



Introduction to **STATISTICS**

by Jim Mirabella

Table of Contents

Chapter One: Graphing & Descriptive Statistics

What is Statistics?	T	1-1
Why Do We Learn Statistics?.....	R	1-1
Descriptive vs. Inferential Statistics	A	1-1
Basic Terminology.....	V	1-1
Types of Variables	I	1-2
Levels of Measurement	S	1-2
Organizing and Graphing Categorical Data	-	1-3
Organizing and Graphing Numerical Data	K	1-5
Descriptive Statistics	E	1-6
Central Tendency	N	1-7
Dispersion	D	1-8
Percentiles and Quartiles	A	1-10
Discussion Questions	L	1-11
Practice Problems	L	1-11
Assigned Problems	L	1-12

Chapter Two: Probability & the Normal Distribution

Basics of Probability	A	2-1
Solving Probability Problems	L	2-2
The Normal Distribution	I	2-4
The Central Limit Theorem	C	2-7
Discussion Questions	E	2-9
Practice Problems	E	2-9
Assigned Problems		2-10

Chapter Three: Confidence Intervals & Sample Sizes

Sampling	4	3-1
Confidence Intervals	0	3-6
Sample Size	5	3-12
Discussion Questions	1	3-15
Practice Problems	T	3-15
Assigned Problems	S	3-16

Table of Contents

(continued)

Chapter Four: Hypothesis Testing for One Sample

The Logic of Hypothesis Testing	T	4-1
Z-tests vs. t-tests	R	4-6
Closing Notes	A	4-6
Discussion Questions	V	4-7
Practice Problems	I	4-7
Assigned Problems	S	4-8

Chapter Five: Hypothesis Testing for Two or More Samples

Choosing the Correct Test	K	5-1
T-test for Independent Samples	E	5-2
T-test for Paired Samples	N	5-5
Analysis of Variance	D	5-6
Discussion Questions	A	5-8
Practice Problems	L	5-8
Assigned Problems	L	5-9

Chapter Six: Chi Square Testing

Categorical Analysis	,	6-1
Goodness of Fit		6-1
Crosstabulations	A	6-5
Discussion Questions	L	6-9
Practice Problems	I	6-9
Assigned Problems	C	6-10

Chapter Seven: Correlation & Regression

Correlation	T	7-1
Regression	4	7-5
Multiple Regression	0	7-6
Discussion Questions	5	7-10
Practice Problems	1	7-10
Assigned Problems	T	7-11

Chapter Eight: Data Analysis Project

Appendix: Pivot Table Tool Tutorial

CHAPTER ONE

GRAPHING & DESCRIPTIVE STATISTICS

What is Statistics?

We use numbers every day to analyze the world or our piece of it, whether it be related to sports, finance, weather, politics, etc. Statistics is essentially a compilation of tools and techniques used for describing, organizing and interpreting these numbers; in other words, it is a means for converting data into information. In the business world, there are four main uses of statistics → to summarize business data, to draw conclusions from that data, to make reliable forecasts, and to make decisions important to the business.

Why Do We Learn Statistics?

While it would be ideal if everyone could strive to be a doer of statistics, the aim here is for you to become an adept user of statistics. Rather than trust headlines without substance, you might be able to see through the numbers and question results intelligently. You won't just trust the media or politicians blindly. You can play fantasy sports with greater success. You can make business decisions with facts and figures that you can interpret and not just with gut instinct. Above all else, you will think differently and see the world in a way few people do.

Descriptive vs. Inferential Statistics

Statistics can be broken down into two distinct branches. The first part of the course focuses on **descriptive statistics**, which is the collecting, summarizing, presenting and analyzing of data sets. Here we use tables, charts and graphs to display and picture the data, and we compute statistics to describe the data with single numbers. The second part of the course focuses on **inferential statistics**, which involves taking the results from a data set and drawing inferences to the larger data sets that it came from. This is where the power of statistics truly comes alive. It is here that conclusions are drawn, decisions are made, and forecasting models are created.

Basic Terminology

A **variable** is a characteristic of an item or individual that is measured. If you break up the word, you will note that a variable must be *able to vary*. Ages, heights, weights, genders are variables because they each have varying values, but if we were studying only those who wear

contact lenses, that would be a constant and not a variable. For a given variable, an **observation** is the simplest form of data; it is a single value in a data set. When we have all possible observations of a variable, we have a **population**. It is the population that we typically wish to evaluate with our analysis, and we do so by utilizing a subset of that population, known as a **sample**. The numerical measures we compute from a population are known as **parameters**, while the numerical measures we compute from a sample are known as **samples**. To help remember this, note that *parameter* and *population* begin with “p”, while *sample* and *statistic* begin with “s.”

Types of Variables

Variables can be classified as either **categorical** or **numerical**. A **categorical** variable is essentially qualitative and has non-numerical values, like yes / no, or red / white / blue, or republican / democrat, or freshman / sophomore / junior / senior. Most survey questions have categorical responses, and the analysis of such data is limited to counting the frequency of each value and computing their relative percentages. A **numerical** variable is quantitative in nature and its values are a measurement; these variables can be further subdivided into **discrete** and **continuous** variables. **Discrete numerical** variables are finite integers, such as the number of children in a household, or the number of books on a shelf or the number of cars in a parking lot. **Continuous numerical** variables can have any numerical value along a continuum, and can be fractions or have many decimal places; with continuous data, it is not unusual to see no repetition in the data set. For example, if you measured the times of runners in the NYC marathon, you will see as many different times as there are runners, with a rare tie, and the times are computed in fractions of a second. Likewise if we measured the heights or weights of people, while we tend to round off weight to the nearest pound and height to the nearest half-inch, if measured precisely, you would rarely if ever see a tie. And because of this lack of repetition, analysis of continuous data must be handled differently from discrete data, as shown in the later section on organizing and graphing numerical data.

Levels of Measurement

Before beginning to analyze data, you must determine the appropriate analytical techniques to use, and that question is answered by knowing the **level of measurement** of the variables in question. There are four levels of measurement → **nominal**, **ordinal**, **interval** and **ratio**. A **nominal** variable, also known as categorical, is essentially a naming variable. It has no measurement tied to it whatsoever. Most of the time its values have no numbers (e.g., yes / no, male / female, Caucasian / African-American / Asian / etc.) but sometimes it can have numerical values which are truly categorical in nature and have no quantitative meaning (e.g., a zip code, an area code, or even on surveys where codes are used and maybe 1 = female and 0 = male). You would never average the zip codes in a class to determine where the typical student lives, but you might count how many live in each zip code to understand the commutes for the entire class or how the school attracts potential students by location. Whether the values have numbers or just text, nominal data can only be analyzed at the simplest level (i.e, you can count the occurrences of each value and compute percentages too).

The next level up from nominal data is **ordinal**, which includes a ranking or order in that we know one value is greater than another which is greater than another, but it is not apparent as to how much they differ. Surveys often use Likert scales (i.e., *Strongly Agree – Agree – Neutral – Disagree – Strongly Disagree*) and you are asked to mark 1 for *Strongly Disagree* up to 5 for *Strongly Agree*. We know that *Strongly Agree* is higher than *Agree* and *Disagree* is higher than *Strongly Disagree*, but the gaps are not consistent, and they vary by respondent. Even letter grades at school are ordinal → while an A is always higher than a B, and a C is always higher than a D, grades of 90, 89, 79 and 60 should illustrate that the A (90) is only slightly higher than the B (89), while the C (79) is significantly higher than the D (60), and so the ranking is clear but the gaps are inconsistent.

The two highest levels of measurement are **interval** and **ratio**, both of which maintain the ranking / order of magnitude but now they also have consistent gaps in the data. The only difference between the two is that **ratio** data also has an absolute zero, meaning that zero on its scale truly means “the absence of”, and as a result, we can do ratio comparisons on the data. 200 pounds is twice as heavy as 100 pounds and four times as heavy as 50 pounds (note the ratio comparisons), and this is possible because 0 pounds means the absence of weight, so weight in pounds is a ratio variable. A golf score of +2 (i.e., 2 over par) is 1 stroke worse than +1 but it is not twice as bad, and that is because 0 (i.e., par) is not the absence of strokes, but is instead just another value on the scale; thus golf scores are interval data. Common ratio variables include height, weight, age and income, while common interval variables include golf scores and temperatures in Fahrenheit or Celsius. Truthfully, when it comes to statistical analysis, it doesn’t matter if the variable is interval or ratio, as you can perform the same techniques on both (with the exception of a geometric mean, which is rarely used and which will not be part of this course). Some software packages even simplify matters further by using the term **scale** to represent **interval** or **ratio** data, so we would really only have nominal, ordinal and scale data. Since the intent here is to determine the correct statistical technique, it seems a sensible idea.

Organizing and Graphing Categorical Data

When we wish to analyze categorical / nominal data, we are limited in what we can do. Essentially we can count how many of each value we have, and compute the percentage of each value, and then we can graph the results in a bar chart or pie chart. Just because it is a limitation doesn’t mean it isn’t valuable; in fact, one advantage is that you know exactly what you need to do with the data.

Suppose we had a sample of 25 students and we had a variable that represented the student’s class year. Using the **Categorical_Data_Analysis.xls** file provided (shown in Figure 1), we would copy the 25 values into the *Data* column (after deleting whatever data is there as the default). Then in the *Values* column we would list all of the unique values in the data set (and you can see there are only four class years to list → Freshman, Sophomore, Junior and Senior, spelled exactly as it is spelled in the data set). The *Frequency* column counts the total occurrences of each value, and it shows 4 Freshmen, 5 Sophomores, 7 Juniors and 9 Seniors, for a total of 25 students as the sample size. The *Relative Frequency* column computes the percentages of the sample size represented by each value, and the 4 Freshmen in the sample of 25 accounts for $4 / 25 = 16\%$ of the sample. Similarly, 20% of the sample are Sophomores, 28% are Juniors and 36% are Seniors, for a total of 100%.

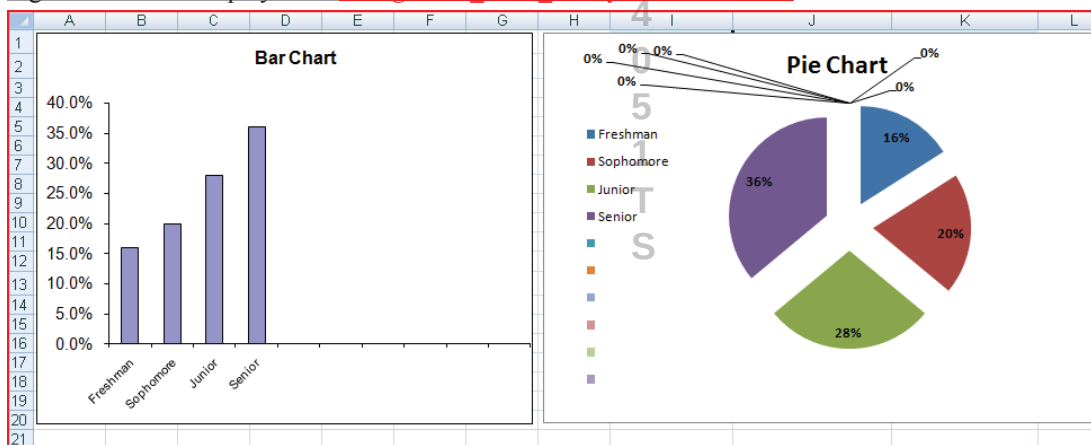
This is a complete Frequency table of the class years in our data set and we can now begin to make analytical statements about the data that we couldn't do before (i.e., if you look at the list under the *Data* column, what can you really say?). Things to note might be that there are increasing numbers of students at the higher class levels, or that there are as many Seniors as there are Freshmen and Sophomores combined. Note that we cannot give reasons for this, as it is only a small sample and we don't know much about the population. It might be tempting to make a comment about how new enrollments are down or how there are greater numbers of late transfers into the school or how dropout rates are highest at the lowest levels, but we know nothing except what we see. So all we can do is to describe what is there and leave it to experts (i.e., those with a working knowledge of this particular data set) to draw conclusions or make guesses about what might be occurring.

1	A	B	C	D	E	F
2	Observation #	Data		Values	Frequency	Relative Frequency
3	Observation 1	Senior				
4	Observation 2	Sophomore		Freshman	4	16.0%
5	Observation 3	Junior		Sophomore	5	20.0%
6	Observation 4	Senior		Junior	7	28.0%
7	Observation 5	Freshman		Senior	9	36.0%
8	Observation 6	Senior				
9	Observation 7	Junior				
10	Observation 8	Senior				
11	Observation 9	Freshman				
12	Observation 10	Sophomore				
13	Observation 11	Junior				
14	Observation 12	Senior		TOTAL	25	100.0%
15	Observation 13	Freshman				
16	Observation 14	Sophomore				
17	Observation 15	Junior				
18	Observation 16	Senior				
19	Observation 17	Freshman				
20	Observation 18	Sophomore				
21	Observation 19	Junior				
22	Observation 20	Senior				
23	Observation 21	Junior				
24	Observation 22	Sophomore				
25	Observation 23	Junior				
26	Observation 24	Senior				
27	Observation 25	Senior				
28	Observation 26	Senior				

Figure 1: Screen display from Categorical Data Analysis.xls

There was a great song about how a picture paints a thousand words. The “thousand” might be debatable but the point is well-taken. If you click on the *Charts* tab at the bottom of the screen, you will be taken to another sheet in the Excel file where the tabulated data is charted (see Figure 2).

Figure 2: Screen display from Categorical Data Analysis.xls – Charts



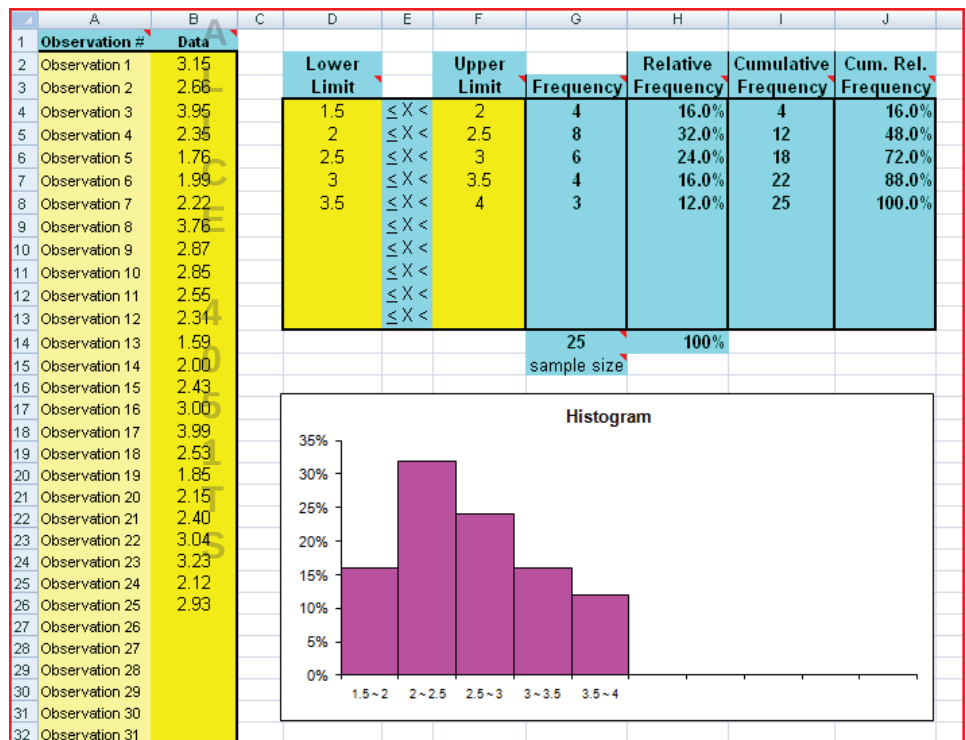
On the left is a Bar Chart in which each bar is one row from the Frequency Table and illustrates the relative frequencies. Note that the Freshman bar lines up with 16% and the Sophomore bar lines up with 20%, and so on. Visually you can see increases in class sizes over the years. On the right is a Pie Chart which treats the whole pie as 100% and each value has a slice of the pie, sized in accordance with the relative frequencies. While it has the disadvantage of getting cluttered at times and not being as visually appealing as bar charts, it has the advantage of being out of 100%, so each slice tells you how much of the whole it truly represents (which the bar chart doesn't do as well). Together the two charts should help to paint a more vivid description of the data, but still be warned not to overstep your bounds and do more than describe what you see (which is the essence of descriptive statistics anyway). If a pattern looks interesting and warrants further investigation, then get more data and keep digging, but don't conclude beyond the scope of the data that you have, especially if you are not extremely knowledgeable about that sample.

Organizing and Graphing Numerical Data

When analyzing numerical variables, first determine if the data is discrete. If it is, you could use the **Categorical_Data_Analysis.xls** file as an option providing there aren't too many different values in the data set. For example, if a subdivision had some homes with 0, 1, 2, 3 or 4 children in each house, it would work just fine to count the totals, compute relative percents, and create bar and pie charts. But if there are a lot of different values (like the ages in years of each house) or if the data is continuous, this would clearly not work. Because of the lack of repetition in continuous data, we cannot count the values as we would with discrete data, so we group the data into ranges and count frequencies within each range. This is known as a Frequency Distribution and is addressed in the **Frequency_Distribution.xls** file.

Suppose we had a sample of 25 students and we had a variable that represented the student's GPA. Using the **Frequency_Distribution.xls** file provided (shown in Figure 3), we would copy the 25 values into the *Data* column (after deleting whatever data is there as the default). Since the values range from 1.59 to 3.99, the first row should have a lower limit of 1.59 or less, and the last row should have an upper limit that exceeds 3.99. It is important that each row have consistent widths (i.e., the difference between the upper and lower limits) or it can be deceiving. For simplicity's sake, let's use intervals of $\frac{1}{2}$ point in width, ranging from 1.50 up to 4.00. Make sure that the upper limit of one row becomes the lower limit of the next row so there are no gaps or overlaps.

Figure 3: Screen display from **Frequency_Distribution.xls**



Note that exceptions can be made on the consistent widths to capture an extreme value at either end. If there were a single student with a 0.25 GPA, for example, you might not want to add a bunch of rows that have frequencies of zero just to capture that one value, so you might make the first range from 0.00 to 2.00, and then continue 2.00 to 2.50, etc. Once the ranges are established, the *Frequency* column computes how many of the observations are in that range. There are 4 observations between 1.50 and 2.00, 8 observations between 2.00 and 2.50, etc. When added together, there are 25 total observations. If the total shown in the Frequency Distribution does not match the total in the data set, it means that the intervals were not set up properly and there are gaps or overlaps, so be sure to fix them. The *Relative Frequency* column computes the percentages of the sample size represented by each interval, and the 4 students with GPAs between 1.50 and 2.00 accounts for $4 / 25 = 16\%$ of the sample. Similarly, 32% of the sample have between 2.00 and 2.50, etc., for a total of 100%. The *Cumulative Frequency* column essentially accumulates the total by adding up the frequencies in that row plus all prior rows; thus a cumulative frequency of 18 in the 2.50 ~ 3.00 row means that 18 students have less than 3.00 ($4 + 8 + 6 = 18$, which are the frequencies of the first 3 rows). This same pattern continues for all rows, with the cumulative frequency of the last row being exactly the same as the sample size. Finally, the *Cumulative Relative Frequency* column computes the percentages of the sample size represented by each interval's cumulative frequency, and the 72% in the 2.50 ~ 3.00 row means that 72% of the sample has less than a 3.00 GPA (which can be computed by dividing 18 into the sample size of 25 OR by adding $16\% + 32\% + 24\%$ from that row and all prior rows).

The graphical display of the frequency distribution is known as a *histogram*. It is essentially a bar chart but, unlike that of the categorical data analysis, this one has the bars touching. Since the data is continuous, there is no distinct gap between the bars (i.e., where one ends, the next begins). The lower and upper limits are displayed beneath each bar so you can visibly see the distribution by ranges. You can see a peak in the GPAs in the 2.00 ~ 2.50 range and then a steady decline as the GPAs increase. Often analysts will try to overlay a bell-shaped curve atop the histogram to see if it appears to be a normal distribution (which will be addressed in a later unit). Essentially what we have done is to take a data set with all or almost all unique values on a continuum and grouped them in such a way that we could draw a picture of the distribution of GPAs, thereby making order out of chaos and getting information from data.

Descriptive Statistics

While graphing is a helpful method for helping to describe a data set (providing the graphs are created properly), a more common approach is to compute descriptive statistics. To describe any data set, it is imperative that there be two different measures – one of central tendency, and one of dispersion. Using the **Descriptive_Statistics.xls** file provided (shown in Figure 4), we could copy data from a sample in the *Data* column and explore several measures of Central Tendency and Dispersion as well as some other useful measures.

Figure 4: Screen display from **Descriptive_Statistics.xls**

1	A	B	C	D	E	F	G	H	I	J
2	Observation #	Data								
3	Observation 1	1								
4	Observation 2	2								
5	Observation 3	4								
6	Observation 4	5								
7	Observation 5	7								
8	Observation 6	9								
9	Observation 7	10								
10	Observation 8	10								
11	Observation 9	12								
12	Observation 10	15								
13	Observation 11	16								
14	Observation 12	18								
15	Observation 13	22								
16	Observation 14	23								
17	Observation 15	25								
18	Observation 16	26								
19	Observation 17	28								
20	Observation 18	32								
21	Observation 19	33								
22	Observation 20	40								
23	Observation 21									

Measures of Central Tendency				Measures of Dispersion			
Mean		16.90		Range		39.00	
Median		15.50		Interquartile Range		18.00	
Mode		10.00		Standard Deviation		11.29	
Midrange		20.50		Variance		127.57	
				Coefficient of Variation		66.83%	

Other Measures			
Sample Size		20	
Minimum		1.00	
Maximum		40.00	
First Quartile		7.00	
Third Quartile		25.00	
Given percentile =		90%	
		Value of percentile =	32.1

Mean +/- 1 standard dev.	5.61	28.19
# Observations within 1 standard dev.		13
Mean +/- 2 standard dev.	-5.69	39.49
# Observations within 2 standard dev.		19
Mean +/- 3 standard dev.	-16.98	50.78
# Observations within 3 standard dev.		20

Central Tendency

Central tendency is the extent to which the data groups around some central value; it is essentially a measure of location and sometimes answers questions regarding what a typical value is. There are three common measures of central tendency – mean, median and mode – and one less common measure – midrange.

The **mean** is the most common measure of central tendency. Commonly referred to as an average, it is the balancing point in the data. If a 10-foot board had a number line drawn on it and bricks placed at different places on it, the mean would be the marking on the number line where a fulcrum would balance the bricks on the board perfectly. A PICTURE OF THIS WOULD BE HELPFUL. The population mean (symbolized by the Greek letter μ , pronounced “mu”) is computed by adding the values in the population and dividing by the size of the population (symbolized by “N”). The sample mean (symbolized by X-bar, which is just a capital X with a line over it) is computed by adding the values in the sample data set and dividing by the sample size (symbolized by “n”). So if a data set consisted of 1, 3, 6, 10, the sample mean would be $(1+3+6+10)/4 = 20/4 = 5.00$. While the mean is susceptible to being pulled in either direction by extreme values (known as **outliers**), it has an advantage in that it is computed from all of the data in the set and it is an excellent predictor / estimator of the true value of the population mean.

The **median** is the middle value or mid-point in the data set that is sorted in ranked order. It tells you that half of the data is above and half is below this special number. Unlike the mean, the median is not susceptible to outliers. If the data set consisted of 1, 3, 6, 8, 8, 10, 99999, the median would be 8 since it is the 4th highest (or 4th lowest) value in the sample of 7 values, and the extremely large 99999 has no influence on the median. When the sample has an odd number of values, the median falls on one of them, but when the sample has an even number of values, the median falls midway between the two center values (so if the data set consisted of 1, 3, 7, 8, 8, 10, the median would be 7.5 which is halfway between the 7 and 8 in the middle).

The **mode** is the most commonly occurring value in the data set. Just like with the median, outliers have no effect on the mode, but unlike the median, it can be located anywhere in the data set (so it is not necessarily central). If the data set consisted of 1, 1, 1, 2, 3, 3, 99999, the mode would be 1 since there are more of it than any other value, but note that it is also the smallest value in the data set.

A less common measure of central tendency is the **midrange**, which is merely the average of the largest and smallest values in the data set. If the data is rather symmetrical, the midrange will be relatively central, but if the data is skewed one way or the other because of an outlier, the midrange will be far from central. If the data set consisted of 1, 2, 3, 4, 5, 99, the midrange would be computed as $(1+99)/2 = 50$. The midrange is the least used of the measures of central tendency, and would only likely be used in two situations – when only the high and low are known or when someone just wants to do a quick-and-dirty computation of the approximate center when there is a lot of data and you just don't have time to do the math. So when looking at temperatures throughout the day, normally the highs and lows are reported instead of minute by minute temperatures, so hearing that the high will be 90 and the low will be 70 would make one think the temperature to be around 80 throughout the day.

Each of the three common measures of central tendency have their advantages and disadvantages. When describing what the typical value is for a situation, you could opt to use any of these measures, but the rule of thumb is to only use the mode if there is repetition of a few values, and then the median if there are outliers which would distort the results, and then the mean in all other cases. If you were describing the typical salary of an employee, you wouldn't likely use the mode since there isn't likely to be repetition (unless there were a lot of employees on the same starting salary or minimum wage), and a few high salaries would distort the mean, so the median makes the most sense. If you were describing the typical number of courses each student in the class is taking, everyone is likely taking 1, 2, 3 or 4 classes, with a rare 5+, so the mode makes sense here (you wouldn't want to use the mean and report that the typical student is taking 2.736 classes, would you?). And if you wanted to describe the typical GPA of the MBA students, you couldn't use the mode since repetition is infrequent, and considering the GPA is bounded in a small range (it cannot exceed 4.00 and cannot dip below the minimum requirement) there is not likely to be outliers to distort results, so the mean is an excellent choice. There can be more than one option for different situations, but it is important to understand how each are used, and to know that people often use the one that best makes their case. Thus it is important to look at all measures of central tendency and decide for yourself the best way to truly describe the data.

Dispersion

Dispersion is the variation or amount of spread of the data around the central value. If you were to ask about the typical price of a hotel room in midtown Manhattan, and the response was "about \$200", that essentially means \$200 give or take some amount. Well the dispersion is essentially a way of measuring how big is the "about". There are several measures of dispersion, and only a couple of known well or used much, and one is better off not being used.

When asked for the range of a data set, we often think to just state the smallest and largest number (e.g., ages range from 18 to 50). But the formal definition of **range** in Statistics is the difference between those two numbers (i.e., $50 - 32 = 18$). This simplest descriptive measure of variation essentially captures the total spread in a set of data, and is most sensitive to outliers since its computation only involves the two most extreme values in the data set.

The **standard deviation** (represented by the Greek letter sigma “ σ ” for populations and by the letter “s” for samples) is a bit more complex to grasp conceptually due to its unappealing formula. But if you look at the formula closely you see that the numerator is essentially asking you to compute the differences between each value in the data set vs. the mean, and the denominator is essentially the sample size; thus the standard deviation computes the average deviation / scatter around the mean. Now some differences are positive and some are negative, and if you added them up, it would always total zero, so we square the differences to make them all positive, and then take the square root to compensate for the squaring. So the standard deviation is essentially the average weighted distance of all the data about the mean. By itself, a standard deviation doesn’t say much about a data set, lending more support to the notion that you should always look at a measure of central tendency jointly with a measure of dispersion. As we will see in later units, the standard deviation combined with the mean is very informative. For example, if you took the mean and added and subtracted one full standard deviation from it, you would find that approximately 68% of the data will likely fall in the range between those two values (note in the display in figure 4, the mean minus one standard deviation equals 5.61 and the mean plus one standard deviation equals 28.19, and 13 of the 20 values in the data set (i.e., 65%) falls in the range between 5.61 and 28.19. Additionally, approximately 95% should fall in the range of the mean $\pm$ 2 standard deviations (the example in figure 4 shows 19/20 or 95% in that range). And approximately 99.7% should fall in the range of the mean $\pm$ 3 standard deviations (the example in figure 4 shows all 100% of the 20 in the sample in that range). It doesn’t always work this perfectly, but with larger data sets, it comes closer to being true. It also means that knowing the standard deviation by itself, you could expect the range to be approximately 6 times its size for large data sets (but for smaller data sets, it is more like 3-4 times its size). This is truly a measure that you get more comfortable with as you use it. For now, understand that it works with means to define a data set.

The **variance** (represented by “ σ^2 ”, or sigma-squared, for populations and by s^2 for samples) is truly a meaningless measure. This is not the same variance used in finance to determine the difference between budget vs. actual expenses. While the standard deviation is in the same units as the mean and median, the standard deviation’s units make no sense since it is squared (after all, if a standard deviation were 3 years or 5 pounds or 7 dollars, then how could you explain a variance of 9 square years or 25 square pounds or 49 square dollars). Because of this confusion with the units, most statistics books just choose to omit the units, and that only adds to the confusion further. So for our purposes, we will not discuss the use of variances further.

The **coefficient of variation** is not really a measure of variation but rather a descriptive measure that gives more meaning to how large the spread truly is. If you were told that the standard deviation of salaries in an organization is \$100,000, you might think it a lot, but if you found out the organization was an NFL franchise, then suddenly it doesn’t sound like much of a spread given

the many salaries in the millions. Likewise a spread of \$5,000 in salaries sounds small, but if it were for factory workers averaging \$15,000, then it is colossal. To compute the coefficient of variation, you merely divide the standard deviation by the mean and convert to a percentage. So if the mean weight of students in a class is 150 pounds and the standard deviation is 15 pounds, then the coefficient of variation = $15 / 150 = 10\%$. This is helpful if you wished to compare to another group to see which has a greater spread, relatively speaking.

To understand the final measure of dispersion, it is important to first address **quartiles**. The median divides the data into two equal halves, which we will call the upper half and the lower half. If you compute the median of the upper half of the data, that value is known as the **upper quartile**, and it separates the top 25% from the lower 75%. If you compute the median of the lower half of the data, that value is known as the **lower quartile**, and it separates the bottom 25% from the upper 75%. The values between the lower quartile and the upper quartile represent the middle 50% of the data, and the size of that range, known as the **interquartile range**, is computed as the upper quartile minus the lower quartile. Unlike the range which is sensitive to even a single outlier, the interquartile range eliminates the top and bottom 25% and all of its outliers, and tells you how spread out the core of the data is. In the display in Figure 4 you can see a range of 39.00 but an interquartile range of 18, so the middle 50% has only an 18-point spread while the entire data set has a 39-point spread.

Percentiles and Quartiles

If you have ever taken a standardized test, you will notice how often the scores were reported as percentiles, on a scale from 00 to 99. Those who don't understand statistics might ask their teachers why the top score is 99, and why not 100. Teachers who are not "in the know" or who just don't feel like explaining statistics to a child might opt for the simplest response → "nobody's perfect!" But that is not why the scores are the way they are. A **percentile** measures the percentage of values in the data set which fall below a certain number. The lowest value in a data set is the 0th percentile because nothing is lower than it, and it is impossible for any value in the data set to be the 100th percentile since it is impossible for a score to be lower than itself (and 100% of the data must be below it). In Figure 4, the largest value would be the 95th percentile since 19/20 fall below it, and in a larger data set it might even reach the 99th percentile. To put it in perspective further, the median is also known as the 50th percentile since 50% fall below it, the upper quartile is also known as the 75th percentile since 75% fall below it, and the lower quartile is also known as the 25th percentile since 25% fall below it. This is a useful measurement since you can exclude outliers and set whatever cutoff you wish (e.g., if on a physical fitness test, it is determined that the bottom 20% won't make the cut, you can compute the 20th percentile to determine what the cut is).

CHAPTER ONE KNOWLEDGE ASSESSMENT

Graphing & Descriptive Statistics

Discussion Questions

DISCUSSION QUESTION 1

CHOOSING THE BEST MEASURE: Suppose you applied to be a waiter or waitress at a local restaurant and asked the manager what a typical dinner shift was like. How might the manager describe the typical size of a party at a table? The mode would be useful, since there are a few repeated values. You might have an occasional table for 20, but, for the most part, parties consist of one to five people. What if you asked for the typical size of a check, so you could estimate your tips? The mode makes no sense, since you would not likely have many repeats, and the mean is easily distorted by that table for 20, but the median tells you that half the time you can expect to earn a certain amount of money in tips. Now suppose you were a real estate agent and you were asked by a client about the “typical” home in a subdivision. Being the astute agent you are, you have gathered the following information on each house in the subdivision: Price, square footage, numbers of bedrooms, number of bathrooms, and age. What statistic (mean, median, or mode) would you use to describe each aspect of the typical home and why? Try to imagine the type of answers you would be giving your client based on your selections for a subdivision that has 100 homes with a wide variety of sizes and prices.

DISCUSSION QUESTION 2

TRUSTING STATISTICS: Do you always trust the statistics you see in print? How about the ones you hear on the news or on your favorite talk show? What makes the statistics credible? For this question, I want you to tell me what personal criteria you would (or should) use to determine if you should believe reported statistics.

Practice Problems: Real Estate

Solutions are provided to practice problems so you can check your work.

Use the Real_Estate.xls file which consists of 100 homes purchased in 2007 and appraised in 2008. It includes variables regarding the number of bedrooms, number of bathrooms, whether the house has a pool or garage, the age, size and price of the home, what the house is constructed from, how far it is to the city center, and the appraisals from two agents.

PRACTICE PROBLEM 1:

Perform a categorical analysis on the construction of the homes. Describe your findings.

PRACTICE PROBLEM 2:

Create a frequency distribution & histogram on the sales price of the homes. Use ranges of 125-150k, 150-175k, ..., 300-325k. Describe your findings.

CHAPTER ONE KNOWLEDGE ASSESSMENT

Graphing & Descriptive Statistics

PRACTICE PROBLEM 3:

Compute the descriptive statistics for the sizes of the homes. Describe your findings.

Assigned Problems: Student Data

Use the Student_Data.xls file which consists of 200 MBA at Whatsamattu U. It includes variables regarding their age, gender, major, GPA, Bachelors GPA, course load, English speaking status, family, weekly hours spent studying.

ASSIGNED PROBLEM 1:

Perform a categorical analysis on the majors of students enrolled in the MBA. Describe your findings.

ASSIGNED PROBLEM 2:

Create a frequency distribution & histogram on the GPAs of the students. Use ranges of 2.50-2.75, 2.75-3.00, 3.25-3.50, 3.50-3.75 and 3.75-4.00. Describe your findings.

ASSIGNED PROBLEM 3:

Compute the descriptive statistics for the weekly hours spent studying. Describe your findings.

T
R
A
V
I
S
-
K
E
N
D
A
L
L
,
A
L
I
C
E

4
0
5
1
T
S