chapter 2-9

Non-Parametric Statistics

Introduction

Parametric statistics such as ANOVA and regression carry the underlying assumption that the data are normally distributed. When the data are not normally distributed, it opens up a whole new set of statistical tests that are called non-parametric statistics. One of the most commonly used non-parametric statistic is the percentile. Unlike the z-score, no assumption is made about the distribution of the data. Percentiles work equally well describing normal or non-normal distributions. The purpose of this chapter is not to review all non-parametric tests, but to provide insight into three commonly used tests that are examples of the 3 main types of hypothesis testing for which inferential statistics are used, as discussed in chapter 2-4.

Is There a Difference?

Chi-square: Analogous to ANOVA, it tests differences in the frequencies of observations of categorical data. A 2x2 table is equivalent to z test between two proportions.

Wilcoxson signed rank test: Analogous to paired t-test.

Wilcoxson rank sum test: Analogous to independent t-test.

Is there a Relationship?

Rank Order Correlation: Analogous to the correlation coefficient tests for relationships between ordinal variables. Both the **Spearman's Rank Order Correlation** (r_s) & **Kendall's Tau** (τ) will be discussed

Can we predict?

Logistic Regression: Analogous to linear regression, it assess the ability of variables to predict a dichotomous (0/1) variable.

Chi-square

Previously in chapter 2-8, on testing differences between means, the t test was discussed as a way to test for differences in means of two continuous variables. When categorical variables are involved, a Chi-square (χ^2) analysis can be carried out. A categorical variable is a qualitative

variable in which cases are classified or categorized into one (and only one) of the possible multiple levels of the variable. For example we often use sex of the subject as a variable that is classified into one of two possible levels, men or women. The chi-square is a test of a difference in the proportion of observed frequencies in categories versus the expected proportions. Figure 2-9.1 shows the results of an analysis on the handedness data in the Grip Strength data set (see Appendix A). Subjects reported they were either right or left hand dominant (fortunately, nobody in this sample reported they were ambidextrous). By default, when a chi-square test is run, the null hypothesis is that there is no difference in the numbers in the categories, or in other words, that there are an equal number of right and left handers.

$$\chi^2 = \frac{(O-E)^2}{E}$$

In our example there were 44 subjects, 6 of whom reported that they were left handers. The observed frequencies are therefore 6 and 38 for left and right handers, respectively. If we are testing whether there are equal numbers of right and left handers then the expected frequencies to be tested against would be 22 and 22. The value of Chi-square would therefore be calculated as

$$\chi^2 = \frac{(6-22)^2}{22} + \frac{(38-22)^2}{22} = 23.273$$

As with any other test statistic we have discussed, there is a specific probability associated with any given value of that test statistic. As you can see in Figure 2-9.1, a χ^2 is associated with a probability of 0.000 when reported to 3 significant figures. In the table (Asymp. Sig.). the number is a little larger than 0, so what it means is that you are more than 99.9% confident that there is a difference in proportion of right and left handers in the population from which this sample was drawn. From the literature however, we know that sample usually contain 10% to 15% of left handers. We can test this by asking for different expected frequencies to be assumed in the test. If for instance we want to test if there are 15% left handers in the sample, then the expected frequencies out of a sample of 44 for left handers would be 6.6 and for right handers 37.4 (do not worry that you can not have 0.6 of a person in reality). The analysis in the right hand column of Figure 2-9.1 shows the results of this analysis. The value of Chi-square (χ^2) would therefore be calculated as

$$\chi^2 = \frac{(6-6.6)^2}{6.6} + \frac{(38-37.4)^2}{37.4} = 0.064$$

This value of χ^2 is associated with a probability of 0.8; therefore we are only 20% confident that the observed and expected frequencies are different, in other words we do not reach our

95% confidence criterion and therefore accept the null hypothesis that there is no difference in proportions. So our sample does conform to the expectation that there are 15% left handers in the population.

NPar Tests

Descriptive Statistics

	N	Mean	Std. Deviation	Minimum	Maximum	Percentiles 25th	50th (Median)	75th
Handedness	44	1.1364	.3471	1.00	2.00	1.0000	1.0000	1.0000

HYPOTHESIS: EQUAL NUMBERS OF RIGHT AND LEFT HANDERS (22 out of 44)

Chi-Square Test
Frequencies

Handedness

	Observed N	Expected N	Residual
Right Handed	38	22.0	16.0
Left Handed	6	22.0	-16.0
Total	44		

Test Statistics

	Handedness
Chi-Square[a]	23.273
df	1
Asymp. Sig.	.000

a. 0 cells (.0%) have expected frequencies less than 5. The minimum expected cell frequency is 22.0.

HYPOTHESIS: 15% LEFT HANDERS (6.6 out of 44)

Chi-Square Test
Frequencies

Handedness

	Observed N	Expected N	Residual
Right Handed	38	37.4	.6
Left Handed	6	6.6	-.6
Total	44		

Test Statistics

	Handedness
Chi-Square[a]	.064
df	1
Asymp. Sig.	.800

a. 0 cells (.0%) have expected frequencies less than 5. The minimum expected cell frequency is 6.6.

Figure 2-9.1: Chi-square analysis of handedness in the Grip Strength Data set.

Two-way Chi-square

A Chi-square analysis can also be applied when two categorical variables are considered simultaneously. In this case, the Chi-square test is a test of independence between the two categorical variables. Chi-square is used to determine if there is a significant difference in the frequency of observations in the categories of the two variables. The null hypothesis set up is that there is no difference in the frequency of observations found in each cell of one variable with respect to the other. If the

		Male	Female	Total
Ex-Smoker	Observed	14	14	28
	Expected	12.6	15.4	
Current Smoker	Observed	12	18	30
	Expected	13.4	16.6	
	Total	26	32	58

Table 2-9.2: Frequency of subjects by smoking and sex categories in Smoking data set.

observed and expected frequencies are similar within each variable, the chi-square test will not be significant. If the observed frequencies deviate considerably from the expected frequencies in one or more categories, the chi-square test will be significant. A significant chi-square test suggests that there is likely to be a real difference across the categories in the population from which the sample was drawn. By example, Table 2-9.2 shows the frequency of observation of smoking category by sex in the Smoking data set (see Appendix A).

There were actually three categories of smokers; Non-Smokers, Ex-Smokers and Current Smokers, although only the latter two groups are considered in this analysis. There were 14 male and 14 female Ex-Smokers and 12 male and 18 female Current Smokers for a total sample size of 58. The expected values are calculated based upon the hypothesis that there is no difference in the proportion of males and females within each smoking category. The expected value for Male Ex-Smokers would therefore be the proportion of males in the total sample (26/58) multiplied by the total number of Ex-Smokers (28).

Expected Number of Male Ex-Smokers	***= 26/58 x 28 = 12.6***
Using the same logic:	
Expected Number of Female Ex-Smokers	***= 32/58 x 28 = 15.4***
Expected Number of Male Current Smokers	***= 26/58 x 30 = 13.4***
Expected Number of Female Current Smokers	***= 32/58 x 30 = 16.6***

Now Chi-square is calculated using the squared differences between observed and expected divided by expected, summed over all four combinations of sex and smoker category. For instance, for male Ex-Smokers the calculation is $(14\text{-}12.6)^2$ / 12.6. The result when this is summed for all four combinations as shown in figure 2-9.2, is a chi-square value of 0.587 with an associated probability of 0.444. Thus, the conclusion is made that there is no significant difference in the distribution of males and females with respect to smoking status.

The second analysis in figure 2-9.3, shows a similar chi-square analysis for smoking status by report of the answer to the question “Do you regularly have itchy eyes? Yes or no?”. Here there appears to be a greater proportion of Ex-Smokers reporting itchy eyes in comparison to Current Smokers. When the results of the Chi-square test are reviewed it shows a calculated Chi-square of 3.807 with an associated probability of 0.051. The numerical value of the probability is interesting. To accept the hypothesis that there is a difference we often use 95% confidence or p = 0.05 as our cut off, thus p should be less than 0.05. Here p = 0.051, so we can not actually say it is less than 0.05, but it is heartbreakingly close. We are 94.9% confident there is a difference in proportions by smoking group and therefore, that Ex-Smokers report itchy eyes, more often than Current Smokers, which incidentally is not an unexpected result, based upon literature in this area.

Case Processing Summary

	Cases					
	Valid		Missing		Total	
	N	Percent	N	Percent	N	Percent
Smoking Category * Sex of Subject	58	100.0%	0	.0%	58	100.0%
Smoking Category * Do you regularly have itchy eyes?	57	98.3%	1	1.7%	58	100.0%

Smoking Category * Sex of Subject

Crosstab

			Sex of Subject		
			Male	Female	Total
Smoking Category	ExSmoker	Count	14	14	28
		Expected Count	12.6	15.4	28.0
		% within Smoking Category	50.0%	50.0%	100.0%
		% within Sex of Subject	53.8%	43.8%	48.3%
		% of Total	24.1%	24.1%	48.3%
	Current Smoker	Count	12	18	30
		Expected Count	13.4	16.6	30.0
		% within Smoking Category	40.0%	60.0%	100.0%
		% within Sex of Subject	46.2%	56.3%	51.7%
		% of Total	20.7%	31.0%	51.7%
Total		Count	26	32	58
		Expected Count	26.0	32.0	58.0
		% within Smoking Category	44.8%	55.2%	100.0%
		% within Sex of Subject	100.0%	100.0%	100.0%
		% of Total	44.8%	55.2%	100.0%

Chi-Square Tests

	Value	df	Asymp. Sig. (2-sided)	Exact Sig. (2-sided)	Exact Sig. (1-sided)
Pearson Chi-Square	.586[b]	1	.444		
Continuity Correction[a]	.251	1	.616		
Likelihood Ratio	.586	1	.444		
Fisher's Exact Test				.598	.308
Linear-by-Linear Association	.575	1	.448		
N of Valid Cases	58				

a. Computed only for a 2x2 table

b. 0 cells (.0%) have expected count less than 5. The minimum expected count is 12.55.

Smoking Category * Do you regularly have itchy eyes?

Crosstab

			Do you regularly have itchy eyes?		
			No	Yes	Total
Smoking Category	ExSmoker	Count	12	15	27
		Expected Count	15.6	11.4	27.0
		% within Smoking Category	44.4%	55.6%	100.0%
		% within Do you regularly have itchy eyes?	36.4%	62.5%	47.4%
		% of Total	21.1%	26.3%	47.4%
	Current Smoker	Count	21	9	30
		Expected Count	17.4	12.6	30.0
		% within Smoking Category	70.0%	30.0%	100.0%
		% within Do you regularly have itchy eyes?	63.6%	37.5%	52.6%
		% of Total	36.8%	15.8%	52.6%
Total		Count	33	24	57
		Expected Count	33.0	24.0	57.0
		% within Smoking Category	57.9%	42.1%	100.0%
		% within Do you regularly have itchy eyes?	100.0%	100.0%	100.0%
		% of Total	57.9%	42.1%	100.0%

Chi-Square Tests

	Value	df	Asymp. Sig. (2-sided)	Exact Sig. (2-sided)	Exact Sig. (1-sided)
Pearson Chi-Square	3.807[b]	1	.051		
Continuity Correction[a]	2.831	1	.092		
Likelihood Ratio	3.844	1	.050		
Fisher's Exact Test				.064	.046
Linear-by-Linear Association	3.740	1	.053		
N of Valid Cases	57				

a. Computed only for a 2x2 table

b. 0 cells (.0%) have expected count less than 5. The minimum expected count is 11.37.

Figure 2-9.3: SPSS Cutput for CROSSTABS of Smoking Status and Sex in the Smoking data set.

Chi-square using SPSS

One Way Chi-square is found as an option under the Nonparametric Tests option of the Analyze menu as shown in Figure 2-9.4. When the option is selected a chi-square dialog box comes up (also in Figure 2-9.5), where the variable or variables chosen to be analyzed can be selected.

Note also that there is an “Expected Values” option on this dialog box. By default the “All categories equal“ selection will be checked. But, if you want to specify exact values for expected values these can also be entered. Remember in the earlier example of there being 15% left handers (6.6 out of 44) in the grip strength data set, this is the dialog box that was used to

produce that analysis. Note that under “Expected Values” the “Values” button is checked and the values 37.4 and 6.6 have been entered. To enter a value type it in the box to the right of “Values” then click on Add. The value will be added into the list below. In this example 37.4 and 6.6 represent 85% and 15% of the sample size of 44, which are the expected values to use for right handers and left handers, respectively. This sets up for Chi-square to test the hypothesis that the predicted counts are significantly different than 15% left handers in the population. This was indeed found not to be the case as the null hypothesis was accepted (Figure 2-9.3). You can put any values you want in, as long as they make sense for your hypothesis. In this case, there was previous literature that stated that the proportion of left handers in a population tends to be 10% to 15%. Many times, however, you want simply to test the hypothesis that the counts are different, so go with the default “All categories are equal” option.

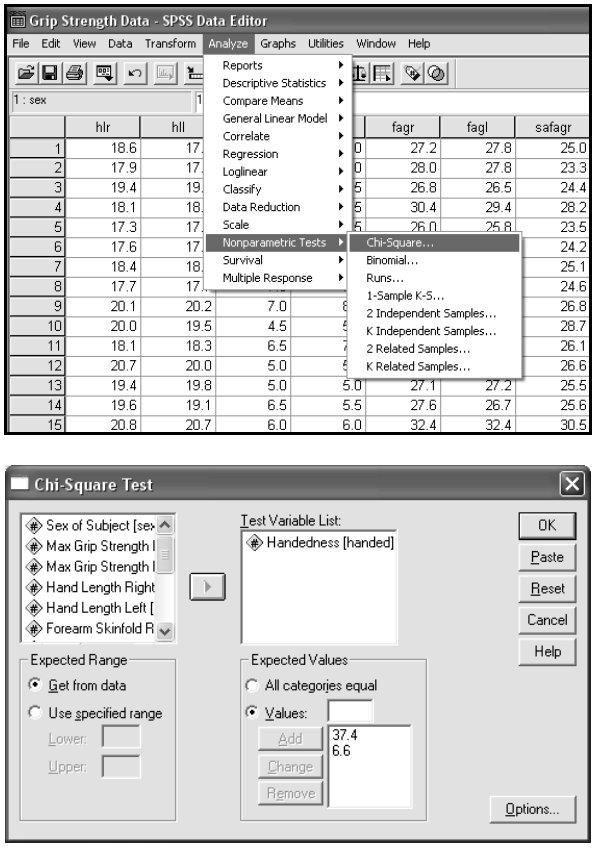

Figure 2-9.4: SPSS dialog boxes for One-way Chi-square

Two-way Chi-square is found by using the “Crosstabs” option under the “Descriptives” option of the “Analyze” menu (Figure 2-9.5). Crosstabs is short for crosstabulation, which simply means that data is put into tables with rows and columns representing categories of the categorical variables. The entries in the cell are summaries in the form of counts or percentages. Multiple comparisons can be specified in the dialog box. In Figure 2-9.5, the Smoking data is used and the analysis is specified as smoking category as the rows and the sex of the subject and the response to the “do you regularly have itchy eyes” question as the columns. This will set up two Chi-square analyses Smoking Category * Sex of Subject and Smoking Category * “Do you regularly have itchy eyes?” The results of these analyses were reported in Figure 2-9.3 earlier. Note that we had a Select Cases in place so that only Ex-Smokers and Current Smokers were included in the analysis. You can see that this appears as a Smokegrp < 1 (FILTER) listed at the bottom of the variable list on the CROSSTABS dialog box.

Before clicking the "OK" button, you will need to click the "Statistics" button to bring up the Crosstabs: statistics dialog box. On that dialog box check the Chi square option, in order to have chi-

square calculated for both comparisons. Click "OK" and return to the Crosstabs dialog box and click on the "Cells" button. As shown in Figure 2-9.5 the Cells dialog box allows a choice of information to be displayed in each cell. The Observed count is a default selection but if you wish to display the expected values that are being used for hypothesis testing select Expected. A useful feature is to be able to include various total, row and column percentages since in reporting results we often talk about the percentage of one category or another. Figure 2-9.3 discussed earlier shows what the output looks like with all percentages selected.

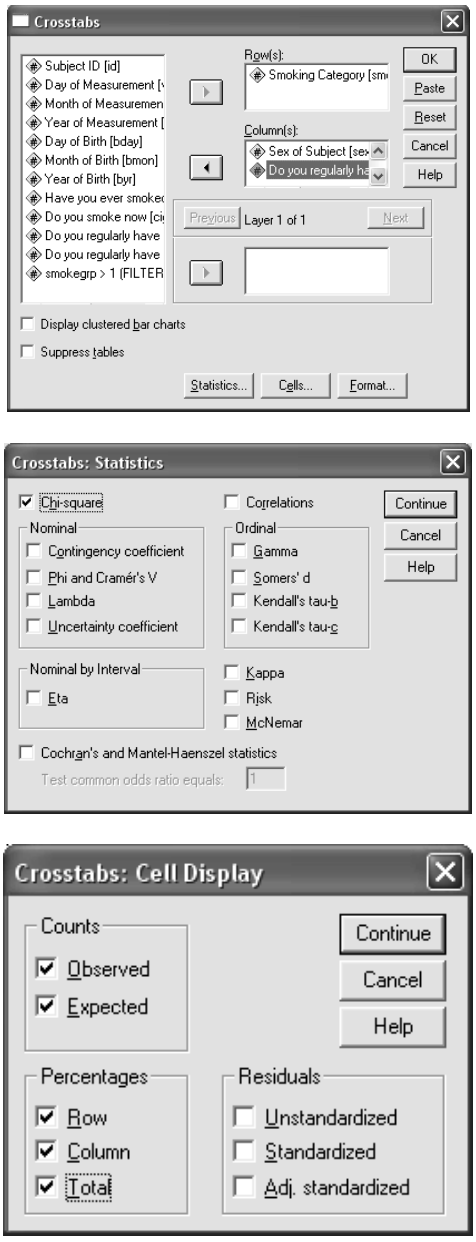

Figure 2-9.5: Crosstabs (two-way Chi-square) dialog boxes

Wilcoxson's Signed Rank Test

The Wilcoxson's Signed Rank test is the nonparametric equivalent of the paired t test where the hypothesis tested is whether the median of the differences between pairs is zero in the population sampled from.

Table 2-9.1 shows the results of this test on data used previously to illustrate the paired t-test (chapter 2-7). The data is the right and left grip strength for the males in the Grip Strength data set (Appendix A).

The steps in calculating Wilcoxson's Signed Rank test are as follows:

- calculate the difference between right and left grip strengths for each subject.
- rank the differences from lowest to highest irrespective of sign and ignoring the zero differences.
- sum the ranks of the negative differences (T-) and the positive differences (T+).
- T is then the smaller value of T- or T+.
- Count the number of non-zero differences (n).
- Look up the critical value of T at p=0.05 (Crit $T_{0.05}$). In order to save space, the table of critical values of T have not been reproduced in this text. However, this is not a problem since SPSS and other statistical software will give the exact probability of T for your hypothesis testing.

Right max Grip Strength	Left max Grip Strength	Difference	Rank
45	53	-8	16
55.5	61	-5.5	15
41.5	46.5	-5	13
44	48	-4	11
65	68	-3	9
60	61.5	-1.5	7
51	52	-1	3
42.1	43	-0.9	1
40	40	0	
66	66	0	
60	59	1	2
54	53	1	4
54	52.9	1.1	5
43	41.9	1.1	6
59	57	2	8
58.1	54.2	3.9	10
48	44	4	12
54	49	5	14
52	44	8	17
54	45	9	18
Sum of ranks of negative differences		**T-**	75
Sum of ranks of positive differences		**T+**	96
T = smaller of T- and T+		**T**	75
n = number of non-zero differences		**n**	18
Critical value of T at p = 0.05		**Crit $T_{0.05}$**	40
T is not less than Crit $T_{0.05}$ therefore no significant difference			

Table 2-9.1: Wilcoxson's Rank Sum Test results for difference in left and right grip strength (N = 20)

The test is then whether T is greater or smaller than Crit $T_{0.05}$. In chapter 2-4 when discussing inferential statistics is was mentioned that almost always the calculated statistic needs to be bigger than the critical value of the test statistic for there to be a significant difference or relationship. This however, is one of the few cases where that is not true, because T needs to be smaller than Crit $T_{0.05}$ in this Wilcoxson's signed rank test. Therefore, in the example in Table 2-9.1, having T = 75 and Crit $T_{0.05}$ = 40, there is no significant difference ($p<0.05$) between right and left grip strengths.

Wilcoxson's Rank Sum Test

This is the non-parametric equivalent of the independent t-test; the hypothesis being tested is whether the difference between medians of the two samples is zero. Table 2-9.2 shows the

results of this test on data used previously to illustrate the independent t-test (chapter 2-7). The data is the right grip strength for the males and females in the Grip Strength data set (Appendix A). In order to reduce sample size a random sample of 22 of the subjects was taken, resulting in 9 males and 13 females.

The steps in calculating the Wilcoxson's Rank Sum Test are as follows:

- Rank the scores from the two groups combined from lowest to highest. Any ties receive the average ranking, e.g. ranking 8th through 11th all were 40 lbs, therefore, the average ranking of 9.5 was assigned to 40 lbs.
- Add up the ranks of the group with the smallest sample size (T).

Sex	Grip Strength	Rank	Sex	Grip Strength	Rank
F	24	1	M	40	9.5
F	28	2	M	41.5	12
F	29	3	M	43	13.5
F	30	4	M	44	16
F	30.5	5	M	48	18
F	34.5	6	M	54	19
F	37	7	M	54	20
F	40	9.5	M	55.5	21
F	40	9.5	M	65	22
F	40	9.5			
F	43	13.5	tied	8th 9th 10th 11th	
F	43.5	15	tied	13th 14th	
F	46	17			
Sum of ranks for sample with smallest sample size (T)					151
Sample sizes of two groups					9 13
Critical range for sum of ranks for sample sizes 9 & 13 at p = 0.05					73 - 134

Table 2-9.2: Wilcoxson's Rank Sum Test results for difference in grip strength between males (N = 9) and females (N=13)

- Look up the critical range of T at p=0.05 (Crit $T_{0.05}$) for sample sizes of 9 and 13. In order to save space, the table of critical ranges of T have not been reproduced in this text. However, this is not a problem since SPSS and other statistical software will give the exact probability of T for your hypothesis testing.

The test is then whether T is above or below the critical range of T for p = 0.05. In chapter 2-4 when discussing inferential statistics is was mentioned that almost always the calculated statistic needs to be bigger than the critical value of the test statistic for there to be a significant difference or relationship. This however, is one of the few cases where that is not true; in fact the T needs to be smaller or bigger than a critical range of T for the Wilcoxson's rank sum test. In the example in Table 2-9.2, T = 151 and the Critical range for sum of ranks for sample sizes 9 & 13 at p = 0.05 is 73 – 134; therefore there is a significant difference ($p<0.05$) between male and female grip strengths.

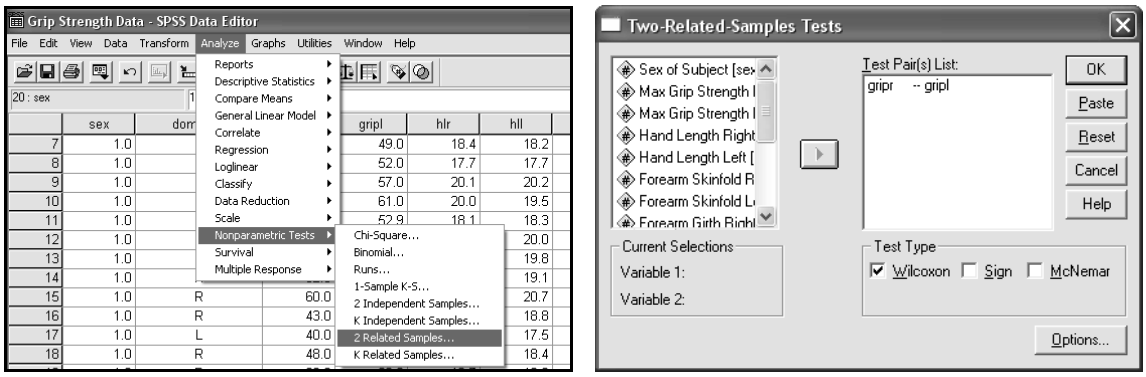

Figure 2-9.6: SPSS Dialog boxes for ANALYZE – 2 RELATED SAMPLES menu (left) and 2 RELATED SAMPLES TESTS (right)

Spearman's Rank Order Correlation (r_s) & Kendall's Tau (τ)

Often you wish to know if there is a relationship between variables but neither of the variables is normally distributed. The calculation of the Pearson correlation coefficient (r) for probability estimation is not appropriate in this situation. Sometimes you can normalize the variables with some transformation as discussed in chapter 2-3. If one of the variables is normally distributed you can still use r, but if both are not normally distributed, then you can use Spearman's Rank Order Correlation Coefficient (r_s) or Kendall's tau (τ). These tests rely on the two variables being rankings. A good example would be judges' rankings on two different tests. These tests would test for a relationship between these rankings. Continuous non-normally distributed variables could also be turned into rankings by simply ordering them from highest to lowest.

Spearman's Rank Order Correlation (r_s)

Spearman's rank correlation is actually calculated using the Pearson product moment correlation coefficient equation, but on the rankings of the scores, rather than the scores themselves. There is also a simplified form of the equation that gives the same answer.

$$r_s = 1 - \frac{6\Sigma d^2}{n(n^2-1)}$$

As an example of the use of r_s, table 2-9.3 shows the rankings of 6 llamas by two judges in a recent llama show. You probably have never heard of llama shows, but just like pedigree dogs you can take your Canadian Livestock Records Corporation registered llama to shows where judges will place them based upon conformation, presence and movement. Figure

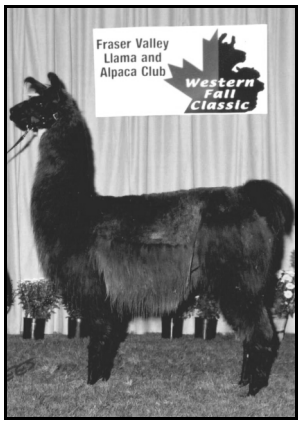

Figure 2-9.7: LJ's Serenade, a Grand Champion female llama.

2-9.7 shows LJ's Serenade, a 5 year-old female llama who has twice been made Grand Champion female at shows. An interesting feature of the Western Fall Classic show is that there are two judges that judge the animals independently and each gives out placings and rosettes. Ideally, the judges place the animals the same way, but the task is very subjective although based upon stated expected characteristics of the breed which each judge is working to. Table 2-9.3 shows the placings by each judge of six llamas competing in one age category. Each judge placed llama #1 first and agreed that llamas #4 and #6 were the worst although in opposite order.

However, their ratings of the 2nd 3rd and 4th place llamas were more mixed. Spearmans rank correlation will quantify how well these ratings agree. This is a paired analysis, and as such the difference in ratings (d) is calculated. This is then squared and these squared values are then summed. The simple Spearman's equation is then applied and the result is r_s = 0.77. As with the Pearson Product Moment Correlation coefficient (r), r_s can be compared to a critical value of r_s for the required probability level.

Llama #	Judge 1	Judge 2	d	d^2
1	1	1	0	0
2	3	4	-1	1
3	4	2	2	4
4	5	6	-1	1
5	2	3	-1	1
6	6	5	1	1
			Σd	Σd^2
			0	8
$r_s = 1 - \frac{6\Sigma d^2}{n(n^2-1)}$		$r_s = 1 - \frac{6 \times 8}{6(6^2-1)}$		$r_s = 0.771$

Figure 2-9.3: Calculation of Spearman's r_s between two judges' show placings of 6 llamas

For samples of more than 10 pairs the probability distribution is similar to that of r, so the table shown previously in chapter 2-5 can be used. For samples of less than 10 pairs the significance levels of r_s are given in table 2-9.4. In our llama judging example, the r_s was found to be 0.77. The sample size was less than 11 so we can use Table 2-9.4. to find that with a sample size of 6 and p=0.05 the critical value of r_s is 0.886, so unfortunately for the judges, we can not say that their placings were significantly related, at the 95% confidence level.

	Probability	
Sample Size	**0.05**	**0.01**
≤ 4	none	none
5	1.000	none
6	0.886	1.000
7	0.750	0.893
8	0.714	0.857
9	0.683	0.833
10	0.648	0.794
≥ 11	Use Table 2-5.?	

Table 2-9.4: Significance levels of r_s in small samples

Kendall's Tau (τ)

Kendall's τ is used to assess the degree of association of rankings. The steps in calculation for τ are as follows:

- Rank the cases by one of the ratings in ascending order.
- Taking each of the second ratings in turncount how many of the ranks below it are smaller than it.
- Sum these counts to get Q.
- Calculate τ using the following equation:

$$\tau = 1 - \frac{4Q}{n(n-1)}$$

Llama #	Judge 1	Judge 2	# number of ranks lower in ranking with smaller values
1	1	1	0
5	2	3	1
2	3	4	1
3	4	2	0
4	5	6	1
6	6	5	

$$Q = 3$$

$$\tau = 1 - \frac{4Q}{n(n-1)} \qquad \tau = 1 - \frac{4 \times 3}{6(6-1)} \qquad \tau = 0.60$$

Figure 2-9.8: Calculation of Kendall's τ between two judges show placings of 6 llamas

Figure 2-9.8 shows the calculation of τ for the llama judging data. The data is sorted by judge 1 and then the rankings of judge 2 are assessed one by one. For llama #1 none have smaller rankings. For llama #5 only llama #3 has a smaller ranking so its count is 1. Moving down, llamas #2 and #4 each have one llama below that has a smaller rank, therefore they each get a count of 1. The counts add up to 3 which is the value used for Q. τ is then calculated, using the formula given above, to be 0.60. If all the pairs were the same, termed complete concordance, then τ would be equal to +1, with complete disagreement being -1. Spearman's rank correlation is more popular and easier to compute but Kendall's tau is preferred by statistician's because of its statistical properties.

Using SPSS to Calculate r_s and τ

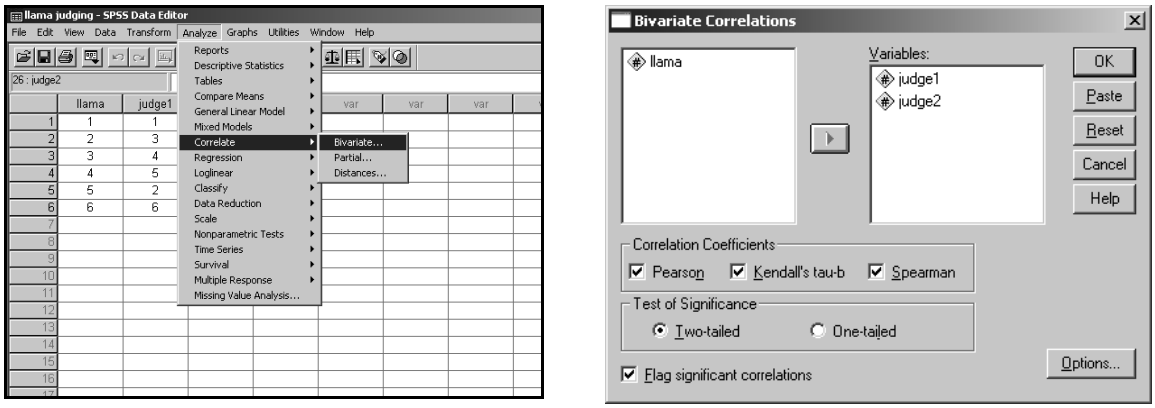

Figure 2-9.9: SPSS dialog boxes for rank correlation coefficients.

Nonparametric tests of association are found under the "Bivariate" option of the "Correlate" option in the "Analyze" menu (Figure 2-9.9). Selecting the "Bivariate" option brings up the dialog box shown in the right hand panel of Figure 2-9.9.

This allows a selection of one or more of the three measures of association; Pearson Correlation Coefficient (r) for normally distributed variables, or the Spearman's (r_s) or Kendall's (τ) for non-normal data. Figure 2-9.10 shows the output for the SPSS correlation analysis for the llama judging data produced from the dialog box shown in figure 2-9.9. The first part is the result of the selection of "Pearson" in the dialog box. This is the parametric correlation and therefore inappropriate for this analysis. The next part is the result of the two nonparametric tests. The results are the same as calculated earlier in the chapter r_s = 0.771 and τ = 0.60. You do not need to look these values up in a table for significance testing since SPSS provides the exact probability associated with each coefficient.

Correlations

Correlations

		JUDGE1	JUDGE2
JUDGE1	Pearson Correlation	1	.771
	Sig. (2-tailed)	.	.072
	N	6	6
JUDGE2	Pearson Correlation	.771	1
	Sig. (2-tailed)	.072	.
	N	6	6

Nonparametric Correlations

Correlations

			JUDGE1	JUDGE2
Kendall's tau_b	JUDGE1	Correlation Coefficient	1.000	.600
		Sig. (2-tailed)	.	.091
		N	6	6
	JUDGE2	Correlation Coefficient	.600	1.000
		Sig. (2-tailed)	.091	.
		N	6	6
Spearman's rho	JUDGE1	Correlation Coefficient	1.000	.771
		Sig. (2-tailed)	.	.072
		N	6	6
	JUDGE2	Correlation Coefficient	.771	1.000
		Sig. (2-tailed)	.072	.
		N	6	6

Figure 2-9.10: SPSS dialog boxes for rank correlation coefficients.

For Kendall's τ the two-tailed probability is 0.091, meaning we are 90.9% confident there is a relationship in the population that this sample was drawn from. When accepting at the 95% level, this is not good enough, so we accept the null hypothesis that there is no relationship

($p<0.05$). With a Spearman's coefficient having a probability of 0.072 we come to the same conclusion, as we are only 92.8% confident there is a relationship in the population that this sample was drawn from.

Logistic Regression

Logistic regression is analogous to linear regression analysis in that an equation to predict a dependent variable from independent variables is produced. The big difference is that the dependent variable (outcome) is categorical for logistic regression versus continuous for linear regression. Although dependent variables with multiple categories can be used, it is most common to use binary (dichotomous) dependent variables, and the discussion in this text will be restricted to such. Binary variables have only two possible values such as a Yes or No answer to a question on a questionnaire, or sex of a subject being man or woman. It is usual to code them as 0 or 1, such that men might be coded as 1 and women coded as 0. If a sample is coded with 1s and 0s, the mean of a binary variable represents the proportion of 1s. For instance, if you have a sample size of 100 with a binary variable Sex coded as men = 1 and women = 0 and there were 80 men and 20 women, then the mean of the variable Sex would be .80 which is also the proportion of men in the sample. The proportion of women would then be 1 – 0.8 = 0.2. The mean of the binary variable, and therefore the proportion of 1s, is labeled P, with the proportion of 0s being labeled Q with Q = P – 1. In parametric statistics, the mean of a sample has an associated variance and standard deviation, so too does a binary variable. The variance is PQ, with the standard deviation being $\sqrt{PQ}$. P not only tells you the proportion of 1s but it also gives you the probability of selecting a 1 from the population. In our example you would have an 80% chance of selecting a man and a 20% chance of selecting a woman if you randomly selected from the population.

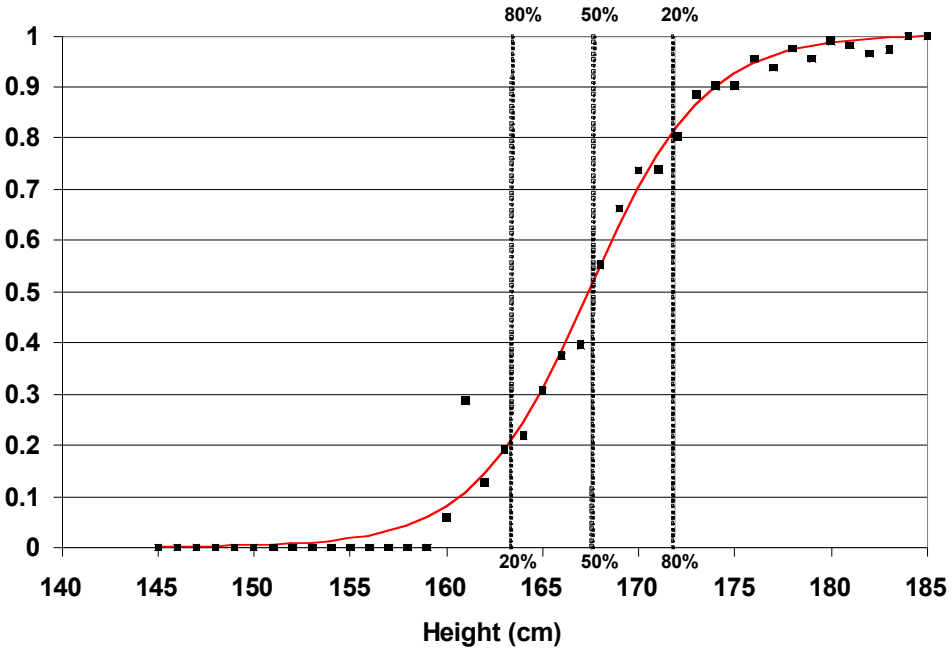

Figure 2-9.11: Logistic curve fitting through rolling means of binary variable sex (1=men, 0=women) versus height category in cm.

On average adult men are taller than adult women, but can you predict whether someone is a man or a woman based upon their height alone. Figure 2-9.11 shows a plot of the mean value (P) of the binary variable Sex (men = 1, women = 0) by 1 cm increments in height for the adults (age ≥ 18 years) from the Canada Fitness Survey data (see Appendix A). This mean is also the proportion of men in each height group. The data produce a sigmoidal curve. The vertical dashed lines show those heights where 80%, 50% and 20% respectively of the sample are men.

P	Q	PQ Variance
0	1	0
.1	.9	.09
.2	.8	.16
.3	.7	.21
.4	.6	.24
.5	.5	.25
.6	.4	.24
.7	.3	.21
.8	.2	.16
.9	.1	.09
1	0	0

Table 2-9.4; Variance for different values of P and Q

There are several reasons why logistic regression should be used rather than ordinary linear regression in the prediction of binary variables:

- Predicted values of a binary variable can not theoretically be greater than 1 or less than 0. This could happen however, when you predict the dependent variable using a linear regression equation. If you make the dependent variable large or small enough this would happen.
- It is assumed in linear regression that the variance of Y is constant across all values of X. This is referred to as homoscedasticity. Remember that the variance of a binary variable is

PQ. Therefore, the variance is dependent upon the proportion at any given value of the independent variable. Table 2-9.4 shows the variance for different values of P and Q. Note that the variance is greatest when 50% are 1s and 50% are 0s. Variance reduces to 0 as P reaches 1 or 0. This variability of variance is referred to as heteroscedasticity.

- Linear regression assumes that the residuals are normally distributed, but this is clearly not the case when the dependent variable can only have values of 1 or 0.

The Logistic Curve

The logistic curve relates the independent variable, X, to the rolling mean of the dependent variable, P ($\bar{Y}$). The formula to do so may be written as either

$$P = \frac{1}{1 + e^{-(a+bX)}}$$

or as

$$P = \frac{e^{a+bX}}{1 + e^{a+bX}}$$

where ***P*** is the probability of a 1 (the proportion of 1s, the mean of ***Y***), ***e*** is the base of the natural logarithm (about 2.718) and ***a*** and ***b*** are the parameters of the model. The value of ***a*** yields ***P*** when ***X*** is zero, and ***b*** adjusts how quickly the probability changes with changing ***X*** a single unit. Because the relation between ***X*** and ***P*** is nonlinear, ***b*** does not have a straightforward interpretation in this model as it does in ordinary linear regression.

Maximum Likelihood

When we were dealing with linear regression the best fitting line was based on the least squares approach. The least sum of squares is referred to as a loss function. The loss function quantifies the goodness of fit of the equation to the data. Unfortunately, the logistic curve we are required to fit in logistic regression is nonlinear. What this means is that we can not use least sum of squares as our loss function, nor indeed is there any mathematically defined loss function that can be used. For logistic curve fitting and other nonlinear curves the method used is called **maximum likelihood**. For the logistic curve to fit we need to find the appropriate values of a and b. In the procedure, values for a and b are picked randomly and then the likelihood of the data given those values of the parameters is calculated. The values are changed and the likelihood compared to see if it has increased. If it has, another series of changes are made to further increase the likelihood of the data. If not, changes are made in the opposite direction to increase the likelihood. Each one of these changes is called an iteration.

The process continues, iteration after iteration, until the largest possible value or Maximum Likelihood has been found. Normally, criteria are set as to the number of iterations allowed or a limit to the increase in likelihood from iteration to iteration.

Odds & log Odds

Suppose we only know a person's height and we want to predict whether that person is a man or a woman. We can talk about the probability of being a man or a woman, or we can talk about the odds of being a man or a woman. Let's say that the probability of being a man at a given height is .90. Then the odds of being a man would be

$$Odds = \frac{P}{1-P} = \frac{0.9}{1-0.9} = 9/1$$

That means that the odds of being a woman would be .11 (.10/.90). This asymmetry is unappealing, because the odds of being a man should be the opposite of the odds of being a woman. We can take care of this asymmetry though the natural logarithm, ln. The natural log of 9 is 2.217 [ln(.9/.1)=2.217]. The natural log of 1/9 is -2.217 [ln(.1/.9)=-2.217], so the log odds of being a man is exactly opposite to the log odds of being a woman.

In logistic regression, the dependent variable is a logit or log odds, which is defined as the natural log of the odds:

$$\log(odds) = \text{logit}(P) = \ln\left(\frac{P}{1-P}\right)$$

In logistic regression, we find $\boldsymbol{logit(P) = a + bX}$. The log odds (logit) is assumed to be linearly related to X, the independent variable. In order to get to probabilities take the log out of both sides of the equation and convert odds to a simple probability:

$$\ln\left(\frac{P}{1-P}\right) = a + bX \qquad \frac{P}{1-P} = e^{a+bX} \qquad P = \frac{e^{a+bX}}{1+e^{a+bX}}$$

The simple probability is this ugly equation that you saw earlier. If log odds are linearly related to X, then the relation between X and P is nonlinear, and has the form of the S-shaped curve you saw in the graph and the function form (equation) shown immediately above.

The Odds Ratio

It was stated earlier that the odds for one group is :

$$Odds = \frac{P}{1-P}$$

Table 2-9.5 shows the results of the odds ratio calculation for the occurrence of heart attack in two samples of patients, one undergoing drug treatment, the other not.

	Heart Attack	No Heart Attack	Probability	Odds
Treatment	3	6	3/(3+6)=0.33	0.33/(1-0.33) = 0.50
No Treatment	7	4	7/(7+4)=0.64	0.64/(1-0.64) = 1.75
			Odds Ratio	1.75/0.50 = **3.50**

Table 2-9.5; Odds ratio calculation for occurrence of heart attack in patients with and without drug treatment

The odds of having a heart attack for the treatment group are 3/6 = 0.5. The probability of a heart attack is 3/(3+6) = 3/9 = .33. The odds from this probability are .33/(1-.33) = .33/.66 = 0.5 The odds for the no treatment group are 7/4 or 1.75. The odds ratio therefore would be 1.75/0.5 = 3.50. This would indicate that the individuals in the no treatment group were 3.5 times more likely to have a heart attack than the treatment group.

Using SPSS to Calculate Logistic Regression

Logistic regression is found as the BINARY LOGISTIC option under the ANALYZE menu as shown in Figure 2-9.12. This will bring up the binary logistic regression dialog box shown in Figure 2-9.13. You select your dependent variable from the list of variables of the left. The dependent variable must be a binary variable. The covariates are those variables that you wish to use as predictors and find odds ratios for.

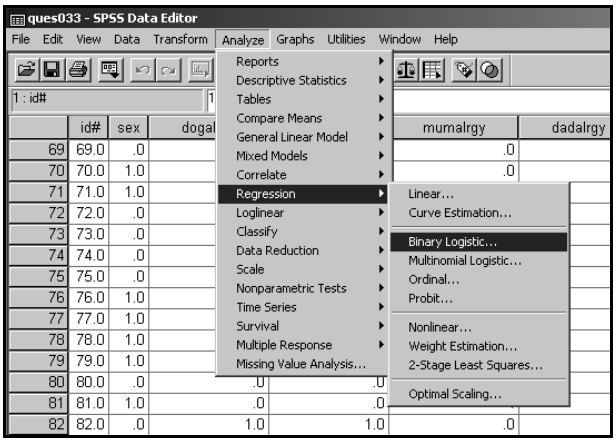

Figure 2-9.12: SPSS dialog box for he BINARY LOGISTIC option under the ANALYZE menu

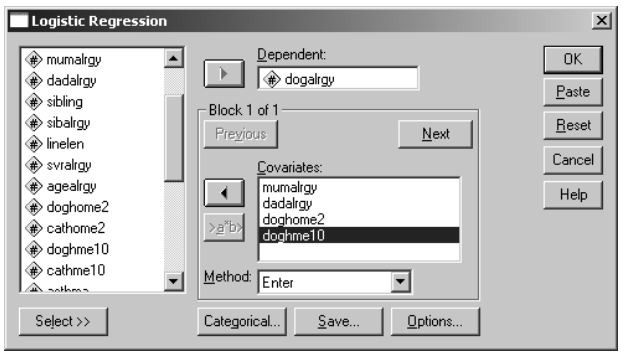

Figure 2-9.13: SPSS dialog box for he BINARY LOGISTIC option under the ANALYZE menu

In this example we are looking at data from a questionnaire where respondents were asked about allergies to cats and dogs, in addition to information on allergies in family members and previous exposure to cats and dogs. This was a questionnaire put together by students as a class assignment to learn about questionnaire design and analysis. They wanted to know if having a dog or cat in the house as a child or having parents with allergy were contributory factors to the respondent having an allergy. This was a small sample survey (n = 169)

Figure 2-9.14 shows the results of one of the analyses on this data. The three variables shown in this analysis in Figure 2-0.14 were:

catalrgy: Do you have an allergy to cats (No = 0, Yes = 1)

mumalrgy: Does your mother have an allergy to cats (No = 0, Yes = 1)

dadalrgy: Does your father have an allergy to cats (No = 0, Yes = 1)

The analysis shows the odds ratios (exp(B)) for the binary logistic regression where the presence of a cat allergy was predicted by whether the mother or father had a cat allergy. The results showed that respondents with mothers who had a cat allergy were 4.457 times more likely to have a cat allergy than those respondents who had mothers who did not have a cat allergy (odds ratio = 4.457, $p = 0.033$, therefore $p < 0.05$ significant). Whereas respondents with fathers who had a cat allergy were not more likely to have a cat allergy than those

respondents who had fathers who did not have a cat allergy (odds ratio = 7.393, p = 0.068, therefore p > 0.05 not significant)

Variables in the Equation

		B	S.E.	Wald	df	Sig.	Exp(B)
Step 1[a]	MUMALRGY	1.494	.702	4.534	1	.033	4.457
	DADALRGY	2.000	1.096	3.329	1	.068	7.393
	Constant	-.056	.297	.035	1	.852	.946

a. Variable(s) entered on step 1: MUMALRGY, DADALRGY.

Figure 2-9.14: Binary Logistic Regression of CATALRGY predicted from MUMALERGY and DADALRGY in student allergy questionnaire data