Week 1 Lecture 1

Class Approach to Statistics

Statistics is basically a set of tools that allow us to get information out of data sets (we will get to the more formal definition below). As such, it can be taught as a math class (focusing on formulas), a logic class (If this, then that), or as a case study (here is the problem, what are we going to do). We have chosen the later – we will be examining statistical tools and approaches as they help us answer a business question.

The question we will focus on involves the Equal Pay Act, specifically the requirement that males and females be paid the same if they are performing equal or equivalent work. So, our business research question is: are males and females paid the same for equal work?

In starting out with our case, we will have a data set that provides a number of variables (measures that can assume different values with different subjects) for each of 50 employees selected randomly from our company. (The company and employee data are fictitious, of course).

For each employee (labeled 1 thru 50 in the ID column), we will have:

- Salary, the annual salary, rounded to the nearest hundred dollars; for example, a salary of 32, 650 would be rounded to 32.7.
- Compa (short for compa-ratio or Comparative ratio) – a measure of how a salary relates to the midpoint of a pay range, found by dividing the salary by the pay range midpoint.
- Midpoint – the middle of the salary range assigned to each grade.
- Age – the employee's age (rounded to the nearest birthday)
- Performance rating – a value between 1 and 100 showing the manager's rating how good the employee performs their job
- Service – the years the employee has been with the company (rounded to the nearest hiring anniversary
- Gender – a numerical code indicating the employee's gender (1 = female, 0 = male)
- Raise – the percent increase in pay of the last performance based increase in salary
- Degree – the educational achievement of the employee (0 = BA/BS, 1 = Master's or more)
- Gender1 – a letter code indicating the employee's gender (F = female, M = male)
- Grade – the employee's pay level – grade A is the lowest (entry level) and grade E is the highest.

During each week, we will examine some of these variables to see if they help us answer the question of males and females receiving equal pay for equal work. In the weekly lectures, we will work with the variable salary. In the homework assignments for weeks 2, 3, and 4; you will have the same questions but work with the variable compa, which – by definition – is an alternate method of looking at pay.

If you have any questions about this description of our course case, please ask them in either *Ask Your Instructor* or in one of the class posts.

Introduction to Statistics

Formally, we can define statistics as "the science of collecting, organizing, presenting, analyzing, and interpreting data to assist in making more effective decisions" (Lind, Marchel, & Wathen, 2008, p. 4). This makes statistics and statistical analysis a subset of both critical thinking and quantitative thinking, both skills that Ashford University has identified as critical abilities for any student graduating with a degree. H. G. Wells, the author, once said that "one day quantitative reasoning will be as necessary for effective citizenship as the ability to read."

In this class, we will focus mostly on the analyzing and interpreting of data that we will assume has been correctly collected to allow us to use it to make decisions with. In doing this, there is a fairly well agreed upon approach to understanding what the data is trying to tell us. This approach will be followed in this class, and involves:

- Identifying what kinds of data we are working with, then
- Developing summary statistics for the data
- Developing appropriate statistical tests to make decisions about the population the data came from.
- Drawing conclusions from the test results to answer the initial research question(s).

Data Characteristics

We all recognized that not all data is the same. Saying we "like" something is quite a bit different than saying, the part weighs 3.7 ounces. We treat these two kinds of data in very different ways.

The first distinction we make in data types involves identifying our data as either qualitative or quantitative. Qualitative data identifies characteristics or attributes of something being studied. They are non-numeric and can often be used for grouping purposes. Some examples include nationality, gender, type of car, etc.

Quantitative data, on the other hand, tend to measure how much of what is being examined exists. Examples of these kinds of variables include, money, temperature, number of drawers in a desk, etc.

Within quantitative data, we can identify *continuous* and *discrete* data types. Continuous data variables can assume any value with limits. For example, depending upon how accurate our measuring instrument is, the temperature, in degrees Fahrenheit, could be 75, 75.3 75.32, 75.3287468…. There are no natural "breaks" in temperature even though we typically only report it in whole numbers and ignore the decimal portion. Height would be another continuous data variable. Discrete data, on the other hand, has only certain values, and shows breaks between these values. The number of drawers in a desk could be 3 or 4, but not 3.56, for example.

The second important approach in defining data is the "*level*" of the data. There exist four distinct levels:

- **Nominal** – these serve as names or labels, and could be considered qualitative. The basic use for this level is to identify distinctions between and among subjects, such as ID numbers, gender identification (Male or Female), car type (Ford, Nissan, etc.). We can basically only count how many exist within each group of a nominal data variable.
- **Ordinal** – these data have the same characteristics as nominal with the addition of being rankable – that is, we can place them in a descending or ascending order. One example is rating something using good, better, best (even if coded 1 = good, 2 = better, and 3 = best). We can rank this preference, but cannot say the difference between each data point is the same for everyone.
- **Interval** – this level of data adds the element of constant differences between sequential data points – while we did not know the difference between good and better or better and best; we do know the difference between 57 degrees and 58 degrees – and it is the same as the difference between 67 and 68 degrees.
- **Ratio** – this level adds a "meaningful" 0 – which means the absence of any characteristic. Temperature (at least for the Celsius and Fahrenheit scales)) does not have a 0 point meaning no heat at all. A scale with a meaningful 0, such as length, has equal ratios – the ratio of 4 feet to 2 feet has the same value as that of 8 feet to 4 feet – both are 2. This cannot be said of temperatures, for example (Tanner & Youssef-Morgan, 2013).

These are often recalled by the acronym **NOIR**.

Knowing what kinds of data we have is important, as it identifies what kinds of statistical analysis we can do.

Equal Pay Question

At the end of each lecture, we will apply the topics discussed to our research question of do males and females receive equal pay for equal work. In this section, we will look at identifying the data characteristics for each of our data variables.

In looking at our first classification of qualitative versus quantitative, we have

Qualitative	Quantitative	
	Continuous	*Discrete*
ID	Compa	Salary
Gender	Age	Midpoint
Gender1	Raise	Performance Rating
Degree		Service
Grade		

Most of these are fairly clear – the variables in the qualitative column merely identify different groups. The continuous variable lists can all – theoretically – be carried out to many decimal points, while those in the discrete list all have distinct values within their range of available values.

The identification for the NOIR classification are shown below.

Nominal	Ordinal	Interval	Ratio
ID	Degree	Performance Rating	Salary
Gender	Grade		Midpoint
Gender1			Service
			Compa
			Age
			Raise

While an argument can be made that Performance Ratings, being basically opinions, are really ordinal data; for this class let us assume that they are interval level as many organizations treat them as such.

An important reason for always knowing the data level for each variable is that we are limited to what can be done with different levels. With nominal scales, we can count the differences. With ordinal scales, we can do some limited analysis of differences using certain tests that are not covered in this course. 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. During our last two weeks, we will look at some techniques for nominal and ordinal data measures.

In Lecture 2, we will start to see what kinds of things we can do with each level of the NOIR characteristics.

If you have any questions about this material, please ask questions in either *Ask Your Instructor* or in the discussion area.

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.