BIO 682
Nonparametric Statistics
Spring 2010

Steve Shuster
http://www4.nau.edu/shustercourses/BIO682/index.htm

Lecture 9

Kruskal-Wallis: Tied Ranks

2. The corrected value of $H_{adj} = H/D$,

a. This serves to *increase* the value of H and make the result more likely to be significant.

b. Why? Uncorrected scores are unnecessarily conservative.

c. An example of how tied ranks makes it more difficult to distinguish between group medians.

Kruskal-Wallis: Example

1. The numbers of beetles on three colors of flowers

White	Yellow	Purple
96	82	115
128	124	149
83	132	166
61	135	147
101	109	

Example With Tied Ranks

White	Yellow	Purple
→ 96	82	115
128	124	149
83	→ 135	166
61	→ 135	→ 135
→ 96	109	

Note tied ranks

Kruskal-Wallis: Example

2. Rank the scores as a *single series* from lowest to highest.

	White	Yellow	Purple
	4	2	7
	9	8	13
	3	10	14
	1	11	12
	5	6	
R_j	22	37	46

Example With Tied Ranks

2. Rank the scores from lowest to highest

	White	Yellow	Purple
	4.5	2	7
	9	8	13
	3	11	14
	1	11	11
	4.5	6	
R_j	22	38	45

tied scores: (4+5)/2 = 4.5; (10+11+12)/3 = 11

Example With Tied Ranks

3. Note that rank scores R_{2-3} have changed:

	White	Yellow	Purple
	4.5	2	7
	9	8	13
	3	11	14
	1	11	11
	4.5	6	
R_j	22	38(37)	45(46)

a. This is because of tied ranks in these columns.

Kruskal-Wallis: Example

	White	Yellow	Purple
	4	2	7
	9	8	13
	3	10	14
	1	11	12
	5	6	
R_j	22	37	46

3. Then use K-W formula to calculate H:

$$H = \frac{12}{N(N+1)} \sum^{a} \frac{R_j^2}{n_j} - 3(N+1)$$

with df = a-1

$$= \frac{12}{14(14+1)} [(22)^2/5 + (37)^2/5 + (46)^2/4] - 3(14+1)$$

= 6.4, df= 2

$H_{[.05;5,5,4]} = 5.64$, $P<0.05$.

And Now,

$$D = 1 - \frac{\Sigma T}{N^3 - N}$$

$\Sigma T = [(2)^3 - 2] + [(3)^3 - 3] = 6 + 24 = 30$

$D = 1 - \{30/[(14)^3 - 14]\} = .989$

Thus,

$H_{adj} = H/D$

$= 6.4/.989 = 6.47$

$P < 0.049$.

Measures of Association

1. Are used to examine the relationship (covariance) between two or more variables.

a. Analogous to regression/correlation analysis.

b. Relationships are based on *ranks* rather than raw/transformed scores.

Measures of Association

1. The two best known are:

a. Spearman's rank order correlation.

b. Kendall's rank order correlation.

1. This latter is useful because it is possible to obtain *partial correlation coefficients.*

2. Useful for path analysis with small data sets

c. Also:

1. Multiple variable procedure: Kendall's coefficient of concordance.

Spearman's r_S: Method

1. Consider a herd of red deer in which male mating success depends on his fighting success relative to other males.

a. What is the relationship between number of fights and mating success?

Spearman's r_S: Method

1. Rank variables X and Y separately from lowest to highest.

Ind.	#Fights	#Mates
A	27	6
B	14	4
C	5	1
D	11	5
E	2	3

Spearman's r_S: Method

2. Calculate the deviations for X and Y (d), then d^2 and Σd^2:

Ind.	#Fights	#Mates	d	d^2
A	5	5	0	0
B	4	3	1	1
C	2	1	1	1
D	3	4	− 1	1
E	1	2	− 1	1
				4 = Σd^2

Spearman's r_S: Method

Then,

If:

$$r_s = 1 - \frac{6\sum^{N} d_i^2}{N^3 - N}$$

$$= 1 - \frac{6(4)}{5^3 - 5} \quad = \quad 1 - .196$$

Spearman's r_S: Result

$r_S = .803$

a. Look up significance in Table Q for N < 25.

TABLE Q
Critical values of r_s, the Spearman rank-order correlation coefficient

N	α .25 / α .50	.10 / .20	.05 / .10	.025 / .05	.01 / .02	.005 / .01	.0025 / .005	.001 / .002	.0005 (one-tailed) / .001 (two-tailed)
4	.600	1.000	1.000						
5	.500	.800	.900	1.000	1.000				
6	.371	.657	.829	.886	.943	1.000	1.000		
7	.321	.571	.714	.786	.893	.929	.964	1.000	1.000
8	.310	.524	.643	.738	.833	.881	.905	.952	.976
9	.267	.483	.600	.700	.783	.833	.867	.917	.933
10	.248	.455	.564	.648	.745	.794	.830	.879	.903
11	.236	.427	.536	.618	.709	.755	.800	.845	.873
12	.224	.406	.503	.587	.671	.727	.776	.825	.860
13	.209	.385	.484	.560	.648	.703	.747	.802	.835
14	.200	.367	.464	.538	.622	.675	.723	.776	.811
15	.189	.354	.443	.521	.604	.654	.700	.754	.786
16	.182	.341	.429	.503	.582	.635	.679	.732	.765
17	.176	.328	.414	.485	.566	.615	.662	.713	.748
18	.170	.317	.401	.472	.550	.600	.643	.695	.728
19	.165	.309	.391	.460	.535	.584	.628	.677	.712
20	.161	.299	.380	.447	.520	.570	.612	.662	.696
21	.156	.292	.370	.435	.508	.556	.599	.648	.681
22	.152	.284	.361	.425	.496	.544	.586	.634	.667
23	.148	.278	.353	.415	.486	.532	.573	.622	.654
24	.144	.271	.344	.406	.476	.521	.562	.610	.642
25	.142	.265	.337	.398	.466	.511	.551	.598	.630

Spearman's r_S: Tied ranks

1. The effect of ties is to reduce the sum of squares Σx^2 *below* $(N^3-N)/12$.

a. The correction for ties is:

$T_i = (t^3 - t)/12$, for the i-th tied rank,

and

$\Sigma x^{2'} = (N^3 - N)/12 - \Sigma T$.

b. Do the same for Σy^2 (corrected = $\Sigma y^{2'}$).

Spearman's r_S: Tied ranks

1. these values are then substituted into originally derived formula for r_S:

$$r_S = \frac{\Sigma x^{2'} + \Sigma y^{2'} - \Sigma d^2}{\sqrt{2(\Sigma x^{2'}\ \Sigma y^{2'})}}$$

Spearman's r_S: Tied ranks

2. For Large samples: (N > 10)

a. $t = r_S \sqrt{[(N-2)/(1- r_S)]}$

b. This value is distributed as Student's t with df = N-2

c. This table is in S&R

Derivation of Spearman's r_S

1. This permits visualization of similarity with Pearson's parametric r.

2. Imagine two sets of variables X_i and Y_i

a. Their relationship can be determined by arranging them in pairs and taking the difference between them:

$$d_i = X_i - Y_i$$

Derivation of Spearman's r_S

$$d_i = X_i – Y_i$$

1. If the relationship is perfect, every $d_i = 0$.

2. Deviations from 0 indicate how good or bad the correlation is.

3. Raw scores are difficult to use because - and + scores could cancel.

a. Thus, d_i^2 provides a better estimate for each pair of the deviation from a perfect correlation.

b. Also, with large d_i's, the larger $\Sigma\, d_i^2$ will be.

Derivation of Spearman's r_S

3. If $x = (\boldsymbol{X} - X_i)$ and $y = (\boldsymbol{Y} - Y_i)$,

Where $\boldsymbol{X} = \Sigma X_i/n_i$ and $\boldsymbol{Y} = \Sigma Y_i/n_i$

a. Then the general expression for a parametric correlation coefficient is:

$$r = \frac{\Sigma xy}{\sqrt{(\Sigma x^2 \, \Sigma y^2)}}$$

b. this expression measures the degree to which two variables are correlated.

To See This,

1. Imagine a variable, y, plotted on itself.
2. The general equation then becomes:

$$r = \frac{\Sigma(y)(y)}{\sqrt{(\Sigma y^2 \, \Sigma y^2)}} = 1$$

For a Nonparametric Solution

1. Assume X_i and Y_i are ranks.
2. Then, sum of these integers is:

$$\Sigma X_i = N(N+1)/2$$

2. Really?

1 + 2 + 3 + 4 + 5 = 15; N = 5

5(5+1)/2 = 30/2 = 15

Also,

3. The sum of their squares is:

$$\Sigma X_i^2 = \frac{N(N+1)(2N+1)}{6}$$

4. Really?

$1^2 + 2^2 + 3^2 + 4^2 + 5^2 = 55; N = 5$

$[5(5+1)][(2)(5)+1]/6 = (30)(11)/6 = 55$

Then,

5. It is clear that the expressions used to calculate r_S are simply what arises from sums of integers or their squares.

Also, since

$\Sigma x^2 = \Sigma(\boldsymbol{X} - X_i)^2 = \Sigma X_i^2 - [(\Sigma X_i)^2]/N,$

i.e., the expression for the sum of the squared deviations from the mean (a way of expressing central tendency in parametric statistics),

So,

1. Using the equivalent nonparametric expression:

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

$= (N^3 - N)/12$

1. and similarly, $\Sigma y^2 = (N^3 - N)/12$

Now,

2. Because

$d = x - y$

Then,

$$d^2 = (x-y)^2 = x^2 - 2xy + y^2$$

for each d_i, so,

$$\Sigma d^2 = \Sigma x^2 + \Sigma y^2 - 2\Sigma xy$$

But,

3. In theory,

$$\boldsymbol{r = \frac{\Sigma xy}{\sqrt{(\Sigma x^2\, \Sigma y^2)}} = r_S}$$

if X_i and Y_i are ranks.

Thus, By Substitution,

4. The expression for Σd^2 becomes:

$$\Sigma d^2 = \Sigma x^2 + \Sigma y^2 - 2\, r_S \sqrt{(\Sigma x^2 \Sigma y^2)}$$

Thus,

$$r_S = \frac{\Sigma x^2 + \Sigma y^2 - \Sigma d^2}{2\sqrt{(\Sigma x^2\, \Sigma y^2)}}$$

and by substitution of $\Sigma x^2 = (N^3 - N)/12 = \Sigma y^2$ into this equation,

We Have,

$$r_S = 1 - \frac{6 \Sigma d_i^2}{N^3 - N}$$