***Analysing frequency data with Chi-Square:***

These days, it seems every major conurbation is prone to attack from villains with super-powers. This is especially true of New York and San Francisco, where they seemingly have to repair their major bridges and skyscrapers on a weekly basis after some Transformer, Silver Surfer or Godzilla has knocked them down. London too has its problems, with Daleks, Cybermen and giant Simon Cowells causing structural damage to Big Ben and other hallowed landmarks.

Open the datafile on my website called "[***STPROB2.sav (SPSS data for problem sheet 2)***](http://www.sussex.ac.uk/Users/grahamh/RM1web/Week%206%202012/STPROB1.sav)***.***It contains all of the data that you need for this set of exercises.

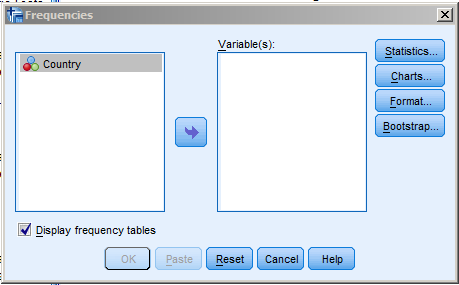
**A worked example of the Chi-Square Goodness of Fit test:**

The data in the column entitled "**Country**" consist of information about attacks by various super-villains during the past five years. Each time a country's name is mentioned in this column, it means that one attack has occurred in that country during the past five years. (Thus there have been 14 attacks in America, 8 in Australia, and so on). Imagine you are the International Monetary Fund, and your job is to loan countries money to pay for repairs. Your null hypothesis (your starting assumption) is that each country has equal need of money because super-villain attacks are equally distributed around the world. However, North America and Britain claim that they deserve more money from the IMF because attacks are not random - they claim that they are especially prone to attacks.

|  |  |
| --- | --- |
| Country | Number of attacks in past 5 years |
| America | 14 |
| Australia | 8 |
| Belgium | 1 |
| Britain | 12 |
| France | 7 |
| India | 6 |
| Japan | 16 |
| South Africa | 1 |
| Sweden | 0 |

(a) First, find out how many attacks have taken place in each country.

To do this, click on **Analyze**, then **Descriptive Statistics**, then **Frequencies**...



Click on the central arrow to put "Country" into "Variable(s). Then click on the "Statistics..." button and put a tick in the box that says "Sum". (These are frequency data, so it would be meaningless to work out a mean or median). Click on "Continue" and then "OK". You should get the following table in the output:

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Country** | | | | | |
|  | | Frequency | Percent | Valid Percent | Cumulative Percent |
| Valid | America | 14 | 21.5 | 21.5 | 21.5 |
| Australia | 8 | 12.3 | 12.3 | 33.8 |
| Belgium | 1 | 1.5 | 1.5 | 35.4 |
| Britain | 12 | 18.5 | 18.5 | 53.8 |
| France | 7 | 10.8 | 10.8 | 64.6 |
| India | 6 | 9.2 | 9.2 | 73.8 |
| Japan | 16 | 24.6 | 24.6 | 98.5 |
| SouthAfrica | 1 | 1.5 | 1.5 | 100.0 |
| Total | 65 | 100.0 | 100.0 |  |

The column headed "Frequency" tells you how often each country was attacked; and "Percent" expresses this as a percentage of the total number of attacks (66).

(The "Valid Percent" values will be identical to the "Percent" values unless you have missing data, in which case the "Valid Percent" column will consist of percentages calculated without talking the missing data into account. "Cumulative percent" is meaningless in this particular example, as our countries do not form categories that can be arranged in any logical sequence. However this column might be useful if it did make sense to order your categories in sequence - for example if they were the number of people scoring "1-5", "6-10", "11-15" and so on).

Looking at the frequencies, it appears that attacks are not equally spread across these eight countries: some countries (America, Australia, Britain, and Japan) seem to be attacked frequently while others (Belgium and South Africa) are attacked less.

There are two possible explanations for this pattern of attacks.

One explanation is that the observed pattern occurred merely by chance: attacks are in fact randomly distributed across the different countries. The true state of affairs is that each country is attacked equally often, and the observed pattern of attacks merely represents random deviations from the true state of affairs - deviations that have occurred purely by chance. This is our **null hypothesis**, that we initially assume to be true unless we can find evidence to the contrary.

The other explanation is that the observed pattern did not occur by chance: attacks are non-random, and some countries are indeed attacked more often than others. This is our **alternative hypothesis**.

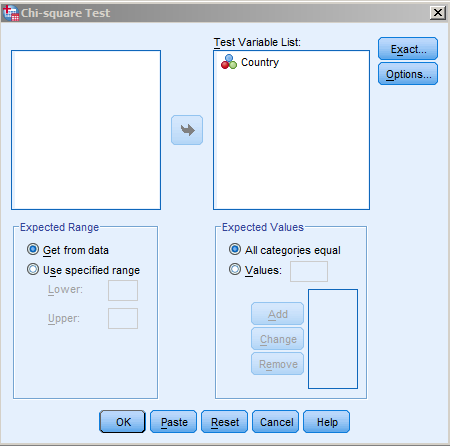
(b) To decide between these alternatives, we can use the **Chi-Square Goodness of Fit test** on these data. This will enable us to see how likely it is to get a pattern of results like the one we have obtained, purely by chance. In its most commonly-used form, the Chi-Square Goodness of Fit test looks at the extent to which the observed category frequencies differ from what you would expect to get if all of the categories were equally likely to occur. It compares an actual distribution to an expected distribution (usually a "rectangular" distribution, so-called because all of the expected frequencies are the same).

Each bar in this graph shows the actual frequency of attacks in each country. The line shows the expected frequency (about 8 attacks in each country - obtained by dividing the total (65 attacks) by the number of categories (8 countries). You can see that attacks are much more frequent than we would expect in some countries, and much less frequent in others.

We will perform a Chi-Square Goodness of Fit test on these data, using expected frequencies based on the null hypothesis that super-villain attacks are equally likely at all of these locations.

Go to Analyze.., then Nonparametric Tests, then Legacy Dialogs, then Chi-Square.

In the dialog box that pops up, put "Country" into "Test Variable List" and then click on "OK".



You should get output that consists of two tables, one containing the frequency of attacks, and the other containing the details of the Chi-Square test:

|  |  |
| --- | --- |
| **Test Statistics** | |
|  | Country |
| Chi-Square | 26.938a |
| df | 7 |
| Asymp. Sig. | .000 |
| a. 0 cells (0.0%) have expected frequencies less than 5. The minimum expected cell frequency is 8.1. | |

|  |  |  |  |
| --- | --- | --- | --- |
| **Country** | | | |
|  | Observed N | Expected N | Residual |
| America | 14 | 8.1 | 5.9 |
| Australia | 8 | 8.1 | -.1 |
| Belgium | 1 | 8.1 | -7.1 |
| Britain | 12 | 8.1 | 3.9 |
| France | 7 | 8.1 | -1.1 |
| India | 6 | 8.1 | -2.1 |
| Japan | 16 | 8.1 | 7.9 |
| SouthAfrica | 1 | 8.1 | -7.1 |
| Total | 65 |  |  |

The frequency table has three columns. The first contains the same raw frequencies as we calculated before. The second shows the expected frequency for each county. The third column shows the "residuals", simply the observed frequency minus its associated expected frequency. Positive residuals mean that the observed frequency is bigger than you would expect to find by chance, and negative residuals mean that the observed frequency is smaller than you would expect to find by chance.

The "Test Statistics" table gives you the value of Chi-Square; the degrees of freedom (which is the number of categories minus one); and the probability of getting a Chi-Square value as large as this one merely by chance. Note that SPSS can only show numbers to four significant digits. ".000" is displayed here because the probability of getting a Chi-Square value as large as this is SO small, that it's gone off the scale - SPSS cannot display it. In practice you would write this as "*p* < .001".

Finally, there is a note to tell you that SPSS has checked one of the assumptions underlying the valid use of Chi-Square: that none of the expected frequencies was less than 5. (It doesn't matter what the *observed* frequencies are).

So, we would write the results of the Chi-Square test something like this:

"A Chi-Square test showed that the observed distribution of attacks differed significantly from what would be expected by chance (*Χ*2(7) = 26.94, *p* < .0001)".

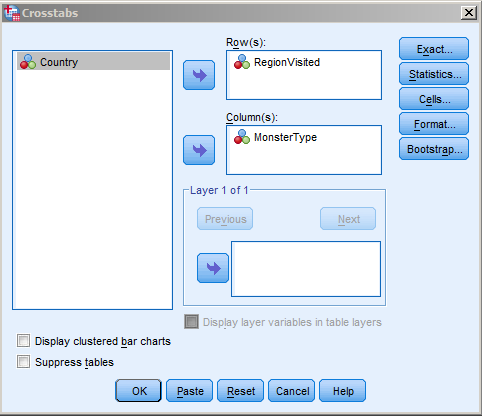
We would conclude that super-villain attacks do not occur randomly: some countries are more prone to attack than others. The Chi-Square merely tells us that the pattern of attacks is significantly different from what we would expect to find if all locations were equally prone to attack. However in this case it's pretty clear that America, Britain and Japan have more attacks than one would expect to find by chance, and Belgium, France, South Africa and Sweden have fewer. The IMF should give more money to the countries that are attacked most, to help them pay for repairs.

**A worked example of the Chi-Square Test of Association:**

Do monsters have regional preferences for where they attack? We could look at this by using Chi-Square to see if there is a statistically-significant association between "type of monster" (Aliens and Predators) and "country" (America, Britain and China). Our *null hypothesis* is that there is no association between type of monster and region visited: Aliens and Predators should crop up roughly equally as often in all three countries. The *alternative hypothesis* is that there is an association between type of monster and region. The Chi-Square test of Association will help us to decide which of these two hypotheses is more plausible.

We have two variables: "type of monster" (with two levels, "Alien" and "Predator") and "region visited (with three levels, America, Britain and China). In the SPSS spreadsheet, we therefore have two columns, "**MonsterType**" and "**RegionVisited**".

Go to "**Analyze..**.", then "**Descriptive Statistics**", then "**Crosstabs**". Put one variable into "Row(s)" and the other variable into "Column(s)".



Click on "Statistics..." and make sure there is a tick next to "Chi-Square". Then click "Continue" and then "OK".

You should get the following output:

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **RegionVisited \* MonsterType Crosstabulation** | | | | |
| Count | | | | |
|  | | MonsterType | | Total |
| Aliens | Predators |
| RegionVisited | America | 5 | 16 | 21 |
| Britain | 3 | 11 | 14 |
| China | 17 | 13 | 30 |
| Total | | 25 | 40 | 65 |

|  |  |  |  |
| --- | --- | --- | --- |
| **Chi-Square Tests** | | | |
|  | Value | df | Asymp. Sig. (2-sided) |
| Pearson Chi-Square | 7.822a | 2 | .020 |
| Likelihood Ratio | 7.961 | 2 | .019 |
| Linear-by-Linear Association | 6.096 | 1 | .014 |
| N of Valid Cases | 65 |  |  |
| a. 0 cells (0.0%) have expected count less than 5. The minimum expected count is 5.38. | | | |

The first table shows us the raw frequencies (the observed numbers of Aliens and Predators visiting each region).

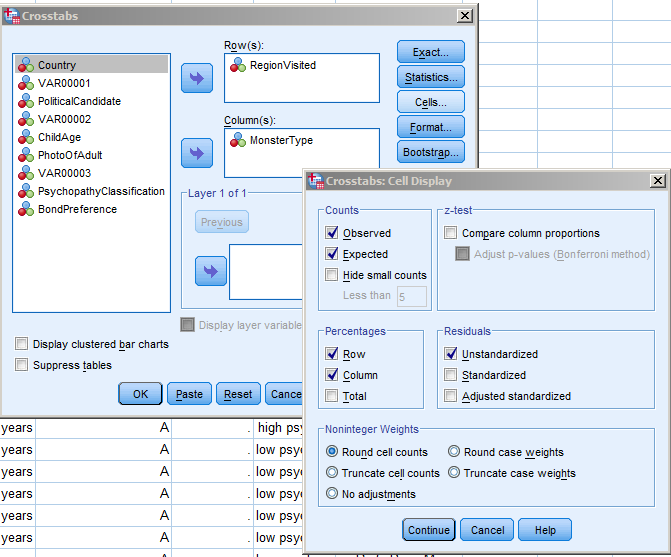
The second table gives the results of a number of tests, but we just want the top row, entitled somewhat confusingly "Pearson Chi-Square" (most people just call it "Chi-Square" and reserve Pearson's name for his correlation test).

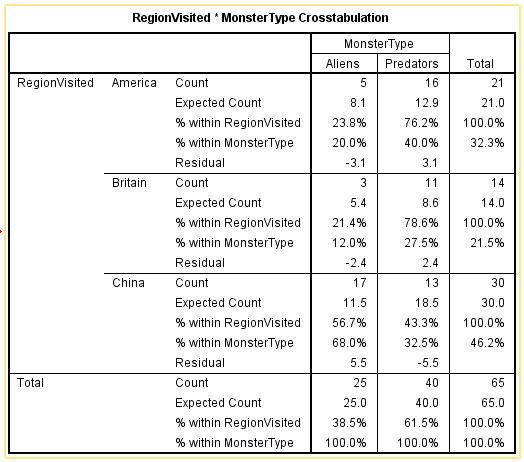
This shows us that the value of Chi-Square is 7.82, with 2 degrees of freedom. A value of Chi-Square as large as this is likely to occur by chance with a probability of .02 - i.e. it's not very likely!

We could write this as follows:

"There was a significant association between type of monster and region visited (*Χ*2(2) = 7.82, *p* = .02). Inspection of the data suggests that Britain and America tended to be visited more often by Predators than by Aliens. China showed a different pattern, with similar numbers of visits by Aliens and Predators (and more visits overall than either Britain or America)".

When you use the "crosstabs" command, SPSS doesn't provide the expected frequencies and the residuals by default. To obtain them, click on "Cells..." and then tick the boxes for the things you want - I like to get raw observed and expected frequencies; observed and expected frequencies expressed as percentages; and "unstandardised residuals" (the differences between the observed and expected frequencies).





You can see from the residuals that for Britain and America, Aliens visit a little less than would be expected by chance, and Predators visit a little more. For China, the pattern goes the other way, with Aliens visiting more than expected, and Predators less.

***Examples for you to try:***

1. A sample of 100 voters are asked which of four candidates they would vote for in an election. The number supporting each candidate is given below. The raw data are in "**PoliticalCandidate**".

Windy Miller Noddy Basil Brush Zippy

41 19 24 16

Do the data suggest that all candidates are equally popular?

2. Children of three ages are asked to indicate their preference for three photographs of adults. Do the data suggest that there is a significant association between age and photograph preference? The data are in "**ChildAge**" and "**PhotoOfAdult".**

Photograph:

A B C

Age of child: 5-6 years: 18 22 20

7-8 years: 2 28 40

9-10 years: 20 10 40

3. Some of the versions of James Bond have been rather bland (e.g. Roger Moore's interpretation of him) while others are much more hard-edged and violent (e.g. Daniel Craig). A psychologist decides to use this as a test of psychopathy, reasoning that people at one end of the psychopathy spectrum will prefer Roger Moore to Daniel Craig, whereas people at the other end will have the opposite preference. Fifty-nine people are given a psychopathy test and rated as "high" or "low". They are also asked whether they prefer Moore or Craig. Is there a significant association between psychopathy and Bond-preference? The data are in "**PsychopathyClassification**" and "**BondPreference"**.

|  |  |  |
| --- | --- | --- |
|  | Bond preference: | |
|  | Roger Moore | Daniel Craig |
| High psychopathy score: | 11 | 15 |
| Low psychopathy score: | 19 | 14 |

Note that this is a 2x2 Chi-Square test of Association, and so you need to decide whether or not to use "Yates' correction for continuity".

**Chi-Square in the 1 d.f. case, and the use of Yates' Correction:**

If you have only 1 d.f., as in the case of a 2x2 contingency table, some textbooks suggest that you apply what's known as "Yates' Correction" to the Chi-Square formula. When the d.f. are very small, (and you can't get much smaller than 1!), the Chi-Square sampling distribution becomes increasingly distorted. As a consequence, the obtained value of Chi-Square tends to overestimate the "true" discrepancy between the observed and expected frequencies. Yates' correction is a simple bodge that makes the formula produce a lower value of Chi-Square than it otherwise would do. This makes the Chi-Square test more "conservative": i.e., it makes it harder to get a statistically significant result. Here is the "corrected" formula:



In English, this means you do the following:

(a) Get the absolute value (i.e., ignore the sign) of each observed frequency minus its accompanying expected frequency. (The vertical lines mean "ignore the sign of").

(b) From each of these absolute values, subtract 0.5.

(c) Divide each value in step (b) by the associated expected frequency.

(d) Add together all of the results of step (c).

(e) Finally, look up the value of Chi-Square in the usual way.

Statisticians are divided about the need for Yates' correction: some say you should use it, others say it doesn't matter, and some say it makes things worse! If you use the SPSS crosstabs command to perform a Chi-Square test on a 2x2 table as in this example), SPSS calculates Chi-Square using both methods. The top line of the output table shows "Pearsons Chi-Square" and this is the *uncorrected* version of Chi-Square (i.e. calculated without using Yates' correction). The line beneath this, entitled "Continuity Correction", gives Chi-Square using Yates' correction. Compare the two results. What happens in this case? Either version of the Chi-Square result is acceptable, as long as you make it clear whether or not you used Yates' correction. Remember - if you do use it, it is *only* to be used when you have just one degree of freedom.

4. We want to test whether short people differ with respect to their leadership qualities (Genghis Khan, Adolf Hitler and Napoleon were all stature-deprived, and it is a little-known fact that Alan Sugar is actually only 2 feet high - how cameras lie!) The following table shows the frequencies with which 23 short people and 21 tall people were categorised as "leaders" or "followers". Is there an association between height and leadership qualities? The data are in "**height**" and "**leadership**".

|  |  |  |
| --- | --- | --- |
|  | Leadership qualities: | |
|  | Leaders | Followers |
| Short: | 15 | 8 |
| Tall: | 5 | 16 |

**Solutions to Problem Sheet 2, Chi-Squared tests:**

**Question 1:**

A Chi-Squared Goodness-of-Fit test is appropriate here. The null hypothesis is that there is no preference for any of the candidates: if this is so, we would expect roughly equal numbers of voters to support each candidate. Our expected frequencies are therefore 100/4 = 25 per candidate.

|  |  |  |  |
| --- | --- | --- | --- |
| **PoliticalCandidate** | | | |
|  | Observed N | Expected N | Residual |
| Windy Miller | 41 | 25.0 | 16.0 |
| Noddy | 19 | 25.0 | -6.0 |
| Basil Brush | 24 | 25.0 | -1.0 |
| Zippy | 16 | 25.0 | -9.0 |
| Total | 100 |  |  |

|  |  |
| --- | --- |
| **Test Statistics** | |
|  | PoliticalCandidate |
| Chi-Square | 14.960a |
| df | 3 |
| Asymp. Sig. | .002 |
| a. 0 cells (0.0%) have expected frequencies less than 5. The minimum expected cell frequency is 25.0. | |

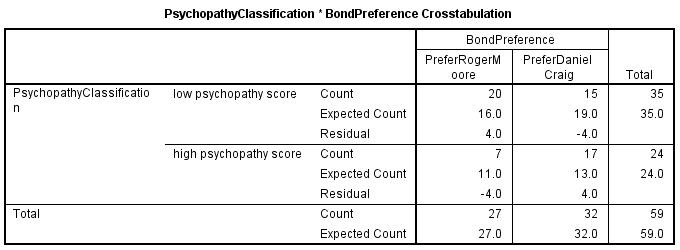
Our obtained value of Chi-Square is 14.96, with 3 d.f. Its probability of occurring by chance is .002 - 2 times in a thousand. It seems more reasonable to conclude that our results are not due to chance, and that the data do indeed suggest voters do not prefer the four candidates equally.

**Question 2:**

|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **ChildAge \* PhotoOfAdult Crosstabulation** | | | | | | | | | |
| Count | | | | | | | | | |
|  | | PhotoOfAdult | | | | | | Total | |
| A | | B | | C | |
| ChildAge | 5-6 years | 18 | | 22 | | 20 | | 60 | |
| 7-8 years | 2 | | 28 | | 40 | | 70 | |
| 9-10 years | 20 | | 10 | | 40 | | 70 | |
| Total | | 40 | | 60 | | 100 | | 200 | |
| **Chi-Square Tests** | | | | | | | | |
|  | | | Value | | df | | Asymp. Sig. (2-sided) | |
| Pearson Chi-Square | | | 29.603a | | 4 | | .000 | |
| Likelihood Ratio | | | 36.327 | | 4 | | .000 | |
| Linear-by-Linear Association | | | 2.834 | | 1 | | .092 | |
| N of Valid Cases | | | 200 | |  | |  | |
| a. 0 cells (0.0%) have expected count less than 5. The minimum expected count is 12.00. | | | | | | | | |

Our obtained Chi-Square value is 29.60, with 4 d.f. It is so unlikely to have occurred by chance that SPSS cannot display the *p*-value (.000, which you could write as "*p* < .0001"). It seems more reasonable to accept the alternative hypothesis, that there is a significant relationship between age of child and photograph preference.

**Question 3:**

Here we have a 2x2 contingency table. Chi-Square is the appropriate test to use, but since we have 1 d.f., we could use the modified version of the formula that includes "Yates' correction for continuity".

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Chi-Square Tests** | | | | | |
|  | Value | df | Asymp. Sig. (2-sided) | Exact Sig. (2-sided) | Exact Sig. (1-sided) |
| Pearson Chi-Square | 4.489a | 1 | .034 |  |  |
| Continuity Correctionb | 3.433 | 1 | .064 |  |  |
| Likelihood Ratio | 4.589 | 1 | .032 |  |  |
| Fisher's Exact Test |  |  |  | .062 | .031 |
| Linear-by-Linear Association | 4.413 | 1 | .036 |  |  |
| N of Valid Cases | 59 |  |  |  |  |
| a. 0 cells (0.0%) have expected count less than 5. The minimum expected count is 10.98. | | | | | |
| b. Computed only for a 2x2 table | | | | | |

If we use the uncorrected version of the formula, we find that there is a significant association between high/low psychopathy and Bond preference, in the direction that we predicted (i.e. low psychopaths tend to prefer Roger Moore, and high psychopaths prefer Daniel Craig). *Χ*2 (1) = 4.49, *p* = .03. However, if we use Yates' correction, the value of Chi-Square is reduced: *Χ*2 (1) = 3.43, *p* = .06. This shows how Yates' correction makes the Chi-Square test slightly more conservative - i.e. it slightly reduces the chances of obtaining a statistically significant value of Chi-Square by making Chi-Square smaller than it would otherwise have been. In psychology, we use the convention that a result is statistically "significant" if it is likely to occur by chance with a probability of .05 or less. Thus, in this case, if we used the uncorrected version of Chi-Square, we would conclude that our observed pattern of preferences was unlikely to have occurred by chance. In contrast, if we used the corrected version, we would conclude that it might well have occurred by chance - that there was no reason to suppose that the observed frequencies had not arisen merely by chance. Which is the correct interpretation? It depends on whether you think Yates' correction should be used with 2x2 tables. Probably the best course of action would be to re-run this study with a bigger sample size, to see if you could replicate the effect.

**Question 4:**

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **height \* leadership Crosstabulation** | | | | | |
|  | | | leadership | | Total |
| leader | follower |
| height | short | Count | 15 | 8 | 23 |
| Expected Count | 10.5 | 12.5 | 23.0 |
| Residual | 4.5 | -4.5 |  |
| tall | Count | 5 | 16 | 21 |
| Expected Count | 9.5 | 11.5 | 21.0 |
| Residual | -4.5 | 4.5 |  |
| Total | | Count | 20 | 24 | 44 |
| Expected Count | 20.0 | 24.0 | 44.0 |

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Chi-Square Tests** | | | | | |
|  | Value | df | Asymp. Sig. (2-sided) | Exact Sig. (2-sided) | Exact Sig. (1-sided) |
| Pearson Chi-Square | 7.591a | 1 | .006 |  |  |
| Continuity Correctionb | 6.013 | 1 | .014 |  |  |
| Likelihood Ratio | 7.860 | 1 | .005 |  |  |
| Fisher's Exact Test |  |  |  | .008 | .007 |
| Linear-by-Linear Association | 7.419 | 1 | .006 |  |  |
| N of Valid Cases | 44 |  |  |  |  |
| a. 0 cells (0.0%) have expected count less than 5. The minimum expected count is 9.55. | | | | | |
| b. Computed only for a 2x2 table | | | | | |

Here, regardless of whether or not we use Yates' correction for continuity, the conclusion is the same: there is a significant association between height and leadership (uncorrected, *Χ*2 (1) = 7.59, *p* = .006; corrected, *Χ*2 (1) = 6.01, *p* = .01. Note that we can only say that there is a relationship between our two variables, not that once *causes* the other. There could be all kinds of explanations for such a relationship.