WEB APPENDIX CHAPTER 5

Mooi, E., Sarstedt, M., & Mooi-Reci, I. (2018). *Market Research. The Process, Data, and Methods using Stata.* Heidelberg: Springer.

Nonparametric tests

In the book, we concentrated on parametric tests, which assume that variables tested are normally distributed. Parametric tests are, however, quite robust against violations of the normality assumption. Specifically, when comparing means, we can often assume that the dependent variable is at least asymptotically normally distributed (in non-statistical terms: close to being normally distributed) when the sample size exceeds 30 and the variable can be expected to be normally distributed. However, some variables are never normally distributed. For example, even for very high sample sizes, income is not normally distributed with a long tail to the right of the distribution. Similarly, a variable indicating whether customers booked a flight in the last month can only take on the values 0 (no) or 1 (yes) and will, hence, never be normally distributed. To cope with potential problems arising from a violation of the normality assumption, we can apply nonparametric tests, which do not assume a specific distribution and can be used on dependent variables measured on a nominal or ordinal scale. Researchers are often interested in examining nominal variables. A useful test is the χ^2 test (pronounced as *chi square*). There are two types of χ^2-tests involving nominal or ordinal data

1.) The *one-sample χ^2-test* (also called χ^2 goodness-of-fit test) can be used to test occurrences in a single variable's categories and compares those against expected occurrences.

2.) The *χ^2-test for independence* (also known as the Pearson Chi-square test) is used to determine whether two nominal (or ordinal) variables are related.

Let's first discuss the one-sample χ^2-test test. Suppose a mobile phone producer plans to launch a new smartphone on the market but is not sure which color to use. 200 people are chosen from the target group, of which 40 indicate "black," 84 "silver," and 76 "red." If we want to establish whether or not these preferences differ significantly from what could be expected *a priori*, then this can be determined using a one-sample χ^2-test. In the simplest case, we could expect all three colors to be equally popular, which means that we expect 66.67 respondents to choose each of these colors.[i] Comparing the sample values with these expected values, we observe a numerical difference between them. The question is whether this difference is attributable to occurring sample variation, or whether it is likely to hold for the population. If the latter is the case, we could conclude that silver is the preferred smartphone color in the population. This is what we are going to explore by means of the one-sample χ^2-test whose null hypothesis states that there is no difference between the expected and observed values.

We can test this null hypothesis using the χ^2-test statistic, which is calculated by collecting observed values for each of the categories and examining the differences between the observed and expected values:

$$\chi^2 = \sum_{i=1}^{k} \frac{(h_i - \tilde{h}_i)^2}{\tilde{h}_i},$$

where h_i is the observed value for category i and $\tilde{h}_i$ is the expected value for the i^{th} category, and k is the total number of categories. In other words, the χ^2-test statistic is the sum of the squared differences between the observed and expected values, divided by the expected values.

As mentioned above, we would expect there to be 200/3=66.67 respondents preferring each of the three colors, which yields the following:

$$\chi^2 = \frac{(40 - 66.67)^2}{66.67} + \frac{(84 - 66.67)^2}{66.67} + \frac{(76 - 66.67)^2}{66.67} = 10.67 + 4.50 + 1.31 = 16.48$$

The test value is not directly interpretable but must be compared to the critical value obtained from the χ^2-statistic with *k-1* (in our example 3 - 1 = 2) degrees of freedom (see Table A3 in the Web Appendix (→Chapters→Additional Material)). As the test statistic value (16.48) is much higher than the critical value (5.991; α=0.05), we can reject the null hypothesis and, thus, conclude that the preferences differ significantly from what could be expected. In this example, we assumed that each category has the same expected frequency (i.e. 66.67). However, we could similarly pre-specify the expected frequencies and test certain assumption regarding the proportions.

Let's now discuss the χ^2-test for independence. In the previous example, we considered only one sample, but researchers are frequently interested in evaluating whether there is a significant relationship between two nominal variables. Suppose that we further differentiated the survey described above by distinguishing between male and female respondents. A possible crosstab is presented in Table A6.1 (please ignore the $\tilde{h}$ values in the table for the time being):

	Male	Female	Σ
Black	28 $\tilde{h}_{11} = 20$	12 $\tilde{h}_{12} = 20$	40
Silver	48 $\tilde{h}_{21} = 42$	36 $\tilde{h}_{22} = 42$	84
White	24 $\tilde{h}_{31} = 38$	52 $\tilde{h}_{32} = 38$	76
Σ	100	100	200

Table A6.1: Crosstab for χ^2-test for independence

This 3x2 crosstab indicates that 28 male respondents prefer the black smartphone, 48 the silver smartphone, and 24 the white smartphone. The last column (row) indicates the column

(row) total indicated by the summation signs (Σ). In this sample, there are 100 males and 100 females. To answer the research question whether there is a relationship between the respondents' gender and their color preferences, we can apply the *χ^2-test for independence*. Specifically, it tests the following hypotheses:

H_0: Preference for color is independent of gender

H_1: Preference for color is dependent of gender

As in the one-sample case, this test examines the degree to which the observed frequencies deviate from the expected frequencies. The expected frequency of a cell $\tilde{h}_{ij}$ (*i* being the index of the first variable with *k* categories and *j* being the index of the second variable with *m* categories) is the column total times the row total divided by the number of observations. This seems complicated but is not. For example, the expected frequency of the cell male/black $\tilde{h}_{11}$ is 100 (column total category "male") times 40 (row total of the category "black"), divided by 200 (the total number of observations), which equals 20. Similarly, the expected frequency of the cell silver/male is $\tilde{h}_{21} = \frac{100 \cdot 84}{200} = 42$, and so on (see Table A6.1 for all cells' expected frequencies).

The computation of the χ^2-test statistic is similar to the example above, with the only exception that we have to append a second summation sign as there are now two nominal variables:

$$\chi^2 = \sum_{i=1}^{k} \sum_{j=1}^{m} \frac{(h_{ij} - \tilde{h}_{ij})^2}{\tilde{h}_{ij}} = \frac{(28-20)^2}{20} + \ \ldots + \frac{(52-38)^2}{38} = 18.43$$

The degrees of freedom are calculated as follows: *df = (k-1)·(m-1)*. This example has 2 degrees of freedom and the associated critical value (for $\alpha = 0.05$) of 5.991 is much smaller than the test value (18.43). Thus, we can assume that there is a significant relationship between the respondents' gender and preference for color.

There are several statistical measures that provide us with information regarding the strength of the association between nominal variables. Their computation only makes sense of course if the χ^2-test renders significant results.

- Fisher's exact test is a non-parametric test used only for 2 x 2 contingency table when the smallest expected value is <5. In a 2 x 2 table, the test assesses the probability of cell frequency given the marginal frequencies based on the hypergeometric probability distribution (McHugh, 2013).
- The *φ (phi) coefficient* is used to measure the strength of association in 2x2 crosstabs. In fact, it is only a correlation coefficient for nominal variables and is computed as follows:

  $$\varphi = \sqrt{\frac{\chi^2}{n}}$$

  If there is no association whatsoever (which is extremely unrealistic) between the variables, φ would be zero. Conversely, a value of 1 implies that the variables are perfectly associated. Generally, a φ value below 0.30 describes a weak association, between 0.30 and 0.49 a moderate one, and above 0.50 a strong association.

- *Goodman and Kruskal gamma* is a measure of rank correlation that captures how closely two pairs of data points are matched when both variables are measured at an ordinal level. When two data points are similar in the orderings of the data, the sorted list of paired observations will be close to +1 and -1 if otherwise.
- *Cramer's V* is a modified version of the *φ (phi)* coefficient and is used in crosstabs larger than 2x2:

$$V = \sqrt{\frac{\chi^2}{n \cdot (r-1)}} = \sqrt{\frac{18.43}{200 \cdot (2-1)}} = 0.30$$

Other measures include the *likelihood-ratio chi*-squared (used when the data set is too small to meet the sample size assumption of the chi-square test that 80% of the cells have expected values of 5 or more), and *Tshuprow's T*. While all these measures are used to measure the strength of association between nominal variables, Stata provides different statistics such as *Kendall's τ-b* (pronounced as *tau*) for ordinal variables. Essentially, these measures use information on the ordering of variables' categories by considering every possible pair of cases in the crosstab. They vary between -1 and +1 and thus distinguish between positive and negative relationships. Higher absolute values denote a stronger degree of association. For more information, see, for example, Fleiss et al. (2003).

Box 6.A2: Measures of the strength of association

While the χ^2-test for independence (as well as Fisher's exact test) helps us explore the relationship between two nominal variables from independent samples, the *McNemar test* allows us to do this when the data stem from two paired samples. More specifically, the NcNemar test is used for dichotomous variables. For example, we might carry out an experiment in which we ask respondents whether they would buy a specific smartphone before and after being exposed to an online banner. The test's null hypothesis is that the number of respondents who changed their response in one direction (i.e. buy instead of not buy) is equal to the number of those who changed in the opposite direction (i.e. not buy instead of buy). The McNemar test compares the observed data to the null expectation, using a goodness-of-fit test and is interpreted like the tests discussed before. More on McNemar tests in Stata can be found here: https://stats.idre.ucla.edu/stata/whatstat/what-statistical-analysis-should-i-usestatistical-analyses-using-stata/.

So far, we looked at tests related to variables that are (at least) nominally scaled, but there are also various tests that are related to ordinal data. We will only look at these tests briefly, as their computation is often rather complex and goes beyond the scope of this book.

An important (nonparametric) test for normality is the *one-sample Kolmogorov–Smirnov (KS) test*. We can use it to test whether or not a variable is normally distributed. Technically, when assuming a normal distribution, the KS test compares the sample scores with an artificial set of normally distributed scores that has the same mean and standard deviation as the sample data. However, this approach is known to yield biased results, which are modified using the Lilliefors correction (1967). The Lilliefors correction takes into consideration that we do not know the true mean and standard deviation of the population. An issue with the KS test with the Lilliefors correction is that it is very sensitive when used on large samples and often rejects the null hypothesis if very small deviations are present. This also holds for

Stata's version of the KS test, which only works well for very large sample sizes (i.e., at least 10,000 observations). Consequently, Stata does not recommend the use of a one-sample KS test (for more, read the information in Stata's help file on the KS test: https://www.stata.com/manuals14/rksmirnov.pdf).

The *Shapiro-Wilk* test also tests the null hypothesis that the test variable under consideration is normally distributed. Thus, rejecting the Shapiro-Wilk test provides evidence that the variable is not normally distributed. It is best used for sample sizes of less than 50. A drawback of the Shapiro-Wilk test however, is that it works poorly if the variable you are testing has many identical values, in which case you should use the Kolmogorov-Smirnov test with Lilliefors correction.

The *Mann-Whitney U-test* is a nonparametric alternative to the independent samples t-test, which can be used if the dependent variable is measured on an ordinal scale. Furthermore, it is commonly applied in situations where the dependent variable is measured on an interval scale but does not follow a normal distribution. Like the t-test, it tests the null hypothesis that the difference in the location of two populations (expressed by the median) is zero. Rather than being based on means, the Mann-Whitney U-test statistic is based on a comparison of the observations' ranks. The corresponding method for paired samples is the *Wilcoxon signed-rank test*, which tests the null hypothesis that two medians stemming from paired samples are identical.

References for this Web Appendix

Fleiss, Joseph L., Bruce Levin and Myunghee C. Paik (2003). *Statistical Methods for Raters and Proportions*, 3rd. edition, New York et al.: Wiley.

Kirkman, T. W. (1996): Statistics to Use, http://www.physics.csbsju.edu/stats/exact_NROW_NCOLUMN_form.html.

Lilliefors, Hubert W. (1967). "On the Kolmogorov-Smirnov Test for Normality with Mean and Variance Unknown," *Journal of the American Statistical Association*, 62 (318), 399–402.

McHugh, Mary L. (2013). "The chi-square test of independence," *Biochem Med (Zagreb)*, 23(2):143–149.

[i] We could likewise take different expected frequencies (e.g. based on experience) and test against these frequencies.