Departamento de Estatística e Matemática Aplicada da UFC
CC0291- Análise Não Paramétrica
Conover:- Teste de Kruskal-Wallis- 25/ 05/2023
Professor: Maurício Mota

1:Introdução

Vamos apresentar como Conover descreve o teste Kruskal-Wallis.

Usamos o software R em várias aplicações do livro.

2: Teste de Kruskal-Wallis.

Dados: Os dados consistem em amostras de tamanhos $n_i,\ i = 1, 2, \ldots, k$ independentes de k populações. Vamos denotar a i-ésima amostra de tamanho n_i por:

$$Y_{i1}, Y_{i2}, \ldots, Y_{in_i}.$$

Vamos fazer um quadro:

Amostra 1	**Amostra 2**	...	**Amostra k**
Y_{11}	Y_{21}	...	Y_{k1}
Y_{12}	Y_{22}	...	Y_{k2}
...	...	...	...
Y_{1n_1}	Y_{2n_2}	...	Y_{kn_k}

Seja Y_{ij} a j-ésima observação da unidade experimental submetida ao i-ésimo tratamento, $i = 1, 2, \ldots, k$ e $j = 1, 2, \ldots, n_i$. Sejam

$$N = \sum_{i=1}^{k} n_i \tag{1}$$

Considere com uma única população e atribua postos aos valores de Y_{ij}. Seja $R(Y_{ij})$ o posto assinalado 'a observação Y_{ij}.

Seja

$$R_i = \sum_{j=1}^{n_i} R(Y_{ij}) \ \ i = 1, 2, \ldots, k, \tag{2}$$

a soma dos postos atribuídos a i-ésima amostra (tratamento).

Calcule R_i para cada tratamento.

Suposições:

1. Todas as amostras são amostras aleatórias são amostras aleatórias de suas respectivas populações.

2. Além da independência dentro de cada amostra, há independência mútua entre as várias amostras

3. A escala de medida é pelo menos ordinal.

4. Ou as funções de distribuição das K populações são idênticas, ou então algumas das populações tendem a valores maiores do que outras populações,

As hipóteses a serem testadas:
H_0 : Todas as k populações tem idênticas Funções de distribuição. H_1 : Pelo menos uma das populações tende a produzir observações maiores do que pelo menos uma das outras populações.

Estatística do teste

$$T = \frac{1}{S^2}\Big(\sum_{i=1}^{k}\frac{R_i^2}{n_i} - \frac{N(N+1)^2}{4}\Big). \tag{3}$$

e

$$S^2 = \frac{1}{N-1}\Big(\sum_{i=1}^{k}\sum_{j=1}^{n_i} R^2(Y_{ij}) - \frac{N(N+1)^2}{4}\Big).$$

Mostre inicialmente que quando não há empates:

$$S^2 = \frac{N(N+1)}{12},$$

e a estatística T se reduz para a equação 5 :

$$T = \frac{N(N+1)}{12}\sum_{i=1}^{k}\frac{R_i^2}{n_i} - 3(N+1).$$

Distribuição Nula

A distribuição exata de T é dada pela tabela $A8$ do Conover para $k = 3$ e $n_i \leq 5$, em geral sua obtenção é muito trabalhosa e é usada uma aproximação para T:

$$T \sim \chi^2(k-1).$$

Rejeitamos H_0 a um nível de significância α se

$$T_{cal} > T_{tab},$$

em que T_{tab} é o percentil de ordem $1-\frac{\alpha}{2}$ da distribuição qui-quadrado com $(k-1)$ graus de liberdade.

Comparações Múltipla Ele nos diz que:

Se e somente se, a hipótese nula é rejeitada podemos usar o seguinte procedimento para determinar quais pares de tratamentos tendem a diferir. Podemos dizer que os tratamentos i e j tendem a diferir se a seguinte inequação é satisfeita:

$$\Big|\frac{R_i}{n_i}\frac{R_j}{n_j}\Big| > t_{1-\alpha/2}\Big(S^2\ \frac{N-T-1}{N-k}\Big)^{1/2}\ \ \Big(\frac{1}{n_i}+\frac{1}{n_j}\Big)^{1/2},$$

em que R_i e R_j são as somas dos postos dos tratamentos i e j, com $i<j$, respectivamente.

Sendo $t_{1-\alpha/2}$ o quantil de ordem $(1-\alpha/2)$ da distribuição t de Student com $(N-k)$ graus de liberdade.

Ele apresenta o seguinte Exemplo: Quatro diferentes métodos de cultivo de milho foram atribuídos aleatoriamente a um grande número de parcelas de terra diferentes e o rendimento por acre foi calculado para cada parcela. A fim de determinar se há uma diferença nos rendimentos como resultado dos métodos usados faça o teste de Kruskal- Wallis.

Método 1	Método 2	Método 3	Método 4
83	91	101	78
91	90	100	82
94	81	91	81
89	83	93	77
89	84	96	79
96	83	95	81
91	88	94	80
92	91		81
90	89		
	84		

Solução: Temos quatro tratamentos que são os métodos de cultivo do milho. O tratamento 1 tem $n_1=9$ observações, o tratamento 2 tem $n_2=10$ observações , o tratamento 3 tem $n_3=7$ observações e o tratamento 4 tem $n_2=8$ observações. No total foram plantados:

$$N=n_1+n_2+n_3+n_4=9+10+7+8=34\ \text{ lotes.}$$

A variável medida foi rendimento do milho por acre.

Só para treinar a notação quanto vale Y_{47}?

A sétima repetição do quarto tratamento vale 80.

As hipóteses a serem testadas :

H_0: Os quatro métodos são equivalentes.

$$H_0 : F_1(y) = F_2(y) = F_3(y) = F_4(y),$$

$F_i(y)$ é a função de distribuição acumulada do método $i, i = 1, 2, 3, 4$.

versus:

H_1 : Alguns métodos de produção de cultivo de milho tendem a fornecer produções maiores que outros métodos.

Vamos resolver pelo software R:

```
>
> ################################################################################
>
>
> k=4
> M1=c(83,91,94,89,89,96,91,92,90)
> n_1=length(M1);n_1
[1] 9
>
> M2=c(91,90,81,83,84,83,88,91,89,84)
>
> n_2=length(M2);n_2
[1] 10
>
> M3=c(101,100,91,93,96,95,94)
>
> n_3=length(M3);n_3
[1] 7
>
> M4=c(78,82,81,77,79,81,80,81)
>
>
>
> n_4=length(M4);n_4
[1] 8
> n=c(n_1,n_2,n_3,n_4);n
[1]  9 10  7  8
> N=sum(n);N
[1] 34
>
```

```
> ####Vamos criar a variável de classificação para definir os tratamentos:
>
> metodo=factor(c(rep(1,n_1),rep(2,n_2),rep(3,n_3),rep(4,n_4)))
> metodo
[1] 1 1 1 1 1 1 1 1 2 2 2 2 2 2 2 2 2 3 3 3 3 3 3 4 4 4 4 4 4 4 4
Levels: 1 2 3 4
>
> Y=c(M1,M2,M3,M4);Y
[1]  83  91  94  89  89  96  91  92  90  91  90  81  83  84  83  88  91  89  84
[20] 101 100  91  93  96  95  94  78  82  81  77  79  81  80  81
> Yo=sort(Y);Yo
[1]  77  78  79  80  81  81  81  81  82  83  83  83  84  84  88  89  89  89  90
[20]  90  91  91  91  91  91  92  93  94  94  95  96  96 100 101
> RY=rank(Y);RY
[1] 11.0 23.0 28.5 17.0 17.0 31.5 23.0 26.0 19.5 23.0 19.5  6.5 11.0 13.5 11.0
[16] 15.0 23.0 17.0 13.5 34.0 33.0 23.0 27.0 31.5 30.0 28.5  2.0  9.0  6.5  1.0
[31]  3.0  6.5  4.0  6.5
>
> table(Y)
Y
77  78  79  80  81  82  83  84  88  89  90  91  92  93  94  95  96 100 101
1   1   1   1   4   1   3   2   1   3   2   5   1   1   2   1   2   1   1
>
> ###Temos os seguintes grupos de empates
>
> ##Grupo 1={81,81,81,81}; Grupo 2={83,83,83}, Grupo 3={84,84}, Grupo 4={89,89,89}
>
> ##Grupo 5={90,90}, Grupo 6={91,91,91,91,91}, Grupo 7={94,94}, Crupo 8={96,96}
>
> g=8 ####Número de grupos de empates
> f_1=3;f_2=3;f_3=2;f_4=3;f_5=2;f_6=5;f_7=2;f_8=2
> fi=c(f_1,f_2,f_3,f_4,f_5,f_6,f_7,f_8);fi
[1] 3 3 2 3 2 5 2 2
>
> Gi=f^3 -f;Gi
[1]  24  24   6  24   6 120   6   6
> SG=sum(Gi);SG
[1] 216
```

```
>
>
> aux=N^3-N;Aux
[1] 39270
>
>
> C= 1-  SG/Aux;C
[1] 0.9944996
>
> Ri=tapply(RY,metodo,sum);Ri
1     2     3     4
196.5 153.0 207.0  38.5
>
>
> ###Vamos calcular S^2
> SR2=sum(RY^2);SR2
[1] 13664
> aux1=(N*(N+1)^2)/4;aux1
[1] 10412.5
> S2=(1/(N-1))*( SR2-aux1);S2
[1] 98.5303
>
>
> aux2=sum(Ri^2/n);aux2
[1] 12937.72
>
> ####O valor de T usando a equação com empates
> Temp_cal=(aux2-aux1)/S2;Temp_cal
[1] 25.62884
>
> ######O valor de T usando a equação sem empates
>
> aux3=12/(N*(N+1));aux3
[1] 0.01008403
> T_cal=aux3*aux2-3*(N+1);T_cal
[1] 25.46437
>
> ######Percebe-se que a diferença não é tanta!!!!!!
```

```
>
> alfa=0.05
>
> T_tab=qchisq(1-alfa,k-1);T_tab
[1] 7.814728
>
> T_cal >T_tab  ##### rejeitar H_0.
[1] TRUE
>
>
> ######Vamos descobrir as diferenças:
>
>
> to=qt(0.975,N-k);to
[1] 2.042272
> Q=(S2*(N-1-Temp_cal))/(N-k);Q
[1] 24.20943
>
> media=tapply(RY,metodo,mean);
> media
1        2        3        4
21.83333 15.30000 29.57143  4.81250
> dmu1_2_est=abs(media[2]-media[1]);dmu1_2_est
2
6.533333
> dmu1_3_est=abs(media[3]-media[1]);dmu1_3_est
3
7.738095
> dmu1_4_est=abs(media[4]-media[1]);dmu1_4_est
4
17.02083
> dmu2_3_est=abs(media[3]-media[2]);dmu2_3_est
3
14.27143
> dmu2_4_est=abs(media[4]-media[2]);dmu2_4_est
4
10.4875
> dmu3_4_est=abs(media[4]-media[3]);dmu3_4_est
```

```
4
24.75893
>
> a12=to*sqrt(S2)*( 1/n[1]+1/n[2]);a12
[1] 4.279664
> a13=to*sqrt(S2)*( 1/n[1]+1/n[3]);a13
[1] 5.148468
> a14=to*sqrt(S2)*( 1/n[1]+1/n[4]);a14
[1] 4.786466
> a23=to*sqrt(S2)*( 1/n[2]+1/n[3]);a23
[1] 4.923223
> a24=to*sqrt(S2)*( 1/n[2]+1/n[4]);a24
[1] 4.561221
> a34=to*sqrt(S2)*( 1/n[3]+1/n[4]);a34
[1] 5.430025
>
>
> a=c(a12,a13,a14,a23,a24,a34)
> Dmu_est=c(dmu1_2_est,dmu1_3_est,dmu1_4_est,dmu2_3_est,dmu2_4_est,dmu3_4_est)
>
> IC1= Dmu_est[1]+ c(-1,1)*a[1];IC1
[1]  2.253669 10.812997
> IC2= Dmu_est[2]+ c(-1,1)*a[2];IC2
[1]  2.589627 12.886563
> IC3= Dmu_est[3]+ c(-1,1)*a[3];IC3
[1] 12.23437 21.80730
> IC4= Dmu_est[4]+ c(-1,1)*a[4];IC4
[1]  9.348206 19.194651
> IC5= Dmu_est[5]+ c(-1,1)*a[5];IC5
[1]  5.926279 15.048721
> IC6= Dmu_est[6]+ c(-1,1)*a[6];IC6
[1] 19.32890 30.18895
> LI=c(IC1[1],IC2[1],IC3[1],IC4[1],IC5[1],IC6[1])
> LS=c(IC1[2],IC2[2],IC3[2],IC4[2],IC5[2],IC6[2])
>
>
> tab=cbind(Dmu_est,a,LI,LS);round(tab,3)
Dmu_est     a     LI     LS
```

```
2   6.533 4.280  2.254 10.813
3   7.738 5.148  2.590 12.887
4  17.021 4.786 12.234 21.807
3  14.271 4.923  9.348 19.195
4  10.488 4.561  5.926 15.049
4  24.759 5.430 19.329 30.189
>
> mod2=kruskal.test(Y ~metodo);mod2 #####Leva em conta os empates!!!!!!!

Kruskal-Wallis rank sum test

data:  Y by metodo
Kruskal-Wallis chi-squared = 25.629, df = 3, p-value = 1.141e-05

>
>
>
```

Conover traz ainda uma aplicação do teste de Kruskal-Wallis em tabela de contingência quanto em que as linhas representam categorias ordenadas e as colunas diferentes populações. Assim não deve haver hesitação em aplicar o teste de Kruskal-Wallis a situações com muitos empates.

Vamos descrever este método:

População	1	2	3	...	k	Total da Linha
Categoria 1	O_{11}	O_{12}	O_{13}	...	O_{1k}	t_1
2	O_{21}	O_{22}	O_{23}	...	O_{2k}	t_2
3	O_{31}	O_{32}	O_{33}	...	O_{3k}	t_2
...	...	...	...	...	...	...
c	O_{c1}	O_{c2}	O_{c3}	...	O_{ck}	t_c
Total de colunas	n_1	n_2	n_3	...	n_k	N

O_{ij} é o número de observações da população que pertence à i-ésima categoria.

Vamos explicar os postos. Temos t_1 elementos na categoria 1, Seus postos variam de 1 a t_1. A soma vale

$$\frac{t_1(t_1+1)}{2}.$$

Assim o posto médio da primeira linha é:

$$\bar{R}_1 = \frac{(t_1 + 1)}{2}.$$

Na segunda linha os elementos teriam postos

$$t_1 + 1, t_1 + 2, \ldots, t_1 + t_2$$

cuja soma vale

$$t_2 \times t_1 + \frac{t_2(t_2 + 1)}{2},$$

todo elemento da segunda linha tem posto:

$$\bar{R}_2 = t_1 + \frac{(t_2 + 1)}{2}.$$

Prosseguindo desta maneira todo elemento da categoria c tem posto igual a:

$$\bar{R}_c = \sum_{i=1}^{c-1} t_i + \frac{(t_c + 1)}{2}.$$

a diferença entre este modelo e as tabela s de contingências comuns é que as categorias (9LINHAS) são ordenadas. assim todas as observações da linha 1 são consideradas iguais mas menores que as observações da linha 2 e assim por diante. Para o cálculo da estatística do teste vamos recomendar este procedimento:

A soma dos postos da população j é dado por:

$$R_j = \sum_{i=1}^{c} O_{ij} \times \bar{R}_i,$$

e calcule S^2 através da seguinte expressão:

$$S^2 = \frac{1}{N-1}\Big(\sum_{i=1}^{c} t_i \, \bar{R}_i^2 - -\frac{N(N+1)^2}{4}\Big).$$

e aí aplicamos

$$T = \frac{1}{S^2}\Big(\sum_{i=1}^{k} \frac{R_i^2}{n_i} - \frac{N(N+1)^2}{4}\Big).$$

Vamos apresentar um exemplo: Três instrutores compararam as notas que atribuíram no semestre anterior para ver se alguns deles tendiam a dar notas mais baixas do que outros. A hipótese nula a ser testada é:

H_0 : Os três instrutores avaliam uniformemente entre si, H_1 : alguns instrutores tendem a dar notas mais baixas que os outros.

Os dados obtidos são:

Nota	$I1$	$I2$	$I3$	Total da Linha	Posto Médio
A	4	10	6	20	10,5
B	14	6	7	27	34
C	17	9	8	34	64,5
D	6	7	6	19	91
F	2	6	1	9	105
Total	43	3 8	28	109	

Solução Assim temos tem $k = 3$ populações que são os instrutores e $c = 5$ linhas que são as notas. Sabemos que:

$$A < B < C < D < F.$$

Além disso

$$t_1 = 20\,; t_2 = 27\,;\ t_3 = 34,\ \ ; t_4 = 19\,; t_5 = 9.$$

Além disso

$$N = \sum_{i=}^{5} t_i = 109.$$

O posto médio da Categoria 1 (**A**) é dado por:

$$\bar{R}_1 = \frac{t_1 + 1}{2} = \frac{21}{2} = 10,5.$$

O posto médio da Categoria 2(**B**) é dado por:

$$\bar{R}_2 = t_1 + \frac{t_2 + 1}{2} = 20 + \frac{28}{2} = 34.$$

O posto médio da Categoria 3(**C**) é dado por:

$$\bar{R}_3 = t_1 + t_2 + \frac{t_3 + 1}{2} = 47 + \frac{35}{2} = 64,5.$$

O posto médio da Categoria 4 (**D**) é dado por:

$$\bar{R}_4 = t_1 + t_2 + t_3 + \frac{t_4 + 1}{2} = 81 + \frac{20}{2} = 91.$$

O posto médio da Categoria 5 (**F**) é dado por:

$$\bar{R}_5 = t_1 + t_2 + t_3 + t_4 + \frac{t_5 + 1}{2} = 100 + \frac{10}{2} = 105.$$

Vamos usar o software R:

```
>
> M=matrix(c(4,14,17,6,2,10,6,9,7,6, 6,7,8,6,1),ncol=3);M
[,1] [,2] [,3]
[1,]    4   10    6
[2,]   14    6    7
[3,]   17    9    8
[4,]    6    7    6
[5,]    2    6    1
> rownames(M)=c("A","B","C","D","F")
> colnames(M)=c("I1","I2","I3")
> M
I1 I2 I3
A  4 10  6
B 14  6  7
C 17  9  8
D  6  7  6
F  2  6  1
> TL=apply(M,1,sum);TL
A  B  C  D  F
20 27 34 19  9
> TC=apply(M,2,sum);TC
I1 I2 I3
43 38 28
>
> aux1=cumsum(TL);aux1
A   B   C   D   F
20  47  81 100 109
> aux1=aux1[-5];aux
A   B   C   D
20  47  81 100
>
> aux1=c(0,aux1);aux1
A   B   C   D
0  20  47  81 100
>
> aux2=(TL+1)/2;aux2
A    B    C    D    F
10.5 14.0 17.5 10.0  5.0
```

```
>
> RM=aux1+aux2;RM
A     B     C     D
10.5  34.0  64.5  91.0 105.0
>
> R1=M[,1]%*%RM;R1
[,1]
[1,] 2370.5
> R2=M[,2]%*%RM;R2
[,1]
[1,] 2156.5
> R3=M[,3]%*%RM;R3
[,1]
[1,] 1468
> ni=TC;ni
I1 I2 I3
43 38 28
>
> N=sum(TC);N
[1] 109
>
> aux3=N*(N+1)^2/4;aux3
[1] 329725
> t=TL;t
A  B  C  D  F
20 27 34 19  9
> t*RM^2
A        B        C        D        F
2205.0  31212.0 141448.5 157339.0  99225.0
> aux4=sum(t*RM^2);aux4
[1] 431429.5
> aux5= aux4-aux3;aux5
[1] 101704.5
> S2= aux5/(N-1);S2
[1] 941.7083
>
> Ri=c(R1,R2,R3);Ri
[1] 2370.5 2156.5 1468.0
```

```
> ni
I1 I2 I3
43 38 28
>
> aux6=sum(Ri^2/ni);aux6
[1] 330027.2
>
> T_cal=(1/S2)*(aux6-aux3);T_cal
[1] 0.3209288
>
> alfa=0.05
> T_tab=qchisq(1-alfa,k-1);T_tab
[1] 5.991465
>
> T_cal> T_tab
[1] FALSE
>
> nd=1- pchisq(T_cal,k-1);nd
[1] 0.8517481
>
>
```

Não há evidência de que os instrutores não tenha, desempenho uniforme com relação às notas.

Teoria a distribuição exata de T é encontrada sob a suposição que todas as observações foram obtidas da mesma população. Este método chamado aleatorização, é também usado para obter a estatística do teste de Mann-Whitney. Se estas suposições são válidas temos que contar de quantas maneiras podemos dividir N elementos em k grupos de tamanhos n_i. Isto pode ser feito de :

$$\frac{N!}{\prod_{i=1}^{k} n_i!}.$$

Cada arranjo desse vai nos fornecer um valor para T e cada arranjo tem probabilidade

$$p = \frac{\prod_{i=1}^{k} n_i!}{N!},$$

de ocorrer.

Com estes valores obtidos finalmente é encontrada a distribuição de probabilidade exata de T.

Vamos fazer um exemplo com $n_1 = 2, n_2 = 1, n_3 = 1$. Logo
teremos:

$$N = n_1 = n_2 + n_3 = 2 + 1 = 1 = 4.$$

Os postos aparecem de :

$$\frac{4!}{2!1!1!} = 12,$$

Por exemplo suponha que o tratamento 1 as duas observações foram de postos $(1, 2)$, o tratamento 2 ficou com a posto 3 e o tratamento 3 ficou com a de posto 4.

Assim a soma de postos de cada tratamento é dada por:

$$R_1 = 1 + 2 = 3, \quad R_2 = 3, \quad R_3 = 4.$$

Portanto

$$\frac{R_i^2 n_i}{=} \frac{9}{2} + \frac{9}{1} + \frac{16}{1} = 29,5.$$

Mas

$$\frac{12}{N(N+1)} = \frac{12}{20} = 0,6.$$

$$T = 0,6 * 29,5 - 3 * (4 + 1) = 17,7 - 15 = 2,7.$$

Procedendo deste jeito montamos o seguinte quadro:

Amostra	Trat 1	Trat 2	Trat 3	T
1	(1,2)	3	4	2,7
2	(1,2)	4	3	2,7
3	(1,3)	2	4	1,8
4	(1,3)	4	2	1,8
5	(1,4)	2	3	0,3
6	(1,4)	3	2	0,3
7	(2,3)	1	4	2,7
8	(2,3)	4	1	2,7
9	(2,4)	1	3	1,8
10	(2,4)	3	1	1,8
11	(3,4)	1	2	2,7
12	(3,4)	2	1	2,7

A função de probabilidade e a função de distribuição acumulada de T são dadas a seguir:

t	$f(t) = P(T = t)$	$F(t) = P(T \leq t)$
0,3	$\frac{1}{6}$	$\frac{1}{6}$
1,8	$\frac{1}{3}$	$\frac{1}{2}$
2,7	$\frac{1}{2}$	1

A aproximação de T para amostra grandes é baseada no fato:

Note que

$$R_i = \sum_{j=1}^{n_i} R(Y_{ij}), \;\; i = 1, 2, \ldots, k.$$

R_i é a soma de n_i variáveis aleatórias e que para n_i grande podemos usar o o teorema do limite central.

Fato 1

$$E(R_i) = \frac{n_i(N+1)}{2}.$$

Vamos resolver este problema com esta situação: Uma urna contem N bolas numeradas de $1, 2, \ldots, N$. Um amostra aleatória de n bolas é retirada sem reposição da urna. Seja

S= soma das numerações das n bolas retiradas. Calcule a média e a variância de S.

Solução Considere

$$S = \sum_{i=1}^{N} i \, I_A(i),$$

A é a nossa amostra de tamanho n e $I_A(i) = 1$ se a bola com a numeração i pertence à amostra. $I_A(i)$ é uma variável de Bernoulli com

$$p = P\Big(I_A(i) = 1\Big) = \frac{\binom{N-1}{n-1}}{\binom{N}{n}} = \frac{n}{N},$$

que é também sua esperança.

Assim,

$$E(S) = \sum_{i=1}^{N} i \, E(I_A(i)) = \sum_{i=1}^{N} i \, \frac{n}{N} = \frac{n}{N} \sum_{i=1}^{N} i = \frac{n}{N} \frac{N(N+1)}{2},$$

$$E(S) = \frac{n(N+1)}{2}.$$

No caso em questão vamos escolher n_i postos dentre os N postos sem empates nem reposição. Logo,

$$E(R_i) = \frac{n_i(N+1)}{2}.$$

Fato 2

$$V(R_i) = \frac{n_i(N+1)(N-n_i)}{12}.$$

Vamos agora calcular a variância de S:

$$V(S) = Cov(S,S) = Cov\Big(\sum_{i=1}^{N} i\ I_A(i), \sum_{j=1}^{N} j\ I_A(j)\Big).$$

Usando propriedades de covariância temos:

$$V(S) = \sum_{i=1}^{N}\sum_{j=1}^{N} ij\ \ Cov(I_A(i), I_A(j)).$$

$$V(S) = \sum_{i=1}^{N} i^2\ Cov(I_A(i), I_A(i)) + \sum_{i=1}^{N}\sum_{j=1, j\neq i}^{N} ij\ \ Cov(I_A(i), I_A(j))$$

Seja

$$S_1 = \sum_{i=1}^{N} i^2\ Cov(I_A(i), I_A(i)) = \sum_{i=1}^{N} i^2\ Var(I_A(i)),$$

Como

$$Var(I_A(i)) = p(1-p) = \frac{n}{N}\Big(1 - \frac{n}{N}\Big) = \frac{n(N-n)}{N^2},$$

logo,

$$S_1 = \sum_{i=1}^{N} i^2\ \frac{n(N-n)}{N^2} = \frac{n(N-n)}{N^2}\ \sum_{i=1}^{N} i^2,$$

$$S_1 = \frac{n(N-n)}{N^2}\ \frac{N(N+1)(2N+1)}{6},$$

$$S_1 = \frac{n\ (N+1)(2N+1)(N-n)}{6N}.$$

Vamos calcular

$$E\Big(I_A(i), I_A(j)\Big) = P(I_A(i) \times I_A(j)) = P\Big(I_A(1) = 1, I_A(j) = 1\Big) = p_{ij},$$

esta probabilidade é a probabilidade de que os elementos i e j pertencem 'a amostra.

$$p_{ij} = \frac{\binom{N-2}{n-2}}{\binom{N}{n}} = \frac{n(n-1)}{N(N-1)},$$

A covariância então

$$Cov(I_A(i), I_A(j)) = E\Big(I_A(i), I_A(j)\Big) - E\Big(I_A(i)\Big) \times E\Big(I_A(j)\Big).$$

$$Cov(I_A(i), I_A(j)) = \frac{n(n-1)}{N(N-1)} - \frac{n^2}{N^2},$$

$$Cov(I_A(i), I_A(j)) = \frac{n}{N}\Big[\frac{(n-1)}{(N-1)} - \frac{n}{N},\Big]$$

$$Cov(I_A(i), I_A(j)) = \frac{n}{N} \times \frac{(nN - N - nN + n)}{(N-1)N}$$

$$Cov(I_A(i), I_A(j)) = -\frac{n(N-n)}{(N-1)N^2}.$$

Considere

$$S_2 = \sum_{i=1}^{N} \sum_{j=1, j\neq i}^{N} ij \;\; Cov(I_A(i), I_A(j)) = -\frac{n(N-n)}{(N-1)N^2} \sum_{i=1}^{N} \sum_{j=1, j\neq i}^{N} ij,$$

$$S_2 = -\frac{n(N-n)}{(N-1)N^2} \sum_{i=1}^{N} i \sum_{j=1, j\neq i}^{N} j,$$

Porém

$$\sum_{j=1, j\neq i}^{N} j = \sum_{j=1}^{N} j - i = \frac{N(N+1)}{2} - i,$$

$$\sum_{i=1}^{N} i\Big(\frac{N(N+1)}{2} - i\Big) = \frac{N(N+1)}{2} \sum_{i=1}^{N} i - \sum_{i=1}^{N} i^2,$$

Mas,

$$= \frac{N(N+1)}{2} \sum_{i=1}^{N} i - \sum_{i=1}^{N} i^2 =$$

$$\frac{N(N+1)}{2} \times \frac{N(N+1)}{2} - \frac{N(N+1)(2N+1)}{6} = .$$

$$= \frac{N^2(N+1)^2}{4} \times -\frac{N(N+1)(2N+1)}{6}$$

$$\frac{3N^2(N+1)^2 - 2N(N+1)(2N+1)}{12} = \frac{N(N+1)\Big[3N^2 + 3N - 4N - 2\Big]}{12}$$

$$\frac{N(N+1)\Big[3N^2-N-2\Big]}{12}=\frac{N(N+1)(N-1)(3N+2)}{12}.$$

Logo

$$S_2=-\frac{n(N-n)}{(N-1)N^2}\ \ \frac{N(N+1)(N-1)(3N+2)}{12}=$$

$$S_2=-\frac{n(N+1)(N-n)(3N+2)}{12N}.$$

$$V(S)=S_1+S_2=\frac{n\ (N+1)(2N+1)(N-n)}{6N}-\frac{n(N+1)(N-n)(3N+2)}{12N}=.$$

$$V(S)=\frac{n(N+1)(N-n)}{12N}\ \ \Big[4N+2-3N-2\Big]=\frac{n(N+1)(N-n)}{12}.$$

O que prova a variância de R_i.

Vamos usar o teorema do limite central

$$Z_i=\frac{R_i-E(R_i)}{\sqrt{Var(R_i)}}\sim\ N(0,1).$$

e

$$Z_i^2=\frac{[R_i-\frac{ni(N+1)}{2}]^2}{\frac{n_i(N+1)(N-n)}{12}}\sim\chi^2(1).$$

Se os R_i são independentes a distribuição da soma

$$T_1=\sum_{i=1}^{k}Z_i^2,$$

pode ser aproximada por uma distribuição de qui-quadrado com k graus de liberdade. No entanto

$$\sum R_i=\frac{N(N+1)}{2},$$

mostra uma dependência entre eles. Por isso a perda de um grau de liberdade.

Para melhorar a convergência Kruskal em 1952 multiplicou T_1 pelo fator $(N-n_i)/N$

chegando ao resultado:

$$T=\sum_{i=1}^{k}\frac{[R_i-\frac{ni(N+1)}{2}]^2}{\frac{n_iN(N+1)}{12}}\sim\chi^2(k-1).$$

Segue daí a explicação do uso da qui-quadrado no teste de Kruskal-Wallis.

0.1 Exercícios e Problemas

1. (Conover:Exerc. 1-pg297) Amostras aleatórias de cada um das três marcas diferentes de lâmpadas (A,B,C) foram testadas para ver quanto tempo as lâmpadas duravam, com os seguintes resultados:

A	B	C
73	84	82
64	80	79
67	81	71
62	77	75
70		

   Esses resultados indicam uma diferença significativa entre as marcas? Se sim, quais marcas parecem ser diferentes?

2. (Conover:Exerc. 1-pg297) Quatro programas de treinamento de empregos foram experimentados em 20 novos funcionários, onde 5 funcionários foram designados aleatoriamente para cada programa de treinamento. Os 20 funcionários foram então colocados sob o mesmo supervisor e, ao final de um determinado período especificado, o supervisor classificou os funcionários de acordo com a capacidade para o trabalho, com as categorias mais baixas sendo atribuídas aos funcionários com a capacidade de trabalho

Programa	Posto
1	4,6,7,2,10
2	1,8,12,3,11
3	20,19,16,14,5
4	5 18,15,17,13,9

   Esses resultados indicam uma diferença significativa na eficácia dos quatro programas de treinamento? Em caso afirmativo quais deles?

3. (Conover:Exerc. 3-pg 298) A intensidade da avaria no solo de uma fazenda causada pela água e pelo vento foi avaliada em diversas fazendas. Ao mesmo tempo foi registrado o tipo de agricultura praticado em cada local. Os resultados obtidos foram:

	Tipo da Fazenda			
	Cultivo Mínimo	Contour	Terrace	Outro
Perda	Número de Fazendas			
Zero	17	19	4	21
Pequena	3	10	4	42
Moderada	0	2	2	34
Severa	0	0	2	6

4. (Conover:Exerc. 4 -pg 298) Três tipos diferentes de rádios, fabricados pela empresa são, todos têm garantia de 1 ano. Um registro é mantido de quantos rádios precisaram ser substituídos, foram reparados ou não foram devolvidos pela garantia.

	A	B	C
Trocados	12	3	6
Reparados	10	8	7
Não Devolvidos	82	96	58

   Foi observada uma diferença significativa entre as confiabilidades dos diferentes tipos de rádio? Em caso afirmativo, quais parecem ser diferentes?

5. (Conover:Exerc. 5 -pg 299) A quantidade de ferro presente no fígado de ratos brancos é medida depois que os animais foram alimentados com uma das cinco dietas por um período de tempo pre-estabelecido. Existem 50 rato brancos semelhantes. 10 animais foram distribuídos aleatoriamente a cada uma das cinco dietas.

Dieta A	Dieta B	Dieta C	Dieta D	Dieta E
2,23	5,59	4,50	1,35	1,40
1,14	0,96	3,92	1,06	1,51
2,63	6,96	10,33	0,74	2,49
1,00	1,23	8,23	0,96	1,74
1,35	1,61	2,07	1,16	1,59
2,01	2,94	4,90	2,08	1,36
1,64	1,96	6,84	0,69	3,00
1,13	3,68	6,42	0,68	4,81
1,01	1,54	3,72	0,84	5,21
1,70	2,59	6,00	1,34	5,12

   As diferentes dietas parecem afetar a quantidade de ferro presente no fígado?

6. (Conover:Exerc. 6 -pg 299) Doze voluntários foram designados para cada um dos três planos de redução de peso. A atribuição dos voluntários aos planos foi aleatória e presumiu-se que os 36 voluntários ao todo se pareceriam com uma amostra aleatória de pessoas que poderiam tentar um programa de redução de peso. Teste a hipótese nula de que não há diferença nas distribuições de probabilidade da quantidade de peso perdida durante os três programas em comparação com a alternativa de que há uma diferença. Os resultados são apresentados como o número de libras perdidas por cada pessoa.

Plano A		Plano B		Plano C	
2	17	17	5	29	5
12	4	15	6	3	25
5	25	3	19	25	32
4	;6	19	4	28	24
26	21	5	9	11	36
8	6	14	7	7	20

7. (Conover:Problema 1 -pg 299) Mostre que as equações 3 e 5 são equivalentes na ausência de empates.

   Seja a equação 3

$$T = \frac{1}{S^2}\Big(\sum_{i=1}^{k}\frac{R_i^2}{n_i} - \frac{N(N+1)^2}{4}\Big),$$

   e

$$S^2 = \frac{1}{N-1}\Big(\sum_{i=1}^{k}\sum_{j=1}^{n_i} R(Y_{ij}) - \frac{N(N+1)^2}{4}\Big).$$

   Mostre inicialmente que quando não há empates:

$$S^2 = \frac{N(N+1)}{12},$$

   e a estatística T se reduz para a equação 5 :

$$T = \frac{N(N+1)}{12}\sum_{i=1}^{k}\frac{R_i^2}{n_i} - 3(N+1).$$

8. (Conover:Problema 2 -pg 299) Ache a distribuição exata da estatística do Teste de Kruskal-Wallis quando H_0 é verdade e $n_1 = 3$,$n_2 = 2$, $n_3 = 1$ e quando não há empates. Compare seus resultados com os quantis dados pela tabela $A8$ do livro do Conover.

9. (Conover:Problema 3 -pg 300) No caso de duas populações independentes diga as razões que você prefere usat o teste de Mann-Whitney no lugar do teste de Kruskal-Wallis?

10. (Conover:Problema 4 -pg 300) Mostre que as equações 10 e 4 são equivalentes. a equação 4 é:

$$S^2 = \frac{1}{N-1}\Big(\sum_{i=1}^{k}\sum_{j=1}^{n_i} R(Y_{ij}) - \frac{N(N+1)^2}{4}\Big).$$

A equação 10 é:

$$S^2 = \frac{1}{N-1}\Big(\sum_{i=1}^{c} t_i\ \bar{R}_i^2 - -\frac{N(N+1)^2}{4}\Big).$$

11. (Conover:Problema 5 -pg 300) suponha a estatística F é calculada usando os postos no lugar das observações originais. Mostre que

$$F = \frac{\frac{T}{N-1}}{\frac{N-1-T}{N-k}},$$

se verifica enter F e T dada pela equação 3. Portanto o teste que rejeita H_0 para valores grandes de T é equivalente ao teste que rejeita H_0 para valores grandes de F, se F é calculada com os postos.