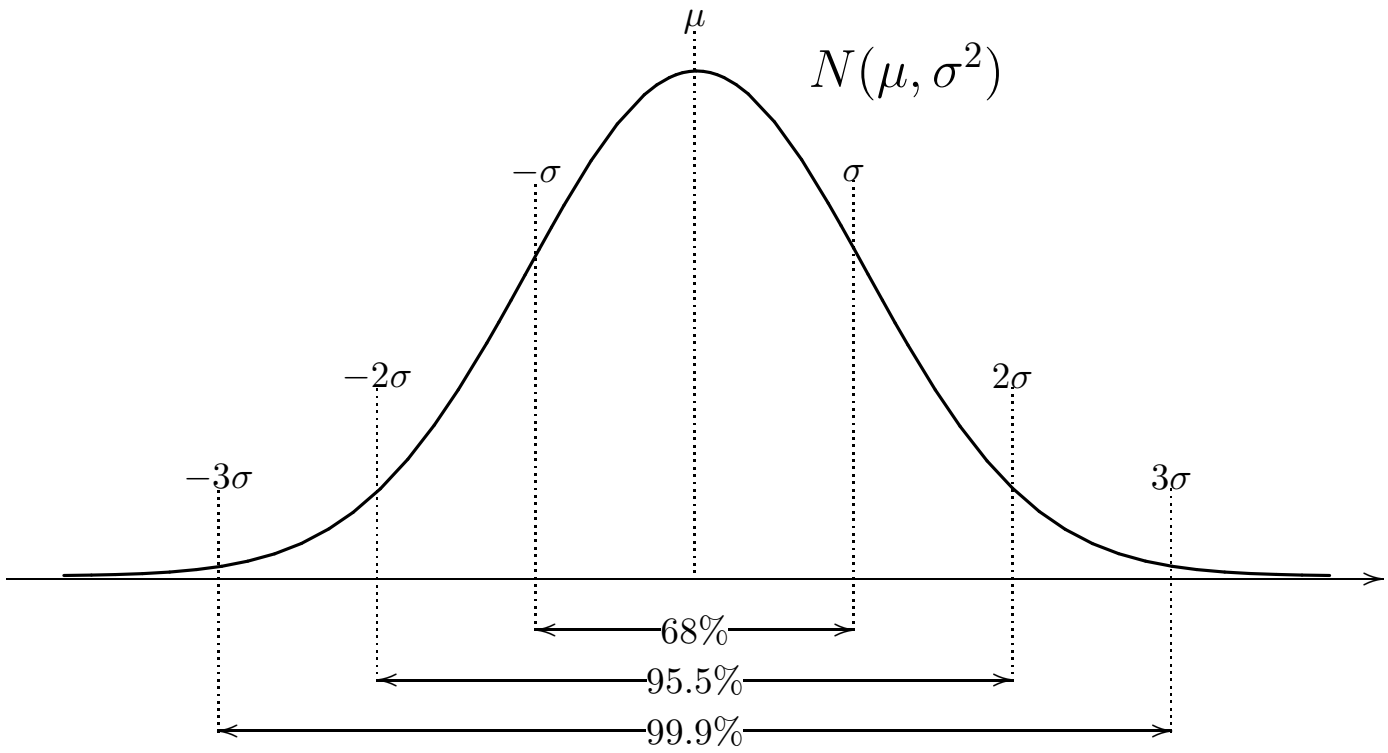


Normal ou Gaussiana, $X \frown N(\mu, \sigma^2)$

$$f(x) = \frac{1}{\sqrt{2\pi}} exp\left\{-\frac{(x-\mu)^2}{2\sigma^2}\right\}, \quad x \in \mathbb{R}, \quad \mu \in \mathbb{R}, \quad \sigma > 0.$$

$$E[X] = \mu, \quad Var[X] = \sigma^2.$$




μ
$N(\mu, \sigma^2)$
$-\sigma$
σ
-2σ
2σ
-3σ
3σ
68%
95.5%
99.9%

Propriedades

1. Se $X \frown N(\mu, \sigma^2)$, então

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

Esta distribuição denomina-se **Normal standard** ou **Normal centrada e reduzida**.

$$\Phi(x) = P(Z \leq x).$$

2. Se $X \frown N(\mu, \sigma^2)$ e $Y = aX + b$ então

$$Y \frown N(a\mu + b, a^2\sigma^2).$$

3. Se $X_i \frown N(\mu_i, \sigma_i^2)$ independentes, $i = 1, \ldots, n$, então

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

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

$$\sum_{i=1}^{n} X_i \frown N(n\mu, n\sigma^2) \;\; e \;\; \bar{X} = \frac{1}{n} \sum_{i=1}^{n} X_i \frown N(\mu, \sigma^2/n).$$

> *Devemos ter o cuidado de não confundir os dados com as abstracções que utilizamos para os analisar.*
>
> William James (1842-1910)

O objectivo da 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**.

Amostragem aleatória simples - os dados são recolhidos aleatoria e independentemente uns dos outros.

Amostra aleatória (a.a.) - conjunto de observações, $X_1, X_2, \ldots, X_n$, independentes e identicamente distribuídas com distribuição F_X.

Estatística - função da amostra que não depende de parâmetros desconhecidos.

O objectivo usual é inferir sobre a forma ou os parâmetros da distribuição F_X.

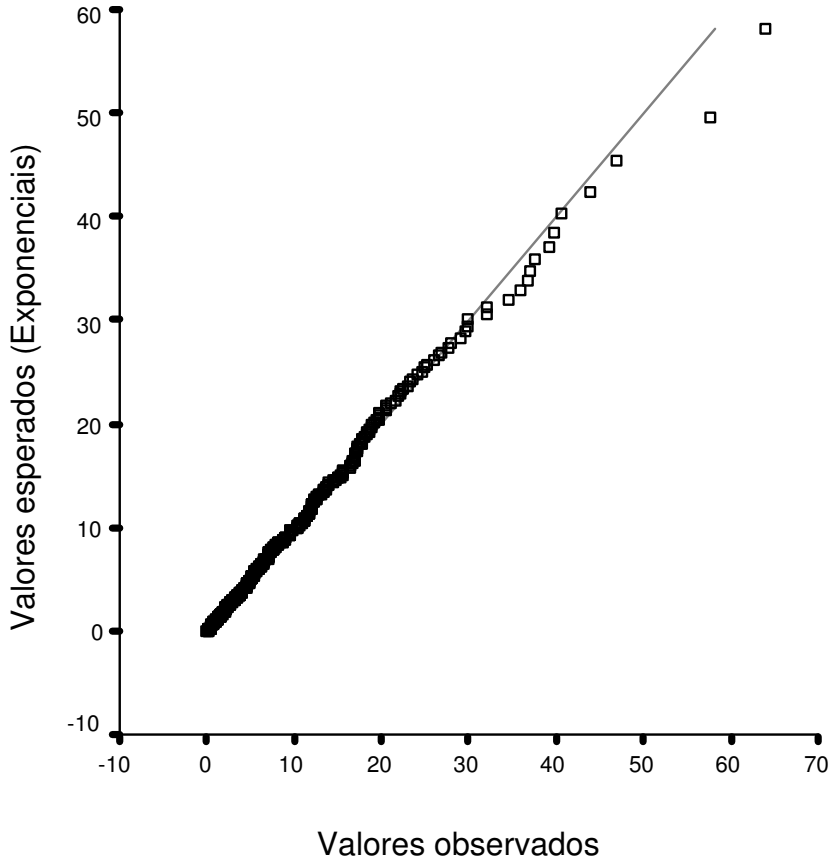
Se estivermos interessados na **forma** podemos começar por comparar o histograma (ou gráfico de frequências) com os gráficos de $f(x)$ das distribuições usuais.

Seguidamente podemos construir gráficos de quantis (QQ-plot) ou de probabilidades (PP-plot). Estes gráficos também são designados papel de probabilidades.

Um **QQ-plot** é um gráfico de dispersão que confronta os quantis da amostra com os quantis de uma distribuição específica (usual). Se a amostra tiver sido retirada de uma população com aquela distribuição o gráfico deve assemelhar-se a um conjunto de pontos mais ou menos sobre uma recta. Caso contrário deverão surgir zonas de não-linearidade no gráfico.

Um **PP-plot** é um gráfico semelhante que em vez de confronta quantis confronta probabilidades, $F_n(x)$ contra $F(x)$. A análise é semelhante ao QQ-plot.

Existem outros procedimentos para inferir sobre a forma de uma distribuição (a ver mais a diante).


Valores esperados (Exponenciais)
60
50
40
30
20
10
0
-10
-10
0
10
20
30
40
50
60
70
Valores observados

Uma vez decidida a forma podemos estar interessados em inferir sobre os parâmetros.

Estimativa pontual de um parâmetro desconhecido - valor obtido a partir da amostra (através de uma estatística) que se destina a fornecer valores aproximados do parâmetro.

Estimador - estatística que fornece estimativas pontuais.

Habitualmente representa-se um estimador de um parâmetro colocando um acento circunflexo sobre a letra que o representa. $(\hat{\mu}, \hat{\sigma}, \hat{\theta})$

Um estimador é uma v.a. e como tal tem uma distribuição que o caracteriza - **distribuição por amostragem**.

Propriedades dos estimadores

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

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

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

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

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

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

Como encontrar estimadores?

Intuitivamente sabemos que podemos considerar $\bar{X}$ como estimador de $\mu = E[X]$ assim como s^2 como estimador de $\sigma^2 = Var[X]$. Por vezes não estamos interessados em μ e σ mas sim noutros parâmetros que se podem relacionar com μ e σ.

Existem dois métodos muito utilizados para obter estimadores: método dos momentos e método da máxima verosimilhança. (Nota: Os estimadores obtidos por diferentes métodos podem diferir)

Capítulo 4: Intervalos de confiança e Testes de hipóteses paramétricos

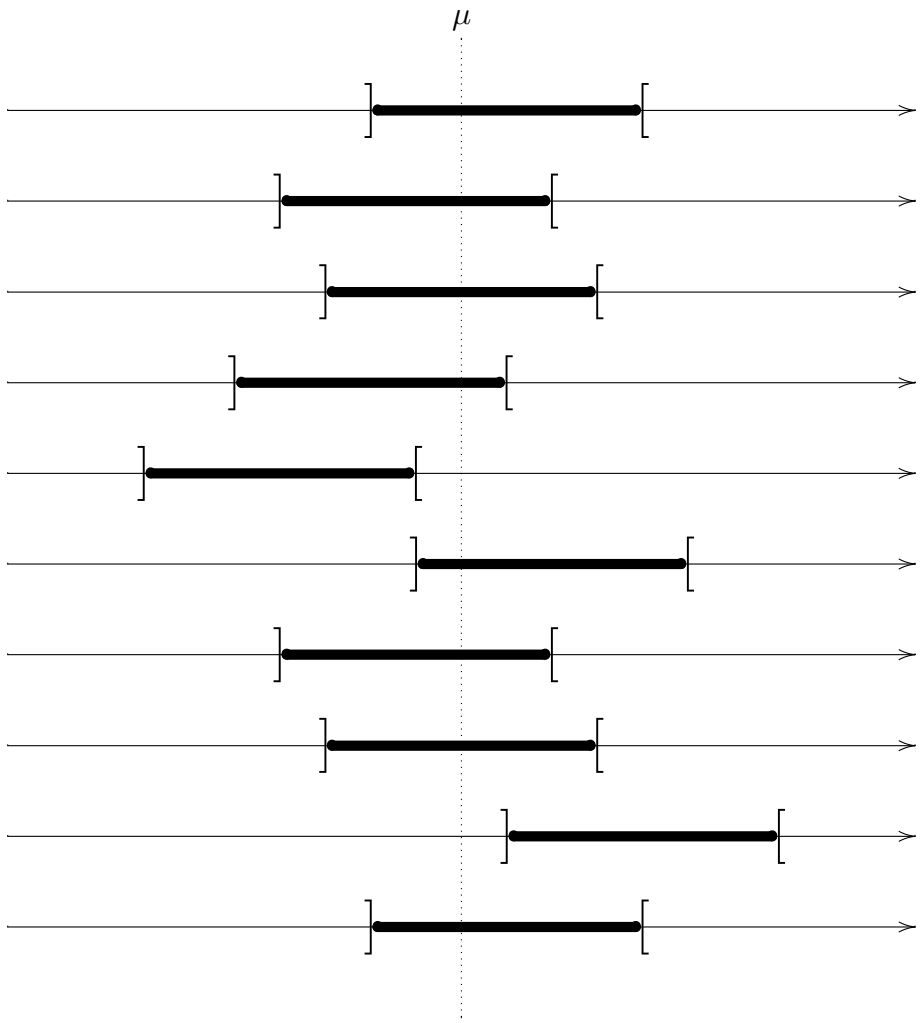
Uma estimativa pontual de um parâmetro não contém informação sobre a precisão do valor obtido. A variância e o EQM fornecem alguma informação. Uma forma mais completa de abordar a questão consiste em construir estimativas na forma de intervalos e conhecer a probabilidade de o intervalo conter o verdadeiro valor do parâmetro.

Um intervalo de confiança para um parâmetro θ, a um níivel de confiança $1-\alpha$, é um intervalo aleatório (θ_1, θ_2) tal que

$$P(\theta_1 < \theta < \theta_2) = 1 - \alpha, \quad \alpha \in (0, 1).$$

(Normalmente α é um valor muito reduzido por forma a termos confianças elevadas.)

Para cada amostra que se observa obtém-se (em geral) um intervalo de confiança diferente para o mesmo parâmetro. Quando dizemos que um intervalo tem confiança $1-\alpha$ estamos a dizer que se observarmos muitas amostras distintas, os intervalos que se obtêm contêm o verdadeiro valor do parâmetro $(1-\alpha) * 100\%$ das vezes.


μ

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

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

$$(\bar{X} - z_{1-\alpha/2}\frac{\sigma}{\sqrt{n}}, \bar{X} + z_{1-\alpha/2}\frac{\sigma}{\sqrt{n}}).$$

Quanto maior o nível de confiança maior a largura do intervalo.

Quanto maior a variância, maior a largura do intervalo,

Quanto maior a amostra, menor a largura do intervalo.

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

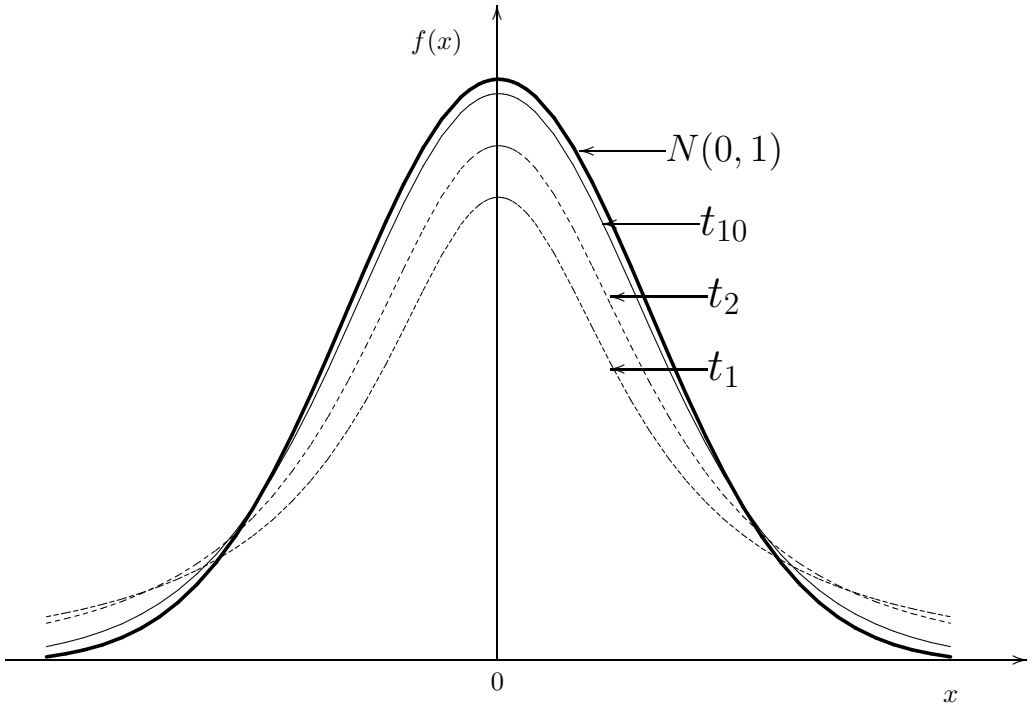
O intervalo de confiança para μ quando a variância é conhecida foi derivado do facto de

$$\frac{\bar{X} - \mu}{\sigma/\sqrt{n}} \frown N(0,1).$$

Se o valor de σ é desconhecido tentamos substitui-lo por uma estimativa S_c ou S. Qual será então a distribuição (de amostragem) da variável

$$\frac{\bar{X} - \mu}{S_c/\sqrt{n}} \frown ?$$

$$T = \frac{\bar{X} - \mu}{S_c/\sqrt{n}} \frown t_{n-1}.$$


$f(x)$
$N(0,1)$
t_{10}
t_2
t_1
0
x

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

$$(\bar{X} - t_{1-\frac{\alpha}{2},n-1}\,\frac{S_c}{\sqrt{n}}, \bar{X} + t_{1-\frac{\alpha}{2},n-1}\,\frac{S_c}{\sqrt{n}}),$$

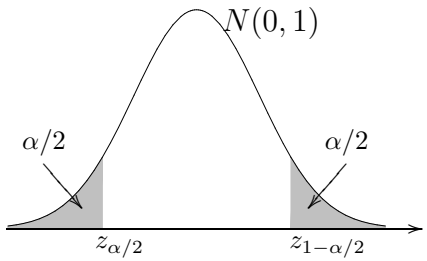
onde $t_{1-\frac{\alpha}{2},n-1}$ representa o quantil de ordem $1-\frac{\alpha}{2}$ da distribuição t de Student com $n-1$ graus de liberdade.

Estes intervalos têm maior largura do que se o valor de σ^2 fosse considerado conhecido, reflectindo a incerteza acrescida pelo desconhecimento deste parâmetro.

$X_1, X_2, \ldots, X_n$ é uma a.a. de dimensão n com distribuição Normal(μ, σ^2), σ conhecido.

$$\bar{X} \text{ estima } \mu$$
$$\bar{X} \frown N(\mu, \sigma^2/n)$$
$$Z = \frac{\bar{X}-\mu}{\sigma/\sqrt{n}} \frown N(0,1).$$

Se $z_{1-\alpha/2}$ representar o quantil de ordem $1-\alpha/2$ de uma Normal standard,

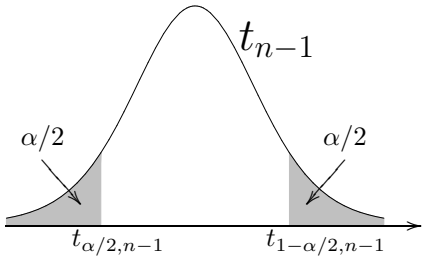


$$P(-z_{1-\alpha/2} < Z < z_{1-\alpha/2}) = 1-\alpha \Leftrightarrow$$
$$P(-z_{1-\alpha/2} < \frac{\bar{X}-\mu}{\sigma/\sqrt{n}} < z_{1-\alpha/2}) = 1-\alpha \Leftrightarrow$$
$$P(\bar{X} - z_{1-\alpha/2}\frac{\sigma}{\sqrt{n}} < \mu < \bar{X} + z_{1-\alpha/2}\frac{\sigma}{\sqrt{n}}) = 1-\alpha.$$

$X_1, X_2, \ldots, X_n$ é uma a.a. de dimensão n com distribuição Normal(μ, σ^2), σ desconhecido.

$$\bar{X} \text{ estima } \mu \text{ e } S_c \text{ estima } \sigma$$
$$\bar{X} \frown N(\mu, \sigma^2/n)$$
$$T = \frac{\bar{X}-\mu}{S_c/\sqrt{n}} \frown t_{n-1}.$$

Se $t_{1-\frac{\alpha}{2},n-1}$ representar o quantil de ordem $1-\frac{\alpha}{2}$ de uma t_{n-1},

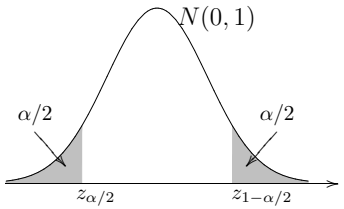


$$P(-t_{1-\frac{\alpha}{2},n-1} < T < t_{1-\frac{\alpha}{2},n-1}) = 1-\alpha \Leftrightarrow$$
$$P(-t_{1-\frac{\alpha}{2},n-1} < \frac{\bar{X}-\mu}{S_c/\sqrt{n}} < t_{1-\frac{\alpha}{2},n-1}) = 1-\alpha \Leftrightarrow$$
$$P(\bar{X} - t_{1-\frac{\alpha}{2},n-1}\frac{S_c}{\sqrt{n}} < \mu < \bar{X} + t_{1-\frac{\alpha}{2},n-1}\frac{S_c}{\sqrt{n}})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

$X_1, X_2, \ldots, X_n$ é uma a.a. de dimensão n com distribuição Normal(μ, σ^2), σ conhecido.

$$\hat{\mu} = \bar{X} \text{ estima } \mu$$
$$\bar{X} \frown N(\mu, \sigma^2_{\bar{X}})$$
$$\sigma_{\hat{\mu}} = \sigma/\sqrt{n}.$$
$$Z = \frac{\hat{\mu} - \mu}{\sigma_{\hat{\mu}}} \frown N(0,1).$$

Se $z_{1-\alpha/2}$ representar o quantil de ordem $1-\alpha/2$ de uma Normal standard,

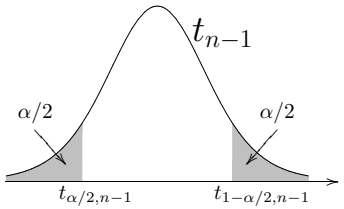


$$P(-z_{1-\alpha/2} < Z < z_{1-\alpha/2}) = 1-\alpha \Leftrightarrow$$
$$P(-z_{1-\alpha/2} < \frac{\hat{\mu} - \mu}{\sigma_{\hat{\mu}}} < z_{1-\alpha/2}) = 1-\alpha \Leftrightarrow$$
$$P(\hat{\mu} - z_{1-\alpha/2}\sigma_{\hat{\mu}} < \mu < \hat{\mu} + z_{1-\alpha/2}\sigma_{\hat{\mu}}) = 1-\alpha.$$

$X_1, X_2, \ldots, X_n$ é uma a.a. de dimensão n com distribuição Normal(μ, σ^2), σ desconhecido.

$$\hat{\mu} = \bar{X} \text{ estima } \mu \text{ e } \hat{\sigma} = S_c \text{ estima } \sigma$$
$$\bar{X} \frown N(\mu, \sigma^2_{\bar{X}})$$
$$\sigma_{\hat{\mu}} = \sigma/\sqrt{n}.$$
$$T = \frac{\hat{\mu} - \mu}{\hat{\sigma}_{\hat{\mu}}} \frown t_{n-1}.$$

Se $t_{1-\frac{\alpha}{2},n-1}$ representar o quantil de ordem $1-\frac{\alpha}{2}$ de uma t_{n-1},



$$P(-t_{1-\frac{\alpha}{2},n-1} < T < t_{1-\frac{\alpha}{2},n-1}) = 1-\alpha \Leftrightarrow$$
$$P(-t_{1-\frac{\alpha}{2},n-1} < \frac{\hat{\mu} - \mu}{\hat{\sigma}_{\hat{\mu}}} < t_{1-\frac{\alpha}{2},n-1}) = 1-\alpha \Leftrightarrow$$
$$P(\hat{\mu} - t_{1-\frac{\alpha}{2},n-1}\,\hat{\sigma}_{\hat{\mu}} < \mu < \hat{\mu} + t_{1-\frac{\alpha}{2},n-1}\,\hat{\sigma}_{\hat{\mu}}) = 1-\alpha.$$

Um IC para μ com σ conhecido é dado por $(\hat{\mu} - z_{1-\frac{\alpha}{2}}\sigma_{\hat{\mu}}, \hat{\mu} + z_{1-\frac{\alpha}{2}}\sigma_{\hat{\mu}})$	Um IC para μ com σ desconhecido é dado por $(\hat{\mu} - t_{1-\frac{\alpha}{2},n-1}\ \hat{\sigma}_{\hat{\mu}}, \hat{\mu} + t_{1-\frac{\alpha}{2},n-1}\ \hat{\sigma}_{\hat{\mu}})$

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

Dadas duas a.a's independentes $X_1, \ldots, X_n, Y_1, \ldots, Y_m$ provenientes de populações Normais $N(\mu_X, \sigma_X^2)$, $N(\mu_Y, \sigma_Y^2)$ respectivamente, podemos considerar a estatística

$$\bar{X} - \bar{Y} \frown N(\mu_X - \mu_Y, \frac{\sigma_X^2}{n} + \frac{\sigma_Y^2}{m}) \Leftrightarrow$$

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

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

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

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

Se não conhecermos as variâncias teremos de assumir que são iguais para podermos obter a distribuição exacta das variáveis em causa.

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

$$\left(\bar{X} - \bar{Y} - t_{1-\frac{\alpha}{2},n+m-2} \sqrt{\frac{1}{n}+\frac{1}{m}} \sqrt{\frac{(n-1)S_{c_X}^2+(m-1)S_{c_Y}^2}{(n+m-2)}}, \right.$$

$$\left. , \bar{X} - \bar{Y} + t_{1-\frac{\alpha}{2},n+m-2} \sqrt{\frac{1}{n}+\frac{1}{m}} \sqrt{\frac{(n-1)S_{c_X}^2+(m-1)S_{c_Y}^2}{(n+m-2)}} \right) .$$

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 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 chamam-se amostras emparelhadas.

O procedimento a seguir é o seguinte:

Dadas duas amostras aleatórias emparelhadas $(X_1, \ldots, X_n)$, $(Y_1, \ldots, Y_n)$ provenientes de populaccões Normais consideram-se as diferenças

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

onde μ_D é igual à diferença das médias das populações e σ_D^2 representa a variância das diferenças D_i.

A variável

$$T = \frac{\bar{D} - \mu_D}{S_{c_D}/\sqrt{n}} \frown t_{n-1}$$

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

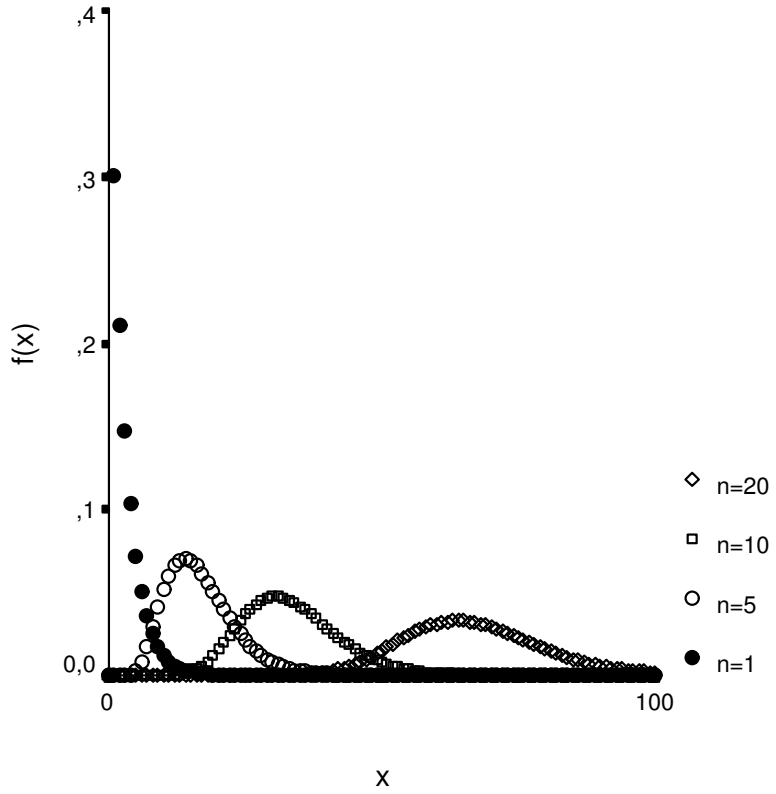
Um intervalo de confiança para a diferença de médias $\mu_X - \mu_Y = \mu_D$ de duas populações Normais com variâncias desconhecidas, obtido a partir de duas amostras emparelhadas, a um nível de confiança $1 - \alpha$, é dado por

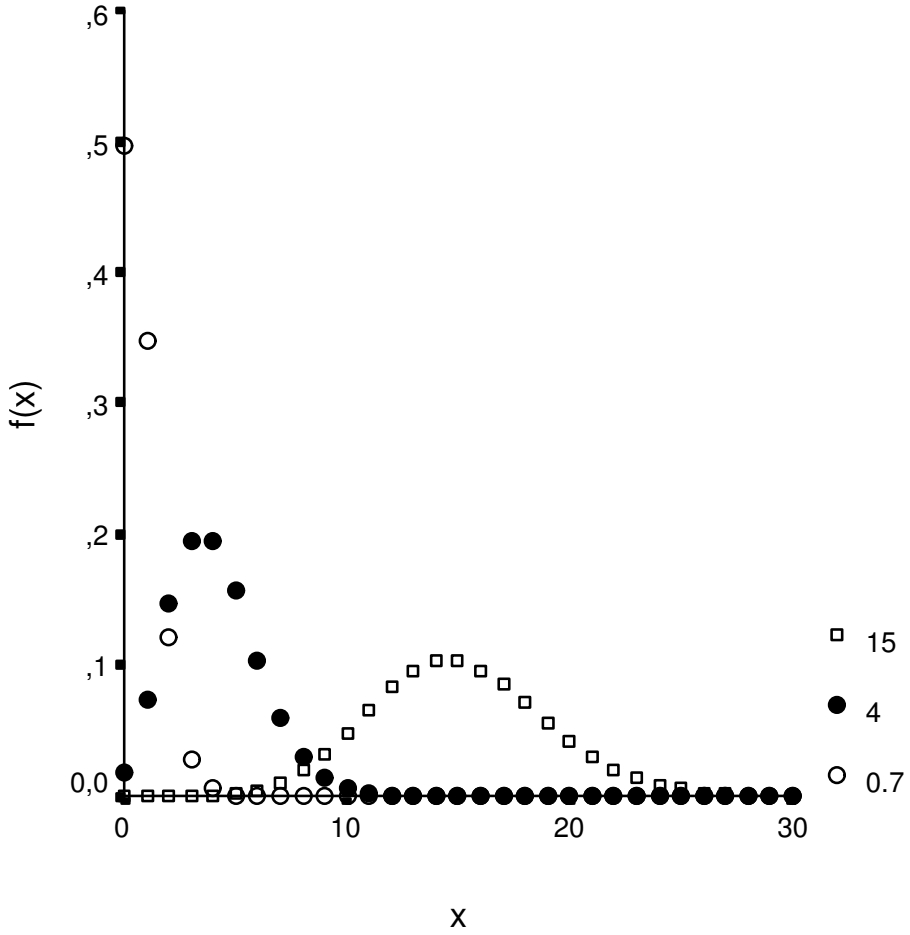
$$\left(\bar{D} - t_{1-\frac{\alpha}{2},n-1} \frac{S_{c_D}}{\sqrt{n}}, \bar{D} + t_{1-\frac{\alpha}{2},n-1} \frac{S_{c_D}}{\sqrt{n}} \right).$$

Teorema do Limite central

Seja $X_1, X_2, \ldots, X_n$ uma sucessão de va's iid com $E[X_i] = \mu$, $Var[X_i] = \sigma^2$, $i = 1, \ldots, n$. Então,

$$\frac{\sum_{i=1}^{n} X_i - n\mu}{\sigma\sqrt{n}} \overset{\circ}{\frown} N(0,1),\ n \to \infty.$$




,6
,5
,4
,3
,2
,1
0,0
f(x)
0
10
20
30
x
15
4
0.7

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

Duma forma geral, conhecendo a variância duma distribuição e considerando válidas as condições do Teorema do Limite Central (n elevado)

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

pelo que podemos obter o mesmo intervalo de confiança para μ:

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

$$(\bar{X} - z_{1-\alpha/2}\frac{\sigma}{\sqrt{n}}, \bar{X} + z_{1-\alpha/2}\frac{\sigma}{\sqrt{n}}).$$

Esta aproximação será tanto melhor quanto maior a dimensão da amostra.

Quando não se conhece a variância σ^2 é usual substiuir σ por S_c ou S e utilizar o mesmo intervalo. Chama-se a atenção que este procedimento só deve ser utilizado em grandes amostras.

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

$$(\bar{X} - z_{1-\alpha/2}\frac{S_c}{\sqrt{n}}, \bar{X} + z_{1-\alpha/2}\frac{S_c}{\sqrt{n}}).$$

Esta aproximação será tanto melhor quanto maior a dimensão da amostra.

Testes de Hipóteses

Objectivo: Testar a validade de um modelo; testar se um modelo mudou em relação a resultados do passado; testar se os modelos para descrever populações distintas devem ou não ser diferentes.

Exemplos: • Testar se uma distribuição é Normal com média $\mu = 5$ ou com outra média $\mu \neq 5$.

• Testar se a distribuição da população de onde provêm os dados é de Poisson.

As hipóteses: Num teste de hipóteses há sempre duas hipóteses:

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

Exemplos:

1. $H_0 : \mu = 5 \quad vs \quad H_1 : \mu \neq 5.$ (População Normal)
2. $H_0 : \mu = 3 \quad vs \quad H_1 : \mu < 3.$ (População Normal)
3. $H_0 : \mu = 3 \quad vs \quad H_1 : \mu > 3.$ (População Normal)
4. $H_0 : \mu = 4 \quad vs \quad H_1 : \mu = 7.$ (População Normal)
5. $H_0 : \mu > 1 \quad vs \quad H_1 : \mu \leq 1.$ (População Normal)
6. $H_0 : \mu_X = \mu_Y \quad vs \quad H_1 : \mu_X \neq \mu_Y.$ (População Normal)
7. $H_0 : \sigma^2 = 1 \quad vs \quad H_1 : \sigma^2 > 1.$ (População Normal)
8. $H_0 : X \frown Poisson \quad vs \quad H_1 : X \frown$ outra distribuição.

Tipos de hipóteses : As várias hipóteses podem ser **simples** ou **compostas**. Uma hipótese simples apenas contempla uma possibilidade.

Iremos apenas considerar testes em que H_0 é simples.

Tipos de testes: Os testes podem ser **unilaterais** ou **bilaterais**.

São unilaterais os testes dos exemplos 2, 3, 4, 5 e 7.

São bilaterais os testes dos exemplos 1 e 6.

A realização de um teste de hipóteses consiste em aceitar ou não a hipótese nula H_0, de acordo com uma regra de decisão baseada numa estatística (da amostra). A esta estatística chama-se **estatística de teste** e costuma-se representar por T.

O conjunto de todos os possíveis valores que a estatística de teste pode assumir é dividido em dois subconjuntos - **região de rejeição** (ou **região crítica**) e **região de não rejeição**. Se T pertencer à região crítica rejeita-se H_0 a favor de H_1. Caso contrário não se rejeita H_0.

(Dada a incerteza associada a um teste de hipóteses não se costuma dizer que se aceita H_0, mas sim que não se rejeita H_0.)

Tipos de erros:

		Hipótese verdadeira	
		H_0	H_1
Decisão do teste	Rejeito H_0	Erro de tipo I	$\surd$
	Não rejeito H_0	$\surd$	Erro de tipo II

$$\boldsymbol{P(\text{Erro de tipo I}) = \alpha} \qquad \boldsymbol{P(\text{Erro de tipo II}) = \beta.}$$

Tamanho do teste ou Nível de significância:

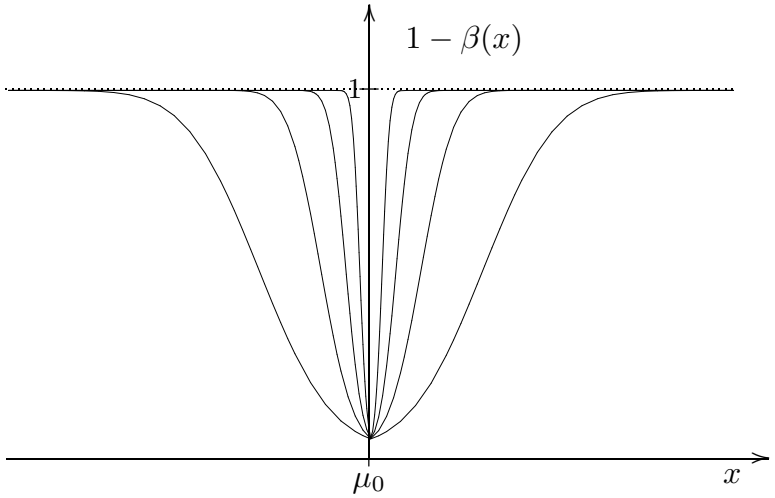
$$\alpha = P(\text{Erro de tipo I}) = P(\text{rejeitar } H_0 | H_0 \text{ verdadeiro}).$$

Potência do teste:

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

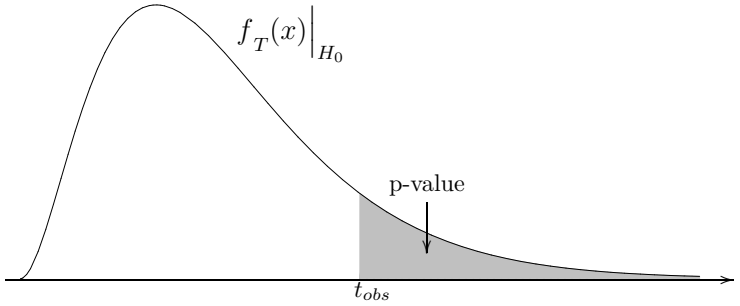
O ideal seria $\alpha = \beta = 0$. Na prática procura-se minimizar β após se ter fixado o valor de α (reduzido). Esta minimização baseia-se na escolha do melhor teste entre os conhecidos. Nós iremos apresentar apenas os melhores testes para cada situação. Notar que ao aumentarmos a dimensão da amostra conseguimos reduzir β, para um α fixo.

Só com uma hipótese nula simples é que podemos predefinir α. Se a hipótese nula for composta teremos´um valor de α para cada valor possível do parâmetro em H_0. De igual forma, β assume pode assumir diferentes valores se a hipótese alternativa for composta. Neste caso dizemos que temos uma função potência e não apenas uma potência.

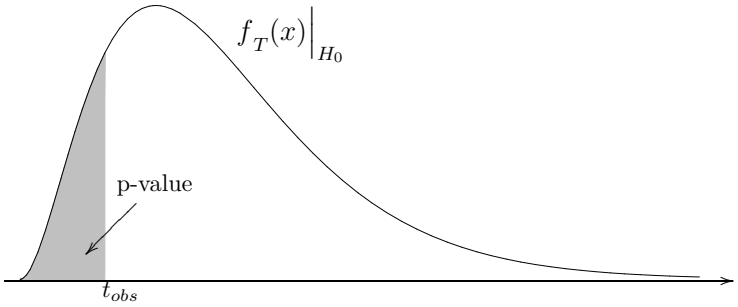


***p-value* do teste** é a probabilidade de observar um valor da estatística de teste tanto ou mais afastado que o valor observado na amostra, assumindo que H_0 é verdadeira. Equivalentemente podemos definir o *p-value* como sendo o menor tamanho do teste que conduz à rejeição de H_0, para uma dada amostra observada. O *p-value* é muito utilizado quando se fazem testes de hipóteses através de software estatístico.

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



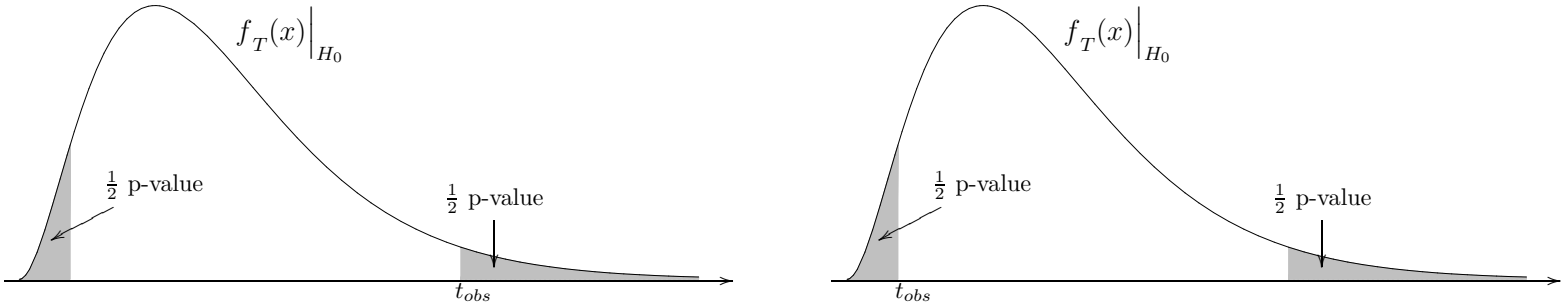
- Quando a região de rejeição é da forma $T < c$ (rejeitar para valores reduzidos da estatística de teste), o *p-value* é igual a $P(T < t_{obs}|H_0)$.



- 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

$$\begin{cases} 2P(T < t_{obs}|H_0) & \text{se } t_{obs} \text{ for reduzido} \\ 2P(T > t_{obs}|H_0) & \text{se } t_{obs} \text{ for elevado} \end{cases}.$$

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.



Procedimentos para a realização de um teste de hipóteses de tamanho α:

1-Procedimento com base na região de rejeição

1. Identificar o parâmetro de interesse e especificar as hipóteses H_0, H_1.
2. Escolher uma estatística de teste, T, com distribuição conhecida (admitindo que H_0 é verdadeira).
3. Identificar a região de rejeição.
4. Calcular t_{obs} que é o valor que T assume para os dados observados.
5. Tomar uma decisão.
6. Concluir.

2-Procedimento alternativo com base nos intervalos de confiança (válido apenas para testes bilaterais)

1. Identificar o parâmetro de interesse e especificar as hipóteses H_0, H_1.
2. Construir um intervalo de confiança para o parâmetro.
3. Rejeitar H_0 se o valor do parâmetro especificado em H_0 não pertencer ao intervalo de confiança. (O intervalo de confiança fornece a região de não rejeição do teste.)

3-Realização de um teste com base no p-value (útil em aplicações computacionais)

1. Identificar o parâmetro de interesse e especificar as hipóteses H_0, H_1.
2. Escolher uma estatística de teste, T, com distribuição conhecida (admitindo que H_0 é verdadeira).
3. Identificar a forma da região de rejeição sem especificar valores críticos.
4. Determinar t_{obs} para a amostra em causa.
5. Determinar o *p-value* do teste

$$\text{Rejeitar } H_0 \text{ se } p\text{-}value < \alpha.$$

$$\text{Não rejeitar } H_0 \text{ se } p\text{-}value \geq \alpha.$$

6. Concluir.

Procedimento para transformação de p-*values* bilaterais em unilaterais.

Por vezes o software estatístico apenas fornece *p-value* bilaterais. Se o teste em causa for unilateral há que saber transformar o *p-value*.

- se a(s) amostra(s) aponta(m) no sentido da hipótese alternativa deve-se dividir o *p-value* por 2 e tomar esse valor como o *p-value* do teste unilateral, $p\text{-}value_{uni} = p\text{-}value_{bil}/2$;
- se a(s) amostra(s) não aponta(m) no sentido da hipótese alternativa, então o *p-value* do teste unilateral é igual a $p\text{-}value_{uni} = 1 - p\text{-}value_{bil}/2$.

Escolha de estatísticas de testes em Populações Normais

1. **Teste para a média μ, variância conhecida,** $H_0 : \mu = \mu_0$.

$$T = \frac{\bar{X} - \mu_0}{\sigma/\sqrt{n}} \underset{sob\ H_0}{\frown} N(0,1)$$

2. **Teste para a média μ, variância desconhecida,** $H_0 : \mu = \mu_0$ **(SPSS)**.

$$T = \frac{\bar{X} - \mu_0}{S_c/\sqrt{n}} \underset{sob\ H_0}{\frown} t_{n-1}$$

3. **Testes para a comparação de médias,** $H_0 : \mu_X - \mu_Y = 0$

   - **variâncias conhecidas (amostras independentes)**.

$$T = \frac{\bar{X} - \bar{Y}}{\sqrt{\frac{\sigma_X^2}{n} + \frac{\sigma_Y^2}{m}}} \underset{sob\ H_0}{\frown} N(0,1)$$

- **variâncias desconhecidas (amostras independentes) (SPSS).**

$$T = \frac{\bar{X} - \bar{Y}}{\sqrt{\frac{1}{n} + \frac{1}{m}}\sqrt{\frac{(n-1)S^2_{c_X} + (m-1)S^2_{c_Y}}{(n+m-2)}}} \underset{sob\ H_0}{\frown} t_{n+m-2}$$

- **variâncias desconhecidas (amostras emparelhadas) (SPSS).**

$$T = \frac{\bar{D}}{S_{c_D}/\sqrt{n}} \underset{sob\ H_0}{\frown} t_{n-1}.$$

4. **Teste para a variância $H_0 : \sigma^2 = \sigma_0^2$.**

$$T = \frac{(n-1)S_c^2}{\sigma_0^2} \underset{sob\ H_0}{\frown} \chi^2_{n-1}.$$

5. **Teste para o desvio padrão $H_0 : \sigma = \sigma_0$.**

   Efectua-se o teste para a variância correspondente.

6. **Teste para a comparação de variâncias $H_0 : \sigma_X^2/\sigma_Y^2 = 1$ (amostras independentes) (SPSS noutras análises, por exemplo ANOVA).**

$$T = \frac{S_{c_X}^2}{S_{c_Y}^2} \underset{sob\ H_0}{\frown} F_{(n-1),(m-1)}.$$

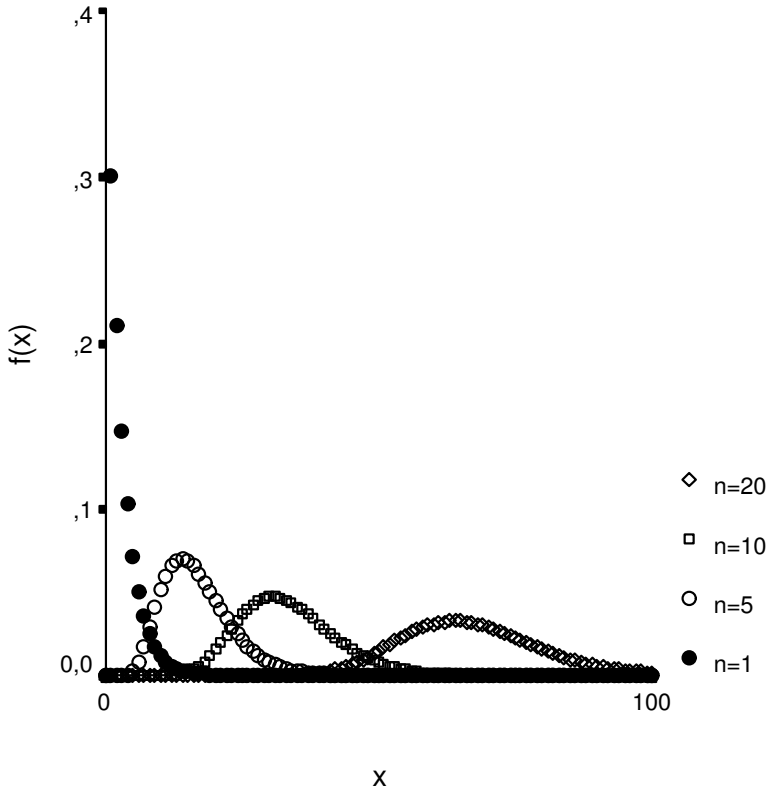
Testes de hipóteses em populações que verificam o TLC

Testes para a média μ

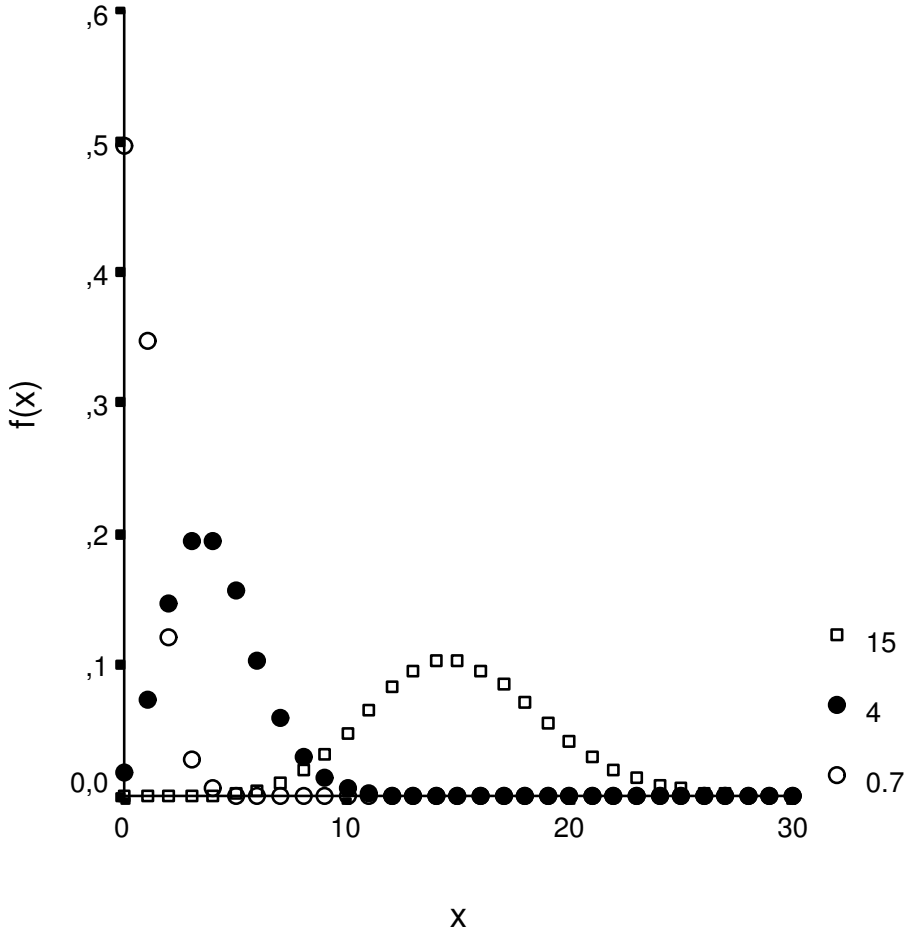
Teorema do Limite central

Seja $X_1, X_2, \ldots, X_n$ uma sucessão de va's iid com $E[X_i] = \mu$, $Var[X_i] = \sigma^2$, $i = 1, \ldots, n$. Então,

$$\frac{\sum_{i=1}^{n} X_i - n\mu}{\sigma\sqrt{n}} = \frac{\bar{X} - \mu}{\sigma/\sqrt{n}} \overset{\circ}{\frown} N(0,1),\ n \to \infty.$$



Exemplo: soma de v.a.'s Geométricas



Exemplo: v.a.'s de Poisson

Seja $(X_1, \cdots, X_n)$ uma a.a. proveniente de uma população X com determinada função de distribuição F (cuja forma analítica é desconhecida) com média μ e variância σ^2.

Como $\frac{\bar{X}-\mu}{\sigma/\sqrt{n}} \overset{\circ}{\frown} N(0,1)$ a construção de um IC ou realização de um teste de hipóteses são perfeitamente análogas ao realizado em contexto Normal. No entanto, apenas podemos dizer que a confiança do intervalo é aproximadamente $1-\alpha$ ou que o tamanho do teste é aproximadamente $\alpha \in (0,1)$.
Num teste de hipóteses com $H_0 : \mu = \mu_0$

a estatística de teste a utilizar é

$$T = \frac{\bar{X} - \mu_0}{\sigma/\sqrt{n}} \underset{sob\ H_0}{\overset{\circ}{\frown}} N(0,1) \qquad \text{ou} \qquad \frac{\bar{X} - \mu_0}{S_c/\sqrt{n}} \overset{\circ}{\frown} N(0,1)$$

O SPSS não realiza estes testes mas, como para n elevado (>>100) os quantis da t de Student confundem-se com os da Normal standard, podemos utilizar o teste fornecido pelo SPSS (one-sample T test ou 2-samples T test) como alternativa.

Testes não paramétricos

Testes não paramétricos são testes de hipóteses que não requerem muitos pressupostos sobre a distribuição subjacente aos dados.

- **Vantagens dos testes não paramétricos:**
  - Se a dimensão da amostra é muito pequena, pode não haver alternativa senão o recurso a testes não paramétricos, a não ser que a distribuição exacta da população seja conhecida.
  - Os testes não paramétricos requerem usualmente poucos pressupostos acerca dos dados e podem ser mais relevantes para uma determinada situação prática.
  - Estão disponíveis testes não paramétricos para analisar dados com uma ordem inerente bem como dados cujos valores numéricos se equiparam a *ranks*. Isto significa que, poderá ser possível afirmar que a presença de determinada característica se acentua mais ou menos num indivíduo do que noutro sem ser necessário quantificar "quão mais ou menos".
  - Os métodos não paramétricos podem ser aplicados a dados categorizados, ou seja, que são medidos numa escala nominal. Nenhuma técnica paramétrica se aplica a tais dados.

- **Desvantagens dos testes não paramétricos:**
  - Se todos os pressupostos de um modelo estatístico paramétrico forem satisfeitos e as hipóteses de interesse puderem ser testadas usando testes paramétricos, estes gozarão de preferência sobre testes não paramétricos por serem mais potentes.
  - Ao contrário dos testes paramétricos que têm sido sistematizados de tal modo que testes diferentes são simplesmente uma variação de um tema central, os testes não paramétricos são alicerçados em propriedades empíricas.

- **Testes não paramétricos no SPSS - menu Analyse/Nonparamentric Tests**

  Sempre que possível deve seleccionar-se a opção `exact` com um tempo de execução limitado (1 minuto) pois caso contrário a maioria dos testes serão efectuados através de distribuições aproximadas.

Sub-menus disponibilizados:

Qui-squared Teste para dados categóricos ou agrupados em classes.

Binomial Teste a uma proporção.

Runs Teste de aleatoridade de sequências.

1-Sample KS Teste de ajustamento para averiguar se a população de onde foram retirados os dados tem determinada distribuição (por exemplo Normal).

2 Independent Samples Alternativa ao teste t para comparação de médias em amostras independentes (**Mann-Whitney U**) e outros testes diferentes.

K Independent Samples Alternativa à ANOVA a um factor clássica (**Kruskal-Wallis**) e outros testes.

2 Related samples Alternativas ao teste t para comparação de médias em amostras emparelhadas (**Sign Test** e **Wilcoxon**) e outros testes diferentes.

K Related Samples Alternativa à ANOVA a um factor com observações repetidas (**Friedman**) e outros testes.

Teste Binomial

Serve para testar se a proporção de indivíduos (na população) com determinada característica é significativamente diferente (ou superior ou inferior)a um certo valor p_0 $(0 < p_0 < 1)$.

$$H_0 : p = p_0 \quad vs \qquad H_1 : \ p \overset{>}{\underset{<}{\neq}} p_0,$$

Estatística de teste: Numa amostra de dimensão n contam-se os indivíduos com a referida característica, X.

$$X \underset{sob\ H_0}{\frown} B(n, p)$$

Binomial Test

		Category	N	Observed Prop.	Test Prop.	Exact Sig. (1-tailed)
Idade em anos	Group 1	<= 60	11	,69	,7	,550[a]
	Group 2	> 60	5	,7		
	Total		16	1,0		

a. Alternative hypothesis states that the proportion of cases in the first group < ,7.

Binomial Test

		Category	N	Observed Prop.	Test Prop.	Exact Sig. (2-tailed)
Sexo	Group 1	Masculino	9	,56	,50	,804
	Group 2	Feminino	7	,44		
	Total		16	1,00		

Quando a proporção em teste é 0.5 o SPSS fornece op-value bilateral. Caso contrário fornece o unilateral e especifica qual a hipóteses alternativa.

Teste dos sinais

Este teste serve para testar valores de quantis de um distribuição.

$$H_0 : x_p = x_0 \quad vs \qquad H_1 : \; x_p \overset{>}{\underset{<}{\neq}} x_0,$$

Em geral utiliza-se para o quantil 0.5, ou seja para a mediana e passa a ser uma alternativa não-paramétrica ao teste t para a média ou comparação de médias (amostras emparelhadas).

$$H_0 : \mu = \mu_0 \quad vs \qquad H_1 : \; \mu \overset{>}{\underset{<}{\neq}} \mu_0,$$

Os pressupostos do teste são os seguintes:

- Os dados disponíveis para análise constituem uma amostra aleatória de n observações, $X_1, \ldots, X_n$;
- A variável de interesse é medida numa escala que seja, pelo menos *ordinal*;
- A f.d. F da população X é *contínua* (para garantir que o quantil x_p é um valor único). Este pressuposto é por vezes contornado.

No SPSS, este teste surge apenas para a mediana e para amostras emparelhadas.

Frequencies

		N
MED80 - Peso inicial	Negative Differences[a]	11
	Positive Differences[b]	5
	Ties[c]	0
	Total	16

a. MED80 < Peso inicial

b. MED80 > Peso inicial

c. Peso inicial = MED80

Test Statistics[b]

	MED80 - Peso inicial
Exact Sig. (2-tailed)	,210[a]
Exact Sig. (1-tailed)	,105
Point Probability	,067

a. Binomial distribution used.

b. Sign Test

Teste para a mediana μ - Wilcoxon signed-rank test

Teste para a mediana ou comparação de medianas em amostras emparelhadas. (No SPSS este teste só aparece para comparação de duas amostras emparelhadas.)

Pressupostos:

- Os dados disponíveis para análise constituem uma realização de uma a.a.;
- A variável de interesse é medida numa escala, pelo menos, *ordinal*;
- A f.d. F da população X é *contínua* e *simétrica relativamente à sua mediana*.

Nota: Se a distribuição for simétrica e tiver média finita a mediana é igual à média.

$$H_0: \ \mu = \mu_0 \quad vs \quad H_1: \mu \overset{>}{\underset{<}{\neq}} \mu_0 \quad \text{(uma amostra)}$$

$$H_0: \ \mu_D = 0 \quad vs \quad H_1: \mu_D \overset{>}{\underset{<}{\neq}} 0 \quad \text{(duas amostras emparelhadas)}$$

onde $\mu_D = \mu - \mu_0$.

Ranks

		N	Mean Rank	Sum of Ranks
MED80 - Peso inicial	Negative Ranks	11[a]	9,91	109,00
	Positive Ranks	5[b]	5,40	27,00
	Ties	0[c]		
	Total	16		

a. MED80 < Peso inicial

b. MED80 > Peso inicial

c. Peso inicial = MED80

Test Statistics[b]

	MED80 - Peso inicial
Z	-2,121[a]
Asymp. Sig. (2-tailed)	,034
Exact Sig. (2-tailed)	,032
Exact Sig. (1-tailed)	,016
Point Probability	,001

a. Based on positive ranks.

b. Wilcoxon Signed Ranks Test

Teste para a comparação de medianas (amostras independentes) - Mann-Whitney U, ou Wilcoxon rank-sum ou Wilcoxon Mann-Whitney test

$$H_0: \ \mu_X - \mu_Y = 0 \quad vs \quad H_1: \mu_X - \mu_Y \overset{>}{\underset{<}{\neq}} 0$$

Pressupostos do teste:

- A variável de interesse é medida numa escala susceptível de ser ordenada com subsequente atribuição de ordens ou *ranks*.
- Os dados disponíveis para análise são compostos por uma realização de duas a.a.'s provenientes de duas populações de interesse.
- As duas amostras, $(X_1, ..., X_n)$ e $(Y_1, ..., Y_m)$, são independentes.
- As f.d.'s das populações X e Y são *contínuas*.
- As distribuições na génese das amostras são idênticas no que respeita à forma. Todavia, não é imperativo que sejam normais.

Ranks

	Sexo	N	Mean Rank	Sum of Ranks
Peso inicial	Masculino	9	12,00	108,00
	Feminino	7	4,00	28,00
	Total	16		

Test Statistics[b]

	Peso inicial
Mann-Whitney U	,000
Wilcoxon W	28,000
Z	-3,339
Asymp. Sig. (2-tailed)	,001
Exact Sig. [2*(1-tailed Sig.)]	,000[a]
Exact Sig. (2-tailed)	,000
Exact Sig. (1-tailed)	,000
Point Probability	,000

a. Not corrected for ties.

b. Grouping Variable: Sexo

Teste de ajustamento - teste de Kolmogorov-Smirnov

Pressupostos do teste: a amostra provém de uma distribuição contínua.

$$H_0 : F(x) = F_0(x), \text{ para todo o } x \quad vs \qquad H_1 : \ F(x) \neq F_0(x), \text{ para algum } x$$

Estatística de teste:

$$D = \sup_{x \in \mathbb{R}} |F_n(x) - F(x)|$$

One-Sample Kolmogorov-Smirnov Test

		Peso inicial
N		16
Normal Parameters[a,b]	Mean	89,98
	Std. Deviation	15,183
Most Extreme Differences	Absolute	,156
	Positive	,156
	Negative	-,138
Kolmogorov-Smirnov Z		,625
Asymp. Sig. (2-tailed)		,830
Exact Sig. (2-tailed)		,775
Point Probability		,000

a. Test distribution is Normal.

b. Calculated from data.

Teste de Kolmogorov-Smirnov para duas amostras

Hipóteses:

$$H_0 : F_X(x) = F_Y(x), \text{ para todo o } x \quad vs \qquad H_1 : F_X(x) \neq F_Y(x), \text{ para algum } x$$

Estatística de teste:

$$D = \max_{x \in \mathbb{R}} |F_{X,n}(x) - F_{Y,m}(x)|$$

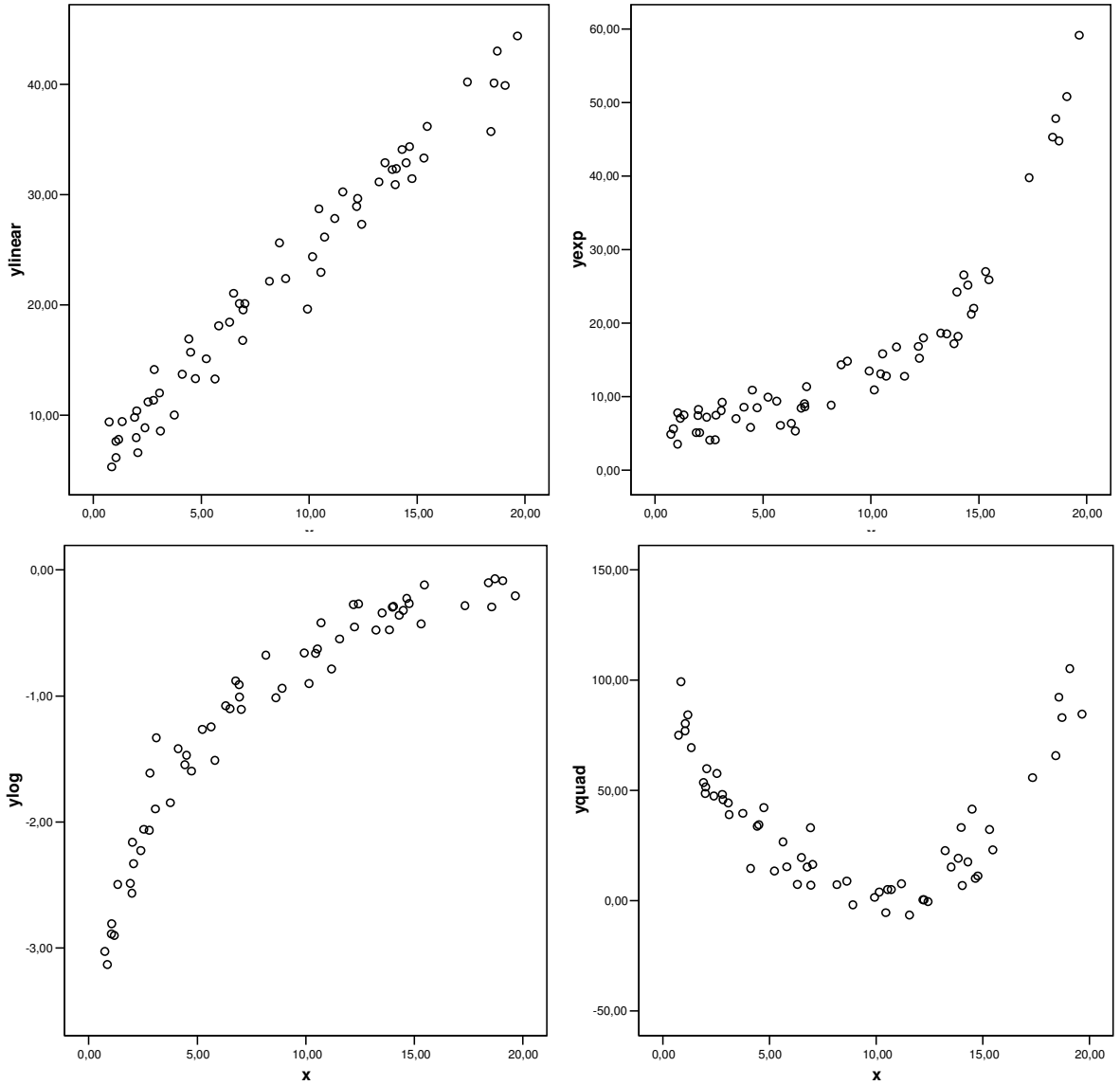
Associação entre variáveis

Questões de interesse:

Será que duas variáveis são independentes ou pelo contrário dependentes? E se forem dependentes, qual o tipo e grau de dependência?

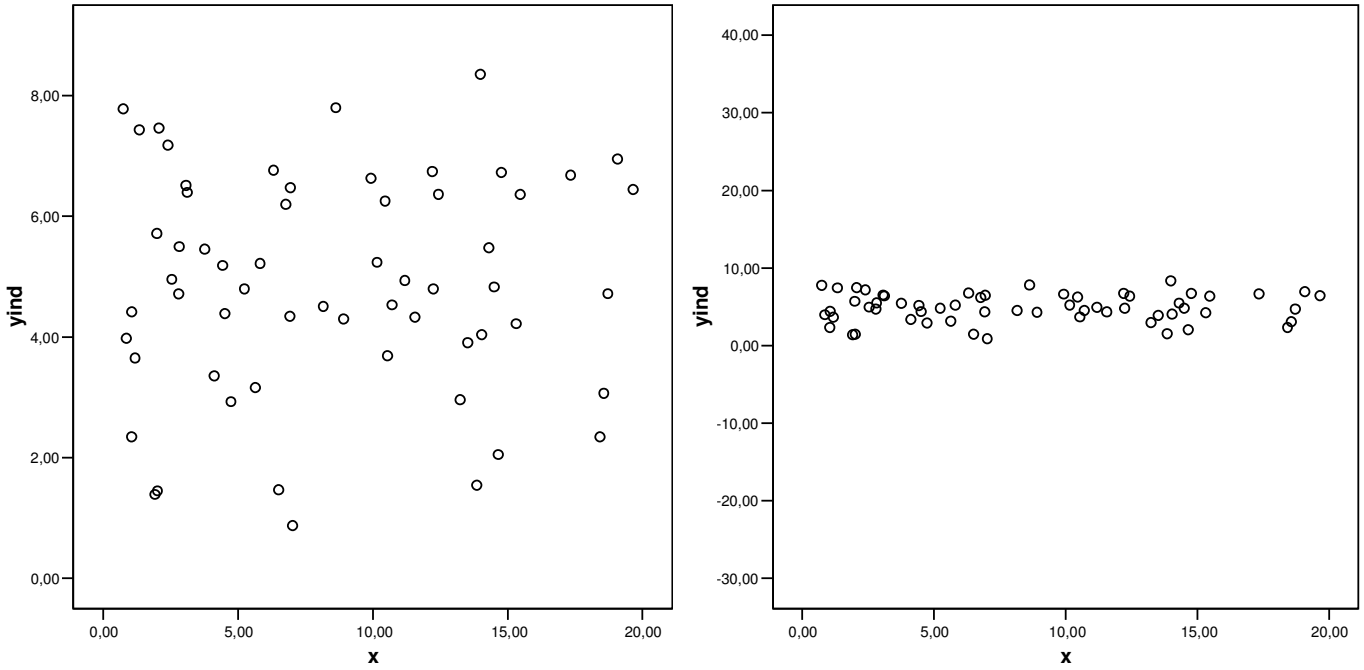
Medir o grau de dependência é mais ambicioso do que simplesmente testar a existência de alguma associação entre variáveis. É obviamente de interesse poder medir o grau de associação entre dois conjuntos de observações obtidos a partir de um dado conjunto de unidades experimentais (indivíduos por exemplo). Mas, talvez seja mais importante podermos dizer se uma certa associação observada nos dados indica ou não uma associação na população de onde foram retirados.

Formas de associação entre variáveis numéricas: lineares, exponenciais, logarítmicas ou quadráticas.

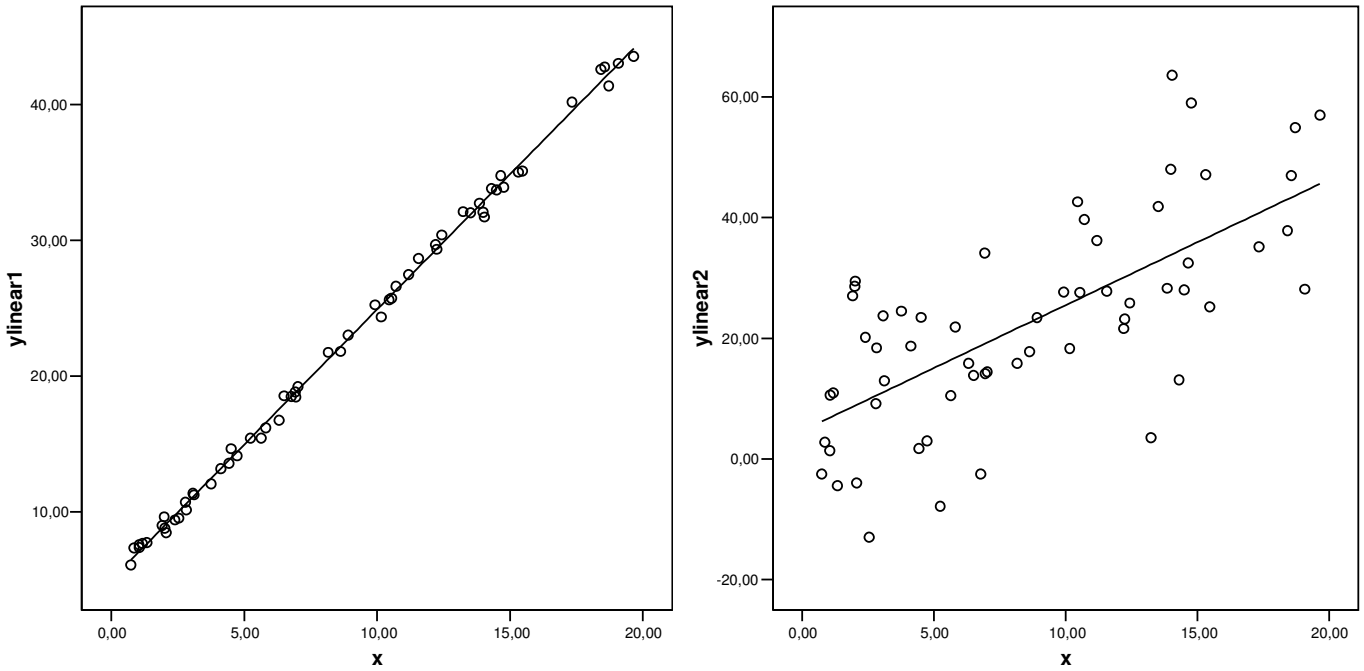

ylinear
40,00
30,00
20,00
10,00
0,00
5,00
10,00
15,00
20,00
yexp
60,00
50,00
40,00
30,00
20,00
10,00
0,00
ylog
0,00
-1,00
-2,00
-3,00
x
yquad
150,00
100,00
50,00
0,00
-50,00

Primeiro passo: construção de diagramas de dispersão.

Quando duas variáveis são independentes, o diagrama de dispersão respectivo apresenta uma mancha de pontos aleatória (ou quando muito) um conjunto de pontos dispostos sobre uma recta horizontal.



Se a relação entre duas variáveis for linear, ao confrontarmos duas amostras num diagrama de dispersão devemos esperar observar um conjunto de pontos que se dispõem aproximadamente sobre uma recta. Por vezes os desvios em relação à recta são mínimos, mas noutras os pontos apresentam bastante dispersão tornando difícil a identificação da dita relação linear.

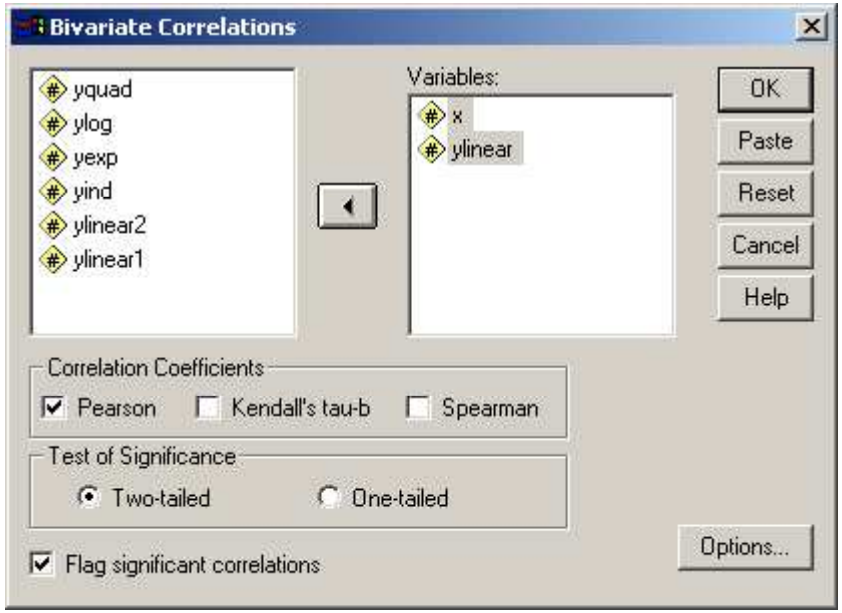


Segundo passo: calcular medidas de associação.

Último passo: realizar um teste de hipóteses para averiguar se os valores das medidas de associação observados nos dados são significativos, ou seja se podemos estatisticamente concluir a favor de uma associação na população.

Medidas de associação para dados numéricos ou ordinais

No SPSS os coeficientes de associação (correlação) para dados numéricos ou ordinais podem ser obtido através do menu `Analyse / Correalte / Bivariate`.



Neste menu podem-se seleccionar mais do que duas variáveis, caso em que o SPSS fornece uma tabela de correlações para todas as combinações de pares de variáveis. O SPSS fornece também o p-value dos testes ao significado dos coeficientes, para cada par de variáveis.

1 - O coeficiente de correlação de Pearson (*Pearson product-moment correlation coefficient*)

Dadas duas amostras de observações medidas numa escala de intervalos ou razões, podemos medir o grau de associação **linear** através da estatística

$$r = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2}\sqrt{\sum_{i=1}^{n}(y_i - \bar{y})^2}}.$$

r pertence ao intervalo $[-1, 1]$. Se $r = 1$ temos uma recta perfeita com declive positivo. Se $r = -1$ temos uma recta perfeita com declive negativo. Se as variáveis são independentes $r \simeq 0$.

Uma interpretação usual: r^2 mede a percentagem de variabilidade de uma das variáveis explicada pela outra.

Podemos testar se duas variáveis são correlacionadas através das hipóteses:

$$H_0 : \ \rho = 0 \quad vs \quad H_1 : \ \rho \neq 0$$

onde ρ representa o coeficiente de correlação da população onde foram retirados os dados.

Pressupostos do teste

1. ambas as populações de onde foram retirados as amostras têm distribuição Normal,
2. a relação entre as variáveis é de forma linear, caso exista.

No SPSS o coeficiente de Pearson pode ser obtido através do menu `Analyse / Correalte / Bivariate`.

2 - O coeficiente de correlação de Spearman (*Spearman rank-order coefficient*)

Aplica-se a duas variáveis medidas apenas numa escala ordinal, ou que apresentam uma relação não linear mas monótona (se uma aumenta a outra tem sempre tendência a aumentar (ou a diminuir)). Aplica-se ainda quando não são satisfeitos os requisitos to teste ao coeficiente de Pearson (variáveis não Normais).

Dadas duas amostras de observação ordenáveis, substitui-se cada um dos seus valores pela sua ordem de ordenação, em inglês *rank*. O coeficiente de Spearman não é mais do que o coeficiente de Pearson aplicado aos *ranks*.

$$r_s = 1 - \frac{6\sum_{i=1}^{n} d_i^2}{n^3 - n}$$

onde d_i representa a diferença de *ranks* correspondentes a cada par de observações x_i, y_i.

Tal como no caso do coeficiente de Pearson é possível testar as hipóteses

$$H_0: \ \rho = 0 \quad vs \quad H_1: \ \rho \neq 0.$$

Tal como para o coeficiente de Pearson, no SPSS o coeficiente de Spearman pode ser obtido através do menu `Analyse / Correalte / Bivariate`.

3- O coeficiente de correlação de Kendall

Uma alternativa ao coeficiente de Spearman é o coeficiente de Kendall (*Kendall's tau coeffficient*) que se aplica nas mesmas condições.

Uma diferença muito importante entre os dois coeficientes (Kendall e Spearman) reside na sua interpretação e na impossibilidade de comparar directamente valores provenientes de ambos. Embora o objectivo comum seja o de medir associação, a forma de o fazer é distinta.

O coeficiente de Kendall é muitas vezes descrito como uma medida de concordância entre dois conjuntos de classificações relativas a um conjunto de objectos ou experiências.

$$T = \frac{\#\text{concordâncias} - \#\text{discordâncias}}{\text{número total de pares}}$$

Tal como para os coeficientes de Pearson e Spearman é possível efectuar um teste de hipóteses para averiguar se a associação é significativa.

No SPSS o coeficiente de Kendall pode ser obtido através do menu `Analyse / Correalte / Bivariate`.

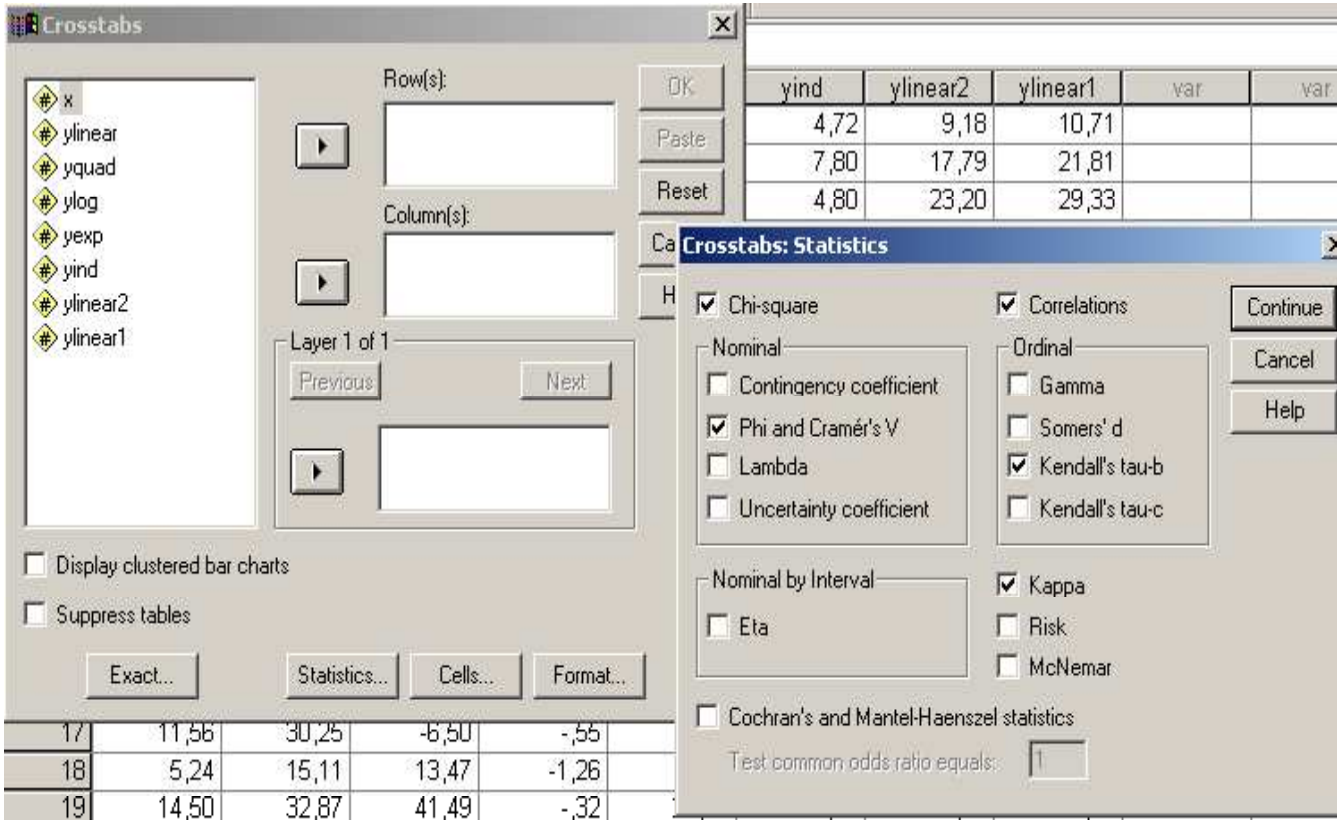
Medidas de associação para dados categóricos

Dados apresentados em **tabelas de contingência**. Por exemplo:

Sexo	Patologia		Total
	Presente	Ausente	
Feminino	30	20	50
Masculino	15	35	50
Total	45	55	100

As medidas de associação e respectivos testes de hipóteses para dados organizados em tabelas de contingência estão disponíveis no SPSS através do menu `Analyze / Descriptive Statistics / Crosstabs`.

Primeiramente há que introduzir os dados da tabela de contingência e seleccionar o menu `Data / Weight cases` por forma a atribuir pesos correspondentes às frequências observadas para cada célula.


Crosstabs
x
ylinear
yquad
ylog
yexp
yind
ylinear2
ylinear1
Row(s):
Column(s):
Layer 1 of 1
Previous
Next
OK
Paste
Reset
Display clustered bar charts
Suppress tables
Exact...
Statistics...
Cells...
Format...
yind
ylinear2
ylinear1
var
4,72
9,18
10,71
7,80
17,79
21,81
4,80
23,20
29,33
Crosstabs: Statistics
Chi-square
Correlations
Nominal
Contingency coefficient
Phi and Cramér's V
Lambda
Uncertainty coefficient
Ordinal
Gamma
Somers' d
Kendall's tau-b
Kendall's tau-c
Nominal by Interval
Eta
Kappa
Risk
McNemar
Cochran's and Mantel-Haenszel statistics
Test common odds ratio equals:
Continue
Cancel
Help
17 11,56 30,25 -6,50 -,55
18 5,24 15,11 13,47 -1,26
19 14,50 32,87 41,49 -,32

1- O teste do χ^2

$$H_0: \text{ as variáveis são independentes} \quad vs \quad H_1: \text{ as variáveis são dependentes.}$$

Estatística de teste:

$$X^2 = \sum_{\text{todas as células}} \frac{(O_i - E_i)^2}{E_i},$$

onde E_i representa a frequência esperada e O_i a observada.

Quando o número de observações é elevado a distribuição da estatística X^2 é aproximadamente a do χ^2 e daí o nome do teste.

Rejeita-se a hipótese de independência entre as variáveis quando o valor da estatística de teste é superior a um certo valor crítico (reflectindo grandes desvios entre as frequências observadas e esperadas).

Resumindo:

O teste do χ^2 aplica-se sempre que quisermos averiguar a existência de dependência entre duas variáveis de tipo categórico.

Requisitos do teste: As frequências esperadas em cada classe não devem ser inferiores a 5 unidades sempre que o número total de observações é $n \leq 20$. Se $n > 20$ não deverá existir mais do que 20% das células com frequências esperadas inferiores a 5 nem deverá existir nenhuma célula com frequência esperada inferior a 1.

Inconvenientes do teste:

1. Uma vez que a distribuição da estatística de teste é apenas aproximada (assintótica), para amostras pequenas o valor do *p-value* poderá conter um erro apreciável. No caso de tabelas 2×2 e sempre que $n \leq 20$ deve-se recorrer ao **teste de Fisher** que fornece valores exactos para os *p-values* do teste.
2. Devido à natureza discreta da contagem das frequências o valor da estatística do χ^2 vem acrescida de um erro. No caso de tabelas 2×2 deve-se utilizar uma **correcção à continuidade** (fornecida pelo SPSS).

Inconvenientes da estatística do χ^2 enquanto medida de associação

A estatística X^2 utilizada no teste do χ^2 é uma medida de associação entre duas variáveis já que assume valores próximos de zero quando as variáveis são independentes e valores elevados (positivos) quando existe dependência. No entanto, ao contrário do que acontecia com os coeficientes de assimetria, esta medida não está limitada ao intervalo $[0, 1]$ e o seu valor máximo depende do número total de observações.

Coeficientes de associação para dados categóricos que se assemelham aos coeficientes de correlação:

1 - O coeficiente de Cramér

O coeficiente de Cramér é uma medida de associação entre duas variáveis medidas numa escala categórica. Portanto pode ser aplicado em situações onde a informação se encontra distribuída por categorias nominais não ordenáveis.

$$C = \sqrt{\frac{X^2}{n(l-1)}}$$

onde n representa o número total de observações, l representa o mínimo entre o número de linhas e colunas da tabela de contingência, e X^2 é o valor da estatística do teste de χ^2.

A partir do valor do coeficiente de Cramér também é possível efectuar um teste às hipóteses

$$H_0: \text{ as variáveis são independentes} \quad vs \quad H_1: \text{ as variáveis são dependentes.}$$

Vantagens do coeficiente de Cramér:

o seu valor está limitado ao intervalo $[0, 1]$.

quando as variáveis são totalmente independentes $C = 0$.

quanto maior a associação maior o valor do coeficiente.

o coeficiente pode ser determinado em situações onde mais nenhum coeficiente (dos já expostos) pode ser aplicado.

ao contrário da estatística X^2, o coeficiente pode ser aplicado para comparar tabelas de contingência de dimensão diferente ou baseadas em amostras de dimensão diferente.

Desvantagens do coeficiente:

quando $C = 1$ pode não haver associação perfeita entre as duas variáveis. A associação só é perfeita se o número de linhas for igual ao número de colunas.

o coeficiente de Cramér está sujeito aos mesmos pressupostos do teste do qui-quadrado se pretendermos testar o seu significado.

este coeficiente não deve ser comparado directamente com outros. Se os dados forem ordinais podemos calcular o coeficiente de Cramér mas não devemos comparar directamente o seu valor com o valor do coeficiente de Pearson. Embora o coeficiente aumente com o grau de associação as diferenças na magnitude não têm uma interpretação directa.

2 - O coeficiente Φ

O coeficiente Φ é muito semelhante ao coeficiente de Cramér e foi proposto inicialmente apenas para tabelas de contingência 2×2. Neste caso o teste de independência que se pode efectuar pode ser baseado no teste exacto de Fisher fornecendo valores mais exactos que os do coeficiente de Cramér.

Para tabelas 2×2 com conteúdo representado pelas letras $\begin{matrix} A & B \\ C & D \end{matrix}$ o coeficiente é dado por

$$R_\phi = \frac{|AD - BC|}{\sqrt{(A+B)(C+D)(A+C)(B+D)}}$$

Se retirarmos o módulo do numerador obtemos um coeficiente que pode assumir valores negativos detectando assim um sentido na associação entre as duas variáveis.

No que respeita a vantagens e desvantagens do coeficiente, elas são idênticas às do coeficiente de Cramér.