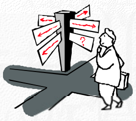


# Testes Não Paramétricos

## KAI (k Amostras Independentes)

Prof. Lorí Viali, Dr.
http://www.mat.ufrgs.br/viali/
viali@mat.ufrgs.br



# Testes para k amostras Independentes

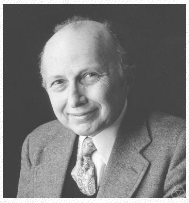
Prof. Lorí Viali, Dr. – UFRGS – Instituto de Matemática - Departamento de Estatística – Curso de Estatística

## Os testes

- O teste de Kruskal-Wallis (Análise de variância de uma classificação por postos)
- O teste qui-Quadrado

Prof. Lorí Viali, Dr. – UFRGS – Instituto de Matemática - Departamento de Estatística – Curso de Estatística

## O teste de Kruskal-Wallis



William Henry **Kruskal**
(1919 - 2005)

William Allen **Wallis**
(1912 - 1998)

Prof. Lorí Viali, Dr. – UFRGS – Instituto de Matemática - Departamento de Estatística – Curso de Estatística

## Objetivo

O teste de Kruskal-Wallis é utilizado para decidir se k amostras independentes podem ter sido extraídas de populações diferentes.

Prof. Lorí Viali, Dr. – UFRGS – Instituto de Matemática - Departamento de Estatística – Curso de Estatística

Os valores amostrais diferem entre si e deve-se decidir se essas diferenças amostrais significam diferenças efetivas entre as populações, ou se representam apenas variações casuais.

Prof. Lorí Viali, Dr. – UFRGS – Instituto de Matemática - Departamento de Estatística – Curso de Estatística

O teste supõe que a variável em estudo tenha distribuição contínua e exige mensuração no mínimo ao nível ordinal.



## Metodologia

Cada um dos **n** valores é substituído por um posto. Isto é, os escores de todas as $k$ amostras combinadas são dispostos em uma única série de postos. Ao menor escore é atribuído o posto 1, ao seguinte o posto 2 e assim por diante até o maior posto que é n = número total de observações.



Feito isso, determina-se a soma dos postos em cada amostra (coluna). A prova então testa se estas somas são tão diferentes entre si, de modo que não seja provável que tenham sido todas retiradas de uma mesma população.



Tratamentos

| $T_1$ | $T_2$ | ... | $T_k$ |
|---|---|---|---|
| $X_{11}$ | $X_{12}$ | ... | $X_{1k}$ |
| $X_{21}$ | $X_{22}$ | ... | $X_{2k}$ |
| $X_{31}$ | $X_{32}$ | ... | $X_{3k}$ |
| ... | ... | ... | ... |
| $X_{n1}$ | $X_{n2}$ | ... | $X_{nk}$ |



## Suposições

No artigo original de Kruskal-Wallis de 1952 apenas suposições gerais foram estabelecidas:

(i) As observações são independentes;

(ii) Dentro de cada amostra as observações são da mesma população e

(iii) As $k$ populações são da mesma forma e contínuas.



## Hipóteses

$H_0$: Os $k$ tratamentos não diferem entre si;

$H_1$: Pelos menos dois tratamentos diferem entre si.

## A estatística teste

Se as k amostras forem de uma mesma população ($H_0$ é V) então a estatística de Kruskal-Wallis tem distribuição conhecida (Tabela O) se as amostras forem pequenas ($n < 5$) ou Qui-Quadrado com $gl = k - 1$, desde que os tamanhos das k amostras não sejam muito pequenos (5 ou mais elementos).



A estatística amostral é:

$$H = \frac{12}{n(n+1)} \sum_{j=1}^{k} \left( \frac{(\sum R_j)^2}{n_j} \right) - 3(n+1)$$



## Onde:

$k$ = número de amostras;

$n_j$ = número de elementos na amostra "j";

$R_j$ = soma dos postos do tratamento (amostra ou coluna) "j";

$n = \sum n_j$ = número total de elementos de todas as amostras combinadas;



Alguns autores recomendam que se existe um número grande de empates o valor de H dever ser corrigido. A correção consiste em aumentar levemente o valor de H. A seguinte equação é utilizada para obter o valor da correção de H:

$$C = 1 - \frac{\sum_{i=1}^{s}(t_i^3 - t_i)}{n^3 - n}$$

Onde:

$S$ = número de grupos empatados;

$t_i$ = número de valores empatados.



O valor de H corrigido será igual então a:

$H_C = H/C$



## Decisão:

Rejeitamos $H_0$ se $H \geq h$, onde $P(H \geq h) = \alpha$.

A tabela O fornece os limites de h para $n_i \leq 6$ e $k = 3$.

À medida que os $n_i$ crescem a distribuição de H sob $H_0$ tende para a $\chi^2$ com $k - 1$ graus de liberdade.

# E x e m p l o



Verificar a influência do Fator "Idade" sobre a variável "tempo, em dias, para conseguir um emprego", considerando as seguintes amostras:



| Acima de 40 anos | Entre 25 e 40 | Abaixo de 25 |
|---|---|---|
| 63 | 33 | 25 |
| 20 | 42 | 31 |
| 43 | 27 | 6 |
| 58 | 28 | 14 |
| 57 | 51 | 18 |
| 71 | 64 | 33 |
| 45 | 12 | |
| | 30 | |



Tem-se n = 21 (total de informações). Então o maior posto será 21.



| | Postos (1) | Postos (2) | Postos (3) |
|---|---|---|---|
| 1 | 5 | 2 | 1 |
| 2 | 14 | 7 | 3 |
| 3 | 15 | 8 | 4 |
| 4 | 17 | 9 | 6 |
| 5 | 18 | 11,5 | 10 |
| 6 | 19 | 13 | 11,5 |
| 7 | 21 | 16 | |
| 8 | | 20 | |
| $\Sigma R_j$ | 109 | 86,5 | 35,5 |
| Média | 15,57 | 10,81 | 5,92 |



A variável teste será:

$$H = \frac{12}{n(n+1)} \sum_{j=1}^{k} \left( \frac{(\Sigma R_j)^2}{n_j} \right) - 3(n+1) =$$

$$= \frac{12}{21(21+1)} \left( \frac{109^2}{7} + \frac{86,5^2}{8} + \frac{35,5^2}{6} \right) - 3(21+1) \quad =$$

$$= 73,834 - 66 = 7,834$$

O grau de liberdade é:

$v = k - 1 = 3 - 1 = 2$

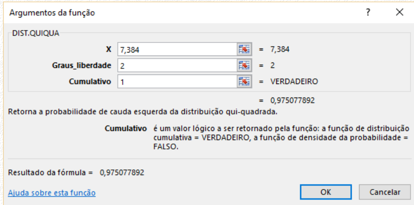
Como ocorreu apenas um conjunto de empates e com apenas dois valores, não vale a pena utilizar a correção para H.

$$C = 1 - \frac{\sum_{i=1}^{s} (t_i^3 - t_i)}{n^3 - n}$$

Contudo se isto fosse feito o novo valor de H, isto é, Hc = 7,839.



O qui-quadrado tabelado será:




Assim o valor-p deste resultado será

1 – 0,9751 = 2,49%



## Conclusão

A 5% de significância é possível afirmar que o fator "idade" tem influência sobre o "tempo para encontrar trabalho".



# S o l u ç ã o
# S P S S



## Resultados SPSS

### Kruskal-Wallis Test

| Controle | n | Mean Rank |
|---|---|---|
| 0 | 7 | 15,57 |
| 1 | 8 | 10,81 |
| 2 | 6 | 5,92 |
| Total | 21 | |

| | Tempo |
|---|---|
| Chi-Square | 7,839 |
| df | 2 |
| Assyp. Sig. | 0,020 |



# E x e r c í c i o

Uma indústria de pneus testou a distância de frenagem, em pista molhada, das suas cinco opções de pneus. As distâncias observadas, em metros, estão na tabela. Verifique se existe diferença entre as marcas.



| | A | B | C | D | E |
|---|---|---|---|---|---|
| **1** | 45,8 | 47,6 | 40,9 | 44,5 | 44,2 |
| **2** | 43,3 | 47,9 | 44,2 | 52,7 | 51,8 |
| **3** | 48,1 | 45,4 | 43,0 | 54,2 | 50,6 |
| **4** | 46,0 | 43,0 | 39,1 | 49,4 | 43,9 |
| **5** | 47,2 | 42,4 | 42,1 | 44,8 | 44,5 |
| **6** | | | | 50,0 | 50,3 |



| | A | B | C | D | E |
|---|---|---|---|---|---|
| 1 | | | | | |
| 2 | | | | | |
| 3 | | | | | |
| 4 | | | | | |
| 5 | | | | | |
| 6 | | | | | |
| Total | | | | | |
| Média | | | | | |



# Comparações Múltiplas



Quando a diferença for significativa, isto é, existem tratamentos que diferem é possível identificar qual ou quais pares diferem por intermedio das comparações múltiplas.



Para $j = 1, 2, ..., k$, sejam $R_j$ a soma dos postos e $r_j = R_j/n_j$ a média do postos do tratamento $j$. Dados $r_i$ e $r_j$ diremos que os dois tratamentos diferem se $|r_i - r_j|$ for significativamente grande.

Para comparar todos os tratamentos dois a dois é necessário fazer $k(k - 1)/2$ comparações. Nesse caso, é razoável adotar uma probabilidade maior de erro. Em geral 10% ou até um pouco mais.



Para um dado determinamos se dois tratamentos diferem adotando o seguinte procedimento:

Determinar o valor z que corresponde à probabilidade $\alpha' = 2.\alpha/k(k-1)$ da cauda superior da normal padrão.



Para cada para i, j com $i \neq j$ calcula-se:

$$z_{ij} = \frac{r_i - r_j}{\sigma_{ij}}$$

Onde:

$$\sigma_{ij} = \sqrt{\frac{n(n+1)}{12}\left(\frac{1}{n_i} + \frac{1}{n_j}\right)}$$



Diremos que o tratamento i produz medidas menores que o j se $z_{ij} < -z$; medidas maiores do que j se $z_{ij} > z$ e medidas da mesma ordem de grandeza se $-z \leq z \leq z$.



Retornando ao exercício dos pneus, onde queremos comparar as cinco marcas, duas a duas, com respeito ao desempenho na frenagem. Vamos supor que estamos dispostos a tolerar uma significância de 10%.



Então $\alpha' = \alpha/k(k-1) = 1\%$ e $z = 2,326$.

Os postos médios são:

15,0 12,1 4,2 20,1 16,8

Ao invés de compara $z = 2,326$ com o quociente $z_{ij} = \frac{r_i - r_j}{\sigma_{ij}}$ vamos comparar as diferenças $r_i - r_j$ com os produtos $z\sigma_{ij}$.

Como os tamanhos das amostras são diferentes temos $\sigma_{ij}$ diferentes. Então:

| Tamanhos das amostras | $\sigma_{ij}$ | $z\sigma_{ij}$ |
|---|---|---|
| 5 e 5 | 5,02 | |
| 5 e 6 | 4,81 | |
| 6 e 6 | 4,58 | |



# O teste Qui-Quadrado



## O teste qui-quadrado

O teste $\chi^2$ de "k" amostras independentes pode ser utilizado para verificar a dependência ou independência entre as variáveis sendo consideradas.



O teste é uma extensão direta do qui-quadrado para duas amostras independentes. Em geral, o teste é o mesmo, tanto para duas, como para k amostras independentes.



## Hipóteses e Cálculo

$H_0$: As variáveis são independentes

$H_1$: As variáveis são dependentes

A variável teste é:

$$\chi^2_v = \frac{\sum_{i=1}^{k}\sum_{j=1}^{l}\left(O_{ij}-E_{ij}\right)^2}{E_{ij}}$$



## Expressão alternativa

A variável teste é:

$$\chi^2_v = \frac{\sum_{i=1}^{k}\sum_{j=1}^{l}\left(O_{ij}-E_{ij}\right)^2}{E_{ij}} = \frac{\sum_{i=1}^{k}\sum_{j=1}^{l} O_{ij}^2}{E_{ij}} - n$$

Onde:

$r$ = número de linhas da tabela;

$L$ = número de colunas da tabela;

$O_{ij}$ = frequência observada na interseção da linha $i$ com a coluna $j$.

$E_{ij}$ = número de casos esperados na interseção da linha $i$ com a coluna $j$.



Onde:

$\chi^2_v$ é a estatística teste;

$n = \sum_{i=1}^{k} \sum_{j=1}^{L} O_{ij}$ = tamanho da amostra;

$E_{ij} = np_{ij}$ são as frequências esperadas de cada célula $ij$ da tabela.



$p_{ij}$ é a probabilidade de ocorrer uma observação na célula $ij$. Se as variáveis são supostamente independentes ($H_0$ é Verdadeira), então $p_{ij}$ = $p_{i.}p_{.j}$, onde $p_{i.}$ é a probabilidade marginal correspondente à linha "i" e $p_{.j}$ é a probabilidade marginal correspondente a coluna $j$.



Como não se conhecem as probabilidades marginais, elas devem ser estimadas através das correspondentes frequências relativas. Então:

$$E_{ij} = n\,p_{ij} = n\,p_{i.}.p_{.j} = n.\frac{f_{i.}}{n}.\frac{f_{.j}}{n} = \frac{f_{i.}f_{.j}}{n}$$



$$f_{i.} = \sum_{j=1}^{L} f_{ij} \quad e \quad f_{.j} = \sum_{i=1}^{k} f_{ij}$$



# Teste de Jonckheere-Terpstra

*Jonckheere-Terpstra Test. This test for differences among several independent samples is more powerful than the Kruskal-Wallis H or median tests. However, it requires that the independent samples be ordinally arranged on the criterion variable (ex., city samples arranged by welfare caseload per 10,000 population, where this is the variable of interest).*



*The J-T test tests the hypothesis that as one moves from samples low on the criterion to samples high on the criterion, the within-sample magnitude of the criterion variable increases.*

*Correção para empates*

[http://www.ens.gu.edu.au/stats/aes3121/lectures/exam_ho.htm](http://www.ens.gu.edu.au/stats/aes3121/lectures/exam_ho.htm)



KRUSKAL, William Henry; WALLIS, William Allen. Use of ranks in one-criterion variance analysis. **Journal of the American Statistical Association**, v. 47, n. 260, p. 583–621, 1952.