BUS308 – Week 1 Lecture 1

Statistics

Expected Outcomes

After reading this lecture, the student should be familiar with:

1. The basic ideas of data analysis.
2. Key statistical concepts and terms.
3. The basic approach for this class.
4. The case focus for the class.

What we are all about

Data, measurements, counts, etc., is often considered the language of business. However, it also plays an important role in our personal lives as well. Data, or more accurately, the analysis of data answers our questions. These may be business related or personal. Some questions we may have heard that require data to answer include:

1. On average, how long does it take you to get to work? Or, alternately, when do you have to leave to get to work on time?
2. For budget purposes, what is the average expense for utilities, food, etc.?
3. Has the quality rejection rate on production Line 3 changed?
4. Did the new attendance incentive program reduce the tardiness for the department?
5. Which vendor has the best average price for what we order?
6. Which customers have the most complaints about our products?
7. Has the average production time decreased with the new process?
8. Do different groups respond differently to an employee questionnaire?
9. What are the chances that a customer will complain about or return a product?

Note that all of these very reasonable questions require that we collect data, analyze it, and reach some conclusion based upon that result.

Making Sense of Data

This class is about ways to turn data sets, lots of raw numbers, into information that we can use. This may include simple descriptions of the data with measures such as average, range, high and low values, etc. It also includes ways to examine the information within the data set so that we can make decisions, identify patterns, and identify existing relationships. This is often called data analysis; some courses discuss this approach with the term "data-based decision making." During this class we will focus on the logic of analyzing data and interpreting these results.

What this class is not

This class is not a mathematics course. I know, it is called statistics and it deals with numbers, but we do not focus on creating formulas or even doing calculations. Excel will do all

of the calculations for us; for those of you who have not used Excel before, and even for some who have, you will be pleasantly surprised at how powerful and relatively easy to use it is.

It is also not a class in collecting the data. Courses in research focus on how to plan on collecting data so that it is fair and unbiased. Statistics deals with working on the data after it has been collected.

Class structure

There are two main themes to this class. The first focuses on interpreting statistical outcomes. When someone says, the result is statistically significant with a p-value of 0.01; we need, as professionals, to know what it means. As you move higher into business and other professional positions, you will probably hear others report on studies using this kind of language. (Data analysis is becoming increasing more common in business.)

The second thread focuses on how to take some data and generate statistical reports using Excel. Excel is a fairly common PC program that is part of Microsoft's Office suite of tools, and as such many businesses have it available for professionals and managers. Even if you just do a quick analysis of some data, this program is tremendously useful.

This class does not have a text, but rather provides the material you need in three lectures each week. The first lecture is an overview, it provides a structure about what the week's focus is all about. The second lecture focuses on understanding the statistical tools being presented; how to read the outputs and how to understand and interpret what they are telling us. The third lecture for each week focuses on Excel and presenting the steps needed to generate the statistical output.

Unlike other classes, we have three weekly discussions; one related to each of the lecture segments. The intent is for you to read a lecture and then go to the discussion thread for a couple of days. Then go read the next lecture, discuss it for a couple of days, and then finish with the last lecture. This chunking of material is designed to let the information "sink in" before moving to new things.

Introducing Statistical Analysis

Data analysis

Data analysis, whether statistical, financial, operational, etc., often appears to be a set of unrelated tools and techniques that are somehow applied sequentially to get an answer to some question. Indeed, most textbooks present statistical analysis this way; introduce a topic, provide some examples, present practice exercises, and then on to the next topic with new examples and exercises that often have nothing to do with what was previously presented.

This approach, while common in many numerical and even qualitative courses, often leaves students with an incomplete idea of how everything fits together. We are trying a different approach in this class and will be using a single case/situation to demonstrate the interconnectedness of our tools.

Data analysis, and particularly statistical analysis, is much like solving a mystery. Those who work with these tools are like the detectives we see on TV shows. In general, the process involves a situation (or crime) presents itself and the team goes to work. Initially, they look at the "big picture" to try and understand the basics of the situation. After that, the focus shifts to specific details as they examine suspects, look for and verify alibis, find links between different individuals and activities; often this part of the investigation seems uncoordinated and even a bit chaotic with little obvious links to the overall situation. But, finally everything is pulled together and all the various threads form a conclusion and they "nab the culprit."

So, to tie what the TV detectives do with what we, as data analysts, will be doing, take a look at the following. Hopefully, this will relate the familiar crime story to data analysis.

- The "crime" we focus on presents itself as some outcome; results of a manufacturing process, customer satisfaction ratings differences, financial outcomes, etc.; that we do not fully understand.
- The "witnesses" we look at are the different data measurements we have.
- Our "questions" are just that – questions about what we want to find out from the data.
- Our "theory of the crime" focuses on how we think the data is related to our questions.
- The "alibis" are the data summaries and test outcomes that show if particular data is related to the outcome or not.
- The "false leads" are data measures that are not actually helpful in answering our questions.
- The "person(s) of interest" or suspects are the specific measurements or counts that could influence pay levels. These include grade level, seniority, performance appraisal rating, gender, raise, and education/degree level
- And, finally, the "guilty party" is the data that is related to any illegal pay discrepancies we uncover.

Just as with any of our favorite shows, we need to take all of these elements and work through them to come up with the answers to our questions; and, often, an understanding of why the issue exists at all.

The Crime

In this course, we will have a single "crime" to focus on. This issue will form the basis for the lectures each week and the assignments. We will be looking at a Human Resources issue: are males and females getting equal pay for doing equal work?

As background, The Federal Equal Pay Act requires companies to pay males and females the same pay if they are doing (substantially) the same work. We will be taking the role of data analysts (AKA detectives) in a company that has received some evidence that they are going to have a Federal audit of their pay practices due to a complaint on unfair pay practices. Our "job," the basis of the class assignments, is to determine if we do or do not comply with the Equal Pay Act.

HR professionals often examine pay issues from two points of view. One is the actual salary an employee makes, a very obvious approach. This is the approach that you will take as

you do the weekly assignments. Each assignment and each question require you to focus on the actual salaries paid to the employees. What differences do we see, how consistent is the data, what impacts salary outcomes, etc.?

The second approach is more relative in nature and deals with a compensation measure called the compa-ratio (comparison-ratio). This measure compares an employee's salary to the midpoint of the salary grade; this is done simply by dividing the employee's salary by the midpoint of the salary grade the employee is in. (For those not familiar with salary grades, they are groups of jobs within a company that generally require the same skill and knowledge levels and are paid within the same salary range. The midpoints of these ranges are considered to be the market rate, or average pay that companies in the community pay for similar work.) Examining compa-ratios lets HR see how employees are distributed around the midpoint without focusing on the actual different salaries involved. It provides a second way to examine how males and females are paid without worrying about the grades they are in. This approach will be used in the weekly lectures to provide both an example of how to do each homework problem and a way of providing a different view on the equal pay question.

So, each week we will be looking at the pay practices of "our company" in two ways. The lectures and the weekly assignments will each deal with the same questions but will do so with a different measure of pay. In the homework, you will be asked to form tentative conclusions each about the equal pay question using the insights from BOTH the lecture examples of compa-ratio and the salary-based results from your homework problems.

One additional point, the data used in the weekly lectures will be slightly different than the data set you will be working with. We can consider these differences to be minor, as if the lecture uses a different sample of employees, but one that is consistent with the sample used for the homework. The conclusions reached in each week's homework should use the findings from both the lecture's examples and the homework problems. The actual reason for the difference is that students in the past have copied answers from websites and other students and handed them in as their own original work. So, to keep this from happening, the class data set you will be working with changes periodically, so previous answers will not be correct. It does not make sense to redo the lectures every time the data changes, so the lecture's salary and compa-ratio data is comparable but not identical.

Getting Started

In real life on the job or with assignments we often, as do TV detectives, have an overwhelming amount of data that we need to sift through to get to our clues; and then interpret the clues to get the information we need to answer our questions about what happened with the process or outcome we are looking at. The information that we are first presented with is typically a bunch of numbers that measure, count, and code lots of things. Note we have three kinds of data we will deal with:

- Measures tell us how <u>much</u> exists; such as a salary measure would tell us how many dollars someone is paid.

- Counts tell us how many exist, such as counting how many employees have a master's degree.
- Codes tell us about a characteristic; for example, we might code being male as M and being female as F. However, we could also use 0 for male and 1 for female. These numbers do not mean one gender is somehow 'better" or "higher" than the other, they merely show a difference. They are identifiers. More about this latter.

So, as data detectives, we approach any question by finding numbers (measures, counts, and codes) that somehow relate to the issue and the question we need to answer about the situation. Once we have this data, we need to sort thru it to find the clues that need to be examined to understand the situation or outcome. For this class, clues are what we get once we have done some statistical work on the data. This work, as we will see throughout the class, starts with relatively simply summaries – average values for different groups or things, measures of how consistent things are, etc. These summary measures become our first clues. And, just as with any good detective story, not all the clues are meaningful and some are not immediately apparent. The detective/analyst needs to find out what happened what the clues mean to understand and "solve" the crime.

Before we start with the data and how to tease clues from it, we need to understand a couple of concepts:

- *Population*: includes all of the "things" we are interested in; for example, the population of the U.S. would include everyone living in the country.
- *Sample*: involves only a selected sub-group of the population; for example, those selected for a national opinion poll.
- *Random Sample*: a sample where every member of the population has an initial equal chance of being selected; this is the only way to obtain a sample that is truly representative of the entire population. Details on how to conduct a random survey are covered in research courses; we will assume the data we will be working with comes from a random sample of a company's exempt employees. Note: an exempt employee, AKA salaried employee, does not get overtime pay for working more than 40 hours in a week ("exempt from overtime requirements").
- *Parameter*: a characteristic of the population; the average age of everyone in the US would be a parameter.
- *Statistic*: a characteristic of a sample; the average age of everyone you know who attends school would be a statistic as the group is a sub-group of all students.
- *Descriptive Statistics*: measures that summarize characteristics of a group.
- *Inferential Statistics*: measures that summarize the characteristics of a random sample and are used to infer the value of the population parameters.
- *Statistical test*: a quantitative technique to make a judgement about population parameters based upon random sample outcomes (statistics).

The Case

Our class, as a group of data analysts/detectives, will play the role of helping a Human Resources Department (in our assumed "company") prepare for an audit from the government about our pay practices. This routine audit will focus on the question of equal pay for equal work, as required by both State and Federal statutes. Specifically, these require that if males and females are performing substantially the same job (equal work), then they should be paid equally.

Of course, nothing is quite that simple. The laws do allow some differences based on company policies calling for pay differences due to performance, seniority, education, and – with some companies – functional areas. Our company does have policies saying we pay for organizational level (different kinds of work, which are represented by job grades), performance (as measured by the performance rating), seniority, experience, and educational achievements.

Our first step is to decide upon some questions that need to be answered, as questions lead to the data we need to collect. The overall question, also known as the Research Question, is simply: "Do males and females receive equal pay for equal work?" This just means that if a male and female are doing the same work for the company, are they paid the same? As straightforward as this question seems, it is very difficult to answer directly. So, after brainstorming, secondary or intermediate (more basic) questions have been identified as needing to be answered as we build our case towards the final answer. Some of these secondary questions (which will be address throughout the course) include:

- Do we have any measures that show pay comparisons between males and females?
- Since different factors influence pay, do males and females fare differently on them; such as age, service, education, performance ratings, etc.?
- How do the various salary related inputs interact with each other? How do they impact pay levels?

These general questions lead to our collecting data from a random sample of employees. Note that a random sample (covered in research courses) is the best approach to give us a sample that closely represents the actual employee population. The sample consists of 25 males and 25 females. The following data was collected on each employee selected:

- Salary, rounded to the nearest $100 dollars and measured in thousands of dollars, for example an annual salary of $38,825 is recorded as 38.8.
- Age, rounded (up or down) to the age as of the employee's nearest birthday.
- Seniority, rounded (up or down) to the nearest hiring anniversary.
- Performance Appraisal Rating, based on a 100-point scale.
- Raise – the percent of their last performance merit increase.
- Job grade – groups of jobs that are considered substantially similar work (for equal work purposes) that are grouped into classifications ranging from A (the lowest grade) through F (the highest grade). Note: all employees in this study are exempt employees – paid with a salary and not eligible for overtime payments. They are considered middle management and professional level employees.

- Midpoint – the middle of the salary range assigned to each Job Grade level. The midpoint is considered to be the average market rate that companies pay for jobs within each grade.
- Degree – the educational achievement level, coded as 0 for those having a Bachelor's degree and 1 for those having a Master's degree or higher.
- Gender – coded as M for Males, and F for Females, also coded 0 (Males) and 1 (Females) for use in an advanced statistical technique introduced in Week 4.

In addition to these collected measures, the HR Compensation Department has provided the compa-ratio for each employee. The Compa-ratio is defined as the salary divided by the employee's grade midpoint. For example, an employee with a salary of $50,000 and a company salary range midpoint of $48,000 would have a Compa-ratio of 50/48 = 1.042 (rounded to three decimal places). Employees with a Compa-ratio greater (>) 1.0 are paid more than the market rate for their job, while employees with a Compa-ration less than (<) 1.0 are paid less than the prevailing market rate. Compensation professionals use Compa-ratios to examine the spread and relative pay levels of employees while the impact of grade is removed from the picture.

Here is the data collected that will be used in the lecture examples and discussions.

ID	Salary	Compa-ratio	Mid	Age	Perf App.	Service	Gender	Raise	Deg.	Gender 1	Grade
1	58	1.017	57	34	85	8	0	5.7	0	M	E
2	27	0.870	31	52	80	7	0	3.9	0	M	B
3	34	1.096	31	30	75	5	1	3.6	1	F	B
4	66	1.157	57	42	100	16	0	5.5	1	M	E
5	47	0.979	48	36	90	16	0	5.7	1	M	D
6	76	1.134	67	36	70	12	0	4.5	1	M	F
7	41	1.025	40	32	100	8	1	5.7	1	F	C
8	23	1.000	23	32	90	9	1	5.8	1	F	A
9	77	1.149	67	49	100	10	0	4	1	M	F
10	22	0.956	23	30	80	7	1	4.7	1	F	A
11	23	1.000	23	41	100	19	1	4.8	1	F	A
12	60	1.052	57	52	95	22	0	4.5	0	M	E
13	42	1.050	40	30	100	2	1	4.7	0	F	C
14	24	1.043	23	32	90	12	1	6	1	F	A
15	24	1.043	23	32	80	8	1	4.9	1	F	A
16	47	1.175	40	44	90	4	0	5.7	0	M	C
17	69	1.210	57	27	55	3	1	3	1	F	E
18	36	1.161	31	31	80	11	1	5.6	0	F	B
19	24	1.043	23	32	85	1	0	4.6	1	M	A
20	34	1.096	31	44	70	16	1	4.8	0	F	B
21	76	1.134	67	43	95	13	0	6.3	1	M	F
22	57	1.187	48	48	65	6	1	3.8	1	F	D

23	23	1.000	23	36	65	6	1	3.3	0	F	A
24	50	1.041	48	30	75	9	1	3.8	0	F	D
25	24	1.043	23	41	70	4	0	4	0	M	A
26	24	1.043	23	22	95	2	1	6.2	0	F	A
27	40	1.000	40	35	80	7	0	3.9	1	M	C
28	75	1.119	67	44	95	9	1	4.4	0	F	F
29	72	1.074	67	52	95	5	0	5.4	0	M	F
30	49	1.020	48	45	90	18	0	4.3	0	M	D
31	24	1.043	23	29	60	4	1	3.9	1	F	A
32	28	0.903	31	25	95	4	0	5.6	0	M	B
33	64	1.122	57	35	90	9	0	5.5	1	M	E
34	28	0.903	31	26	80	2	0	4.9	1	M	B
35	24	1.043	23	23	90	4	1	5.3	0	F	A
36	23	1.000	23	27	75	3	1	4.3	0	F	A
37	22	0.956	23	22	95	2	1	6.2	0	F	A
38	56	0.982	57	45	95	11	0	4.5	0	M	E
39	35	1.129	31	27	90	6	1	5.5	0	F	B
40	25	1.086	23	24	90	2	0	6.3	0	M	A
41	43	1.075	40	25	80	5	0	4.3	0	M	C
42	24	1.043	23	32	100	8	1	5.7	1	F	A
43	77	1.149	67	42	95	20	1	5.5	0	F	F
44	60	1.052	57	45	90	16	0	5.2	1	M	E
45	55	1.145	48	36	95	8	1	5.2	1	F	D
46	65	1.140	57	39	75	20	0	3.9	1	M	E
47	62	1.087	57	37	95	5	0	5.5	1	M	E
48	65	1.140	57	34	90	11	1	5.3	1	F	E
49	60	1.052	57	41	95	21	0	6.6	0	M	E
50	66	1.157	57	38	80	12	0	4.6	0	M	E

(Note that this table can be copied into an Excel file if you would like to duplicate the examples provided in the lectures.)

What kind of data do we have?

Just as all clues and information uncovered on mystery shows are not equally valuable, or even useful; not all data is equally useful in answering questions. But, all data has some value. As we look at this data set, it is clear that not all the data is the same. We have some measures (salary, seniority, etc.) but we also have some labels (ID, for example merely identifies different employees in the data set, and is not useful for much else). We have some data that are clearly codes, gender and degree for example. In general, our data set can be sorted into four kinds of data, nominal, ordinal, interval, and ratio (NOIR):

- *Nominal:* these are basically names or labels. For example, in our data set we see Gender1 labeled as M and F (for males and females). Other examples of nominal data include names of cars (Ford, Chevrolet, Dodge, etc.), cities and states, flowers, etc. Anything where the name/label just indicates a difference from something else that is similar is nominal level data. Now, we can "code" with words and letters (such as Male or M) but we can also code them with using 0 and 1(for male and female) as we do with the Gender variable. Regardless of one looking like a label (letters) and one looking like a measurement (numbers), both of these are simply ways to label males and females – they indicate a difference between the groups only – not that one is somehow higher than the other (as we typically think of 1 as higher or more than 0).
  Nominal level data are used in two ways. First, we can count them; for example, how many males and females exist in the group? Second, we can use them as group labels to identify different groups, and list other characteristics in each group; a list of all male and female compa-ratios will be quite helpful in our analysis, for example.
- *Ordinal:* these variables add a sense of order to the difference, but where the differences are not the same between levels. Often, these variables are based on judgement calls creating labels that can be placed in a rank order, such as *good, better, best*. The grade and degree variables in our data set are ordinal. We cannot assume that the amount of work to get the higher degree or higher job grade is the same for all differences. Note: Even though we only show education as bachelor and graduate, we could include no high school diploma, high school diploma on the low end and doctoral degree and professional certification on the upper end.
- *Interval:* these variables have a constant difference between successive values. Temperature is a common example – the difference between, for example, 45 and 46 degrees is the same amount of heat as between 87 and 88 degrees. Note: Often, analysts will assume that personal judgement scores such as Performance Appraisal ratings or responses on a questionnaire scale using scores of 1 to 5 are ordinal as it cannot be proven the differences are constant. Other researchers have suggested that these measures can be considered interval in nature for analysis purposes. We will consider performance appraisal ratings to be interval level data for our analysis purposed.
- *Ratio:* these are interval measures that have a 0 point that means none. For example, while 0 dollars in your wallet means no money, a temperature of 0 degrees does not mean no heat. Ratio level variables include salary, compa-ratio, midpoint, age, service, and raise – even if our measurements do not go down to 0, each measure does have a 0 point that would mean none.

These differences are important, as we can do different kinds of analysis with each level, and attempting to use the wrong level of data in an approach will result in misleading or wrong outcomes. Within our data set our variables fit into these groups.

- Nominal: ID, Gender, Gender1 (merely labels showing a difference)
- Ordinal: Grade, Degree (can be ordered from low to high, ex Grade A is the lowest and Grade F is the highest grade.)

- Interval: Performance Rating (Constant difference between values, but no meaningful 0 score)
- Ratio: Salary, Compa-ratio, Midpoint, Seniority, Age, Raise (All have a 0 point that means none)

Wow – a lot of background material. But, now that we have this covered, we can get to actually looking at our data. As suggested above, the first question we need to ask is "do we have any measures that show pay comparisons between our males and females?"

Now, we move on to some specific ways we can use to see if the company is guilty of not paying males and females the same for doing equal work in lectures two and three for the week.

Summary

This class is about uncovering the secrets hidden within data sets. As data detectives (AKA data analysts), we need to develop both the tools and the logic to examine data sets and use them to answer business questions.

This class, will use a single business question: Are males and females paid the same for doing equal work? Each week we will look at different tools and techniques to summarize and make sense out of the data set assigned to your class.

We looked at a lot of what we could call "background" in this lecture. Information that is needed to understand what we are doing but which does little more than set-up the problem. This included the lecture's data set and definitions of the variables to be used, some statistical concepts that help identify what we are doing and the kinds of data that we are using.

Next Steps

Please respond to Discussion Thread 1 for this week with your initial response and responses to others over a couple of days before moving on to reading the second lecture for the week. Ask questions and please share what is unclear. Statistics has been compared to learning a new language, we need to understand what the terms mean and how they apply.

Please ask your instructor if you have any questions about this material.