Introdução

Vamos apresentar agora o Teste de Kruskal-Wallis do livro de James Higgins , "An Introduction to Modern Nonparametric Statistics, nas páginas 86 a 96.

1 Teste de Kruskal-Wallis

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

Seja R_{ij} o respectivo posto de Y_{ij} considerando todas as N observações.
Ele supõe inicialmente que não há empates. Portanto

$$\sum_{i=}^{k} \sum_{j=1}^{n_i} R_{ij} = \sum_{l=1}^{N} l = \frac{N(N+1)}{2}.$$

A média geral dos postos é dada por:

$$\bar{R} = \frac{N+1}{2}\ .$$

Para ajudar na análise ele apresente a seguinte tabela:

Tratamentos	Postos	Tamanho Amostral	Posto Médio
...	...	...	
1	$R_{11}, R_{12}, \ldots, R_{1n_1}$	n_1	$\bar{R}_1$
2	$R_{21}, R_{22}, \ldots, R_{2n_2}$	n_2	$\bar{R}_2$
...	...	...	
k	$R_{k1}, R_{k2}, \ldots, R_{kn_k}$	n_1	$\bar{R}_k$

A estatística do teste é:

$$KW = \frac{1}{12N(N+1)} \sum_{i=}^{k} n_i \left(\bar{R}_i - \frac{N+1}{2} \right)^2.$$

O termo

$$\sum_{i=}^{k} n_i \left(\bar{R}_i - \frac{N+1}{2} \right)^2,$$

é a soma de quadrados dos tratamentos aplicada aos postos. O fator

$$C = \frac{1}{12N(N+1)}$$

é um fator escala que nos permite usar a distribuição qui-quadrado com $k-1$ graus de liberdade para aproximar a distribuição da estatística envolvendo as $N!$ permutações.

Vamos mostrar agora que esta estatística é a mesma apresentada por Humberto de Campos:

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

em que R_i é a soma dos postos atribuídos ao tratamento i.

Prova
Seja

$$A = \sum_{i=1}^{k} n_i \left(\bar{R}_i - \frac{N+1}{2} \right)^2,$$

Desenvolvendo o quadrado temos:

$$A = \sum_{i=1}^{k} n_i \left(\bar{R}_i^2 - 2\,\frac{N+1}{2}\bar{R}_i + \frac{(N+1)^2}{4} \right)^2,$$

logo,

$$A = \sum_{i=1}^{k} \left(n_i \bar{R}_i^2 - (N+1) n_i\, \bar{R}_i + \frac{(N+1)^2}{4} \right)^2,$$

Separando os somatórios temos:

$$A = \sum_{i=1}^{k} n_i \bar{R}_i^2 - \sum_{i=1}^{k} (N+1) n_i \bar{R}_i + \sum_{i=1}^{k} \frac{(N+1)^2}{4} n_i,$$

Mas,

$$\sum_{i=1}^{k} (N+1) n_i \bar{R}_i = (N+1) \sum_{i=1}^{k} n_i \bar{R}_i = (N+1) \sum_{i=1}^{k} n_i \frac{\sum_{j=1}^{n_i} R_{ij}}{n_i},$$

$$\sum_{i=1}^{k} (N+1) n_i \bar{R}_i = (N+1) \sum_{i=1}^{k} \sum_{j=1}^{n_i} R_{ij} = (N+1) \frac{N(N+1)}{2} = \frac{N(N+1)^2}{2}.$$

Por outro lado,

$$\sum_{i=1}^{k} \frac{(N+1)^2}{4} n_i = \frac{(N+1)^2}{4} \sum_{i=1}^{k} n_i = \frac{(N+1)^2}{4} \times N = \frac{N(N+1)^2}{4}$$

Somando estas duas parcelas temos:

$$B = -\frac{N(N+1)^2}{2} + \frac{N(N+1)^2}{4} = -\frac{N(N+1)^2}{4} = -\frac{3N(N+1)^2}{12}.$$

Multiplicando por:

$$\frac{12}{N(N+1)}$$

temos

$$\frac{12}{N(N+1)}\, B = -3(N+1).$$

Vamos analisar o primeiro termo:

$$\sum_{i=1}^{k} n_i\, \bar{R}_i^2 = \sum_{i=1}^{k} n_i \frac{R_i^2}{n_i^2} = \sum_{i=1}^{k} \frac{R_i^2}{n_i}.$$

Logo

$$KW = \frac{12}{N(N+1)} \left[\sum_{i=1}^{k} \frac{R_i^2}{n_i} \right] - 3(N+1) = H.$$

Ele também apresenta o modelo de análise de variância com um fator de classificação

$$Y_{ij} = \mu_i + \epsilon_{ij},$$

em que

$$\epsilon_{ij} \sim N(0, \sigma^2),$$

Elas são independentes e identicamente distribuídas para $i = 1, 2, \ldots, k$ e $j = 1, 2, \ldots, n_i$.
Assim

$$Y_{ij} \sim N(\mu_i, \sigma^2).$$

Note que: Para $i = 1, 2, \ldots, k$

$$\bar{Y_i} = \frac{\sum_{j=1}^{n_i} Y_{ij}}{n_i} \sim N\Big(\mu_i, \frac{\sigma^2}{n_i}\Big).$$

Elas são independentes entre si e independentes de S_i^2
Por outo lado temos

$$V_i = S_i^2 = \frac{(n_i - 1)S_i^2}{\sigma^2} \sim \chi^2(n_i - 1).$$

As variáveis são independentes.
Um quadro resume bem a situação:

Tratamentos	Observações	Tamanhos Amostrais	Médias	Variâncias
1	$Y_{11}, Y_{12}, \ldots, Y_{1n_1}$	n_1	$\bar{Y}_1$	S_1^2
2	$Y_{21}, Y_{22}, \ldots, Y_{2n_2}$	n_2	$\bar{Y}_2$	S_2^2
...	...	...	...	
k	$Y_{k1}, Y_{k2}, \ldots, Y_{kn_k}$	n_k	$\bar{Y}_k$	S_1^k

O nosso objetivo é desenvolver um teste para:

$$H_0 : \mu_1 = \mu_2 = \ldots = \mu_k = \mu.$$

Se H_0 é verdade vamos estimar μ por

$$\bar{Y} = \frac{\sum_{i=1}^{k}\sum_{j=1}^{n_i} Y_{ij}}{N}.$$

A expressão

$$SQT = \sum_{i=1}^{k}\sum_{j=1}^{n_i} \Big(Y_{ij} - \bar{Y})^2,$$

ela representa o erro quadrático cometido ao estimarmos todas as observações pela média geral. Mas se realmente há diferenças entre os tratamentos as médias dos tratamentos variam bastante de um para outro tratamento. Vamos analisar este efeito:

$$SQT = \sum_{i=1}^{k}\sum_{j=1}^{n_i} \Big(Y_{ij} - \bar{Y})^2,$$

Vamos começar a brincadeira:

Vou somar e subtrair $\bar{Y}_i$

$$SQT = \sum_{i=1}^{k}\sum_{j=1}^{n_i} \Big(Y_{ij} - \bar{Y}_i + \bar{Y}_i - \bar{Y})^2,$$

Vamos elevar ao quadrado:

$$SQT = \sum_{i=1}^{k}\sum_{j=1}^{n_i} \big(Y_{ij} - \bar{Y}_i\big)^2 + \sum_{i=1}^{k}\sum_{j=1}^{n_i} \big(\bar{Y}_i - \bar{Y}\big)^2 + 2\sum_{i=1}^{k}\sum_{j=1}^{n_i} \big(\bar{Y}_i - \bar{Y}\big)\big(Y_{ij} - \bar{Y}_i\big) \ ,$$

Mas,

$$\sum_{i=1}^{k}\sum_{j=1}^{n_i} \big(\bar{Y}_i - \bar{Y}\big)\big(Y_{ij} - \bar{Y}_i\big) = \sum_{i=1}^{k} \big(\bar{Y}_i - \bar{Y}\big) \sum_{j=1}^{n_i} \big(Y_{ij} - \bar{Y}_i\big) = 0,$$

pois,

$$\sum_{j=1}^{n_i} \big(Y_{ij} - \bar{Y}_i\big) = \sum_{j=1}^{n_i} Y_{ij} - \sum_{j=1}^{n_i} \bar{Y}_i = \sum_{j=1}^{n_i} Y_{ij} - n_i\bar{Y}_i = n_i\bar{Y}_i - n_i\bar{Y}_i = 0.$$

Além disso:

$$\sum_{i=1}^{k}\sum_{j=1}^{n_i} \big(\bar{Y}_i - \bar{Y}\big)^2 = \sum_{i=1}^{k} \ n_i\big(\bar{Y}_i - \bar{Y}\big)^2.$$

A soma de quadrados total se torna:

$$\sum_{i=1}^{k}\sum_{j=1}^{n_i} \Big(Y_{ij} - \bar{Y})^2 = \sum_{i=1}^{k} \ n_i\big(\bar{Y}_i - \bar{Y}\big)^2 + \sum_{i=1}^{k}\sum_{j=1}^{n_i} \big(Y_{ij} - \bar{Y}_i\big)^2.$$

Logo

$$SQTotal = SQTrat + SQRes.$$

$$SQTrat = \sum_{i=1}^{k} \ n_i\big(\bar{Y}_i - \bar{Y}\big)^2,$$

se o experimento for balanceado $n_i = r, \ \ i = 1, 2, \ldots, k$

$$SQTrat = r \times \sum_{i=1}^{k} \ \big(\bar{Y}_i - \bar{Y}\big)^2,$$

A soma de quados residual é dada por:

$$\sum_{i=1}^{k}\sum_{j=1}^{n_i} \big(Y_{ij} - \bar{Y}_i\big)^2 = \sum_{i=1}^{k} (n_i - 1)S_i^2.$$

Analisando SQTrat vemos que ela leva as diferenças entre as médias dos tratamentos e a média geral. Percebemos claramente isto na fórmula. Mas mesmo a estimativa de $\bar{Y}_i$ sendo mesmo que a média geral ainda tem uma parte desta variação que fica não explicada pelo nosso modelo. Esta é a soma de quadrados residual.

A distribuição de

$$W = \frac{\sum_{i=1}^{k}\sum_{j=1}^{n_i}\left(Y_{ij}-\mu\right)^2}{\sigma^2} = \sum_{i=1}^{k}\sum_{j=1}^{n_i}\frac{\left(Y_{ij}-\mu\right)^2}{\sigma^2} = \sum_{i=1}^{k}\sum_{j=1}^{n_i} Z_{ij}^2 \sim \chi^2(N).$$

Na disciplina de Planejamentos usa-se $N = n$.

A quantidade

$$QMTrat = \frac{SQTrat}{k-1},$$

é chamada de quadrado médio de tratamentos. Se H_0 é verdade

$$V_1 = \frac{SQTrat}{\sigma^2} \sim \chi^2(k-1).$$

A quantidade

$$QMRes = \frac{SQRes}{N-k} =,$$

é chamada de quadrado médio dos resíduos

$$V_2 = \frac{SQRes}{\sigma^2} \sim \chi^2(N-k).$$

Além disso

$$F = \frac{\frac{V_1}{k-1}}{\frac{V_2}{N-k}} = \frac{QMTrat}{QMRes} \sim F(k-1, N-k).$$

Rejeitar H_0 a um nível de significância α se

$$F_{cal} = \frac{QMTrat}{QMRes} > F_tab,$$

em que

$$P(F(k-1, N-k)) \geq F_{tab} = \alpha.$$

O nível descritivo será

$$nd = P\Big(F(k-1, N-k) \geq F_{cal}\Big).$$

Costuma-se apresentar este teste em um quadro chamado de Quadro de Análise de Variância que será mostrado a seguir:

Fonte de Variação	gl	SQ	QM	F
Entre(Tratamentos)	k-1	SQTrat	QMTrat	
Dentro(Residual)	N-k	SQRes	$QMRes$	
Total	SQTotal	N-1	QMT	

Se rejeitarmos H_0 precisamos detectar que tratamentos são responsáveis por isso.

Isto é feito através de contrastes colocados em forma matricial:

$$H_0 : C\beta = M,$$

em que $\beta = (\mu_1, \mu_2, \ldots, \mu_k)^t$ e M é um vetor de constantes. Suponha ainda que a característica ou posto de C é p. Seja $s^2 = QMRes$.

A estatística do teste será:

$$Q = \frac{(Cb - C\beta)'[C(X'X)^{-1}C']^{-1}(Cb - C\beta)}{ps^2},$$

onde p é o posto da matriz C e s^2 é o quadrado médio residual e b é o vetor de médias amostrais dos k tratamentos. Se H_0 é verdade,

$$Q = \frac{(Cb - M)'[C(X'X)^{-1}C']^{-1}(Cb - M)}{ps^2} \sim F(p, glres).$$

Se todas as observações são selecionadas aleatoriamente de populações normais independentes com variâncias iguais, então a estatística F tem uma distribuição de Fisher-Snedecor com $(k-1)$ graus de liberdade no numerador e $(N-k)$ graus de liberdade no denominador. O nível descritivo é calculado como seu auxílio. No entanto se não tivermos populações normais então é preferível usar um teste de Permutação que agora será descrito:

Exemplo 1: Três conservantes e um controle foram comparados em termos de sua capacidade de inibir o crescimento de bactérias. As amostras são tratadas com um dos três conservantes ou deixadas sem tratamento para o controle. As contagens de bactérias são feitas 48 horas após a aplicação. Os dados apresentados são os logaritmos dessa contagens. A análise não paramétrica em termos de postos pode ser feita com as contagens reais ou as trans formadas.

A contagem real de bactérias não satisfaria as premissas da análise de variância. As variâncias seriam desiguais entre os tratamentos. O objetivo de fazer a transformação logarítmica é obter dados que atendam às suposições de análise de variância padrão.

Controle	4,302	4,017	4,049	4,176		
Conservante 1	2,201	3,190	3,250	3,276	3,292	3,267
Conservante 2	3,397	3,552	3,630	3,578	3,612	
Conservante 3	2,699	2,929	2,785	2,176	2,845	2,913

Soluçâo: Vamos fazer pelo R.
Considere os tratamentos:

$$t_1 = Controle;\ \ t_2 = Conservante\ 1\ \ t_3 = Conservante\ 2;\ \ t_4 = Conservante 3.$$

Vamos fazer o teste de Análise de variância para testar

$$\mu_1 = \mu_2 = \mu_3 = \mu_4.$$

Passos para elaborar um teste F de permutação:

1. Obtenha para os dados originais o valor da estatística F e chame-a de F_{obs}.
2. Obtenha todas as permutações possíveis das N observações entre os k tratamentos com $n_i,\ i = 1, 2, \ldots, k$ no tratamento i. Existem

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

de tais possibilidades. Se não for possível gerar todas as permutações, então selecione um amostra aleatória de R permutações.

3. Para cada permutação, calcule o valor da estatística F.

4. Obtenha o nível descritivo da seguinte maneira:

$$\hat{\alpha} = nd = \frac{A}{B},$$

em que A, número de permutações em F maiores ou iguais a F_{obs}.

O nível descritivo é aproximado se ele for baseado na amostra aleatória. Perceba que sempre será um teste unilateral à direita.

Segundo, Higgins, o número de permutações é grande mesmo para tamanhos amostrais modestos Por exemplo se há $k = 3$ tratamentos e com cinco repetições para cada um deles teremos:

$$B = 756756 \quad \text{possibilidades,}$$

```
> n_1=5;n_2=5;n_3=5
>
> N=n_1+n_2+n_3;N
[1] 15
>
>
> B=factorial(15)/(factorial(5)*factorial(5)*factorial(5));B
[1] 756756
```

Precisamos descrever um procedimento para agilizar o nosso teste aproximado:

Ele apresenta um exemplo que vamos descrever agora: Ele quer comparar o valor do nível descritivo do teste padrão da **ANOVA** com o obtido pelo teste de permutação. Os dados vem de três populações normais independentes com médias 15,25 e 30 , respectivamente,e mesmo desvio padrão 9.

Tratamento 1	6,08	22,29	7,51	34,36	23,68
Tratamento 2	30,45	22,71	44,52	31,47	36,81
Tratamento 3	32,04	28,03	32,74	23,84	29,64

Vamos fazer pelo software R:

```
> T1=c(608,2229,751,3436,2368)/100
> T2=c(3045,2271,4452,3147,3681)/100
> T3=c(3204,2803,3274,2384,2964)/100
> n_1=5;n_2=5;n_3=5;N=15
>
> rep(1,n_1)
[1] 1 1 1 1 1
> Pop=factor(c(rep(1,n_1),rep(2,n_2),rep(3,n_3)))
> Y=c(T1,T2,T3);Y
[1]  6.08 22.29  7.51 34.36 23.68 30.45 22.71 44.52 31.47 36.81 32.04 28.03
[13] 32.74 23.84 29.64
```

```
> Pop
[1] 1 1 1 1 1 2 2 2 2 2 3 3 3 3 3
Levels: 1 2 3
> mod1=aov(Y~Pop); summary(mod1)
Df Sum Sq Mean Sq F value Pr(>F)
Pop          2  554.6  277.31   3.781 0.0533 .
Residuals   12  880.0   73.33
---
Signif. codes:  0 ?***? 0.001 ?**? 0.01 ?*? 0.05 ?.? 0.1 ? ? 1
>
```

O valor da estatística $F(2,12)$ observado foi $F_{obs} = 3,78$ O nível descritivo foi

$$nd = 0,0533$$

en quanto o valor exato obtido por Higgins foi $0,0513$.

Comentário Sobre a Relação entre os testes-F da Anova e das Permutações: Em vários estudos científicos , as unidades experimentais não são retirados aleatoriamente de uma população maior, mas, em vez disso, são unidades que estão disponíveis no momento do estudo. Para evitar viés, as unidades são designadas aleatoriamente para serem tratamentos. Por causa da aleatorização, o teste F das permutações pode ser usado para testar diferenças entre os tratamentos, embora as inferências se aplicariam apenas a essas unidades do estudo. No entanto, pesquisadores, muitas vezes aplicam a ANOVA na situação, como se as observações fossem selecionadas aleatoriamente a partir de populações normalmente distribuídas. Vamos apresentar a justificativa para fazer isso.

A estreita concordância entre os valores p no exemplo anterior sugere que a distribuição da estatística F do permutação pode ser aproximada pela distribuição F usual com os graus de liberdade apropriados. Para ilustrar isso, obtivemos 10000 valores da estatística F de permutações selecionadas aleatoriamente dos dados do exemplo. A próxima tabela compara os percentis da distribuição F obtida com aqueles da distribuição $F(2,12)$. Temos uma boa concordância exceto nas caudas da distribuição.

Percentil	80	85	90	95	97,5	99
Permutação	1,8	2,2	2,8	3,8	4,8	5,9
Distribuição F	1,8	2,2	2,8	3,9	5,1	6,9

Exemplo 2 Pimentel Gomes (1978) apresenta um exemplo de um experimento(fictício) de alimentação de porcos, em que se usaram quatro rações (A, B, C, D), cada uma fornecida a cinco animais, escolhidos acaso. Os aumentos de peso observados, em kg,foram:

A	B	C	D
35	40	39	27
19	35	27	12
31	46	20	13
15	41	29	28
30	33	45	30

Houve diferença, no comportamento das quatro rações, quanto ao ganho de peso dos animais?

Solução Temos 4 populações:
Y_1 ganho de peso usando a ração A.
Y_2 ganho de peso usando a ração B.
Y_3 ganho de peso usando a ração C.

Y_4 ganho de peso usando a ração D.

Suposições:As variáveis são independentes normais com a mesma variância.

$$Y_1 \sim N(\mu,\sigma^2)\ ; Y_2 \sim N(\mu,\sigma^2)\ Y_3 \sim N(\mu_3,\sigma^2);\ \ Y_4 \sim N(\mu_4,\sigma^2).$$

$$H_0 : \mu_1 = \mu_2 = \mu_3 = \mu_4.$$

Vamos fazer direto no R;

Vamos fazer um exemplo com empates.

Exemplo 2 Pimentel Gomes (1978) apresenta um exemplo de um experimento9fictício) de alimentação de porcos, em que se usaram quatro rações (A, B, C, D), cada uma fornecida a cinco animais, escolhidos acaso. os aumentos de peso observados, em kg,foram:

A	B	C	D
35	40	39	27
19	35	27	12
31	46	20	13
15	41	29	28
30	33	45	30

Houve diferença, no comportamento das quatro rações, quanto ao ganho de peso dos animais?

Solução A nossa variável analisada Y=ganho do peso é uma variável quantitativa. As condições do teste estão satisfeitas:

a. As observações são todas independentes.

b. Dentro de uma dada amostra, todas as observações são provenientes da mesma população.

c. As k populações são aproximadamente da mesma forma e contínuas

Vamos ordenar nossa variável Y:

```
> A=c(35,19,31,15,30);B=c(40,35,46,41,33);C=c(39,27,20,29,45);D=c(27,12,13,28,30)
> Y=c(A,B,C,D);Y
[1] 35 19 31 15 30 40 35 46 41 33 39 27 20 29 45 27 12 13 28 30
> Yo=sort(Y);Yo
[1] 12 13 15 19 20 27 27 28 29 30 30 31 33 35 35 39 40 41 45 46
>
```

Temos 3 grupos de empates: $G_1 27, 27$ com $f_1 = 2$, $G_2 30, 30$ com $f_2 = 2$e
$G_3 35, 35$ com $f_3 = 2$. Vamos dar os postos considerando uma única população:
Vamos agora definir nossa covariável de classificação:

```
> n_1=length(A);n_2=length(B);n_3=length(C);n_4=length(D)
>
> n_1;n_2;n_3;n_4
[1] 5
[1] 5
[1] 5
```

```
[1] 5
>
>
> N=n_1+ n_2+n_3+n_4;N
[1] 20
>
> Trat=factor( c(rep(1,n_1),rep(2,n_2),rep(3,n_3),rep(4,n_4)))
> cbind(Trat,Y,R)
      Trat  Y    R
[1,]    1 35 14.5
[2,]    1 19  4.0
[3,]    1 31 12.0
[4,]    1 15  3.0
[5,]    1 30 10.5
[6,]    2 40 17.0
[7,]    2 35 14.5
[8,]    2 46 20.0
[9,]    2 41 18.0
[10,]   2 33 13.0
[11,]   3 39 16.0
[12,]   3 27  6.5
[13,]   3 20  5.0
[14,]   3 29  9.0
[15,]   3 45 19.0
[16,]   4 27  6.5
[17,]   4 12  1.0
[18,]   4 13  2.0
[19,]   4 28  8.0
[20,]   4 30 10.5
>
```