FREQUENCY ANALYSIS AND TESTING PARAMETRIC ASSUMPTIONS
BI 311
Homework 05

Introduction:

As you learned in a previous homework, **inferential statistics** allow you to make conclusions about the statistical significance of different types of data comparisons. If we want to learn whether two (or more) populations are different, or if two (or more) variables are correlated, or if the frequency distribution of two (or more) variables differs from a random distribution, we use inferential tests to determine a p-value, the probability of making a **type I error**.

This homework will introduce you to a common type of inferential statistics known as frequency (or Chi Square) analysis. We will also introduce some of the assumptions associated with common types of parametric tests (e.g. t-tests, ANOVA and regression). One important assumption is that the frequency distribution of a continuous variable is normally distributed. We will introduce methods to test this assumption

Skills:

1. Learn how to use EXCEL to estimate a Chi Square statistic.
2. Learn how to use EXCEL to create a frequency histogram.
3. Learn how to use descriptive statistics to test for assumptions of normality.

Materials:

- A Handbook of Biological Investigation, Ambrose et al. 2007
- Access to WORD and EXCEL

EXERCISE 1-FREQUENCY ANALYSIS

Background:

Frequency analysis (also known as Chi-square Test for Independence) allows you to test for differences between two or more data sets that are nothing more than counts of observed events. Anything you can count can be tested with frequency analysis. Pages 96-98 in Chapter 8 of Ambrose et al. (2007) walk you through examples of how to use Chi-Square to test for independence.

To introduce you to this statistical test, I've created a Chi-square template in an EXCEL file labeled *Wood Frog Data* available on moodle. The study was conducted in Alaska. In the worksheet labeled *alder versus breeding activity*, I have compiled the number of times we observed wood frogs (*Rana sylvatica*) breeding in ponds with alder versus without alder vegetation. Alder is a dominant shrub in the Alaskan tundra and because the wood frog prefers areas with a forest canopy, we wanted to test the hypothesis that wood frogs would be found more frequently in ponds that had alder vegetation around the perimeter versus ponds that did not have alder.

I constructed a Chi-square template in EXCEL on the *alder versus breeding* worksheet. Note that there are 3 tables: Table 1 includes the observed data (the counts of sites with breeding activity that did/did not have alder), table 2 includes the expected frequencies, and table 3 is simply the difference between the observed and the expected squared and then divided by the expected. Note: expected frequencies for each cell = (the row total) x (the column total) / (the grand total). Because I have populated tables 2 and 3 with formulas, I can simply type in the observed data in table 1 and my template will automatically fill in table 2 and 3. The box below table 3 automatically calculates the Chi-Square statistic as the sum of all the values in table 3, the degrees of freedom, and I've inserted an EXCEL function called the CHITEST function that provides me with the p-value. Again, these values are automatically calculated once I enter the values in my observed table.

Note that the p-value is less than 0.05. Therefore, we found that the presence of alder and the observation of breeding activity were <u>not</u> independent of each other. In other words, the probability that I observed breeding activity depended on the presence or absence of alder. I reject the null that there is no association and accept the alternative hypothesis that the two variables are associated with each other. A quick check of the associated figure reveals that ponds with breeding activity were more likely (than expected by chance) to have alder and vice versa. Remember how the data were interpreted when you complete your homework.

Laboratory Procedure:

1. You can use my template for estimating Chi Square by simply typing in observed values for any 2x2 frequency analysis. For example, type in the data on page 98 of Ambrose et al. (2007) for the comparisons of IQ with marital status for men (note: you only have to type in the 4 observed frequencies; the row and column totals will automatically add up). Are the two variables independent?
2. In the EXCEL file labeled *wood frog data* is a second worksheet (see the tabs at the bottom) that has the observations for ponds of different depths versus the observations of breeding activity. Calculate a Chi-square statistic and associated p-value for this data. Note: because this data set has 3 levels of pond depth, the tables will be a 2 x 3. You can either create an entirely new Chi-square template on the new worksheet or, you can add a row to tables 1, 2 and 3 on the first worksheet.

3. Are the two variables independent ($p>0.05$) or not ($p<0.05$)? In other words, does the depth of the pond influence whether or not we observed breeding activity at a pond?
4. Save your Chi-square values, degrees of freedom and p-values for the depth analysis to submit as part of your homework (see below).

EXERCISE 2- USE EXCEL TO CREATE FREQUENCY HISTOGRAMS

Background:

Chapter 4 of Ambrose et al. (2007) provides an introduction to frequency histograms. Frequency histograms are useful because they can help determine which type of inferential statistic to use to test for differences between two statistical groups. With normally distributed data, many types of parametric tests (e.g. t-test and ANOVA) can be applied. With skewed data, it is better to use non-parametric tests (e.g. Friedman's and Kruskal-Wallis tests).

If your data sets are small as in chapter 4 of Ambrose et al. (2007) it is very easy to construct a frequency histogram by hand. It quickly becomes tedious when you have dozens or hundreds of samples. This exercise is designed to teach you how to construct a histogram using EXCEL. Learning to use this tool will allow you to construct histograms of your own data.

Laboratory Procedure:

Open the EXCEL file titled *Scrub Lizard Demographics* available on moodle. This practice data set will allow you to create frequency distributions for male versus female individuals of the Florida Scrub Lizard.

You may need to activate the *Data Analysis ToolPak* for EXCEL to be able to create a histogram. Select the *DATA* tab and look for an option for *Data Analysis* to the far right. If there is one, skip to section B below. If not, you need to activate the tool pack by following the directions in Section A.

Part A: Activating Data Analysis Toolpak.

1. Select the round FILE menu in the upper left corner of the screen.
2. Select *Excel Options*, then *Add Ins* and *Analysis ToolPak*.
3. Select *Go* and check the upper box for the *Analysis ToolPak* and *OK*.
4. EXCEL will ask you if you would like to install the tool pack: select *Yes* and wait.
5. Now when you select the *DATA* tab, *Data Analysis* should appear on your tool bar at the top.

Part B: Creating a Frequency Histogram

1. In cell I1 type "Female Home Range".
2. Type in the following numbers for cells I2 through I11: 100, 300 500, 700, 900, 1100, 1300, 1500, 1900, 2300. These are the categories (bin range) of home range sizes (m^2) for which we wish count frequencies.
3. Select the *DATA* tab, then *Data Analysis* then *Histogram*.
4. Select the red arrow to the right of the *Input Range*.
5. Highlight (by right clicking and dragging) cells B2 through B38 to select all the female home ranges for the data input and select *Enter*.
6. Select the red arrow to the right of *Bin Range* and highlight cells I1 through I11 and select *Enter*.
7. Select the *Labels* option, and the radio button for *Output Range*.
8. Select the red arrow to the right of *Output Range* and highlight cell J1, select *Enter* and select *OK*. The analysis tool should fill cell K2-K12 with a count of the number of females with home ranges falling into each size category defined by cells I2-I11.
9. Now to make a figure: Highlight cells K2-K12 and select the *INSERT* tab, then *Column* graph and then select a *2D Column* graph.
10. Choose *Select Data* from the *DESIGN* tab tool bar (note: you must have the chart highlighted to see the *DESIGN* tab), then *Edit* Horizontal (category) Axis Labels.
11. Select the red arrow and highlight cells J2-J12 and select *Enter*.
12. Select *OK* twice and the home ranges should appear on the X axis.
13. Select the *LAYOUT* tab and then *Axis Titles* and type in appropriate labels for the axes: "Female Home Range" for the X axis and "Frequency" for the Y axis.
14. You can delete the series legend: whenever you have only one set of data, there is no need to label the series.
15. Repeat steps 1-14 for the male home ranges.

EXERCISE 3- USING A HISTOGRAMS AND DESCRIPTIVE STATISTICS TO TEST THE ASSUMPTION OF NORMALITY

Background:

You have already been introduced the fundamentals of inferential statistics and practiced Chi-squared tests of independence useful for comparing frequency counts. However, often you have two data sets made of continuous measurements, not counts, and you

would like to test for differences or correlations between the data sets. Parametric statistics are powerful analytical tools that allow such testing as long as data within each data set are normally distributed.

Parametric tests are considered "robust" to small deviations from normality and often data can be transformed to meet parametric assumptions. Far too often, researchers do not test assumptions of normality before conducting their analysis. This can lead to type I errors. Just as you constructed a cumulative mean figure to test for sufficient sample size before you continued in your field study, you should get in the habit of conducting a preliminary analysis that tests for and perhaps corrects assumptions of normality. There are 3 tests that are simple and effective at assessing the normality of you data: 1) constructing a frequency histogram and visually inspect the figure for departures from normality, 2) compare your median and mean estimates, and 3) using a test of skewness.

Laboratory Procedure:

Open the EXCEL file named *Scrub Lizard Demographics* available on moodle and move to the worksheet labeled *Home Range*. The first two columns of data are home range estimates for female and male scrub lizards that you used for exercise 2. We would like to know whether the home range sizes differ for males versus females. However, to use a parametric test, we must first make certain that the data is normally distributed.

Examine your frequency histogram for both the male and female home range estimates for scrub lizards. Do they look normally distributed or close? Now, use the *Data Analysis* option (under the Data tab) to calculate *Descriptive Statistics* for both male and female home range sizes. Remember the rule of thumb that if your median and mean are within 3 units of each other, your data are more than likely close enough to normal for you to proceed. Are the means and medians for the males within 3 units? Is the same true for the females?

Note that one of the parameters produced in your descriptive output is called *Skewness*. This is a measure of how non-normal a set of continuous data are. If the data are perfectly normal, the skewness will equal zero. If data are positively skewed (fewer observations in the right tail of the distribution compared to the left tail) the skewness value will be positive and vice versa. As you can see, the skewness values for females and males are 1.34 and 1.06 respectively indicating positive skewness. But is the skewness significant? Remember our parametric tests are robust to a small amount of skewness and thus, skewness doesn't have to be exactly zero.

We can test for significant skewness by calculating the standard error of skewness as follows:

SE Skewness = $\sqrt{6}$/sample size.

For females, $\sqrt{6}/37 = 0.40$ and for males $\sqrt{6}/46 = 0.36$. Now we multiply these values by 2 to get 0.81 and 0.72 respectively. A simple test for significant skewness states that if the absolute value of skewness is larger than twice the standard error of skewness then it is significant. Note that in our case, $1.34 > 0.81$ and $1.06 > 0.72$ so we can conclude that the data for both females and males is significantly skewed.

If either of our data sets are significantly skewed we need to address the problem. There are two possible solutions. First, we could skip using a parametric test and instead use a non-parametric test. The only problem is that non-parametric tests are less powerful and may lead to **a type II error**. Another option is to transform the data so that it better fits a normal distribution.

Note that positive skewness means that we have many smaller values and fewer and fewer large values. In this case, there are more individuals with small home range sizes than with larger home range sizes. A common solution for positive skewness is to transform the data by taking the log of each value. Because of the nature of logs, the larger values are transformed relatively more than are the smaller values. For example, the $\log_{10}$ of 10 is 1 while the $\log_{10}$ of 100 is 2. The number 10 gets transformed from 10 to 1, a difference of only 9, while the 100 changes to 2, a difference of 98.

You can log transform using any base value. If you have only minor skewness (as in our case) you can transform to the $\log_e$. Columns E and F have the $\log_e$ values of home range sizes for females and males. I created these with a simple formula (=Ln(A1)) and filled in the cells.

Repeat the calculations for descriptive statistics using the transformed (Ln) data. What are your new skewness statistics for females and males? Are the log transformed values of the data significantly skewed? Is it now ok to use parametric analysis on the log transformed data?

HOMEWORK SUBMISSION

Each person should create a WORD file named *Homework 5*, answer the following questions, and submit their file to me via email.

1. Provide your Chi Square value, degrees of freedom and p-value for the analysis of pond depth versus wood frog breeding activity in exercise 1.
2. What did you conclude with respect to pond depth? Is breeding activity independent or dependent on the depth of a pond?
3. Paste your frequency histograms (for female and male) from exercise 2 into your WORD file. Be sure to add appropriate axis labels.
4. Do your male and female frequency histograms look normally distributed or skewed?
5. Is your median value for home range size within 3 units of the mean for female home range size? Male home range size?
6. After transforming home range size, are the log transformed values significantly skewed for female and male home range size?