Week 1 Lecture 2

In Lecture 1, we focused on identifying the characteristics – quantitative, qualitative, discrete, continuous, NOIR – of the data. In this section, we will take a look at how we can summarize a data set with descriptive statistics, and how we can ensure that these *descriptive* statistics can be used as *inferential* statistics to make inferences and judgments about a larger population. We are moving into the second step of the analysis approach mentioned in Lecture 1.

Descriptive Statistics

Once we understand the kinds of data we have, the natural reaction is want to summarize it – reduce what may be a lot of data into a few measures to make sense of what we have. We start with summary descriptions, the principle types focus on *location, variability, and likelihood*. (Note, we will deal with likelihood, AKA *probability*, in Lecture 3 for this week.)

For nominal data, our analysis is limited to counting how many exist in each group, such as how many cars by car company (Ford, Nissan, etc.) are in the company parking lot. However, we can also use nominal data as a group name to form different groups to examine, in this case we do nothing with the actual data label, but do some analysis with the data in each group. An example related to our class case: we can group the salary data values into two groups using the nominal variable gender (or gender1).

With ordinal scales, we can do some limited analysis of differences using certain tests; most of which are not covered in this course. We can also use ordinal data as grouping labels, for example we could do some analysis of salary by educational degree.

Both interval and ratio scales allow us to do both inferential and descriptive analysis (Tanner & Youssef-Morgan, 2013). Most of the statistical tools we will cover in this class require data scales that are at least interval in nature.

Location measures. When working with interval or ratio level variables, the first measure most researchers look are indications of *location* – **mean, median, mode**. The *mean* is the numerical average of the data – simply add the values and divide by the total count. The *median* is the middle of the data set; rank order the values form low to high or high to low, and pick the value that is in the middle. This is easy if we have an odd number of values, we can find the middle exactly. If we have an even number of variables, the middle is the average of the middle two values. For example, in this data set: 2, 3, 4, 5, 6, we have five values and the median is 4. However, in this data set: 2, 3, 4, 5, we have only four values and the median is the average of the middle 2 numbers = (3 + 4)/2 = 7/2 = 3.5. Finally, the *mode* is the most frequently occurring value; as such, it may or may not occur. And, there may be more than one mode in any data set.

Generally, the mean is the most useful measure for a data set, as it contains information regarding all the values. It is the location measure that is used in many statistical tests. The symbol for the mean of a population is μ – called mu – while we use $\overline{x}$ – sometimes typed as x-bar – for the sample mean.

Variation measures. After finding our mean (or other center measure), we generally want to know how consistent the data is – that is, is the data bunched around the center, or is it spread out. The more spread out a data is, the less any single measure accurately describes all of the data. Looking at the consistency (or lack of consistency) in a data set will often give us a different understanding of what is going on. A simple example, if we have two departments in a company that each averaged 3.0 on a question in a company morale survey, we might be tempted to say they were the same. However, if we looked at the actual scores and saw that one department had individual scores of 3, 3, 3, 3, 3, and 3 while the other department's scores were 5, 5, 5, 1, 1, and1 we can now see that the groups are quite a bit different. The mean alone did not provide enough information to interpret what was going on in each group.

We have 3 general measures of *variation* – **range**, **standard deviation**, and **variance**. *Range* is simply the difference between the largest and smallest value (largest – smallest = range).

Standard deviation and *variance* are related values. The *variance* is a somewhat awkward measure to initially understand. To calculate it, we first take the difference between each value and mean of the entire group. This outcome will have both positive and negative values, and if we add them together we would get a result of 0. So, to eliminate the negative values, we square each outcome. Then we get the sum these squared values and divide it by the count. (Note: this is the same as the mean of the squared differences.) For example, the variance of this data set (2, 3, 4) would be:

- Mean = (2 + 3 + 4)/3 = 9/3 = 3
- Variance = ((2 - 3)^2 + (3 – 3)^2 + (4 – 3)^2)/3 = ((-1)^2 + (0)^2 + (1)^2))/3 = (1 + 0 + 1)/3 = 2/3 = 0.667.

This gives us an awkward measure – the variance of something measured in inches, for example, would be measured in inches squared – not a measure we all use on a daily basis.

The *standard deviation* changes this awkward measure to one that makes more intuitive sense. It does so by taking the positive square root of the variance. This would give us, for our inches measure a result that is expressed in inches. The standard deviation is always expressed in the same units as the initial measure. For our example above with the variance of 0.667, the standard deviation would be the square root of 0.667 or 0.817. Both the variance and standard deviation require data that is at least interval in nature. The standard deviation is about 1/6 of the range, and is considered the average difference from the mean for all of the data values in the set (Tanner & Youssef-Morgan, 2013).

Technical point – both the variance and the standard deviation have two different formulas, one for *populations* and one for *samples.* The difference is that with the sample formula, the average is found with the (count -1) rather than the full count. This serves to increase the estimate, since the data in a sample will not be as spread out as in the population (unlikely to have the extreme largest and smallest value). The symbol for the population standard deviation is σ, while the sample standard deviation symbol is s. In statistics, since we deal with samples, we use the sample formulas – to be discussed below.

The nice thing about descriptive statistics is that Excel will do all of the math calculations for us, we just need to know how to interpret our results.

For a video discussion of descriptive statistics take a look at Descriptive statistics from the Kahn Academy - https://www.khanacademy.org/math/probability/descriptive-statistics.

Research Question Example

Now that we have identified the data types for each of our variables, we need to develop some descriptive statistics – particularly for those at the interval and ratio level. In our discussion and example of salary, we will be using a salary sample of 50 that does not exactly match the data that is available in your data set. It is not significantly different, and should be considered to come from a different sample of the same population. The results will be accurate enough to consider them in answering our equal pay for equal work question for the sample results provided to the class.

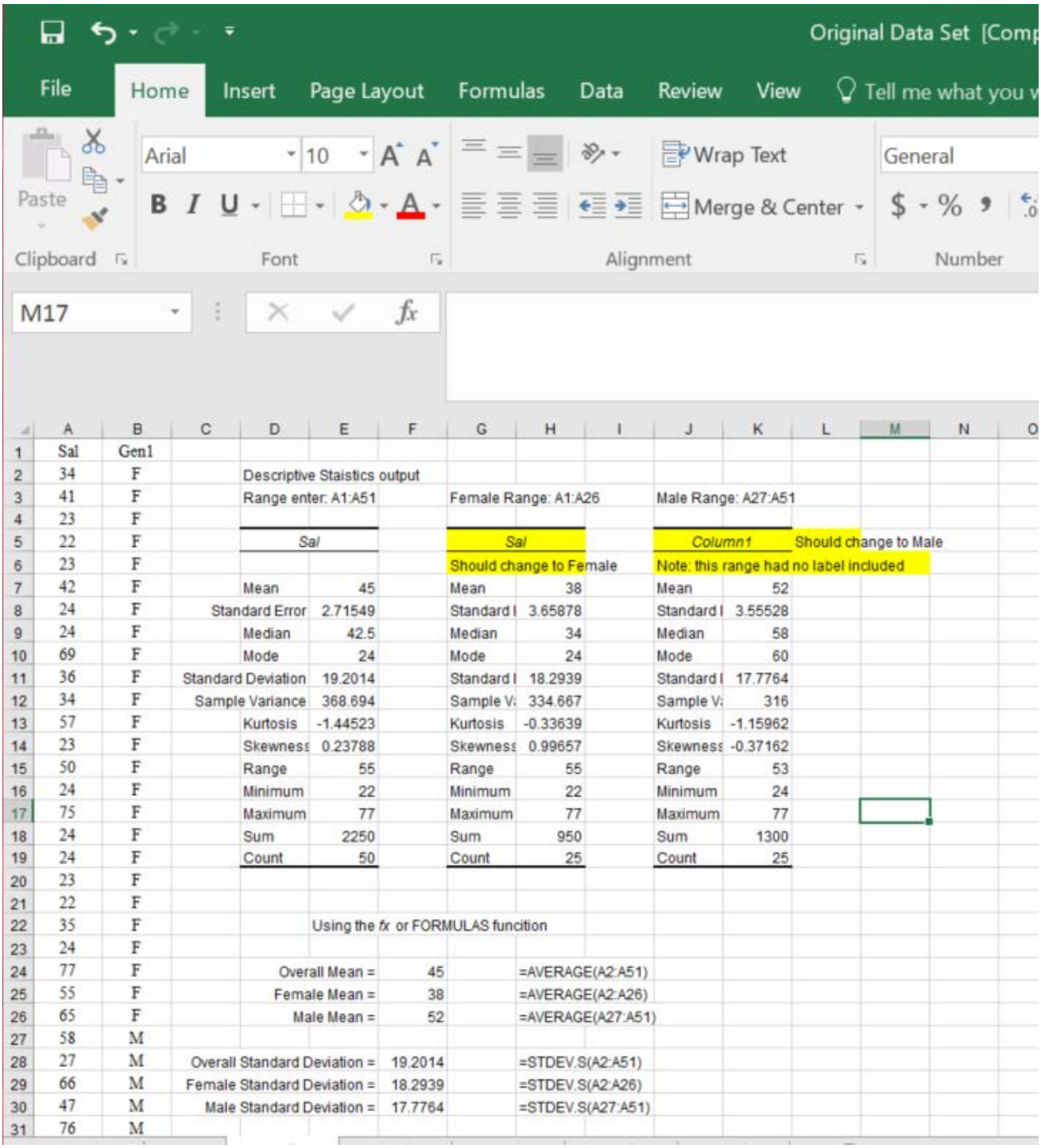
Equal Pay Question. The obvious first question to ask is what is the overall average salary, and what is the average for the males and females separately? This descriptive statistic should also be accompanied with the standard deviation of each group to examine group diversity. (Reminder: the salary results presented each week will not exactly match the results from this class' data set if you choose to duplicate the results presented in this lecture. The results are statistically close enough to use to answer our assignment question on equal pay.)

The related question concerns the standard deviation of each of the three groups (entire sample, males, and females) – what is the standard deviation for each group.

In setting up the data for this, copy the salary data column (B1:B51) and paste it on a new sheet. This is a recommend practice – never do analysis on the raw data set so that relationships between various columns are not compromised. Then copy the Gender column (M and F) and paste it beside the salary data. Using Excel's sort function, sort the two columns (at the same time) using Gender as the sort key. This will give you the salary data grouped by males and females.

The screen shot below displays the results using both the Descriptive Statistics option found in the Data Analysis list and the =Average and = Stdev.s functions found in the *fx* or Formulas – statistics section.

Note a couple of things about the Descriptive Statistics output. First, since for both the overall and female groups, the input range included the label *Sal*, this was shown at the top. The male range did not have a label, so Column 1 was automatically used. We can use Descriptive Statistics for any number of contiguous columns in the input range box. For reporting purposes, we should change the Sal and Column 1 labels to Overall, Female, and Male.



The second issue about the descriptive statistics output is that it contains much more information than we were looking for. This is a good tool for an overall look at a data set.

Looking at the *fx* values and those from the Descriptive Statistics output, we can see that the means and standard deviations are identical for each group – so, it does not matter which approach you use.

Now, looking at the actual statistical values, we see that the overall all salary mean (45) lies in the middle between the lower female mean (38) and the upper male mean (52) – overall means will always be flanked by sub-group means, but the differences will not also be equidistant.

The standard deviations, on the other hand are much closer together with the overall (19.2) being somewhat larger than either the female (18.3) or male (17.8). This is also somewhat common – the variation in the entire group is generally a bit larger than for the sub-groups.

While we did not specifically ask for it, we can also note that the range in each group is very close 22 – 77 for overall and females and 24 – 77 for males.

So, what can we say at this point? It appears that males and females have about the same range and standard deviations for salaries, but that females appear to average less than the males. However, at this point, we cannot say anything about our equal pay for equal work question as the Salaries have not been divided into equal work groups. So, at this point we have some interesting information, but no conclusive results yet.

References

Lind, D. A., Marchel, W. G., & Wathen, S. A. (2008). *Statistical Techniques in Business & Finance.* (13th Ed.) Boston: McGraw-Hill Irwin.

Tanner, D. E. & Youssef-Morgan, C. M. (2013). *Statistics for Managers.* San Diego, CA: Bridgeport Education.