Lab 4: Analyzing Proportions

Lab Objectives:

- Use SPSS to conduct binomial testing
- Use SPSS to compute a complete binomial distribution
- Use Excel to calculate confidence intervals for proportions using the Agresti-Coull method
- Practice using SPSS and Excel

Exercise 1: Binomial testing in SPSS

The binomial distribution

The binomial distribution describes the probability of a given number of successes out of *n* trials, where each trial independently has a *p* probability of success. It is given by the following equation:

$$\Pr[X\ successes] = \binom{n}{X} p^X (1-p)^{n-X}$$

.

The **binomial test** compares the observed number of successes in a data set to that expected under a null hypothesis. Today we will learn how to conduct this test in SPSS. We will also use additional data sets to practice using both SPSS and Excel.

Ear turn

Do people typically use a particular ear preferentially when listening to strangers? To answer this question, a researcher approached and spoke to strangers in a noisy nightclub, and an observer recorded whether the person turned their left or right ear toward the researcher. This data is found in the file *Earturn.sav*

- Open *Earturn.sav*
- In order to use this data in the binomial test, we have to change the ear data to numeric. We will use a value of 1 for right ear, and a value of 2 for left ear. Beginning with the data view, change all right ears to a value of 1 and all left ears to a value of 2. Note that since we have a relatively small data set this is fairly easy to do quickly. However, there is also a "replace" function under the Edit menu, if you would like to try that.
- Switch to variable view, and change the data to **numeric** (under type), and add the appropriate **values** for right ear and left ear.
- Select Analyze, **Nonparametric tests**, Legacy dialog, and Binomial.
- In the window that pops up, add Ear to your test variable list.

- **Note** that you can change the test proportion according to the null hypothesis. The null hypothesis for the ear turned is that there is no difference between left and right ears, so 0.5 is appropriate. However, this may change for other tests.

- Examine the output. It shows the observed proportion for each group (right ear and left ear), the test proportion that was used, and the exact significance (P-value) for a two-tailed test.

- Return to the data and change one right ear to a left ear. Re-run the Binomial test. Is the difference still significant? If so, return to the data and change another right ear to a left ear and re-run the binomial test until you find the point at which the proportion is no longer significantly different from the null hypothesis.

➢ Answer the first question on Sakai.

Spermatogenesis genes

A study of genes involved in spermatogenesis was carried out to test the hypothesis that spermatogenesis genes should occur disproportionately often on the X chromosome, which contains 6.1% of the total genes. The researchers found that 40% of the analyzed spermatogenesis genes were on the X chromosome. This data is found in the file *Spermatogenesis.sav*

- Open *Spermatogenesis.sav*

- First we will visualize this data. Select Analyze, Descriptive statistics, Frequencies.

- Add the variable Chromosome. Under charts select bar chart. Under format check that the order is "ascending values"

- Check in your frequency table that the percent of genes on the X is 40%. Use the bar chart to visualize the difference across the chromosomes.

- Return to the data view. Remember that we have to change the data to numeric in order to use the binomial test. We will change the variable "on X". Use a value of 1 for YES and a value of 2 for NO. Change the data using data view first.

- Switch to variable view, and change the data to numeric (under type), and add the appropriate values for YES and NO.

- The SPSS Binomial Test has a somewhat odd feature: the test proportion we enter applies to the category that's first encountered in the data. So the hypothesis that's tested depends on the order of the cases. Because our test proportion applies to the genes on the X (rather than *not* on the X), we need to move the genes on the X to the top of the data file. Highlight the "onX" column, right-click, and select Sort Ascending.

- Select Analyze, Nonparametric tests, Legacy dialog, and Binomial test.

- In the window that pops up, add onX to your test variable list.

- Note that this time you should change your test proportion. What is the test proportion of the null hypothesis in this case? Change it to the appropriate value then click OK.
- View the output. Note that the P-value provided in this case is **one-tailed**. When the test proportion is changed, SPSS returns a one-tailed P-value rather than two-tailed. We will have to multiply this value by 2 to get the actual two-tailed P-value. Since the value is too small for us to see, highlight the cell, select cell properties, and increase the decimals to 8 places to see the actual value for the one-tailed test. Verify that it is 9.9×10^{-7}.
- Now imagine we are interested in the proportion of spermatogenesis genes on the Y chromosome. Add an additional variable in the third column called "onY". Enter values of 1 (Yes) for the genes that are on the Y, and values of 2 (No) for the genes that are not on the Y.
- Add the appropriate labels in the variable view.
- The Y chromosome contains only 1.2% of the total genes in the mouse genome. Run the binomial test to determine if the proportion of spermatogenesis genes observed on the Y-chromosome is significantly different than expected under the null hypothesis. Increase the number of decimal places in the P-value cell to 5, so you have a clear view of the value.
- Make note of the observed proportion of genes on the Y chromosome and the P-value for a **two-tailed** test.

➢ Answer question 2 on Sakai.

Lost wallets

In a study in Scotland, researchers left a total of 240 wallets around Edinburgh, as though the wallets were lost. Each contained contact information including an address. Data on the number of wallets that were returned is found in the file *wallets.sav.*

- Open *wallets.sav*
- Note that in this file the data are already encoded with numeric values.
- Conduct a binomial test to determine if the number of wallets returned is significantly different from that expected under the null hypothesis that half of the wallets would be returned.

➢ Answer questions 3 & 4 on Sakai.

Exercise 2: Computing a binomial distribution in SPSS

One function of SPSS that we have not yet examined is its use to compute various statistical values as new variables in your data sheet. On test 1, there was an example in which 15 participants chose between two types of wine (cheap & expensive). If there was no preference between the two types of wine, then we would expect the proportion of individuals preferring each type of wine to be

approximately 0.5 (50%). The binomial distribution provides the probability distribution for the number of successes in a fixed number of trials, when the probability of success is the same in each trial. Applied to our example, the binomial distribution can tell us the probability of having a particular number of participants (e.g. 4, 8, 14) prefer the expensive wine over the cheap wine.

- Open a new data file in SPSS
- In variable view, label variable 1 "NumPick". Set to numeric & scale.
- In data view, enter the values from 0 to 15.
- Select transform, compute variable.
- For target variable, type in "Binom".
- Under function group, select "PDF & Noncentral PDF"
- Under functions and special variables, select "Pdf.Binom".
- Note that a box of text shows up describing the function, and how it should be used and entered. Press the up button to move it into the numeric expression box.
- The first question mark needs to be replaced with the number of successes. Because this is indicated by our NumPick variable column, highlight NumPick and click the right arrow to move it into the numeric expression.
- The next question mark needs to be replaced with the total n for the sample. So type 15 to replace this question mark.
- The final question mark needs to be replaced with the proportion of successes expected according to the null hypothesis. In this example, it should be 0.5.
- Click OK. A new variable labelled Binom should appear in your data sheet. This variable shows the probability of having 0 – 15 participants prefer one type of wine over the other, given a null hypothesis of no preference (p = 0.5).

Heterozygous crosses

In a cross between two heterozygotes, 25% of the offspring is expected to be homozygous recessive. In pea plants, green pod colour is dominant, and yellow pod colour is recessive. In a cross between two heterozygous pea plants you obtain 20 offspring. Compute the binomial distribution for obtaining X yellow plants out of n = 20.

➢ Answer questions 5 & 6 on Sakai.

Exercise 3: 95% Confidence intervals for proportions

To practice the Agresti-Coull method for calculating the 95% confidence interval of a proportion, we will use Excel. We will need the following equations:

$$p' = \frac{X + 2}{n + 4}$$

$$p' - 1.96\sqrt{\frac{p'(1-p')}{n+4}} < \boldsymbol{p} < p' + 1.96\sqrt{\frac{p'(1-p')}{n+4}}$$

We will use these formulas and Excel to calculate the 95% confidence interval using two examples.

Murphy's Law

In a test of Murphy's Law, pieces of toast were buttered on one side and then dropped. Murphy's Law predicts that they will land butter-side down. Out of 9821 slices of toast dropped, 6101 landed butter-side down.

- Open a new empty spreadsheet in Excel.
- In Excel label the cells in column A as shown below. In column B, insert the appropriate values for X and n from the Murphy's Law experiment. For the remainder of the cells, enter formulas that reference the appropriate cells or values so that you work through each step of calculating the 95% confidence interval. For example, for X+2 enter the formula =B1+2. This way we will be able to apply the same formulas to different examples.

  X
  N
  X+2
  n+4
  p
  p'
  p'(1-p')
  divided by n+4
  square root
  multiplied by 1.96
  lower limit
  upper limit

- The cells for lower limit and upper limit should provide the 95% confidence interval for the Murphy's Law experiment.

Surgical infections

Out of 67410 surgeries tracked in a study in the UK, 2832 were followed by surgical site infections.

- Enter the appropriate values from the surgical data into C1 and C2. Then highlight cells B3 to B12 and drag the formulas into column C to determine the 95% CI for surgical site infections. If you completed the formulas properly in column B, this should apply all of the formulas to the surgical data. You may wish to save this Excel file for future use!

If a 95% confidence interval does not include the value provided in the null hypothesis, then the null hypothesis will usually be rejected. This makes sense, as it indicates that the value of the null hypothesis is not one of the most plausible values given the data. Let's check this using SPSS.

Buttered toast

- Open the file *butter.sav,* which contains the data from the Murphy 's Law experiment described above in SPSS format.
- Note that in this file the data are already encoded with numeric values. Phew!
- Run a binomial test on the data.

➢ Answer the next three questions on Sakai.

Exercise 4: SPSS Practice

Country data

A variety of population statistics collected from 209 different countries is found in the file *Countries.sav*.

- Open the file *Countries.sav*
- DPT is a combination vaccine against diphtheria, pertussis and tetanus. Use the chart builder to plot a histogram for the Immunization DPT variable, which contains data on the % of children ages 12 – 23 months who received this vaccine in different countries. Open the graph and change the binning to a custom interval width of 10.
- Histograms can also be plotted using descriptive statistics, frequencies. Use this function to generate statistics (mean, standard deviation, and standard error) and histograms for personal computers per 100 people, and internet users per 100 people. Uncheck the box "display frequency table". Change the binning to a custom interval width of 10 again. Activate the data labels to view the number of countries in each bin.
- If we are interested in the descriptive statistics (values) only, we may choose to use the function descriptive statistics, descriptives. Use this function to produce a table containing mean, standard

deviation, minimum, maximum, and SE mean for the three variables examining life expectancy at birth (male, female, and total).

- The variable population_female indicates the percent of the total population that is female. Calculate mean and 95% confidence interval for this variable. μ

➢ Answer the remaining questions on Sakai.