**POPS Project Statistical Help**

Most of you will have completed a Statistic Course before taking Biology 2425. Please go back and review the helpful materials from that course. However, some of you have no completed a statistic course and/or took it so long ago that you do not remember the information. This simple guide was made to help you review or gain the statistical concepts needed to usefully analyze your POPS Project data. In addition: feel free to use much more complex analysis that these – there is a lot of power in correct data analysis, go ahead and show off your advanced knowledge!

Lists of raw data alone are not often useful for recognizing relationships between variables related to human health. Simple descriptive statistics, including the mean, standard deviation, median, mode, and range, allow scientists to more easily appreciate major trends in data. Data also becomes more meaningful when organized into a table, which is also the first step in creating a graph. As you learned in the previous lab, graphing is a technique that allows visual examination of data and is another way to reveal trends in lists of data. Last week’s lab also allowed you to practice correctly positioning the independent and dependent variables as well as to recognize trends in the data.

Computer programs, such as Excel©, are commonly used to organize data into tables and then to analyze and graph that data. Data that is not well organized can be misleading to the researcher and sometimes data is deliberately analyzed incorrectly to appear to back up a fallacious claim. Learning how to use computer programs to organize, analyze and graph data is an important skill not only for scientists but also for educated people. This skill will also be invaluable in helping you tell the difference between real data and fallacious claims.

In order to determine cause and effect relationships, scientists record observations about how changes in one variable cause another variable to change. A variable is the factor or characteristic being measured and can differ or vary in amounts or types. Examples of variables include height, weight, femur length, eye color, attitude, and health status. A list of data collected about variables often is not useful by itself; therefore, descriptive statistics are used to summarize the data and give scientists a clearer view of the data in order to identify trends or patterns in the data.

Examples of descriptive statistics with their definitions are listed below. The mean, median, and mode are used to determine the “center” of the values in a data set. The range and standard deviation indicate the spread of the data set values around the “center”.

* Mean: The mean is often called the average. It is calculated by adding up all the values in a data set and then dividing by the total number of values. The mean is very sensitive to unusual values (outliers) and is easily distorted. For example, consider the data set with the following values: 4, 7, 3, 8, 10, 8. The average is 6.7. But what if the last value was mistakenly entered as 80 instead of 8? Now the average is 18.7!

* Median: The median is the middle value of the data set. Half of the values are lower than the median and half are higher than the median. In order to determine the median, the values must be listed in numerical order. If there is an odd number of values, the median is the ((n + 1)/2)th value (n = total number of values). If there is an even number of values, the median is the average of the two middlemost values, the (n/2)th and [(n/2)+1]th values. For example, the data set 5, 10, 6, 9, 12 contains an odd number of values. First, we must arrange the values in numerical order: 5, 6, 9, 10, 12. The median is the ((5+1)/2) or 3rd value, or 9. It is easy to see by looking at this data set that the value 9 is in the middle.

Now calculate the median of the data set from the mean example above. The data set listed in numerical order is 3, 4, 7, 8, 8, 10 and contains an even number of values. The median is the average of the 6/2 or 3rd value and the [(6/2) +1] or 4th value. The median is (7+8)/2, which equals 7.5. Note that if we used the data set from the mean example with 80 instead of 8, the median is still 7.5! The median is not as sensitive to outliers.

Remember, median = middle value.

* Mode: The mode is the value that occurs the most frequently in a [data set](http://en.wikipedia.org/wiki/Data_set). *If all values in a data set only occur once, then there is no mode*. Remember, mode = most often.

The mode for the data set 4, 8, 3, 8, 10, 7 is 8 because that value occurs the most often.

What is the mode for the data set 5, 10, 6, 9, 12? In this case, there is no mode.

* Range: The range is the difference between the highest and lowest values. It indicates the spread of the values in the data set. Since the range only considers the extreme values in the data set, it is also highly influenced by outliers.

For the data set 3, 4, 7, 8, 8, 10, the range is 10-3 = 7.

If one of the 8s was mistakenly recorded as an 80 in the data set above, then the range is 80-3 = 77!

* Standard deviation: The standard deviation is a measure of the variability or spread of the data around the mean. When comparing two data sets collected on the same variable (ex. respiratory rate), the data set with the higher standard deviation has a greater amount of variability among the values; the values are more spread out. In contrast, the data set with the smaller standard deviation has values that are more similar and closer to the mean. For example, respiratory rates in breaths per minute were recorded for two different groups of students.

Group 1: 16, 15, 14, 17, 18 Mean = 16 bpm Standard deviation = 1.6 bpm

Group 2: 10, 10, 20, 25, 15 Mean = 16 bpm Standard deviation = 6.5 bpm

Even though the two groups have the same mean, group 2 has a higher standard deviation, indicating that the values are more variable and spread out.

Many important physiological variables, such as blood pressure, height, and weight, have a pattern of data value distribution called a normal distribution. For example, if you randomly sampled a large group of people and plotted their blood pressure values, a curve that is symmetrical about the mean blood pressure is obtained. This curve is called a normal curve or bell-shaped curve (see Figure 1). If the data is distributed in the form of a normal or bell-shaped curve, then the following conclusions can be made:

* approximately 68% of the values lie within +/- 1 standard deviation from the mean
* approximately 95% of the values are within +/- 2 standard deviations from the mean
* approximately 99% of the values lie within +/- 3 standard deviations from the mean



 Figure 1. Normal or bell-shaped curve with standard deviations

The standard deviation is defined as the square root of the sum of the value deviations from the mean divided by one less than the sample size! This can be represented by the formula:



where are the observed values of the sample *i*.

The following example of how to calculate standard deviation will allow you to understand the above definition and formula better. Please follow along with the example, so that you will be able to calculate standard deviation for the other data sets.

Researchers recorded the number of precancerous skin lesions found on six patients who had a skin cancer mass removed 10 years ago. The data set is as follows: 1, 3, 4, 6, 9, and 19 precancerous skin lesions.

* Calculate the mean for the data set.

Mean = (1+3+4+6+9+19) / 6 = 42 / 6 = 7 Mean = 7 precancerous skin lesions

* Subtract the mean from every number to get the list of deviations. It is OK to get negative numbers here.

list of deviations: -6, -4, -3, -1, 2, 12

* Next, square the resulting deviations.

 squares of deviations: 36, 16, 9, 1, 4, 144

* Add up all of the resulting squares to get their total sum.

sum of squared deviations: 36+16+9+1+4+144 = 210

|  |  |  |
| --- | --- | --- |
| Data Value (xi) (# of precancerous skin lesions) | Data Value –Mean(xi - $\overbar{x}$) (= deviation) | Deviation Squared |
| 1 | -6 | 36 |
| 3 | -4 | 16 |
| 4 | -3 | 9 |
| 6 | -1 | 1 |
| 9 | 2 | 4 |
| 19 | 12 | 144 |
| Total $ \sum\_{i=1}^{N}(x\_{i}-\overbar{x})^{2}$ |  | 210 |

Table 1: calculation of standard deviations from numbers of precancerous skin lesions.

* Divide the sum of squared deviations by one less than the number of values.

\_\_

210 / 5 = 42

Then take the square root of this number: √42 = 6.48 precancerous skin lesions = standard deviation

If this data formed a normal curve (which it does not), we could conclude that 68% of the values will fall in the range of +/- 1 standard deviations from the mean. That is, 68% of the values in the data set are within the range of 7 – (1 x 6.48) to 7 + (1 x 6.48) or 0.5 to 13 precancerous skin lesions.

**Graphing**

Graphing is a technique that allows visual examination of data, and, in addition to descriptive statistics, is another useful way to reveal trends in lists of data. You will be using graphs to examine data during the remainder of this laboratory exercise. Here are some guidelines for presentation of data in graphs or tables:

1. In a paper, each graph or table must be identified by a figure or table reference number. Besides this number should be a short but clear description of what the table/graph/figure illustrates. This figure number should be referenced in the prose of the paper itself
2. The graph or table must have a brief, clear, descriptive title.
3. In tables, the column and row headings should clearly identify the variable and units of measurement.
4. In graphs, the axes should be clearly labeled and the units of measurement denoted. NOTE: always graph time (the independent variable) on the horizontal (x) axis. Note: look back at Lab 1 and review the differences between the independent and dependent variables
5. A graph or table, along with its descriptive title, should be able to stand alone without the reader having to read additional text to understand it (hence numbers 1-4)

One of the most commonly used graphs is the Scatterplot (scatter graph and/or scatter chart). A scatterplot graphically represents 2 variables on a Cartesian coordinate system. The pattern of data dots can be analyzed for types of correlations. Often straight or curved lines are drawn between the data. To use this function, you will need to create a spreadsheet with the two variables – both the dependent and independent – listed in adjacent columns. Select the data and then go to “insert” “charts” “scatter” and then choose the type of scatterplot that you feel will illustrate well your data type. In figure 2, below, age is the independent variable and weight is the dependent variable. Independent data points are indicated with dots (you can also have the program draw a line or curve between the data point to illuminate any continuity between data). A trendline is graphed in a dashed line. What kind of correlation does this graph suggest between these two different variables?

Figure 2: Age versus weight for 25 Physiology Lab students at Salt Lake Community College.

Another very common graph type is the histogram. Histograms visually display the frequency distribution of data. The absolute frequency is the number of times a value occurs in a data set. It is plotted on the y-axis while the dependent variable (the variable you are studying) is plotted on the x-axis. For example, researchers collected serum cholesterol values from a large sample of men, ages 25-34. The serum cholesterol values are divided into intervals of equal widths, which allows for comparisons among the intervals. The frequencies for each interval are presented in Table 2 and the corresponding histogram is seen in Figure 3.

|  |  |  |
| --- | --- | --- |
| Cholesterol Level (mg/100 ml) | Ages 25-34 | Ages 55-64 |
| Number of Men | Relative Frequency (%) | Number of Men | Relative Frequency (%) |
| 80-119 |  13 |  1.2 |  5 | 0.4 |
| 120-159 | 150 | 14.1 | 48 | 3.9 |
| 160-199 | 442 | 41.4 | 265 | 21.6 |
| 200-239 | 299 | 28.0 | 458 | 37.3 |
| 240-279 | 115 | 10.8 | 281 | 22.9 |
| 280-319 |  34 |  3.2 | 128 | 10.4 |
| 320-359 |  9 |  0.8 |  35 |  2.9 |
| 360-399 |  5 |  0.5 |  7 |  0.6 |
| Total | 1067 | 100.0 | 1227 | 100.0 |

Table 2: Absolute and relative frequencies of serum cholesterol levels for 2294 U.S. males, 1976-1980

Data from National Center for Health Statistics, Fulwood R, Kalsbeek W, Rifkind B, Russell-Briefel R, Muesing R, LaRosa J, and Lippel K. Total serum cholesterol levels of adults 20-74 years of age: United States, 1976-1980. *Vital and Health Statistics*, Series 11, Number 236, May 1986.

The relative frequency for an interval is calculated by dividing the absolute frequency by the total number of values. The resulting number is then multiplied by 100%. This gives the proportion of values that fall into an interval rather than the absolute number. For example, the relative frequency in the 160-199 mg/100 ml interval for men ages 25-34 is (442/1067) x 100% = 41.4%. The relative frequency is used to compare two or more data sets where the number of values obtained is not equal. For example, serum cholesterol values were also collected for 1227 men aged 55-64 years. Because the group sizes are not the same, one cannot compare the absolute frequencies of the two groups. However, the relative frequencies can be compared, and one can see that the older men tend to have higher serum cholesterol levels than the younger men.

### Figure 3: Absolute frequencies of serum cholesterol levels for 1067 U.S. Males, Aged 25-34 years, 1976-1980

 As seen in the previous examples, management of data often requires sorting the data, constructing a well-organized table, and calculating statistical parameters. Computerized spreadsheets can be invaluable when performing these tasks. One of the most popular spreadsheets available is Microsoft’s Excel©. The following describes how to enter data and formulas into an Excel spreadsheet. You will be asked, each week, in lab to create and use spreadsheets to organize and analyze your data.

For help with Statistical Analysis using Excel:

<https://www.excel-easy.com/data-analysis.html>

For Creating Charts on Excel:

<https://www.youtube.com/watch?v=TfkNkrKMF5c>