Ensaios Não Paramétricos, – TESTE DE AJUSTAMENTO

- Inferência paramétrica e não paramétrica
- Ideia base $\rightarrow$ Propor uma distribuição para o fenómeno e testar a proposta
- A proposta pode corresponder a uma hipótese simples ou a uma hipótese composta.
  - Hipótese simples: $f_0(x)$ é completamente especificada

    Exemplos: Poisson com média igual a 10, binomial 5 e 0.3
  - Hipótese composta: $f_0(x)$ não é completamente especificada, isto é depende de parâmetros desconhecidos e escreve-se $f_0(x \mid \theta_1, \theta_2, ..., \theta_k)$

    Exemplos: Poisson com média desconhecida, binomial 5 e p
- Problema: Como testar $H_0 : X \sim f_0(x)$?

  Uma possível solução $\rightarrow$

  Teste do Qui-quadrado à bondade do ajustamento.

  Teste de Kolmogorov, Kolmogorov-Smirnov

TESTE DO QUI-QUADRADO À BONDADE DO AJUSTAMENTO

1. Introdução

- O objectivo é efectuar testes do tipo $H_0 : f(x) = f_0(x)$ sendo $f_0(x)$ uma função densidade ou de probabilidade conhecida.
- A hipótese H_0 é dita simples quando $f_0(x)$ não envolve parâmetros desconhecidos. Quando tal não acontece H_0 é dita composta;
- Primeiro passo na inferência não paramétrica.

2. Distribuição multinomial

- Observa-se uma amostra casual simples de dimensão *n* referente a uma variável qualitativa que pode assumir *m* valores diferentes (por exemplo a côr de um electrodoméstico que é vendido em 5 cores diferentes);
- Alguma notação:
    - $p_j = \Pr(X = j)$ - Probabilidade da variável aleatória X assumir o valor j no universo ($j = 1,2,...,m$)
    - $\sum_{j=1}^{m} p_j = 1$ logo $p_m = 1 - (p_1 + p_2 + ... + p_{m-1})$
    - $p_j > 0 \qquad (j = 1,2,...,m)$
- Considerando uma amostra casual simples de dimensão *n* seja N_j o número de elementos da amostra para os quais $X = j$. Tem-se então
    - $\sum_{j=1}^{m} N_j = n$ logo $N_m = n - (N_1 + N_2 + ... + N_{m-1})$, isto é apenas se vai dispor de $(m-1)$ variáveis independentes
    - O vector aleatório $(N_1, N_2, ..., N_{m-1})$ vai ter uma distribuição multinomial cuja função de probabilidade é dada por

$$f(n_1,n_2,...,n_m) = \Pr(N_1 = n_1, N_2 = n_2,..., N_m = n_m)$$
$$= \frac{n!}{n_1!n_2!...n_m!}\, p_1^{n_1}\, p_2^{n_2} ... p_m^{n_m}$$

com $p_j > 0$, $\sum_{j=1}^{m} p_j = 1$ e $n_j \geq 0$, $\sum_{j=1}^{m} n_j = 1$

- Neste quadro, testar $H_0 : f(x) = f_0(x)$ equivale a testar $H_0 : p_j = p_{0j}$ para $j = 1,2,...,m$ sendo os p_{oj} conhecidos.
- O teste proposto por Pearson no início do século XX baseia-se no seguinte teorema
- Teorema

Seja $(N_1, N_2, ..., N_{m-1})$ um vector aleatório com distribuição multinomial de parâmetros $(n, p_{01}, p_{02}, ..., p_{0(m-1)})$. A variável $Q = \sum_{j=1}^{m} \frac{(N_j - n\,p_{oj})^2}{n\,p_{oj}}$ tem como distribuição assintótica ($n \to \infty$) uma $\chi^2_{(m-1)}$.

- Demonstração: versão rigorosa em Cramér (1946) *Mathematical Methods of Statistics*
- Versão "light" tirada de Murteira (1990) e inspirada em Fisher.
  - Assumindo que os N_j ($j = 1,2,...,m$) são independentes e que cada um deles tem distribuição de Poisson, isto é, $N_j \sim \text{Po}(n\,p_{oj})$, mostra-se que a função de probabilidade da amostra é uma multinomial de parâmetros $(n, p_{01}, p_{02}, ..., p_{0(m-1)})$. A simplificação de Fisher consiste em estudar apenas uma das situações que conduzem à multinomial.

$$f(n_1,n_2,...,n_m \mid \sum N_i = n) = \Pr(N_1 = n_1, N_2 = n_2,..., N_m = n_m \mid \sum N_i = n)$$
$$= \begin{cases} 0 & \sum n_i \neq n \\ \dfrac{\Pr(N_1 = n_1, N_2 = n_2,..., N_m = n_m)}{\Pr\left(\sum N_i = n\right)} & \sum n_i = n \end{cases}$$

Ora

$$\Pr(N_1 = n_1, N_2 = n_2, ..., N_m = n_m) = \prod_{i=1}^{m} \Pr(N_i = n_i)$$

$$= \prod_{i=1}^{m} \frac{e^{-n\,p_{oj}}\,(n\,p_{oj})^{n_j}}{n_j!}$$

$$= e^{-n}\,n^n \prod_{i=1}^{m} \frac{{p_{oj}}^{n_j}}{n_j!}$$

$$\Pr\left(\sum N_i = n\right) = \frac{e^{-n}\,n^n}{n!}$$

Logo

$$f(n_1, n_2, ..., n_m \mid \sum N_i = n) = \begin{cases} 0 & \sum n_i \neq n \\ \dfrac{\Pr(N_1 = n_1, N_2 = n_2, ..., N_m = n_m)}{\Pr\left(\sum N_i = n\right)} & \sum n_i = n \end{cases}$$

$$= \begin{cases} 0 & \sum n_i \neq n \\ \dfrac{n!}{n_1! n_2! ... n_m!} \prod_{i=1}^{m} {p_{oj}}^{n_j} & \sum n_i = n \end{cases}$$

- Com base na Poisson é mais fácil de demonstrar o teorema. $N_j \sim \text{Po}(n\,p_{oj})$ logo quando $n \to \infty$ com p_{oj} fixo, $n\,p_{oj} \to \infty$ e portanto $\dfrac{N_j - n\,p_{oj}}{\sqrt{n\,p_{oj}}} \overset{o}{\sim} n(0;1)$, ou seja, $\dfrac{\left(N_j - n\,p_{oj}\right)^2}{n\,p_{oj}} \overset{o}{\sim} \chi^2_{(1)}$
- Somando agora e tendo em consideração que apenas temos $m-1$ variáveis independentes, $\sum_{j=1}^{m} \dfrac{\left(N_j - n\,p_{oj}\right)^2}{n\,p_{oj}} \overset{o}{\sim} \chi^2_{(m-1)}$

- Quando $m = 2$, a multinomial reduz-se à binomial e a demonstração torna-se simples. $N_1 \sim b(n, p_{01})$ e $N_2 = n - N_1$, $p_{02} = 1 - p_{01}$. Pelo teorema de De Moivre-Laplace, $\frac{N_1 - n\,p_{01}}{\sqrt{n\,p_{01}(1-p_{01})}} \overset{o}{\sim} n(0;1)$ ou seja $\frac{(N_1 - n\,p_{01})^2}{n\,p_{01}(1-p_{01})} \overset{o}{\sim} \chi^2_{(1)}$. Basta então mostrar que $Q = \frac{(N_1 - n\,p_{01})^2}{n\,p_{01}(1-p_{01})}$. Ora

$$
\begin{aligned}
Q &= \frac{(N_1 - n\,p_{01})^2}{n\,p_{01}} + \frac{(N_2 - n\,p_{02})^2}{n\,p_{02}} \\
&= \frac{p_{02}\,(N_1 - n\,p_{01})^2 + p_{01}\,(N_2 - n\,p_{02})^2}{n\,p_{01}\,p_{02}} \\
&= \frac{p_{02}\,(N_1 - n\,p_{01})^2 + p_{01}\,((n - N_1) - n(1-p_{01}))^2}{n\,p_{01}\,p_{02}} \\
&= \frac{p_{02}\,(N_1 - n\,p_{01})^2 + p_{01}\,(-N_1 + n\,p_{01})^2}{n\,p_{01}\,p_{02}} \\
&= \frac{p_{02}\,(N_1 - n\,p_{01})^2 + p_{01}\,(N_1 - n\,p_{01})^2}{n\,p_{01}\,p_{02}} = \frac{(p_{01} + p_{02})\,(N_1 - n\,p_{01})^2}{n\,p_{01}(1-p_{01})} \\
&= \frac{(N_1 - n\,p_{01})^2}{n\,p_{01}(1-p_{01})}
\end{aligned}
$$

- Estabelecido o teorema, torna-se fácil proceder ao teste quer em termos de um valor de referência Q_α quer do valor-p tendo presente que a região de rejeição se irá naturalmente situar na aba direita da distribuição.
- Tenha-se também presente que quando se aplica o teste em amostras finitas é necessário ter alguns cuidados para que a aproximação se mantenha aceitável, nomeadamente que $n\,p_{0j} \geq 5$ ($j = 1,2,...,m$) ou mesmo $n\,p_{0j} \geq 10$.

- **Exemplo** [Murteira *et al*.(2007)]: Um aspirador é vendido em cinco cores: verde (A_1), castanho (A_2), encarnado (A_3), azul (A_4) e branco (A_5). Num estudo de mercado, analisou-se uma amostra casual de 300 vendas recentes:

A_1	A_2	A_3	A_4	A_5	Total
88	65	52	40	55	300

Pretende testar-se a hipótese de que os consumidores não manifestam preferência por qualquer das cores.

1º Passo – Identificar o problema e formular H_0.

$H_0 : p_{01} = p_{02} = p_{03} = p_{04} = p_{05} = 0.2.$ (equivale a testar uma distribuição uniforme discreta)

2º Passo – Definir a estatística de teste (classes a considerar e graus de liberdade) e calcular o seu valor.

Sendo $n = 300$ e dado H_0 a frequência esperada em cada classe será $n\,p_{0j} = 60 > 10$ para $j = 1,2,...,m$, não havendo portanto de reagrupar classes. Pode-se assim trabalhar com as 5 classes a que correspondem 4 graus de liberdade para a qui-quadrado

Moda-lidades	**Freq. Obs.** (n_j)	**Freq. esp.** (np_{0j})	$\frac{(n_j - np_{0j})^2}{np_{0j}}$
A_1	88	60	13.07
A_2	65	60	0.42
A_3	52	60	1.07
A_4	40	60	6.67
A_5	55	60	0.42
Total	300	300	21.65

Para $\alpha = 0.05$ tem-se $Q_{0.05} = 9.49$ logo rejeita-se H_0. Calculando o valor-*p*=0.00023, induziria a mesma conclusão.

3. Distribuição sem parâmetros desconhecidos

- Corresponde a H_0 ser uma hipótese simples
- Para efectuar o teste apresentado, a ideia é passar da distribuição dada para uma multinomial, definindo-se as *m* classes, tendo naturalmente presente os problemas de dimensão $n\,p_{0j} \geq 5$ que se apresentaram.
- Assim em vez de se testar $H_0 : f(x) = f_0(x)$, vai testar-se $H_0' : p_j = p_{0j}$ para $j = 1,2,...,m$. Para tal, divide-se o domínio de *X* em *m* sub domínios e calcula-se a probabilidade de cada um deles.
- Se este processo tem a vantagem de permitir uma aplicação quase universal do teste, ele apresenta no entanto alguns problemas ligados à perda de informação:
    - Como $H_0 \Rightarrow H_0'$, a rejeição de H_0' leva a concluir pela rejeição de H_0. Já a não rejeição de H_0' nos deixa numa situação mais desconfortável. Existindo inúmeras distribuições diferentes que conduzem à mesma multinomial, a identificação da distribuição de origem sai do âmbito da estatística.
    - Exemplificar e mostrar que quando o número de classes *m* aumenta, a importância do problema vai-se atenuando.
    - De qualquer forma é importante combinar a construção das *m* classes com a existência de um número esperado mínimo de elementos por classe.
- No caso contínuo a solução mais acertada consiste geralmente em construir classes equiprováveis em termos da distribuição $f_0(x)$, o que permite maximizar *m* sujeito à condição $n\,p_{0j} \geq 5$ ($j = 1,2,...,m$). Alguns programas de computador tendem a construir classe de amplitude semelhante (excepção feita, por vezes, às classes extremas) mas depois têm de agregar classes para respeitar a condição. Existem também algumas adaptações que se fazem em função da natureza discreta ou contínua da variável em estudo. Comece por analisar-se o caso contínuo.

Exemplo - Variável contínua e classes pré-construídas - Um estudo sobre o tempo de vida em dias de uma amostra de 1000 tubos electrónicos deu o resultado que se encontra nas duas primeiras colunas do quadro que se segue.

Tempo de vida	Freq. obs.	Freq. esp.
$X < 150$	543	527.63
$150 \le X < 300$	258	249.20
$300 \le X < 450$	120	117.73
$450 \le X < 600$	48	55.61
$600 \le X < 750$	20	26.27
$X \ge 750$	11	23.52
Total	1000	1000.00

O fabricante afirma que o tempo de vida dos tubos, X, tem distribuição exponencial com média $\mu = 200$. Suportam os dados esta hipótese?

1º Passo – Identificar o problema, formular a hipótese que se pretende testar e aquela que vai ser efectivamente testada.

Hipótese que se quer testar

$$H_0 : X \sim f_0(x) = \frac{1}{200}\exp\left\{-\frac{x}{200}\right\} = 0.005\, e^{-0.005x} \quad (x > 0)$$

Hipótese que se vai testar:

$$H_0' : p_{01} = 0.52763 \quad p_{02} = 0.24920 \quad p_{03} = 0.11773$$
$$p_{04} = 0.05561 \quad p_{05} = 0.02627 \quad p_{06} = 0.02352$$

estando as classes pré-definidas, já que

$$p_{01} = P(X < 150) = \int_0^{150} 0.005\, e^{-0.005x} dx = \left(1 - e^{-0.75}\right) \approx 0.52763$$

$$p_{02} = P(150 \le X < 300) = \int_{150}^{300} 0.005\, e^{-0.005x} dx = \left(e^{-0.75} - e^{-1.50}\right) \approx 0.2492$$

…

$$p_{06} = P(X > 750) = \int_{750}^{\infty} 0.005\, e^{-0.005x} dx = e^{-3.75} \approx 0.02352$$

2º Passo – Definir a estatística de teste (classes a considerar e graus de liberdade)

Calcula-se a última coluna do quadro (frequências esperadas) e verifica-se que não existe necessidade de reagrupar classes. Calcula-se então

$$Q_{\text{obs}} = \frac{(543-527.63)^2}{527.63} + \frac{(258-249.2)^2}{249.2} + \cdots + \frac{(11-23.52)^2}{23.52} = 10.0004,$$

e efectua-se o teste (valor-p=0.075).

Exemplo - Variável contínua e dados brutos - Considere-se a seguinte amostra ordenada de 50 observações

-2.18359	-0.97763	-0.38132	-0.08452	1.095023
-2.11793	-0.84724	-0.37024	-0.03248	1.19835
-1.84691	-0.77351	-0.36549	0.028117	1.276474
-1.74248	-0.73648	-0.36288	0.134853	1.342642
-1.69043	-0.6902	-0.32699	0.244257	1.661456
-1.6124	-0.65491	-0.32272	0.538948	1.733133
-1.52157	-0.56792	-0.30023	0.675138	1.918916
-1.44419	-0.5238	-0.23418	0.757611	1.972212
-1.27768	-0.51321	-0.18616	0.865673	2.194502
-1.0867	-0.40405	-0.08528	0.902191	2.375655

Pretende-se testar se se pode considerar que esta amostra provém de um universo com distribuição normal estandardizada.

1º Passo – Identificar o problema, formular a hipótese que se pretende testar.

$$H_0 : X \sim n(0;1)$$

2º Passo – Construir as classes

Querendo-se maximizar o número de classes (m) e manter a restrição de um número **esperado** mínimo de observações em cada classe, optou-se por 10 classes equiprováveis. Assim a classe 1 terá limite inferior em $-\infty$ e limite superior a_1 tal que $P(X < a_1) = 0.1$, a classe 2 irá de a_1 até a_2 tal que $P(X < a_2) = 0.2$ e assim por diante até à classe 10. Feitas as contas, as classes encontram definidas na 1ª coluna do quadro que se apresenta no passo 4.

3º Passo – Com base nas classes definidas no passo anterior o teste que se vai fazer é

$$H_0' : p_{01} = p_{02} = \cdots = p_{0(10)} = 0.1$$

4º Passo – Calcular a estatística de teste e efectuar o teste

O quadro resume os aspectos a considerar

Tempo de vida	Freq. obs.	Freq. esp.
$(-\infty,-1.2816]$	8	5
$(-1.2816,-0.8416]$	4	5
$(-0.8416,-0.5244]$	5	5
$(-0.5244,-0.2533]$	10	5
$(-0.2533,0]$	5	5
$(0,0.2533]$	3	5
$(0.2533,0.5244]$	0	5
$(0.5244,0.8416]$	3	5
$(0.8416,1.2816]$	5	5
$(1.2816,\infty)$	7	5
Total	50	50.00

$$Q_{\text{obs}} = \frac{(8-5)^2}{5} + \frac{(4-5)^2}{5} + \cdots + \frac{(7-5)^2}{5} = 14.4,$$

e efectua-se o teste (valor-p=0.1088).

Observe-se que o número mínimo de elementos em cada classe diz-respeito ao número esperado e não ao número observado.

Exemplo – Variável discreta

Determinada empresa seguradora baseia o seu sistema de prémios para certo risco na premissa de que o número de sinistros por apólice tem distribuição de Poisson de parâmetro $\lambda = 0.2$. Recolhida uma amostra de 1000 apólices referentes ao ano anterior observou-se:

Nº sinistros por apólice	0	1	2	3
Nº apólices	800	175	21	4

A amostra põe em causa a premissa da seguradora?

1º Passo - Identificar o problema e formular a hipótese que se pretende testar

$H_0 : X \sim \text{Po}(0.2)$

2º Passo – Construir as classes

Sendo a variável discreta, segue-se tanto quanto possível os valores que *X* assume.

Como

$P(X = 0 \mid \lambda = 0.2) \approx 0.8187 \quad P(X = 1 \mid \lambda = 0.2) \approx 0.1637 \quad P(X = 2 \mid \lambda = 0.2) \approx 0.0164 \quad P(X = 3 \mid \lambda = 0.2) \approx 0.0011$

$P(X = 4 \mid \lambda = 0.2) \approx 0.0001$ ...

Como se dispõe de 1000 observações, é fácil ver que os valores 0, 1, e 2 podem ser autonomizados, o mesmo não acontecendo com os valores superiores a 2. Opta-se então por reagrupar os valores ≥ 3 a que corresponde $P(X \geq 3 \mid \lambda = 0.2) \approx 0.0102$. Não tendo esta classe dimensão suficiente (esperam-se 1.02 observações) é-se levado a definir apenas 3 classes: 0, 1 e ≥ 2.

3º Passo – Com base nas classes definidas o teste que se vai fazer é

$$H_0' : p_{01} = 0.8187 \quad p_{02} = 0.1637 \quad p_{03} = 0.0176$$

4º Passo – Calcular a estatística de teste e efectuar o teste

O quadro resume os aspectos a considerar

Sinistros/ Apólice	Freq. Obs.	Freq. Esp.	$\frac{(n_j - n\,p_{0j})^2}{n\,p_{0j}}$
0	800	818.731	0.4285
1	175	163.746	0.7734
≥ 2	25	17.523	3.1903
Total	1000	1000.000	4.3923

Obtendo-se um valor-*p* de 0.1112.

4. Distribuição com parâmetros desconhecidos

- Corresponde a H_0 ser uma hipótese composta, isto é $H_0 : f(x) = f_0(x \mid \theta_1, \cdots, \theta_k)$ sendo os *k* parâmetros desconhecidos.
- Para construir as partição que permite passar de $H_0 : f(x) = f_0(x \mid \theta_1, \cdots, \theta_k)$ para vai testar-se $H_0' : p_j = p_{0j}$ para $j = 1,2,...,m$, torna-se necessário estimar os parâmetros desconhecidos já que

$$p_{0j} = \int_{t_{j-1}}^{t_j} f_0(x \mid \theta_1, \cdots \theta_k)\, dx = p_j(\theta_1, \cdots \theta_k)$$

- **Teorema.** Suponha-se que:
  a. $p_j(\theta_1, \cdots \theta_k) > 0$, $j = 1,2, \cdots, m$
  b. As derivadas parciais de 1ª e 2ª ordem , $\frac{\partial p_j}{\partial \theta_r}$ e $\frac{\partial^2 p_j}{\partial \theta_r \partial \theta_s}$ ($j = 1,2, \cdots, m$, $r, s = 1,2, \cdots, k$) existem e são contínuas
  c. A matriz Jacobiana, $m \times k$, de elemento genérico $\left[\frac{\partial p_j}{\partial \theta_r}\right]$ ($j = 1,2, \cdots, m$, $r = 1,2, \cdots, k$) tem característica *k*

  Então, se os parâmetros $\theta_1, \cdots, \theta_k$ são estimados por máxima verosimilhança a partir da multinomial, isto é com os dados classificados, e se H_0' é verdadeira, a estatística

$$Q = \sum_{j=1}^{m} \frac{(N_j - n\, p_j(\hat{\hat{\theta}}_1, \hat{\hat{\theta}}_2, \cdots, \hat{\hat{\theta}}_k))^2}{n\, \hat{p}_j(\hat{\hat{\theta}}_1, \hat{\hat{\theta}}_2, \cdots, \hat{\hat{\theta}}_k)}$$

  tem distribuição assimptótica ($n \to \infty$) do qui-quadrado com $m - k - 1$ graus de liberdade.
  Demonstração: Bastante trabalhosa. Ver Cramer (1946)
- Explicar a intuição por trás da diminuição dos graus de liberdade;
- Estimação por MV com os dados agregados
- A ideia é maximizar a função de verosimilhança agregada e não a função de verosimilhança habitual. Assim, em vez de se maximizar $L = \prod_{i=1}^{n} f_0(x_i \mid \theta_1, \cdots, \theta_k)$ vai proceder-se à maximização de $L^* = \prod_{i=1}^{m} p_j(\theta_1, \cdots, \theta_k)$. Para melhor entender do que se trata retome-se o exemplo anterior com a Poisson.

- O recurso a $L = \prod_{i=1}^{n} f_0(x_i \mid \theta_1, \cdots, \theta_k)$ conduz ao bem conhecido estimador de MV, $\hat{\lambda} = \overline{X}$, e portanto a nossa estimativa será $\hat{\lambda} = 0.229$.
- Com base em $L^* = \prod_{i=1}^{m} p_j(\theta_1, \cdots, \theta_k)$ tem-se

$$L^* = \left(\Pr(X=0)\right)^{800} \times \left(\Pr(X=1)\right)^{175} \times \left(\Pr(X \geq 2)\right)^{25}$$
$$= \left(e^{-\lambda}\right)^{800} \times \left(\lambda e^{-\lambda}\right)^{175} \times (1 - e^{-\lambda} - \lambda e^{-\lambda})^{25}$$
$$= \lambda^{175}\ e^{-975\lambda}\ (1 - e^{-\lambda} - \lambda e^{-\lambda})^{25}$$

Logo $\ln L^* = 175 \ln \lambda - 975\lambda + 25 \ln (1 - e^{-\lambda} - \lambda e^{-\lambda})$

ou $\left(\ln L^*\right)' = \frac{175}{\lambda} - 975 + 25 \frac{\lambda e^{-\lambda}}{1 - e^{-\lambda} - \lambda e^{-\lambda}}$

Não existindo solução explicita recorre-se a um método numérico para maximizar L^* (ou anular a derivada), aqui, $\hat{\hat{\lambda}} = 0.226964$.

- Como é bastante mais trabalhoso estimar os parâmetros pela MV com os dados classificados do que com a amostra original e obtendo-se na grande maioria das situações estimativas muito semelhantes aceita-se que se recorra à MV. No entanto quando assim a verdadeira distribuição assimptótica passa situar-se entre uma qui-quadrado com $m-k-1$ graus de liberdade e uma qui-quadrado com $m-1$ graus de liberdade. A regra prática: estimar os parâmetros por MV e utilizar a χ^2 c/ $m-k-1$ gr. lib.
- As considerações feitas nas secções anteriores mantêm-se.
- Retome-se então o exemplo com a Poisson

Sinistros/ Apólice	Freq. Obs.	Freq. Esp. (1)	$\frac{(n_j - n\,p_{0j})^2}{n\,p_{0j}}$	Freq. Esp.(2)	$\frac{(n_j - n\,p_{0j})^2}{n\,p_{0j}}$
0	800	796.95	0.0117	795.33	0.0274
1	175	180.88	0.1911	182.13	0.2791
≥ 2	25	22.18	0.3608	22.54	0.2682
Total	1000		0.5635		0.5748

(1) MV c/ dados classificados , (2) MV c/ dados n class. Neste caso a conclusão seria igual: $\chi^2(1)$, val-*p*= 0.4484 ou 0..4528.

TESTE DE KOLMOGOROV (Kolmogorov-Smirnov uma amostra)

- Objectivo: Testar $H_0 : F(x) = F_0(x) \qquad -\infty < x < \infty$ contra $H_1 : F(x) \neq F_0(x)$
  sendo $F_0(x)$ completamente especificada e contínua.
  Como se vai ver o teste é válido para pequenas amostras
- Ideia base: Comparar a função de distribuição $F_0(x)$ com a função de distribuição empírica.
- Função de distribuição empírica
  - Função de distribuição empírica da amostra observada $(x_1, x_2, \cdots, x_n)$

$$\hat{F}_n(x) = \frac{1}{n} \#\{x_i \leq x\} = \frac{1}{n} \sum_{i=1}^{n} I_{(-\infty;x]}(x_i), \qquad -\infty < x < +\infty$$

    - Função de distribuição de tipo discreto
    - A função de distribuição empírica da amostra observada não depende de nenhuma hipótese feita sobre o universo. Apenas depende das observações.
    - Exemplificar admitindo a amostra ordenada:

      (0.20; 0.35; 0.44; 0.65; 0.90)

  - Definição da função de distribuição da amostra genérica $(X_1, X_2, \cdots, X_n)$

$$F_n(x) = \frac{1}{n} \#\{X_i \le x\} = \frac{1}{n} \sum_{i=1}^{n} I_{(-\infty;x]}(X_i), \qquad -\infty < x < +\infty$$

ou, de forma alternativa, sendo $(X_{(1)}, X_{(2)}, ..., X_{(n)})$ a amostra ordenada

$$F_n(x) = \begin{cases} 0 & x < X_{1:n} \\ i/n & X_{(i)} \le x < X_{(i+1)} \quad i = 1,2,\cdots,n-1 \\ 1 & X_{(n)} \le x \end{cases}$$

- Enquanto $\hat{F}_n(x)$ é uma função de distribuição de tipo discreto, $F_n(x)$ é, para cada $x \in \Re$, uma variável aleatória função da amostra genérica $(X_1, X_2, \cdots, X_n)$, isto é, uma estatística com função de probabilidade dada por

$$\Pr\left[F_n(x) = \frac{i}{n}\right] = \binom{n}{i} [F(x)]^i \, [1 - F(x)]^{n-i} \qquad i = 0,1,\cdots,n$$

- Complementarmente ao caso anterior, a função de distribuição da amostra genérica não depende das observações mas apenas da distribuição em vigor no universo.

- **Preliminares** para o teste:
  - Considerem-se as estatísticas:

$$D_n = \sup_{-\infty<x<\infty} \left|F_n(x) - F_o(x)\right|, \quad D_n^+ = \sup_{-\infty<x<\infty} \left[F_n(x) - F_o(x)\right], \quad D_n^- = \sup_{-\infty<x<\infty} \left[F_o(x) - F_n(x)\right]$$

donde $D_n = \max(D_n^+, D_n^-)$

  - Mas como (ilustrar com um gráfico)

$$D_n^+ = \max\left(\max_{i=1,2,...n}\left[i/n - F_0(X_{(i)}); 0\right]\right), \quad D_n^- = \max\left(\max_{i=1,2,...n}\left[F_0(X_{(i)}) - (i-1)/n; 0\right]\right)$$

já que, definindo $X_{(0)} = -\infty$ e $X_{(n+1)} = \infty$, vem

$$F_n(x) = i/n \qquad X_{(i)} \le x < X_{(i+1)} \qquad i = 0,1,...,n$$

e portanto

$$D_n^+ = \sup_{-\infty<x<\infty}\left[F_n(x) - F_o(x)\right] = \max_{i=0,1,...,n} \sup_{X_{(i)}<x<X_{(i+1)}}\left[\frac{i}{n} - F_o(x)\right]$$

$$= \max_{i=0,1,...,n}\left[\frac{i}{n} - \inf_{X_{(i)}<x<X_{(i+1)}} F_o(x)\right] = \max_{i=0,1,...,n}\left[\frac{i}{n} - F_o(X_{(i)})\right]$$

$$= \max\left(\max_{i=1,...,n}\left[\frac{i}{n} - F_o(X_{(i)})\right]; 0\right)$$

Para D_n^- o raciocínio é semelhante.

- **Teorema 1** – As estatísticas D_n^+ e D_n^- (e consequentemente D_n) têm uma distribuição que não depende de $F_0(x)$ - "Distribution free" – demonstração fora do âmbito do curso
- A distribuição destas estatísticas encontra-se tabelada para alguns valores de n e para os níveis de significância usuais.
- **Teorema 2** – Seja F_0 uma distribuição contínua qualquer.

  Então para $z \geq 0$,

  $$\lim_{n\to\infty} \Pr(\sqrt{n}\, D_n \leq z) = 1 - 2\sum_{i=1}^{\infty} (-1)^{i-1} e^{-2i^2 z^2} .$$

Esta distribuição também se encontra tabelada. Demonstração fora do âmbito do curso

- **Procedimento** de **teste bilateral** (2-sided test)
  1. Calcular $D_{n,obs}$ o valor observado de D_n fazendo

     $$D^-_{n,\text{obs}} = \max\left(\max_{i=1,2,\ldots n} \left[F_0(x_{(i)}) - (i-1)/n; 0 \right] \right)$$

     e fazer $D_{n,obs} = \max(D^+_{n,obs}, D^-_{n,obs})$
  2. Verificar na tabela se se rejeita, com significância α, a hipótese H_0. Região de rejeição na aba direita

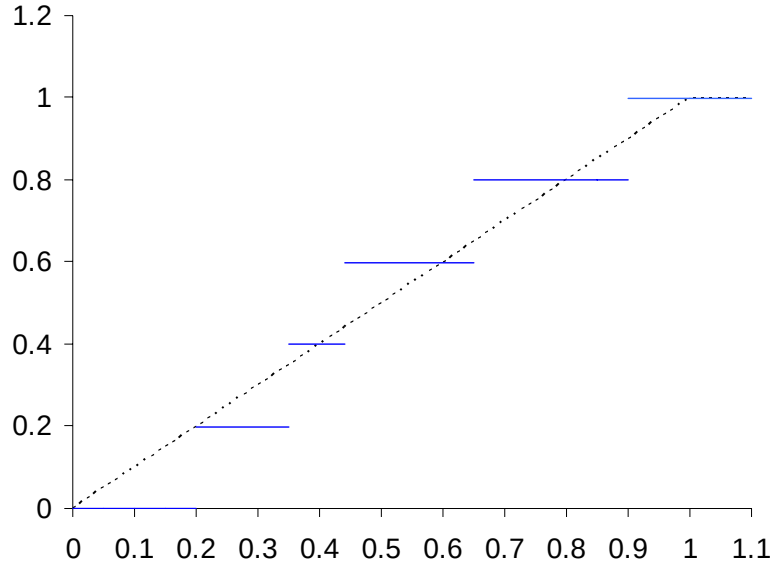
  Ver no exemplo:

- **Exemplo**: Testar se a amostra dada provém de uma uniforme (0 ;1)

i	i/n	$(i-1)/n$	x_i	$F_0(x_i)$	$D^+_{n,\text{obs}}$	$D^-_{n,\text{obs}}$
1	0.2	0	0.20	0.2	0	0.2
2	0.4	0.2	0.35	0.35	0.05	0.15
3	0.6	0.4	0.44	0.44	0.16	0.04
4	0.8	0.6	0.65	0.65	0.15	0.05
5	1	0.8	0.90	0.9	0.1	0.1

$D^+_{n;obs} =$ **0.16** **0.2** $= D^-_{n;obs}$

$H_0 : X \sim U(0;1)$, $D_{n,obs} = \max(D^+_{n,obs}, D^-_{n,obs}) = 0.20$, Tabela a 5% $\rightarrow$ 0.563. Não se rejeita H_0.



Construção de bandas de confiança

- O teste de Kolmogorov permite construir bandas de confiança para a função de distribuição da população, isto é, para cada x podemos enquadrar o valor de $F(x)$, desconhecido, com um determinado grau de confiança.
- Seja então $D_{n;\alpha}$ o menor valor de D_n tal que $\Pr(D_n \geq D_{n;\alpha}) = \alpha$ ou seja o quantil $1-\alpha$ e defina-se
  - $U(x) = \min(F_n(x) + D_{n;\alpha};1) = \begin{cases} F_n(x) + D_{n;\alpha} & F_n(x) + D_{n;\alpha} \leq 1 \\ 1 & F_n(x) + D_{n;\alpha} > 1 \end{cases}$
  - $L(x) = \max(F_n(x) - D_{n;\alpha};0) = \begin{cases} F_n(x) - D_{n;\alpha} & F_n(x) - D_{n;\alpha} \geq 0 \\ 0 & F_n(x) - D_{n;\alpha} < 0 \end{cases}$
- Mostra-se então que $\Pr\big(L(x) \leq F(x) \leq U(x)\big) \geq 1-\alpha$, $\forall x$ - Corresponde ao intervalo aleatório
- Aplicando a uma amostra concreta, obtêm-se as bandas de confiança substituindo $F_n(x)$ por $\hat{F}_n(x)$

- No exemplo: $D_{n;\alpha} = 0.563$ com $n = 5$ e $\alpha = 0.05$ logo

$$\hat{L}(x) = \begin{cases} \max(0-0.563;0) = 0 & x < 0.2 \\ \max(0.2-0.563;0) = 0 & 0.2 \le x < 0.35 \\ \max(0.4-0.563;0) = 0 & 0.35 \le x < 0.44 \\ \max(0.6-0.563;0) = 0.037 & 0.44 \le x < 0.65 \\ \max(0.8-0.563;0) = 0.237 & 0.65 \le x < 0.90 \\ \max(1-0.563;0) = 0.437 & x \ge 0.9 \end{cases} = \begin{cases} 0 & x < 0.44 \\ 0.037 & 0.44 \le x < 0.65 \\ 0.237 & 0.65 \le x < 0.90 \\ 0.437 & x \ge 0.9 \end{cases}$$

$$\hat{U}(x) = \begin{cases} \min(0+0.563;1) = 0.563 & x < 0.2 \\ \min(0.2+0.563;1) = 0.763 & 0.2 \le x < 0.35 \\ \min(0.4+0.563;1) = 0.9631 & 0.35 \le x < 0.44 \\ \min(0.6+0.563;1) = 1 & 0.44 \le x < 0.65 \\ \min(0.8+0.563;1) = 1 & 0.65 \le x < 0.90 \\ \min(1+0.563;1) = 1 & x \ge 0.9 \end{cases} = \begin{cases} 0.563 & x < 0.2 \\ 0.763 & 0.2 \le x < 0.35 \\ 0.9631 & 0.35 \le x < 0.44 \\ 1 & x \ge 0.44 \end{cases}$$

E assim, por exemplo, para $x = 0.7$ vem o intervalo $(0.237;1)$ para $F(0.7)$ que não é particularmente informativo como seria de esperar dada a reduzida dimensão da amostra.

Generalizações

1. **Teste unilateral** (1-sided)
   $H_1 : F(x) > F_0(x)$ ou, alternativamente, $H_1 : F(x) < F_0(x)$
   Adaptar o procedimento e consultar a tabela adequada
2. Que fazer quando a distribuição não se encontra completamente especificada?
   - O teste mostra-se conservador (aceita-se H_0 mais do que se deveria)
   - Existem adaptações ara distribuições particulares, pró exemplo o teste de Lilliefors para universos normais ou com distribuição exponencial
3. Aplicação a universos discretos – O teste é "conservador" – A evitar.
4. **Teste de Kolmogorov-Smirnov (2 amostras**)
   Recolhida uma amostra de dimensão *n* de uma população e outra de dimensão *m* de outra população, pretende-se testar se a distribuição é a mesma nos 2 universos. As amostras são tiradas de forma independente uma da outra → Procedimento semelhante com outras tabelas;

- **Nota Final**
  Para distribuições específicas, nomeadamente para a normal, existem outros testes de ajustamento possíveis e mais potentes. A vantagem essencial do teste de Kolmogorov reside na sua generalidade.

TABELA DE CONTINGÊNCIA

Observa-se uma amostra à luz de 2 atributos: O primeiro reveste r modalidades $A_1, A_2, ..., A_r$ e o segundo s modalidades $B_1, B_2, ..., B_s$. Na célula (i, j) da tabela de contingência regista-se o número de elementos da amostra que verificam o nível i do atributo A e o nível j do atributo B.

Tabela de contingência $r \times s$ observada

	B_1	B_2	…	B_s	**Totais**
A_1	n_{11}	n_{12}	…	n_{1s}	$n_{1\circ}$
A_2	n_{21}	n_{22}	…	n_{2s}	$n_{2\circ}$
…	…	…	…	…	…
A_r	n_{r1}	n_{r2}	…	n_{rs}	$n_{r\circ}$
Totais	$n_{\circ 1}$	$n_{\circ 2}$	…	$n_{\circ s}$	n

n_{ij} $(i = 1,2,...,r,\ j = 1,2,...,s)$ representa a frequência observada na célula definida por (A_i, B_j).

$n_{i\circ} = \sum_{j=1}^{s} n_{ij}$ $(i = 1,2,...,r)$ $\quad n_{\circ j} = \sum_{i=1}^{r} n_{ij}$ $(j = 1,2,...,s)$

Exemplo: Diferentes tipos de electrodomésticos, A_i's, vendidos em diferentes cores, B_j .

Antes de observar a amostra tem-se, em termos da amostra genérica,

	B_1	B_2	…	B_s	**Totais**
A_1	N_{11}	N_{12}	…	N_{1s}	$N_{1\circ}$
A_2	N_{21}	N_{22}	…	N_{2s}	$N_{2\circ}$
…	…	…	…	…	…
A_r	N_{r1}	N_{r2}	…	N_{rs}	$N_{r\circ}$
Totais	$N_{\circ 1}$	$N_{\circ 2}$	…	$N_{\circ s}$	n

Note-se que:

- n é não aleatório, a dimensão da amostra é fixada.
- As frequências em cada classe são aleatórios (variáveis discretas que assumem os valores $0,1,\cdots,n$.
- $N_{i\circ}$ e $N_{\circ j}$ continuam a ser os totais (aleatórios) em linha e coluna respectivamente.
- Existe uma restrição já que $n=\sum_{j=1}^{s} N_{\circ j} = \sum_{i=1}^{r} N_{i\circ} = \sum_{i=1}^{r}\sum_{j=1}^{s} N_{ij}$.

Teste de Independência do χ^2

- Em termos do universo, as probabilidades (desconhecidas) das células (A_i, B_j) representam-se por,

$$p_{ij} = P(A_i, B_j) \ \ (i = 1,2,...,r\,;\, j = 1,2,...,s), \ \ \sum_{i=1}^{r}\sum_{j=1}^{s} p_{ij} = 1.$$

  As respectivas probabilidades marginais são dadas por,

$$p_{i\circ} = \sum_{j=1}^{s} p_{ij} \ \ (i = 1,2,...,r), \ \ \sum_{i=1}^{r} p_{i\circ} = 1;$$

$$p_{\circ j} = \sum_{i=1}^{r} p_{ij} \ \ (j = 1,2,...,s), \ \sum_{j=1}^{s} p_{\circ j} = 1.$$

- Assumir a independência entre os 2 atributos equivale a assumir $P(A_i, B_j) = P(A_i)P(B_j)$, logo a hipótese em teste vai ser

$$H_0 : \forall (i,j) : p_{ij} = p_{i\circ}\, p_{\circ j} \quad \text{contra} \quad H_1 : \exists (i,j) : p_{ij} \neq p_{i\circ}\, p_{\circ j}.$$

- Assumindo H_0, pode-se estimar p_{ij} a partir de $p_{i\circ}$ e de $p_{\circ j}$. Os estimadores de Máxima Verosimilhança de $p_{i\circ}$ e de $p_{\circ j}$ são dados por $\hat{p}_{i\circ} = \dfrac{N_{i\circ}}{n}$ $(i = 1,2,...,r)$; $\hat{p}_{\circ j} = \dfrac{N_{\circ j}}{n}$ $(j = 1,2,...,s)$.

- A estatística de teste vai avaliar a diferença entre a frequência observada N_{ij} e frequência esperada, assumindo H_0 verdadeiro, isto é, retoma-se a filosofia do teste do qui-quadrado à bondade do ajustamento

$$Q = \sum_{i=1}^{r} \sum_{j=1}^{s} \frac{(N_{ij} - n\,\hat{p}_{i\circ}\hat{p}_{\circ j})^2}{n\,\hat{p}_{i\circ}\hat{p}_{\circ j}} \overset{a}{\sim} \chi^2\left[(r-1)(s-1)\right].$$

- Os graus de liberdade obtêm-se verificando que existem rs células e que se estimaram $(r-1)$ parâmetros referentes ao factor *A* (o último valor está pré fixado) e $(s-1)$ referentes ao factor *B*. Tem-se assim

$$rs - 1 - (r-1) - (s-1) = (r-1)(s-1)$$

A região de rejeição vai situar-se, pelas mesmas razões do que no teste do qui-quadrado à bondade do ajustamento na aba direita da distribuição, mantendo-se a restrição referente ao número mínimo esperado de elementos em cada célula (A_i, B_j), isto é, $n\,\hat{p}_{i\circ}\hat{p}_{\circ j} \geq 5$.

Exemplo – No quadro que se segue apresenta-se uma tabela 3×3 construída considerando os 86441 casamentos realizados em 1977 (que se podem considerar uma amostra dos casamentos realizados durante um período de alguns anos), em Portugal Continental (Anuário Estatístico, INE, 1980). Nela são apresentados, para cada sexo, o estado civil dos cônjuges anterior ao casamento.

A hipótese a testar vai a da existência de independência entre o estado civil de cada cônjuge no momento do casamento.

Cônjuges segundo o estado civil anterior ao casamento

Mulheres	Homens			Totais
	Solteiros	Viúvos	Divorciados	
Solteiras	77670	1573	3115	82358
Viúvas	545	796	350	1691
Divorciadas	1343	416	633	2392
Totais	79558	2785	4098	86441

Solução:

Atributo $A \rightarrow$ estado civil da mulher Atributo $B \rightarrow$ estado civil da homem

$$H_0 : p_{ij} = p_{i\circ}\, p_{\circ j} \quad (i, j = 1,2,3).$$

Calculem-se as frequências esperadas em cada célula na hipótese de os atributos serem independentes, fazendo

$$\hat{p}_{i\circ} = \frac{N_{i\circ}}{n} \quad (i = 1,2,...,r); \; \hat{p}_{\circ j} = \frac{N_{\circ j}}{n} \quad (j = 1,2,...,s)$$

$$n\, \hat{p}_{i\circ}\, \hat{p}_{\circ j} = n\, \frac{n_{i\circ}}{n}\, \frac{n_{\circ j}}{n} = \frac{n_{i\circ} \times n_{\circ j}}{n}$$

o que leva ao quadro

Frequências esperadas assumindo H_0 verdadeiro

Mulheres	Homens			Totais
	Solteiros	Viúvos	Divorciados	
Solteiras	75800.12	2653.45	3904.43	82358
Viúvas	1556.35	54.48	80.17	1691
Divorciadas	2201.53	77.07	113.40	2392
Totais	79558	2785	4098	86441

Calcule-se o valor observado da estatística de teste,

$$Q_{\text{obs}} = \frac{(77670 - 75800.12)^2}{75800.12} + \frac{(1573 - 2653.45)^2}{2653.45} + \cdots + \frac{(633 - 113.4)^2}{113.4} = 16509.74,$$

4 graus de liberdade valor-$p \approx 0$ Rejeita-se H_0

Medidas de associação

- Quando se rejeita a independência pode haver interesse em avaliar a intensidade da associação entre os atributos.
- As medidas de associação mais conhecidas, baseadas na estatística Q, são:

a) **Coeficiente de contingência de Pearson**, $C = \sqrt{\frac{Q}{Q+n}}$,

que verifica a dupla desigualdade, $0 \le C \le \sqrt{(q-1)/q} < 1 \quad q = \min\{r, s\}$,

b) **Coeficiente de Tschuprow**, $T = \sqrt{\frac{Q}{n\sqrt{(r-1)(s-1)}}}$,

em que o máximo é 1 apenas no caso em que $r = s$.

c) **O coeficiente de Cramér**, $V = \sqrt{\frac{Q}{n(q-1)}}$,

que verifica $0 \le V \le 1$ e $V \ge T$.

Exemplo– Para medir a associação entre o estado civil dos cônjuges (exemplo anterior), tem-se,

$$C = \sqrt{16509.7/(16509.7 + 86441)} = 0.400,$$

$$T = \sqrt{16509.7/\left(86441 \times \sqrt{2 \times 2}\right)} = 0.309,$$

$$V = \sqrt{16509.7/(86441 \times 2)} = 0.309. \backslash$$

Como também se havia concluído a associação é forte.

Teste de Homogeneidade do χ^2

- Trata-se de testar se a distribuição de determinada variável aleatória é a mesma em diferentes populações. Uma resposta parcial foi esboçada para variáveis contínuas considerando 2 populações com o teste de Kolmogorov-Smirnov (2 amostras)
- Vamos agora tratar a situação para *s* populações, utilizando a filosofia dos testes baseados no qui-quadrado com as vantagens e inconvenientes associados.
- Assume-se então que se observa determinado fenómeno qualitativo com *r* modalidades em *s* populações, com base em outras tantas amostras casuais simples (independentes umas das outras) tendo a amostra para a população *j* ($j = 1,2,\cdots,s$) dimensão pré fixada $n_{\circ j}$. Em termos da amostra genérica pode-se construir uma tabela de contingência com uma diferença fundamental em relação à situação anterior: **A soma das colunas, $n_{\circ j}$, não é aleatória.**

	Populações				
	B_1	B_2	…	B_s	**Totais**
A_1	N_{11}	N_{12}	…	N_{1s}	$N_{1\circ}$
A_2	N_{21}	N_{22}	…	N_{2s}	$N_{2\circ}$
…	…	…	…	…	…
A_r	N_{r1}	N_{r2}	…	N_{rs}	$N_{r\circ}$
Totais	$n_{\circ 1}$	$n_{\circ 2}$	…	$n_{\circ s}$	n

Exemplo: Venda dos electrodomésticos em diferentes cores, A_i's, em diferentes regiões, B_i's.

- Dizer que existe homogeneidade nas diferentes populações, é dizer que as probabilidades de cada modalidades são identicas em cada população, isto é $p_{i|j} = P(A_i \mid B_j) = p_i$, $i = 1,2,\cdots,r$, $j = 1,2,\cdots,s$, sendo $p_{i|j}$ a probabilidade da variável qualitativa assumir o nível *i* na população *j*. Quando existe homogeneidade, estas probabilidades não dependem da população concreta que se considera.
  Em termos de restrições tem-se:
  Amostra $\rightarrow \sum_{i=1}^{r} N_{ij} = n_{\circ j}$, $j = 1,2,\cdots,s$, sendo N_{ij} o número de vezes que se observa o valor *i* na amostra referente à população *j*. Em cada amostra apenas temos $r-1$ observações independentes;

  Universo $\rightarrow \sum_{i=1}^{r} p_{i|j} = 1$, $j = 1,2,\cdots,s$
- A hipótese H_0 escreve-se então $H_0 : \forall (i,j) : p_{i|j} = p_i$
- Assumindo H_0, pode-se estimar $p_{i|j}$ que é constante para todas as populações ($p_{i|j} = p_i$) a partir da amostra global. Na situação mais geral em que nos situamos os estimadores de máxima verosimilhança para p_i são dados por $\hat{p}_i = \frac{N_{i\circ}}{n}$, sendo $N_{i\circ} = \sum_{j=1}^{s} n_{ij}$.

- A estatística de teste vai avaliar a diferença entre a frequência observada N_{ij} e frequência esperada estimada, $n_{\circ j} \times \hat{p}_i$, assumindo H_0 verdadeira, $Q = \sum_{i=1}^{r} \sum_{j=1}^{s} \frac{(N_{ij} - n_{\circ j} \times \hat{p}_i)^2}{n_{\circ j}\ \hat{p}_i}$
- A sua distribuição assintótica será a χ^2 . Os graus de liberdade obtêm-se da seguinte forma:
    - Em cada uma das s populações temos $(r-1)$ células independentes, logo teremos $s \times (r-1)$ células;
    - A este valor é necessário descontar o número de parâmetros estimados, isto é, $(r-1)$ na situação em que estamos;
    - Os graus de liberdade são então $s \times (r-1) - (r-1) = (s-1) \times (r-1)$
- A região de rejeição situa-se na aba direita da distribuição pelas mesmas razões das situações anteriores.
- Mantém-se a restrição referente ao número mínimo esperado de elementos em cada célula, isto é, $n_{\circ j} \times \hat{p}_i \geq 5$.
- Ter presente que pode ser necessário adaptar o procedimento quando se dispõe de mais informação sobre as populações, por exemplo quando se sabe que estas têm determinada distribuição (ver exemplo 2).

Exemplo 1 – Decidiu-se analisar se 3 tipos de materiais reagiam da mesma forma a um tratamento térmico, Considerando-se como efeitos possíveis do tratamento 3 situações: Destruição, pequenos defeitos ou resistência. Para o estudo decidiu-se proceder ao estudo com base numa amostra casual

simples de dimensão 100 de cada um dos materiais. Observados os resultados do quadro que se segue, que conclusão tirar?

Teste de três tipos de material sujeito a tratamento térmico

Classificação das peças	Material I	Material II	Material III	Totais
Completamente destruídas	25	45	30	100
Pequenos defeitos	40	35	35	110
Resistência perfeita	35	20	35	90
Totais	100	100	100	300

Trata-se de um teste de homogeneidade em que nada se sabe da distribuição do fenómeno (repare-se que os totais das colunas são pré-fixados e não flutuam com a amostragem como no caso do teste de independência).

As **frequências esperadas**, admitindo a hipótese de homogeneidade (iguais em linha):

Frequências esperadas: Teste de três tipos de material sujeito a tratamento térmico

Classificação das peças	Material I	Material II	Material III	Totais
Completamente destruídas	33.33	33.33	33.33	100
Pequenos defeitos	36.67	36.67	36.67	110
Resistência perfeita	30	30	30	90
Totais	100	100	100	300

O valor observado para a estatística de teste vem,

$$Q_{\text{obs}} = \frac{(25-33.33)^2}{33.33} + \frac{(45-33.33)^2}{33.33} + \cdots + \frac{(35-30)^2}{30} \approx 11.955,$$

correspondendo um valor-*p* igual a 0.03 (qui-quadrado com 4 graus de liberdade). Este valor o recomenda que sejam postas as maiores reservas sobre a homogeneidade dos três materiais: há, muito provavelmente, comportamentos diferenciados quando se aplica o referido tratamento.

Exemplo 2 – (com distribuição conhecida) – Suponha que se sabe que sabe que 4 populações têm distribuição binomial de parâmetros 3 e p. Pretende-se testar se as populações são homogéneas. Para tal recolheu-se uma amostra de dimensão 200 de cada uma das 2 primeiras e uma amostra de dimensão 100 das 2 últimas, tendo-se observado:

	Populações				
x	P1	P2	P3	P4	Total
0	110	65	31	35	241
1	62	78	45	49	234
2	27	48	21	16	112
3	1	9	3	0	13
Total	200	200	100	100	**600**

Serão as populações homogéneas?

1ª solução: Ignorar a informação referente à binomial

Valores estimados:

	Populações				
x	P1	P2	P3	P4	Total
0	80.33	80.33	40.17	40.17	241
1	78	78	39	39	234
2	37.33	37.33	18.67	18.67	112
3	4.33	4.33	2.17	2.17	13
Total	200	200	100	100	**600**

Como as células referente $x=3$ não têm o número suficiente de observações, reagrupam-se com o valor 2

	Populações				
x	P1	P2	P3	P4	Total
0	80.33	80.33	40.17	40.17	241
1	78	78	39	39	234
2 e 3	41.67	41.67	20.83	20.83	125
Total	200	200	100	100	**600**

A estatística de teste vem $Q_{obs} \approx 35.14$ para uma $\chi^2_{(6)}$ que leva a rejeitar H_0.

2ª solução: Estimar p por max. verosimilhança, sob H_0, isto é que se trata de uma única amostra.

$$\hat{p} = \frac{\overline{x}}{3} \approx 0.2761$$

Valores estimados:

	Populações				
x	P1	P2	P3	P4	Total
0	75.87	75.87	37.93	37.93	227.60
1	86.81	86.81	43.41	43.41	260.44
2	33.11	33.11	16.56	16.56	99.34
3	4.21	4.21	2.10	2.10	12.63
Total	200	200	100	100	**600**

Como as células referente x=3 não têm o número suficiente de observações, reagrupam-se com o valor 2

	Populações				
x	P1	P2	P3	P4	Total
0	75.87	75.87	37.93	37.93	227.60
1	86.81	86.81	43.41	43.41	260.44
2 e 3	37.32	37.32	18.66	18.66	111.97
Total	200	200	100	100	**600**

A estatística de teste vem $Q_{obs} \approx 41.78$ para uma $\chi^2_{(7)}$ que leva a rejeitar H_0.

Teste do sinal

- Teste não paramétrico, a resposta é completamente independente da f.d., $F(x)$.
- Destina-se a testar a mediana de uma população (ligar com a robustez).
- Quando a distribuição da população é simétrica, e tem média, a mediana é igual à média e portanto o teste do sinal serve indirectamente para o teste de hipóteses sobre a média.
- Ter presente que a robustez do teste (independente da distribuição) tem custos em termos da sua potência.
- Seja μ_e a mediana de $F(x)$ e suponhamos contínua, onde $P(X \leq \mu_e) = F(\mu_e) = 1/2$.
- Hipóteses em teste: $H_0 : \mu_e = \mu_{e0}$ contra $H_1 : \mu_e \neq \mu_{e0}$, com base numa amostra casual $(X_1, X_2, ..., X_n)$
- O Teste
  - Definir $Z_i = X_i - \mu_{e0}$, $i = 1,2,...,n$ e passar de $(X_1, X_2, ..., X_n) \rightarrow (Z_1, Z_2, ..., Z_n)$.
  - Retenha-se o sinal, positivo (+) ou negativo (–), das diferenças Z_i e seja S a estatística que designa o número de sinais positivos no conjunto dos *n* sinais. Formalmente

$$S_i = \begin{cases} 0 & Z_i < 0 \\ 1 & Z_i > 0 \end{cases} \quad \text{(assumindo por agora que } Z_i \neq 0\text{)} \quad \text{e} \quad S = \sum_{i=1}^{n} S_i$$

- Se a hipótese H_0 é verdadeira, $S_i \sim Ber(1/2)$ e $S \sim b(n;0.5)$, havendo tendência para um certo equilíbrio entre o número de sinais (+) e o número de sinais (–). Logo, a hipótese H_0 é posta em causa pelos dados quando S é excessivamente "pequeno" ou excessivamente "grande";
- O teste pode então ser feito pelo *valor-p* ou definindo a região de rejeição.
  - **Definindo a região de rejeição**:
    - um teste de dimensão não superior a α é o que leva a rejeitar H_0 quando, $S \le s_1$ ou $S \ge s_2$, onde s_1 é o maior inteiro tal que $P(S \le s_1 \mid H_0) = \sum_{s=0}^{s_1} \binom{n}{s} \left(\frac{1}{2}\right)^n \le \frac{\alpha}{2}$,

      e s_2 é o menor inteiro tal que $P(S \ge s_2 \mid H_0) = \sum_{s=s_2}^{n} \binom{n}{s} \left(\frac{1}{2}\right)^n \le \frac{\alpha}{2}$.

    - Sendo a região crítica formada pelos dois conjuntos, $\{0,1,...,s_1\}$, $\{s_2, s_2+1,...,n\}$, a probabilidade de cometer um erro de 1ª espécie não é superior a α, já que $P(S \le s_1 \mid H_0) + P(S \ge s_2 \mid H_0) \le \alpha$.
    - Quando $n \ge 20$ pode utilizar-se a aproximação à normal preferivelmente com correcção de continuidade, isto é , $s_1 = \frac{n}{2} - \frac{1}{2} - z_{\alpha/2}\frac{\sqrt{n}}{2}$, $s_2 = \frac{n}{2} + \frac{1}{2} + z_{\alpha/2}\frac{\sqrt{n}}{2}$.

- **Recorrendo ao *valor-p***:
  - É necessário ter presente que o teste é bilateral e que os acontecimentos tão ou mais anómalos do que aquele que se verificou se situam nas duas abas da distribuição binomial.
  - Valor esperado $\rightarrow \; n \times p = n/2$
  - Valor observado $\rightarrow S_{\text{obs}}$
  - Diferença absoluta entre o valor observado e o valor esperado $\rightarrow$ $\delta = |S_{\text{obs}} - n\,p|$
  - Limites a considerar:
    - Inferior: $n\,p - \delta$
    - Superior: $n\,p + \delta$
  - *Valor-p*$= \Pr(S \le np - \delta \mid H_0) + \Pr(S \ge np + \delta \mid H_0)$
  - Suponha-se $n = 15$.
    - Se $S_{\text{obs}} = 4$, vem $\delta = |4 - 7.5| = 3.5$, $n\,p - \delta = 4$, $n\,p + \delta = 11 = 15 - 4$
      logo $\text{valor} - \text{p} = \Pr(S \le 4 \mid H_0) + \Pr(S \ge 11 \mid H_0) \approx 2 \times 0.0592 = 0.1184$

- Se $S_{\text{obs}} = 12$, vem $\delta = |12 - 7.5| = 4.5$, $n\,p - \delta = 3 = 15 - 12$, $n\,p + \delta = 12$
  logo $\text{valor} - \text{p} = \Pr(S \leq 3 \mid H_0) + \Pr(S \geq 12 \mid H_0) \approx 2 \times 0.0176 = 0.0352$

- A adaptação para testes unilaterais não levanta problemas. Ver exemplo 2.
- Caso algum valor da amostra seja igual ao valor postulado para a mediana (universos discretos) não se pode atribuir um sinal a esta observação. Existem várias soluções:
  - Não considerar esta observação (prejudica H_0) $\rightarrow$ **solução habitual**
  - Escolher o valor do sinal por forma a diminuir δ, isto é, aproximando S_{obs} do valor esperado (favorece H_0)
  - Casualizar por "moeda ao ar";

Exemplo 1 – Suponha-se que se dispõe de uma amostra casual 20 observações da precipitação anual em Beja. Pretende testar-se, $\alpha = 0.05$, a hipótese de a mediana da população ser igual a 650 *mm*.

Precipitação anual (*mm*) no distrito de Beja

607.4	345.4	497.6	464.0	809.1	620.0	728.4	809.1	488.8	407.7
602.3	721.8	481.1	513.3	672.0	533.9	592.8	527.4	581.1	384.2

Solução:

- Hipóteses em teste: $H_0 : \mu_e = 650\ mm$ contra $H_1 : \mu_e \neq 650\ mm$.
- $S_{obs} = 5$ já que temos 5 valores observados superiores a 650
- Região crítica utilizando a binomial $\rightarrow$ região crítica de dimensão máxima é formada pelos conjuntos, {0, 1, …,5} e {15, 16, …,20} já que
  - $P(S \leq 5 \mid H_0) \approx 0.0207 \leq 0.025$ e $P(S \leq 6 \mid H_0) \approx 0.0577 > 0.025$
  - $P(S \geq 15 \mid H_0) = 0.0207 \leq 0.025$ e $P(S \geq 14 \mid H_0) \approx 0.0577 > 0.025$.
  - Como $S_{obs} = 5$ pertence à região de rejeição, rejeitamos, para o valor de α dado, H_0
  - Repare-se que a dimensão deste teste é efectivamente de 4.14% e não de 5%.
- Região crítica com aproximando à normal ($z_{\alpha/2} = 1.96$)

$$s_1 = 10 - 0.5 - 1.96\frac{\sqrt{20}}{2} \approx 5.12 \to 5 \text{ (sempre por defeito)}$$

$$s_2 = 10 + 0.5 + 1.96\frac{\sqrt{20}}{2} \approx 14.88 \to 15 \text{ (sempre por excesso)}$$

sendo a conclusão a mesma. A hipótese H_0 é de rejeitar.

- *Valor-p*

$\delta = |5 - 10| = 5$, $n\,p - \delta = 5$, $n\,p + \delta = 15 = 20 - 5$ logo

$$\text{valor} - \text{p} = \Pr(S \le 5 \mid H_0) + \Pr(S \ge 15 \mid H_0) \approx 2 \times 0.0207 = 0.0414$$

ou

$$\text{valor} - \text{p} = 2 \times \Pr(S \le 5 \mid H_0) \approx 2 \times \Phi\left(\frac{5.5 - 10}{\sqrt{10 \times 0.5 \times 0.5}}\right) = 0.0417$$

Exemplo 2 – Retome-se o exemplo anterior e suponha-se que se queria testar, $\alpha = 0.05$, $H_0 : \mu_e \geq 650\ mm$ contra $H_1 : \mu_e < 650\ mm$.

Solução:

- $S_{obs} = 5$ já que temos 5 valores observados superiores a 650
- Região crítica utilizando a binomial
  - região crítica dada por $\{0, 1, \ldots, 5\}$ já que $P(S \leq 5 \mid H_0) \approx 0.0207 \leq 0.025$ e $P(S \leq 6 \mid H_0) > 0.025$
  - Como $S_{obs} = 5$ pertence à região de rejeição, rejeitamos, para o valor de α dado, H_0
  - Repare-se que a dimensão deste teste é efectivamente de 2.07% e não de 5%.
- Aproximando à normal
  - $z_\alpha = 1.645$, logo $s_1 = 10 - 0.5 - 1.645\dfrac{\sqrt{20}}{2} = 5.82$ sendo a conclusão a mesma. Rejeita-se H_0.
- Valor-p

$$\text{valor} - \text{p} = \Pr(S \leq S_{\text{obs}} \mid H_0) = \Pr(S \leq 5 \mid H_0) \approx 0.0207$$

Algumas extensões ao teste do sinal

- Testar outros quantis para além da mediana → adaptação imediata através do parâmetro p da binomial
- **Exemplo 3** – Testar, $\alpha = 0.05$, se o primeiro quartil da distribuição da precipitação é superior ou igual a 500 mm.

  Solução:

  - $H_0 : \xi_{0.25} \geq 500\ mm$ contra $H_1 : \xi_{0.25} < 500\ mm$.
  - $S_{obs} = 13$ já que temos 13 valores observados superiores a 500. Dado H_0, $S \sim b(20; 0.75)$
  - $n\,p = 20 \times 0.75 = 15$, logo $S_{obs} = 13 < 15$
  - Utilizando a binomial
    - $\text{valor} - \text{p} = \Pr(S \leq S_{\text{obs}} \mid H_0) = \Pr(S \leq 13 \mid H_0) \approx 0.2142$
    - região crítica de dimensão máxima é formada pelo conjuntos, {0, 1, …,11} já que 11 é o maior inteiro tal que

      $$P(S \leq 11 \mid H_0) = \sum_{s=0}^{11} \binom{20}{s} 0.75^s\ 0.25^{20-s} = 0.0409 \leq 0.05$$

  - Em qualquer dos casos não se rejeita H_0

Teste do sinal e amostras emparelhadas

- O teste do sinal é bastante utilizado para amostras emparelhadas (X_i, Y_i), $i = 1,2,...,n$, cuja escala de medida é pelo menos ordinal (por exemplo: desgaste do pneu dianteiro direito e do pneu traseiro direito; resposta de pares de sujeitos, um submetido a tratamento e outro funcionando como controlo).
- Neste tipo de amostras os n pares (X_i, Y_i) são independentes entre si mas as variáveis X_i e Y_i, dentro de cada par, são dependentes. A hipótese a testar é do tipo,

$$H_0 : P(X_i > Y_i) = P(X_i < Y_i) \text{ contra } H_1 : P(X_i > Y_i) \neq P(X_i < Y_i),$$

  (H_0 é o *status quo*, não há diferença no desgaste dos pneus, o tratamento não produz efeito, etc.).
- É necessário garantir a consistência da informação na população, isto é , que uma e só uma das três situações é possível:
  - $P(X_i > Y_i) > P(X_i < Y_i)$; $P(X_i > Y_i) < P(X_i < Y_i)$; $P(X_i > Y_i) = P(X_i < Y_i)$.
- Considere-se $Z_i = X_i - Y_i$ $(i = 1,2,...,n)$ e veja-se que o teste é equivalente a $H_0 : P(Z_i > 0) = P(Z_i < 0)$ contra $H_1 : P(Z_i > 0) \neq P(Z_i < 0)$, isto vai testar-se se a mediana da distribuição de Z é nula.

Exemplo 4 – Uma amostra aleatória composta por seis homens que seguiram um programa quinzenal destinado a perderem peso originou os seguintes resultados:

Indivíduos	**1**	**2**	**3**	**4**	**5**	**6**
Peso antes	87	96	91	94	101	94
Peso depois	83	93	91	91	102	90

Será que se pode concluir que o programa é eficiente? **Solução:**

- Seja X_i o peso do indivíduo i antes do tratamento e Y_i o peso do mesmo individuo depois do tratamento;
- O teste é então $H_0 : P(X_i > Y_i) = P(X_i < Y_i)$ contra $H_1 : P(X_i > Y_i) > P(X_i < Y_i)$, em que a hipótese H_1 traduz um tratamento eficiente (origina perda de peso) e H_0 a manutenção do *status quo*.
- De forma equivalente, $H_0 : P(Z_i > 0) = P(Z_i < 0)$ contra $H_1 : P(Z_i > 0) > P(Z_i < 0)$, teste unilateral ($Z$=$X$-$Y$).
- Como existe uma situação de empate apenas se consideram 5 valores na amostra, tendo-se obtido 4 valores positivos.
- Dado H_0, $S \sim b(5\,;\,0.5)$
- $\text{valor} - \text{p} = P(S \geq 4 \mid H_0) \approx 0.1876$ logo não se rejeita H_0 e questiona-se o programa.

Teste de Wilcoxon (ou de ordem-sinal)

- Mesma filosofia do que o teste do sinal mas mais "sofisticada" já que se vai utilizar a informação dada pela ordem das diferenças em relação à mediana
- Pretende-se testar $H_0 : \mu_e = \mu_{e0}$ contra $H_1 : \mu_e \neq \mu_{e0}$, com base numa amostra casual de população com função de distribuição $F(x)$ desconhecida, mas **simétrica**.
- Obtenção da estatística de Wilcoxon
  - Calcular as diferenças $Z_i = X_i - \mu_{e0}$ $(i = 1,2,...,n)$, e os respectivos sinais $S_i = \begin{cases} 0 & Z_i < 0 \\ 1 & Z_i > 0 \end{cases}$ assumindo que não há empates e que $Z_i \neq 0$
  - Ordenar os valores de $|Z_i|$ por ordem crescente bem como os S_i (pela ordem dos $|Z_i|$). Seja j o indicador de ordem ($j = 1$ corresponde ao menor $|Z_i|$, $j = 2$ ao seguinte e assim por diante até $j = n$);
  - A estatística S (teste do sinal) é dada por $S = \sum_{i=1}^{n} S_i = \sum_{j=1}^{n} S_j$ e a estatística de Wilcoxon, T, por $T = \sum_{j=1}^{n} j\, S_j$

Exemplo 5 – Retome-se o exemplo 1 (que se vai assumir ter uma distribuição simétrica) e considerem-se apenas as 10 primeiras observações (1ª linha) para diminuir a dimensão dos quadros (corrigi o valor a "bold" para não haver empates). Vai testar-se $H_0 : \mu_e = 650\ mm$ contra $H_1 : \mu_e \neq 650\ mm$ com base no teste de Wilcoxon.

Calculem-se então as estatísticas *S* e *T*.

- 1ª fase: Calcular $|z_i|$ para ordenar.

i	1	2	3	4	5	6	7	8	9	10
x_i	607.4	345.4	497.6	464.0	809.1	620.0	728.4	**809.2**	488.8	407.7
z_i	-42.6	-304.6	-152.4	-186.0	159.1	-30.0	78.4	159.2	-161.2	-242.3
$\lvert z_i \rvert$	42.6	304.6	152.4	186.0	159.1	30.0	78.4	159.2	161.2	242.3
S_i	0	0	0	0	1	0	1	1	0	0

- 2ª fase: Ordenar $|z_i|$ e construir o quadro ($s_i = 1$ quando $z_i > 0$ e 0 quando $z_i < 0$) eliminando-se as observações que originam $z_i = 0$ (solução habitual)

j	1	2	3	4	5	6	7	8	9	10
$\lvert z_i \rvert$ ord	30.0	42.6	78.4	152.4	159.1	159.2	161.2	186.0	242.3	304.6
S_j	0	0	1	0	1	1	0	0	0	0
$j \times S_j$	0	0	3	0	5	6	0	0	0	0

- 3ª fase: $S_{\text{obs}} = 3$, $T_{\text{obs}} = 14$

- Que distribuição para a estatística de teste?
  - A distribuição exacta da estatística de Wilcoxon, quando H_0 é verdadeira, obtém-se segundo um processo de enumeração e está tabelada, em termos aproximados, para valores de *n* pequenos.
  - Alguns quantis aproximados para a estatística de Wilcoxon

n	5	6	7	8	9	10	11	12	13	14	15	20	25	30
1%	0	0	1	2	4	6	8	10	13	16	20	44	77	121
2.5%	0	1	3	4	6	9	11	14	18	22	26	53	90	138
5%	1	3	4	6	9	11	14	18	22	26	31	61	101	152
10%	3	4	6	9	11	15	18	22	27	32	37	70	114	170
m	15	21	28	36	45	55	66	78	91	105	120	210	325	46

- Como a distribuição de *T* apenas está definida para valores inteiros entre 0 e $m = n(n+1)/2$ e é simétrica, apenas se apresentam os primeiros quantis, obtendo os valores da aba direita por simetria (a última linha do quadro refere o valor máximo de *T* em função de *n*). Assim, o quantil 90% para $n = 5$ será,

$$q_{0.90} = \frac{n(n+1)}{2} - q_{0.05} \approx 15 - 3 = 12\,.$$

- Quando a amostra é grande a distribuição da estatística de Wilcoxon, *T*, pode ser aproximada com base no TLC.
- Para tal é necessário obter $E(T)$ e $\text{var}(T)$ assumindo H_0 verdadeira.
    - S_j $(j = 1,2,...,n)$, são independentes e, como $S_j \sim B(1;0.5)$, $E(S_j) = 0.5$ e $\text{var}(S_j) = 0.25$
    - $E(T) = E\left(\sum_{j=1}^{n} j\, S_j\right) = \sum_{j=1}^{n} j\, E(S_j) = 0.5 \sum_{j=1}^{n} j = \dfrac{n(n+1)}{4}$
    - $\text{var}(T) = \text{var}\left(\sum_{j=1}^{n} j\, S_j\right) = \sum_{j=1}^{n} j^2\, \text{var}(S_j) = \dfrac{1}{4}\sum_{j=1}^{n} j^2 = \dfrac{n(n+1)(2n+1)}{24}$
- Aplicando o Teorema do Limite Central vem

$$T^* = \frac{T - \dfrac{n(n+1)}{4}}{\sqrt{\dfrac{n(n+1)(2n+1)}{24}}} \overset{a}{\sim} N(0,1).$$

Exemplo 6 – Retome-se o exemplo anterior, com as 20 observações (mantém-se a correcção), e teste-se,

$$H_0 : \mu_e = 650\ mm \text{ contra } H_1 : \mu_e \neq 650\ mm .$$

Sendo $n = 20$, vem $T_{\text{obs}} = 43$ para a amostra dada (ver livro ou fazer as contas) e o *valor-p* associado à hipótese H_0 é inferior a 0.02 (ver tabela), já
que, $p_{\text{obs}} = P(T \leq 43) + P(T \geq 210 - 43) = 2 \times P(T \leq 43)$. Assim, a hipótese H_0 é de rejeitar.

Com a aproximação conseguida com a distribuição $N(0,1)$ vem,

$$T^*_{\text{obs}} = \frac{43 - 105}{\sqrt{717.5}} = -2.31,$$

e a conclusão é a mesma.

- Comentários finais
  - População contínua $\rightarrow$ probabilidade nula de se obter $Z_i = 0$ ou de existirem empates. Tal não acontece no entanto em população discretas ou quando se arredondam os números.
  - Valores nulos de Z_i. Não existe unanimidade. A solução mais frequente é eliminar esta observação e trabalhar com uma amostra reduzida (esta solução prejudica H_0).

- Existência de empates. Quando existem valores de $|Z_i|$ iguais (empates), atribui-se a cada um destes um número de ordem igual à média das ordens que lhes caberia. Por exemplo, suponha-se que os valores de $|Z_i|$ são os seguintes: 1.1, 1.1, 2.3, 3.2, 3.2, 3.2. Às observações de valor 1.1 correspondem as ordens 1 e 2, logo atribui-se, a cada uma delas, um número de ordem igual a 1.5 (média dos valores 1 e 2). Para o valor 2.3, não existe empate, logo a ordem continua a ser 3. Finalmente, como os três últimos valores estão empatados e correspondem às ordens 4, 5 e 6, vai atribuir-se a cada um a ordem 5. Representando por r_j o número de ordem da observação j (quando se consideram situações de empate), tem-se: $r_1 = 1.5$; $r_2 = 1.5$; $r_3 = 3$; $r_4 = 5$; $r_5 = 5$; $r_6 = 5$.

  Adaptam-se então os parâmetros da distribuição assintótica

  $$E(T) = E\left(\sum_{j=1}^{n} r_j\, S_j\right) = \sum_{j=1}^{n} r_j\, E(S_j) = \frac{1}{2}\sum_{j=1}^{n} r_j = \frac{n(n+1)}{4}$$ não se altera

  $$\mathrm{Var}(T) = \mathrm{Var}\left(\sum_{j=1}^{n} r_j\, S_j\right) = \sum_{j=1}^{n} r_j^2\, \mathrm{Var}(S_j) = \frac{1}{4}\sum_{j=1}^{n} r_j^2$$ altera-se

- Se se verificar um número significativo de empates ou de valores nulos é preferível optar por outra metodologia de teste.

Notas sobre a distribuição do teste de Wilcoxon (em princípio para não dar)

- Depois de ordenar os $|z_i|$ existem 2^n situações possíveis já que cada valor ordenado pode provir de uma observação inferior à mediana hipotética ou superior a esta mediana;
- As situações são equiprováveis já que se trata da mediana;
- Pode-se então enumerar as situações, calcular o valor de T para cada uma delas obtendo-se assim a distribuição por amostragem de T. Tirando partido da simetria da distribuição apenas é necessário analisar metade das situações
- Exemplifiquem-se 2 situações, uma muito simples ($n = 3$) e outra mais trabalhosa ($n = 5$) mas que permite compara os resultados com a tabela;
- Situação 1 – Seja então w_1 o menor valor de $|z_i|$, w_2 o seguinte e assim por diante até w_n. Quando um valor de w_j provém de uma observação superior à mediana ($S_j = 1$) escreve-se w_j^+ e quando provém de uma observação inferior, w_j^-. Listem-se então as 8 (2^3) situações possíveis, cada uma com probabilidade $1/8$ e calcule-se o valor de T associado com cada uma delas.

Situação	T	Situação	T
$w_1^-,\ w_2^-,\ w_3^-$	0	$w_1^+,\ w_2^+,\ w_3^-$	3
$w_1^+,\ w_2^-,\ w_3^-$	1	$w_1^+,\ w_2^-,\ w_3^+$	4
$w_1^-,\ w_2^+,\ w_3^-$	2	$w_1^-,\ w_2^+,\ w_3^+$	5
$w_1^-,\ w_2^-,\ w_3^+$	3	$w_1^+,\ w_2^+,\ w_3^+$	6

Logo a distribuição por amostragem de T será

t	0	1	2	3	4	5	6
$\Pr(T=t)$	1/8	1/8	1/8	2/8	1/8	1/8	1/8

- Situação 2 – Mantenham-se as convenções da situação anterior e listem-se as 32 situações equiprováveis

Situação	T	Situação	T	Situação	T
$w_1^-, w_2^-, w_3^-, w_4^-, w_5^-$	0	$w_1^-, w_2^+, w_3^-, w_4^+, w_5^-$	6	$w_1^-, w_2^+, w_3^+, w_4^+, w_5^-$	9
$w_1^+, w_2^-, w_3^-, w_4^-, w_5^-$	1	$w_1^-, w_2^+, w_3^-, w_4^-, w_5^+$	7	$w_1^-, w_2^+, w_3^+, w_4^-, w_5^+$	10
$w_1^-, w_2^+, w_3^-, w_4^-, w_5^-$	2	$w_1^-, w_2^-, w_3^+, w_4^+, w_5^-$	7	$w_1^-, w_2^+, w_3^-, w_4^+, w_5^+$	11
$w_1^-, w_2^-, w_3^+, w_4^-, w_5^-$	3	$w_1^-, w_2^-, w_3^+, w_4^-, w_5^+$	8	$w_1^-, w_2^-, w_3^+, w_4^+, w_5^+$	12
$w_1^-, w_2^-, w_3^-, w_4^+, w_5^-$	4	$w_1^-, w_2^-, w_3^-, w_4^+, w_5^+$	9	$w_1^+, w_2^+, w_3^+, w_4^+, w_5^-$	10
$w_1^-, w_2^-, w_3^-, w_4^-, w_5^+$	5	$w_1^+, w_2^+, w_3^+, w_4^-, w_5^-$	6	$w_1^+, w_2^+, w_3^+, w_4^-, w_5^+$	11
$w_1^+, w_2^+, w_3^-, w_4^-, w_5^-$	3	$w_1^+, w_2^+, w_3^-, w_4^+, w_5^-$	7	$w_1^+, w_2^+, w_3^-, w_4^+, w_5^+$	12
$w_1^+, w_2^-, w_3^+, w_4^-, w_5^-$	4	$w_1^+, w_2^+, w_3^-, w_4^-, w_5^+$	8	$w_1^+, w_2^-, w_3^+, w_4^+, w_5^+$	13
$w_1^+, w_2^-, w_3^-, w_4^+, w_5^-$	5	$w_1^+, w_2^-, w_3^+, w_4^+, w_5^-$	8	$w_1^-, w_2^+, w_3^+, w_4^+, w_5^+$	14
$w_1^+, w_2^-, w_3^-, w_4^-, w_5^+$	6	$w_1^+, w_2^-, w_3^+, w_4^-, w_5^+$	9	$w_1^+, w_2^+, w_3^+, w_4^+, w_5^+$	15
$w_1^-, w_2^+, w_3^+, w_4^-, w_5^-$	5	$w_1^+, w_2^-, w_3^-, w_4^+, w_5^+$	10	---	--

Logo a distribuição por amostragem de T será

t	0	1	2	3	4	5	6	7
$\Pr(T=t)$	1/32	1/32	1/32	2/32	2/32	3/32	3/32	3/32
t	8	9	10	11	12	13	14	15
$\Pr(T=t)$	3/32	3/32	3/32	2/32	2/32	1/32	1/32	1/32

- Observar a simetria da distribuição (e das situações descritas no quadro anterior)
- Percentis para $n=5$ no quadro com os valores aproximados:

  $1\% \rightarrow 0$ $2.5\% \rightarrow 0$ $5\% \rightarrow 1$ $10\% \rightarrow 3$

t	0	1	2	3	...
$\Pr(T \le t)$	0.03125	0.0625	0.09375	0.15625	...