Chapter 3

Using Models to Interpret Data[1]

What is this chapter about? It's about taking data, possibly thousands of numbers, and finding a few measures (values) that help you make sense of the data and represent it effectively. The main tools you will use are the mean, the standard deviation, and Pivot Tables (an Excel feature). The **mean** turns out to be the simplest and most commonly used model of data. The **standard deviation** can be thought of as a measure for how closely this model fits the data (or equivalently, how appropriate the mean is in modeling the data). Thus, we have the two basic pieces of a model: the model itself (the mean) and a measure of how well the model fits (standard deviation). Another way to think about this process is that we are taking a huge amount of information (the original data) and compressing it, reducing it to fewer pieces of information that give us a sense of the entire data set. Of course, we lose some of the information in the process, but we gain efficiency and a way of communicating and making decisions that would be extremely difficult using only the data itself. In this sense, the mean is the simplest possible model we can produce: we take all of the numerical data, no matter how numerous, and reduce it to one number for each of the numerical variables in the data set. In order to evaluate the quality of this model for each variable, we then compute the standard deviation of that variable.

Section 3.1 (page 63) of the chapter shows you how to use the mean as a model for the data, and how the standard deviation is a measure of how well this model represents the data. Section 3.2 (page 81) of the chapter shows you how to reduce data that has several variables, some of which are categorical, to several means using an Excel feature called **Pivot Tables**.

- *As a result of this chapter, students will learn*
  - √ What a mean is and how it can be used to model the average or typical data point
  - √ How to use the standard deviation as a tool for determining how well the mean represents the data
  - √ What pivot tables are and how they are useful
- *As a result of this chapter, students will be able to*
  - √ Compute means and standard deviations by hand, with Excel, and with add-ins like StatPro

√ Make a Pivot Table that cross-sections your data in order to help you analyze it

3.1 The Mean As A Model

Consider what we have so far: a lot of information in the form of spreadsheets filled with data that we arranged into variables and observations. But what do we do with all this? Unless you're really special, you probably can't learn a lot from looking at a list of one thousand numbers. You probably know even less from looking at a thousand observations for each of four different variables. Sets of data in business and science are usually larger than this, so we need to think of a more efficient analysis tool. The tool we will use is to build a model of the data. A **model** is a number or formula that represents a set of data - it is not the data itself, but is meant to capture certain important features of the data that would otherwise not be recognizable in a long list of numbers.

Using models help us to understand or simplify a situation. They can also help us make predictions about future events. For example, weather models help us analyze current weather and predict potential future weather patterns. Architecutural models help us visualize the design of a building before we commit it to bricks-and-mortar. In this section we will deal with what is possibly the simplest and most widely used model, called the **mean** of a set of data. Other commonly used models are given by graphs and equations, which we will develop in future chapters, eventually having models that include all sorts of features, like categorical variables.

Rather than look at the entire set of data, we want to look at the data one variable at a time in order to find out what that one variable tells us about the situation about which we collected data. To make things even easier, we want to reduce the data down to one number that represents the typical data point for that variable. In general, a number used to represent an entire variable is called a **statistic**. If that statistic is meant to represent the typical data point, we call it an **average**.

Let's look at an example. Shown below are the fat and protein counts for 10 of the most popular sandwiches sold at Beef n' Buns.

Item	TotalFat	Protein
Super Burger	39	29
Super Burger w/ cheese	47	34
Double Super Burger	57	48
Double Super Burger w/ Cheese	65	53
Hamburger	14	18
Cheeseburger	18	20
Double Hamburger	26	31
Double Cheeseburger	34	35
Double Cheeseburger w/ Bacon	37	38
Veggie Burger	10	14

We can reduce all this data down to the following simplistic model, telling us that the "typical" sandwich has 34.7 grams of fat and 32 grams of protein.

Statistic	Total Fat	Protein
Mean (g)	34.7	32.0

The question we should ask ourselves is how well does the mean represent a given set of data. Looking at the data above, we see that although the typical sandwich has 34.7 grams of fat, there are some that have much higher values than that and some that have much less.

The first step in getting an overall measure for how the data values differ from the mean is to develop a standardized ruler to measure how close the observations are to one another. For example, in a crowd of people, your arm-length is a good measuring stick for "closeness": If someone is less than one arm-length away from you, you would consider them "close". However, this distance is not appropriate when driving down the freeway. A more appropriate measuring stick for this situation would be the length of a car. The Federal Aviation Administration has yet another definition of close: aircraft are not allowed within 1000 feet of each other without declaring a "near miss."

These situations all describe ways of measuring "closeness" that refer to real physical distances. Seldom, however, do managers deal with these kinds of distances. More commonly, they collect data measured in dollars or years. Can we find a way to measure distance that will make sense for almost any situation that managers encounter?

As you've probably guessed, we can. To do so, however, we need to decide where to start measuring from. Most of the time we start measuring at zero, but this may not help very much when looking at sales figures in millions of dollars, especially if none of the figures is near zero. Rather than pick a single fixed place from which to always measure zero, it makes more sense to use a measure of central tendency, namely the mean for the variable. .

Once we have selected the mean as the reference point we can then look at the **deviation** of each observation from the mean: Is each observation above the mean or below the mean? By how much? Thus, we will always be measuring the spread of our data from a central reference point that pertains to that particular set of data.

The measuring tool that we will use to measure the spread of our data is called the **standard deviation**. This number is different for each set of data, but it is calculated through the same formula each time.

	TotalFat (g)	Protein (g)
Mean	34.700	32.000
Standard deviation	18.209	12.561

Looking again at the Beef n' Buns data, we see that while the typical sandwich has 34.7 grams of fat, the majority of sandwiches actually range in fat grams from 26.5 (subtract 18.2 from 34.7) to 52.9 (add 18.2 to 34.7) grams.

You have probably encountered the standard deviation before. If you did, you may have thought that the formula was a little complicated and hard to understand. We are going to take a close look at the formula for standard deviation, because if you understand this formula you will understand a lot about statistics. Although the formula looks difficult, you will quickly learn that every piece of the formula makes sense and has a reason for being there. It wasn't developed by some genius who made the formula up from thin air. The formula was developed as the simplest possible way to find an appropriate measuring stick for any set of data. In fact, the formula for standard deviation is essentially the best way to measure the average deviation of the data from the mean.

3.1.1 Definitions and Formulas

Model A model is a number or formula that represents a set of data; it is not the data itself, but is meant to capture certain important features of the data that would otherwise not be recognizable. Models can be descriptive (used to describe a particular situation or set of data), predictive (used to help understand the likely future outcomes of a situation), or interpretive (designed to help one understand how the current situation came about or where the data came from), and can take the form of numbers, graphs, pictures, equations or descriptions.

Empirical Model An empirical model is based only on data and is used to predict, not explain, a system. An empirical model usually consists of a function that captures the trend of the data

Statistic Any number used to represent some aspect of many observations of a single variable or that relates several variables together. For example, the mean is one way to describe a list of numbers; it reduces the entire list to a single statistic representing the typical data point.

Central tendency A statistic that is intended to provide a measure of what a "typical" data point is for a single variable. The most common measure of central tendency is the arithmetic average, or mean. Others include the median, mode and geometric average.

Mean An average computed by adding all the observations of a variable together and then dividing by the number of observations. This is more properly called the **arithmetic mean**. This is the most commonly used average, and it is the most robust average (it will change the least under repeated sampling of the population). In symbols, the mean of the data $x_1, x_2, x_3, \ldots, x_n$ is

$$\bar{x} = \frac{\sum_{i=1}^{n} x_i}{n} = \frac{x_1 + x_2 + x_3 + \cdots + x_n}{n}$$

To compute the mean in Excel, use the =AVERAGE(cell range) formula.

Sigma, Σ This symbol provides a compact way to represent adding a large number of items together if they follow a pattern. For example, the formula $\sum_{i=1}^{5}(i+2)$ means that we are adding together five objects that look like $i + 2$, that is, each object is a number, i, plus 2. So, the first term in the sum starts at the smallest value of i (in this case, 1) and increments up for each term. So, the nice compact formula really represents a much larger addition problem:

$$\sum_{i=1}^{5}(i+2) = (1+2) + (2+2) + (3+2) + (4+2) + (5+2) = 3 + 4 + 5 + 6 + 7 = 25$$

The sigma notation (the symbol is the uppoercase Greek letter S, for "sum") provides a much cleaner way to write the formula. After, all, if we had to add from $i = 1$ to $i = 10,000$, writing each term out by hand would be tedious and rather pointless.

Deviation The deviation of a data point is its signed distance from the mean. To calculate this for data point x_i simply subtract the mean from the data point: $x_i - \bar{x}$. This deviation will be positive if the observation is larger than the mean and negative if the deviation is smaller than the mean.

Total Variation (SSD) This is the sum of the squares of all the deviations of all the observations in the data. In symbols, this is

$$\sum_{i=1}^{n}(x_i - \bar{x})^2$$

The total variation is always positive (since you are adding a bunch of squares of numbers) or zero (if each observation is equal to the mean).

Sample Standard Deviation This is a sort of average deviation for all the observations in the data. The sample standard deviation for a set of data labeled x is denoted by the symbol S_x. To compute this, we take the total variation in the data (see above), divide by the number of degrees of freedom (usually $n-1$) and then convert back into the right units by taking the square root:

$$S_x = \sqrt{\frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n-1}}$$

Degrees of Freedom (DOF) This is related to several key ideas: the number of observations in your data and whether the data is from a population or from a sample. If the data is from a population, then the number of degrees of freedom is the same as the number of observations. However, if you are taking data from a sample and calculating quantities (such as the mean) that describe the population, then you lose a degree of freedom for each calculation you are inferring about the population. For example, to compute the standard deviation of a sample, you must calculate the (inferred) mean of the population. This costs you one degree of freedom, taking you from n to $n-1$.

3.1.2 Worked Examples

Example 3.1. Computing a mean
For this example, we want to compute the mean of a set of test scores:

55, 60, 67, 70, 78, 81, 84, 88, 90, 95, 99

The mean of the data is given by adding the observations and dividing by the number of observations, $n = 11$. Thus, the mean of this data[2] is

$$\bar{x} = \frac{\sum_{i=1}^{n} x_i}{11} = \frac{55+60+67+70+78+81+84+88+90+95+99}{11} \approx 78.82$$

[2]Notice that the last symbol before the result is a squiggly equal sign ($\approx$). This indicates that the answer has been rounded off.

Thus, we would say that *a typical student received a score of about 79 on this test.*

Example 3.2. Deviations and average deviations
Computing the deviation of a data point from the mean is simple. We just subtract the mean from the data point. The result is a signed number (it could be positive or negative) that tells us how far the data point is from the mean. So in the data for burgers at Beef n' Buns, we can compute the deviation of each burger's fat content from the mean fat content of 34.7 g.

Item	TotalFat	Deviation
Super Burger	39	4.3
Super Burger w/ cheese	47	12.3
Double Super Burger	57	22.3
Double Super Burger w/ Cheese	65	30.3
Hamburger	14	-20.7
Cheeseburger	18	-16.7
Double Hamburger	26	-8.7
Double Cheeseburger	34	-0.7
Double Cheeseburger w/ Bacon	37	2.3
Veggie Burger	10	-24.7

From this, it is clear that the veggie burger has much less fat than the typical Beef n' Buns burger, while the Double Super Burger with Cheese contains considerably more fat than the average sandwich. Also notice that none of the burgers is actually right at the average fat content of 34.7 g; the Double Cheeseburger is close, but a little low. Given this, if we randomly chose a burger to eat, what would we expect its fat content to be? This is really just another way to ask what the average deviation of the fat content is.

Well, this is just an average of the deviations, right? So we can add up the deviations and divide by the number of burgers. Unfortunately, we find that the sum of the deviations is zero, giving an average deviation of zero. But how can this be? Not only do none of the burgers have exactly the mean fat content, but many of them are quite far away from the mean. (Just in case you think we have rigged this example, try it with the protein content of the burgers. Then try it with any list of numbers you want to use - your favorite football team's points per game, for example.)

In fact, the sum of the deviations from the mean is zero, for any set of data points. We can use some algebra to decide whether this conjecture is really true.

$$\begin{aligned}\text{Old Data} &= x_i \\ \text{Old Mean} &= \bar{x} \\ \text{New Data} &= x_i - \bar{x}\end{aligned}$$

Now simply add all these data points up and compute the mean of the new (shifted) data:

$$\text{New Mean} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})}{n} = \frac{\sum_{i=1}^{n} x_i - \sum_{i=1}^{n} \bar{x}}{n} = \frac{\sum_{i=1}^{n} x_i}{n} - \frac{\sum_{i=1}^{n} \bar{x}}{n} = \bar{x} - \frac{n\bar{x}}{n} = 0$$

Notice that this calculation works in general. We did not need to have a specific set of data, or a specific mean, or a specific number of data points. By using algebra we can show that the mean of the deviations of any set of data is zero (and thus, the sum of the deviations is zero). This example shows both the power and beauty of mathematics. Rather than work hundreds of examples and rather than calculate each sum of deviations separately, we now have a powerful understanding of what is beneath the actual data.

Why does this happen? Another way to calculate the mean is to "take from the tall and give to the small". The amount you take from the tall (large data values) is equal to the deviation for that stack. The amount that the small needs is the deviation for that stack, which is a negative number. So the mean is gotten by making all the deviations zero. Thus, the reason for the sum of the deviations equaling zero is related to the fact that some deviations are positive and others are negative. Adding the positive and negative numbers cancels out the deviations completely.

What does this result mean, in practical terms? It means that since the sum of the deviations is always zero, we cannot use the deviations themselves to compute an "average distance from the mean". We must construct a new tool to measure the typical distance of an observation from the mean of the data.

Example 3.3. The Standard Deviation Formula: What it all means
The last example showed that the sum of the deviations is always zero because there is always the same total amount of positive deviation (above the mean) as there is negative deviation (below the mean). What we need is a way to turn all the deviations positive; after all, we are really interested in the average distance of the observations from the mean, not in which direction the observation falls. What ways can we take positive and negative numbers and make them all positive? If you're like most people, you can think of at least two ways:

- Drop the negative sign from the negative deviations (this is the same as taking the **absolute value** of the deviations before you add them together)
- Square all the deviations before you add them together

For technical reasons, mathematicians prefer the second method, squaring all the deviations. If we then add up the squared deviations, we get the **total variation** of the data:

$$\text{Total Variation} = \sum_{i=1}^{n} (x_i - \bar{x})^2.$$

This number by itself is not very useful. First of all, no matter what the units of the original data were, the total variation is never in those units; it's always measured in the square of those units. Thus, if the original data were measured in dollars, the total variation would be in square-dollars, whatever those are. Second, the total variation is the sum of a bunch of squared numbers. If a number bigger than 1 is squared, the result is much larger than the original number. Thus, the total variation is often a huge number. Third, this is a total amount, not an average amount. This leads us to the next step: divide by the number of degrees of freedom to compute an average variation:

$$\text{Average Deviation} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n-1}.$$

This number still suffers from the problems of being in the wrong units and being huge. But this is relatively easy to fix. We got the numbers larger and into the wrong units by squaring them. What's the opposite of squaring a number? Taking the square root! (The squaring function and the square root function are **inverse functions**.) This one simple solution will make the numbers smaller and put them into the proper units. We are then left with the sample standard deviation[3] :

$$S_x = \sqrt{\frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n-1}}.$$

So you can see that although the formula for the standard deviation looks complicated, every piece of it is in the formula for a specific reason. For the data listed in example 2, we can determine the standard deviation. (Remember, the mean is 34.7.)

Item	TotalFat x_i	Deviation $x_i - \bar{x}$	Squared Deviation $(x_i - \bar{x})^2$
Super Burger	39	4.3	18.49
Super Burger w/ cheese	47	12.3	151.29
Double Super Burger	57	22.3	497.29
Double Super Burger w/ Cheese	65	30.3	918.09
Hamburger	14	-20.7	428.49
Cheeseburger	18	-16.7	278.89
Double Hamburger	26	-8.7	75.69
Double Cheeseburger	34	-0.7	0.49
Double Cheeseburger w/ Bacon	37	2.3	5.29
Veggie Burger	10	-24.7	610.09

$$\text{Total Variation} = \sum_{i=1}^{10}(x_i - 34.7)^2 = 2984.1.$$

$$\text{Standard Deviation} = \sqrt{\frac{\sum_{i=1}^{10}(x_i - 34.7)^2}{10-1}} = \sqrt{\frac{2984.1}{9}} \approx \sqrt{331.57} \approx 18.21$$

This says that, on average, most data points (approximately 68% of the data points) are within 18.21 units above and 18.21 units below the mean. This would give a range of $\bar{x} - S_x = 34.7 - 18.21 = 16.49$ up to $\bar{x} + S_x = 52.91$ for most of the data. By counting the

[3]One can also compute the *population standard deviation.* In this case, one must be working with data from the entire population, rather than a sample, so that the mean is not estimated but is in fact the actual mean of the population. Then you keep all n degrees of freedom. Since n is bigger than $n-1$, the population standard deviation is always less than the sample standard deviation, indicating that we have more certainty and less variability in the population statistics.

data points, we see that 6 out of the 10 data points (60%) fall inside this range. Going out to two standard deviations above and below the mean should give us 95% of the data. The lower end of that range would be $\bar{x} - 2(S_x) = 34.7 - 2(18.21) = -1.72$ and the upper end would be $\bar{x} + 2(S_x) = 71.12$. We see that the data has all 10 points (100%) within this range.

In general, if the data is **normally distributed** we expect the following empirical rules to be true:

- Approximately 68% of the data points should be within one standard deviation above and below the mean. That is, 68% of the points are less than one standard deviation from the mean.

- Approximately 95% of the data points should be within two standard deviations of the mean.

- Approximately 99.7% of the data points should be within three standard deviations of the mean.

- The data should be symmetrically distributed, with about equal numbers of data points above and below the mean.

We'll learn more about checking and applying this rule later (section 6.2.4, page 183). For now, it's enough to know it exists. It is also the basis for a popular management system known as "Six Sigma" or 6σ. This refers to the use of the symbol σ (a lowercase Greek "s") for the standard deviation. One of the goals of the Six Sigma method is to minimize the amount of production (or whatever you are involved in that can be measured) that falls outside of three standard deviations above and below the mean. At most 0.3% of all data should fall that far away.

Example 3.4. Using Means and Standard Deviations to Compare Sales Performances
The data below shows the total monthly sales for each branch of Cool Toys for Tots in two different regions of the country, the north-east region and the north-central region. (See file "C03 Tots.xls".) Which of these two regions is performing better?

Sales NE	Sales NC
$95,643.20	$668,694.31
$80,000.00	$515,539.13
$543,779.27	$313,879.39
$499,883.07	**$345,156.13**
$173,461.46	$245,182.96
$581,738.16	$273,000.00
$189,368.10	$135,000.00
$485,344.87	$222,973.44
$122,256.49	$161,632.85
$370,026.87	$373,742.75
$140,251.25	$171,235.07
$314,737.79	$215,000.00
$134,896.35	$276,659.53
$438,995.30	$302,689.11
$211,211.90	$244,067.77
$818,405.93	$193,000.00
	$141,903.82
	$393,047.98
	$507,595.76

One way to answer this question is to compare the mean sales in each region. We find that the northeast region has mean sales of $325,000, and the north-central region has mean sales of $300,000.

Based on this information, we might conclude that the northeast region is doing better. But we must consider whether the mean is a good way to model this data. As a clue, when looking at the northeast region, we notice some of the lowest performing stores in the sample! And notice that there is one store in the north-east region with sales of $818,405.93. This is much higher than the sales for the other stores in either region. This single high value is pulling the mean for the north-east region up, even though the stores in the north-central region are typically doing better, as evidenced by the fact that many of them (almost half) are well above the north-central region's mean. In the northeast, however, the stores performing below the mean are typically far below the mean.

This sensitivity to high or low scores is one of the drawbacks of the mean. This is why the Olympics (and many other sports bodies) drop the high and low scores for a competitor before computing the mean. In later chapters, you'll learn what data points like this are called and gain a powerful graphic tool for determining which data points are likely to have too much influence on the mean.

We already know that the mean of the NE region sales is $325,000 and the mean of the NC region is $300,000. What about the standard deviations for each region?

	Sales NE	Sales NC
Mean	$325,000.00	$300,000.00
Standard Deviation	$217,096.62	$141,771.13

Now we have some useful information. The NE region has a much larger standard deviation than the NC region. In the NC region, though, this smaller standard deviation

indicates a much narrower spread of the data. This means that the stores will perform more similarly to each other, indicating more consistency and more stable, dependable sales results in the long run. The stores in the NE region are, on average, more spread out than those in the NC region. We are likely to have very high and very low sales in this region. Notice the last store on the list for the NE region. It had sales of $818,405.93, which is very high compared to the other stores. This store is a potential **outlier** (a data point so atypical that we should not consider it.) What happens if we remove it from the data?

	Sales NE (without outlier)	Sales NE (with outlier)
Mean	$292,106.27	$325,000.00
Standard Deviation	$178,742.58	$217,096.62

Notice the dramatic change in the results! Although the NE region (minus the outlier) is still slightly more spread out than the NC region, it is nowhere near as spread out as it was. Also, the mean for the NE region without the outlier is now slightly below the mean of the NC region. This shows you how important a single observation of the data can be. If you have outliers in the data, it is a good idea to report the statistics with and without the outliers. In this case, the $292,106.27 value is more representative of the bulk of the stores in the NE region than the $325,000 value, which was exaggerated by the outlier. Also, without the outlier, the stores in the NE region have more similar sales results, indicated by the smaller standard deviation.

3.1.3 Exploration 3A: Wait Times at Beef n' Buns

For this exploration we will look at the waiting times collected by our observational survey at Beef n' Buns. Assume that the following data (C03 waittime.xls) represents the number of seconds the customer waited to receive his/her food at Beef n' Buns.

Part One: Wait times

1. Determine the average wait time for a typical customer.
2. How many customers waited less than the average? How many waited more?
3. Compute the deviations from the mean in a new column to the right of WaitTime.
4. What was the largest deviation from the mean?
5. What is the Total Deviation (or sum of the deviations) for this set of data?
6. Compute the Squared Deviation in a new column to the right.
7. Compute the Sum of the Squared Deviations (SSD) for this set of data.
8. Compute the Standard Deviation as follows:

$$S_x = \sqrt{\frac{\text{Sum of Squared Deviations}}{n-1}} = \sqrt{\frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n-1}}$$

9. Now use the built-in Excel function =STDEV to compute the standard deviation.

WaitTime (sec.)	Competitor (sec.)
90	210
152	0
113	118
239	0
54	185
47	0
72	16
276	43
114	17
74	165
61	23
88	9
84	134
60	26
55	22
100	26
53	36
80	273
92	83
65	186
56	109
57	140
72	48
59	56
103	132
40	183
52	153
50	30
21	72
120	104

Part Two: Comparing the competition

Use Excel to compute the mean and standard deviation for your competitor. Complete the following table:

	Beef n' Buns	Competitor
Mean		
Standard Deviation		

For which of the two fast food restaurants is the mean a better model for customer wait times? Why?

3.1.4 How To Guide

For all of the information below, assume that the spreadsheet shown below is being used. It contains data on sample salaries.

Cell References in Excel

Excel organizes information into sheets. Each worksheet is then organized by columns (labeled by letters) and rows (labeled by numbers). Thus, every cell (rectangle on the worksheet that contains information) has a name, called a cell reference. This cell reference is usually given the way you called out locations on the game Battleship: as a column and a row. For example, in the worksheet shown at the right, the word "SALARY" is in cell A1. The mean of the salary data is in cell D1. Such a reference is called a relative cell reference.

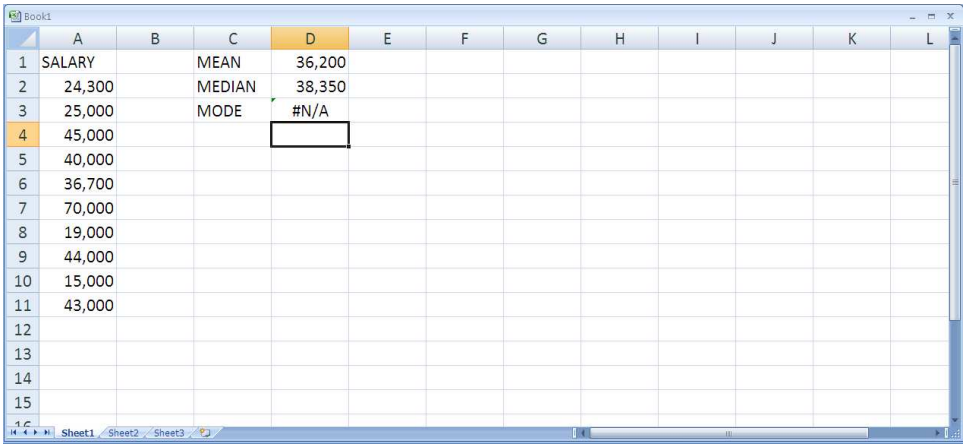

Figure 3.1: Excel worksheet

There are other types of cell references (called absolute cell references) that you will learn about below. The other important thing to know about cell references is that you can easily refer to a block of cells, as long as the block of cells forms some sort of rectangle. For example, to refer to the salaries themselves on the worksheet above, you would refer to all the cells in the rectangle starting in cell A2 and continuing through cell A11. In Excel notation, this entire range of cells is identified by A2:A11.

Computing Means in Excel

Excel uses the function AVERAGE for the mean. To compute the mean of the data in cells A2:A11, we enter the formula

```
= AVERAGE(A2:A11)
```

into any cell on the spreadsheet. If you later move or copy the cell containing this formula, the cell references will be changed since we used relative cell references. This means that the formula will probably not point to the right cells anymore. Also remember that if you

change any of the data in cells A2:A11 the mean will be re-calculated instantly. If, however, you add data outside this range, you will need to change the formula.

For the purposes of the remainder of this guide, assume that there is data in cells A2:A20 with a variable name in cell A1.

Absolute Cell References in Excel

Above, you learned how to refer to any cell or range of cells using the grid system in Excel. If there is data in the cell in column D in row 2, this cell is referred to as D2. However, this type of cell reference (a relative reference) will change if the formula is copied to another cell. Many times (as in the example below of computing deviations) a particular cell reference will need to be absolute. This means that it will not change if the formula is copied. To make a cell reference absolute, place a dollar sign ($) in front of both the column and row. Thus, an absolute reference to cell D2 would look like D2.

As you may have guessed, you can have mixed references also, where either the column or row is absolute. In general, if you don't want part of the reference (either the row or column) to change as you copy the formula, be sure to place a dollar sign in front of it.

When you are typing a cell reference into a formula, you do not have to type the dollar signs to convert them to absolute references. After you type a cell reference in a formula (say you type A2), hit the F4 button along the top row of the keyboard. This converts the current cell reference into an absolute reference (so now you would have A2). If you hit the F4 button again, it is converted to a mixed reference with the row fixed (A$2), hitting it again will convert it to a mixed reference with the column fixed ($A2). Finally, hitting F4 a fourth time will cycle back to a relative reference (A2).

Three dimensional cell references in Excel

In addition to referring to cells by the column and row, Excel allows you to build formulas that include references to cells on other worksheets in the current workbook. Suppose you are entering a formula in 'Sheet 1' of a workbook and there is a number in cell D4 of 'Sheet 2' that you want the formula to look up. Simply typing D4 in the current formula will not work; Excel will simply look up the value in cell D4 of the workbook containing the formula. To get around this, you must use a 3D cell reference. All this involves is including the name of the worksheet in single quotes, followed by the "bang" or exclamation mark symbol (!) and then the normal cell reference. So, in our example, to get a formula in 'Sheet 1' to use the value in cell D4 from 'Sheet 2', you need to type the cell reference exactly in the form

```
'Sheet 2'!D4
```

Computing deviations in Excel

In order to compute the deviations in Excel, we first need the mean of all the data. Let's calculate this with Excel by typing the formula below into cell F1.

```
=average(A2:A20)
```

Now, we will create a new column for the deviations. In cell B1, type "Deviation" so that the column has a label. Now, in cell B2, we want to enter a formula to compute difference between the first data point (in cell A2) and the average (an absolute reference to cell F1). Thus, we enter the formula

```
=A2 - $F$1
```

Now we simply copy this formula (see below) down to the other cells in column B.

Copying Formulas in Excel

There are three different ways to copy formulas in Excel from one cell to another cell or to a group of cells (like a whole column): standard copy and paste commands, dragging the fill handle, or double-clicking the fill handle.

- Method 1: Using Copy and Paste Commands. This method is the most obvious. First select the cell with the formula you want to copy. Copy this using either CTRL+C, the copy button on the toolbar, or the "Edit/Copy" menu command. Now highlight the cell or cells where you want the formula to be placed and paste it in using either CTRL+V, the paste button on the toolbar, or the "Edit/Paste" menu command.
- Method 2: Dragging the Fill Handle. If you want to copy the formula to the column of cells beneath it, or to the row of cells beside it, you can use the fill handle. The fill handle is a tiny black square that appears in the lower right corner of a cell you have selected. If you click on this fill handle and drag down (or right), then, when you release the mouse button, the formula from the first cell (or group of cells!) is copied to all the cells in the area you highlighted by dragging. Be sure that you are clicking on the fill handle, though. You'll know for certain that you are on the fill handle because the cursor will change from a fat plus sign to a skinny plus sign.
- Method 3: Double-clicking the fill handle. In certain circumstances, you can double-click the fill handle and Excel will automatically copy and paste the formula all the way down the column until it reaches the end of the column to the left of the one in which you are pasting the formula. So, if you have a row of salaries in cells A2:A20, and you enter a formula to compute a raise in cell B2, you can copy this formula into cells B2:B20 by double-clicking the fill handle.

  N.B. If the column to the left is empty or has a break in the data (an empty cell), this trick will not fill the column out all the way.

Computing standard deviation in Excel with the built-in formula

There are two different standard deviations in Excel, depending on whether the data is from a sample or a population. To compute the standard deviation of a sample (this is the most commonly used version), use the formula

```
=stdev(range of cells)
```

For the standard deviation of a population, use the formula

```
=stdevp(range of cells)
```

Adding up a list of values

If you have a list of values, you can quickly add them together using the SUM command in Excel. For example, if your values to be added are in cells A2:A26, entering the command

```
=SUM(A2:A26)
```

into cell B2 (or any other cell) will add the values together.

Computing standard deviation in Excel with a user-created formula

One could also create the formula for standard deviation by hand, simply by constructing each piece of the formula separately and then combining them together. In Excel, the square root of a number is computed with the sqrt(number) function. For the standard deviation, we want the square root of the sum of the squared deviations divided by the number of observations minus one. First, we create a column of deviations (see above). Next, create a column of deviations squared (If the deviations are in column B, starting at row 2, then the deviation squared would be calculated with the formula =B2^2 which can then be copied to the rest of the column to compute the other deviations squared. Now, we should have the deviations in cells B2:B20, and the deviations squared in cells C2:C20.) Finally, we construct the entire formula:

```
=sqrt(sum(C2:C20)/(count(C2:C20)-1))
```

As an alternate version, we could have Excel square the deviations and add them together using the "sumproduct" function (mathematically, it is equivalent to a vector dot-product). This formula works directly with the deviations in column B with no need to construct a column for the deviations squared.

```
=sqrt(sumproduct(B2:B20,B2:B20)/(count(B2:B20)-1))
```

Sorting data

Excel makes it relatively easy to sort your data on many variables simultaneously. In order to use this effectively, though, you need to have your data organized as we have discussed in chapter two: your variables (fields) should be the columns and the observations (records) should be the rows. It is also a lot easier if you make sure the first row of the data contains headers (variables names).

To start sorting the data, first select (click on) any cell in the data range. Then go to the data ribbon and select "Sort". This will bring up a dialog box like the one shown. The example shown (figure 3.2) uses the file "C03 EnPact Data.xls". The sort feature automatically assumes that you want to sort on the first column of the data; this is indicated by the line that says "sort by". You can change this, however, by using the pull down menus along that row of information. These let you select a different variable to sort on and different sort orders. You can sort on several variables, by adding more sort conditions using the "Add Level" button. You can delete conditions or add as many as you like. (Excel XP only allowed three sort conditions at a time.) In the upper right-hand side of the dialog box make sure

the "My data has headers" is checked, so that Excel knows what the variable names for each column are.

Say you wanted to sort the data so that employees of the highest job grade are at the top of the data. Then simply select "Job Grade" and "Descending" in the top of the dialog box, then hit "OK".

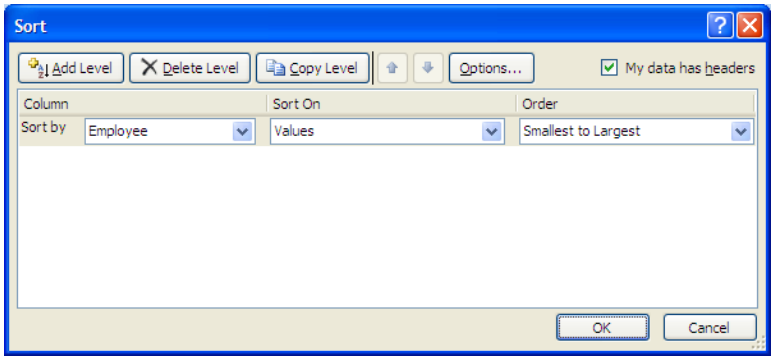

Figure 3.2: The sorting wizard in Excel 2007

If you wanted to sort by job grade and then by gender, you might select "Job Grade" in the first sort condition (and "largest to smallest" for the sort order) and then add another level to the sort and select "Gender" for that level. This will sort the list on two variables. Excel will collect all the employees with JobGrade = 6 at the top of the list, and within that group, the Female employees will be at the top of the list and the Male employees at the bottom. Sorting on three variables is similar. For example, we could take the list produced above and sort first on JobGrade, then on Gender and finally on Education.

Sorting data can be very useful for identifying outliers in the data or other anomalies. For example, if you have data on the diameter of parts being produced by one of your factory machines, and you determine that the mean size of these is 0.45" with a standard deviation of 0.03", sorting the data on the diameter variable would let you quickly find any parts produced that are too far above or below the standards of your company.

Using the Excel 2007 status bar

One nice feature of Excel is called the "status bar" (in Excel XP, this was the Autosum menu, but had far fewer featurs and information than Excel 2007.) If you highlight a set of numbers, any set of numbers, in a workbook, Excel automatically computes some information about those cells and displays the result in the lower right portion of the screen, just below the horizontal scroll bar. This makes it easy to see what is going on with your data without having to enter formulas. By default, the status bar shows information about the average, count, and sum of the numerical values highlighted.

This feature is good for other things as well. If you right-click on the status bar area you will get a menu of other options to have Excel "auto-display". You can select any of the following:

Count	Displays the number of cells highlighted that have data in them (blank cells are ignored)
Average	Computes the mean of the cells highlighted, ignoring blank cells
Minimum	Computes the smallest element of the highlighted cells
Maximum	Computes the largest element of the highlighted cells
Sum	Computes the sum of the highlighted cells
Numerical Count	Counts only the cells containing numerically stored data in the highlighted region

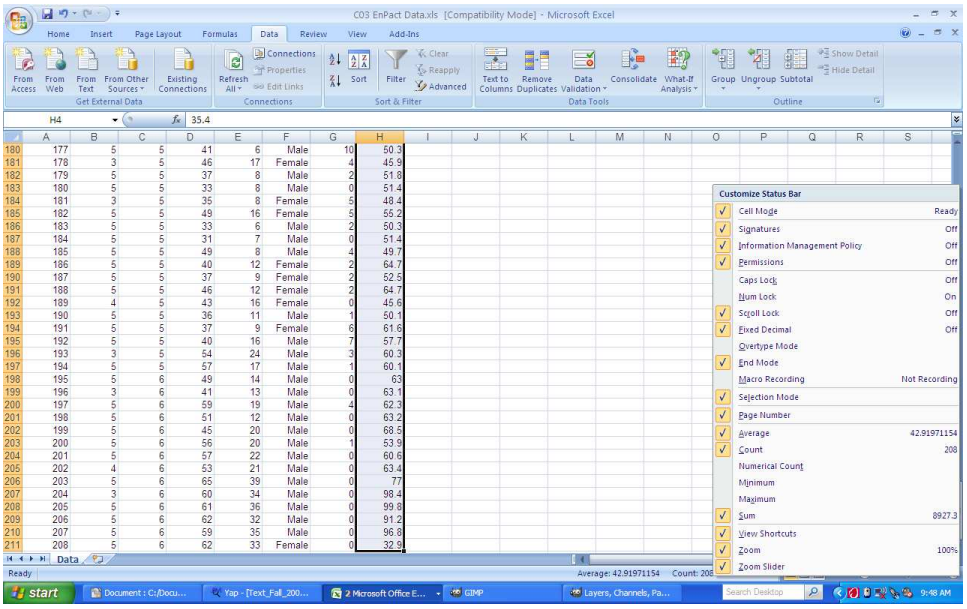

Figure 3.3: The Excel 2007 status bar, showing the options available for display by right-clicking in the status bar region (where the average, sum and count are displayed in the figure)

3.2 Categorical Data and Means

The ideas in the previous section - modeling a set of data by using the mean and treating the standard deviation as the "goodness of fit" for the model - will not work if our data is categorical, because we cannot add, subtract, multiply or divide with categorical data like "Male" and "Female". Yet, we often have both categorical and numerical data and wish to know a variety of things about the underlying situation. Consider shipping records for a company that contain data on each shipment: the total weight, during what shift (morning, afternoon, night) it was loaded, truck size (van, semi, other) and so forth. We might want to profile the data to determine what a typical day-shift load looks like in order to help us plan employee schedules. How might we go about this?

One way would involve going through all the data by hand, adding up the information about just the day shift-related shipments, and then creating a report. However, typical data sets can number in the thousands of observations per variable and such a hand-compiled approach is not practicable. Another way might be to use Excel to sort the data by shift and look at it that way. We could also use the "autofilter" feature of Excel to display only the daytime shift information. But, if we also want to know what a typical afternoon and night shift load are like, we would have to repeat all this again for each shift. An additional drawback is that if the data change (a new month is added, for example) we have to start over.

Fortunately, Excel is designed to easily allow users to cross-section their data in a variety of ways. The main tool for this is called a Pivot Table. These can be used to easily produce a report like the following one.

Average of Count	Location			
Day	Cranny	Hole	Nook	Grand Total
D	26.796	19.115	28.622	24.844
F	19.734	8.770	22.135	16.879
H	16.809	8.592	22.322	15.908
M	16.763	12.056	18.849	15.889
S	19.845	15.697	22.289	19.277
T	16.974	12.260	18.414	15.883
W	13.306	8.743	21.852	14.634
Grand Total	18.604	12.176	22.069	17.616

Table 3.1: Pivot table showing average seating in the areas of Over Easy, by day of week.

This table took about 30 seconds to produce from a set of data with almost 200 observations of each variable. Even for larger sets of data, that's about all it takes. And look at what we can easily learn from this table. One thing that really stands out is that more people sit in the Nook area. This could be for a variety of reasons. For example, this area might have the best window view or be closest to the entrance (or farthest from the entrance). On the other hand, the Hole area doesn't ever seem to have many people in it. It might be too dark or it could be located right next to the bathrooms.

At the same time, we have to ask if these results are significant. Maybe this data is skewed somehow and more typical data from a longer time period (several months) wouldn't display such a large difference between the Hole and the Nook areas? To test this, one normally uses statistical tests called chi-square tests or z-tests. Our approach in this book will be to look for overall patterns rather than test for statistical significance. In this case, there is another important consideration before we read too much into the pivot table: Are all three areas of equal size? If the Hole area only has seating for 20 people, then it is often close to full. At the same time, if the Nook area has seating for 50 people, it is actually not being used much at all. This is one reason why it is often vital to introduce a computed variable into the data. In this case, we should probably run statistics not on the actual counts of people in each area. Instead, we could compute the percent capacity of each area (the number of people divided by maximum capacity of the area) and then run our statistics.

3.2.1 Definitions and Formulas

Pivot Table An Excel tool for allowing you to quickly cross-section and summarize your data. Values can be displayed as means, counts, sums, and standard deviations, percentages (of rows, columns, or totals). The data can be cross-sectioned in up to three variables simultaneously (row, column, and page) and variable ranges can be grouped for easier display.

Cross-sectioning data Term used to refer to taking the data and breaking it down into categories and reporting on each category separately.

Data Mining Process of combing through data, using a variety of tools, such as pivot tables, to find information about the underlying source of the data. For example, companies mine the data on past customer purchases in order to develop targeted advertising and marketing campaigns that address the tendencies of certain types of customers.

Field In a pivot table (or a database) the variables are called fields.

Record In a pivot table (or a database) the observations are called records.

3.2.2 Worked Examples

The following examples deal with data collected at Over Easy using the observational data forms from the last chapter. The data can be found in the file "C03 OverEasy2.xls".

Example 3.5. Reading a one-variable pivot table

Given the data on counts of people at OverEasy for the period shown in the data, it might be nice to see how many people typically sit in each area of the restaurant (Nook, Cranny and Hole). We could go through the data and count these up ourselves, or sort the data and look at it, but in a few seconds, Excel can produce a pivot table doing this for us. Moreover, it is then easy to change the way the data is displayed and cross-sectioned. For

example, the table below shows the average number of people in each area of the restaurant. From this, we see that one area of the restaurant, the Hole section, has many fewer people in it, on average.

Average of Count	
Location	Total
Cranny	18.604
Hole	12.176
Nook	22.069
Grand Total	17.616

However, we do not know how significant these differences are. It would help to know how variable the number of people in each section is. By changing the display to show the standard deviation of the counts of people in each section, we get the table below. Notice that this shows that the overall spread of the data is about 12 people, but that the Hole area is much lower at about 9 people. So not only is the average number of people in that area low, but the spread is much less, indicating that we have consistently fewer people in that area of the restaurant.

StdDev of Count	
Location	Total
Cranny	12.075
Hole	9.042
Nook	12.549
Grand Total	12.046

Realistically, though, we should only be looking at the total count of people in each area if the areas are equal in size. Assuming they are, the above analysis holds; if they are not equal, and if, for example, the Hole area is smaller than the other two, then we would of course expect lower numbers in that area. In the case where the numbers have different maximum values, we need to find a number that is more stable across all the areas of the restaurant. For example, from the observational data, we could create a new variable called "PercentCapacity" that is the actual number of people divided by the maximum capacity of that area in the restaurant. These numbers would all range from 0% to 100%, making them comparable. For now, though, we assume that all the areas are the same size, which leads us to conclude that for some reason, people don't seem to want to use the Hole area of the restaurant.

To find out why this is we would need to collect more data. Perhaps people do not like the decor in that area. Or maybe it's farther from the door, so that people fill up from the front of the restaurant on back. Maybe there's no view out of the windows in that area, or it is too close to the kitchen or bathrooms, making it noisy.

Example 3.6. Pivot tables in two variables

Let us first revisit the example pivot table from the discussion above (table 3.1.) It shows the average number of people in each section of the restaurant, broken down by two

variables: day of the week and location. This lets us explore more patterns in the data and see more clearly what is happening.

Notice that the table shows how much lower the couts are in the Hole area - on every day of the week. And for several days (W = Wednesday, H = Thursday, and F = Friday) the counts in the Hole are about 50% of the counts in either of the other areas. We can also see in the last column that our busiest days tend to be S = Saturday and D = Sunday. Using this information, we could more easily plan how many servers to position in each area each day of the week. But these numbers will clearly change. How much can we expect them to change? The table below displays the standard deviation of the counts, broken down by day and location. Overall, we see a lot of variation in the counts.

StdDev of Count	Location			
Day	Cranny	Hole	Nook	Grand Total
D	16.331	12.153	16.138	15.538
F	11.864	6.031	12.474	12.023
H	10.138	5.804	12.245	11.275
M	10.171	7.860	10.179	9.874
S	11.748	9.791	12.050	11.554
T	10.027	7.956	9.434	9.540
W	7.982	6.175	11.603	10.402
Grand Total	12.075	9.042	12.549	12.046

Another way to view the data in a pivot table is to look at the data not as counts (the default, usually useful only if the data variable is categorical) or as averages or standard deviations, but as a percentage. Most commonly, we would represent the data either as a percentage of the row or the column variable. For example, the table below displays the average counts, but as a percentage of the row variable (day of the week) showing us what percentage of our customers are in each of the three areas of the restaurant. Such as view of the data lets us quickly see - on a common scale - which areas are most and least popular.

Sum of Count	Location			
Day	Cranny	Hole	Nook	Grand Total
D	35.95%	25.65%	38.40%	100.00%
F	38.97%	17.32%	43.71%	100.00%
H	35.22%	18.00%	46.77%	100.00%
M	35.17%	25.29%	39.54%	100.00%
S	34.32%	27.14%	38.54%	100.00%
T	35.62%	25.73%	38.65%	100.00%
W	30.31%	19.92%	49.78%	100.00%
Grand Total	35.20%	23.04%	41.76%	100.00%

These percentages might be useful for many things that the raw numbers would not directly show. For example, by looking at who is where through a percentage, we can make reasonable staffing decisions: if 25% of the customers are in the Hole on Thursday, then 25% of our staff should be in that area. In addition, this also puts all the areas on an equal footing.

Example 3.7. Large pivot tables and grouping

Our data on the counts at Over Easy also includes the time of day. It might be interesting to look at the data using this as one of our variables, but there are a lot of times during the day. A quick two-variable pivot table looking at time and location gives us the rather long table shown below. The length is due to the large number of values for the variable "Time". If our variable had been a continuous numerical variable instead, we could have an even bigger table, with one row for each different value of the variable.

Average of Count	Location			
Time	Cranny	Hole	Nook	Grand Total
500	3.91	2.60	5.95	4.15
530	16.69	10.84	20.01	15.85
600	31.03	20.59	36.51	29.38
630	32.83	22.33	36.94	30.70
700	28.43	18.73	32.63	26.60
730	25.19	16.88	29.48	23.85
800	16.17	10.72	19.43	15.44
830	12.51	8.36	15.90	12.26
900	6.67	4.27	8.99	6.64
930	5.42	3.69	7.63	5.58
1000	6.56	4.52	8.85	6.64
1030	8.71	5.67	11.19	8.52
1100	16.36	10.70	19.87	15.64
1130	20.63	13.17	24.20	19.33
1200	29.29	18.77	33.52	27.19
1230	31.54	19.85	36.02	29.13
1300	31.06	19.27	35.37	28.57
1330	26.52	17.85	30.84	25.07
1400	3.96	2.55	6.00	4.17
Grand Total	18.60	12.18	22.07	17.62

Large tables are generally undesirable. They are harder to read and harder to interpret. Usually, though, when you use a numerical variable as either the column or row variable in a pivot table, there is a way to group the values of the variable to make it easier to read the table. In this case, we could easily group all the times between 500 and 1000 into a "Breakfast" group all the other times into a "Lunch" group.

Average of Count		Location			
Time2	Time	Cranny	Hole	Nook	Grand Total
Group1	500	3.91	2.60	5.95	4.15
	530	16.69	10.84	20.01	15.85
	600	31.03	20.59	36.51	29.38
	630	32.83	22.33	36.94	30.70
	700	28.43	18.73	32.63	26.60
	730	25.19	16.88	29.48	23.85
	800	16.17	10.72	19.43	15.44
	830	12.51	8.36	15.90	12.26
	900	6.67	4.27	8.99	6.64
	930	5.42	3.69	7.63	5.58
Group2	1000	6.56	4.52	8.85	6.64
	1030	8.71	5.67	11.19	8.52
	1100	16.36	10.70	19.87	15.64
	1130	20.63	13.17	24.20	19.33
	1200	29.29	18.77	33.52	27.19
	1230	31.54	19.85	36.02	29.13
	1300	31.06	19.27	35.37	28.57
	1330	26.52	17.85	30.84	25.07
	1400	3.96	2.55	6.00	4.17
Grand Total		18.60	12.18	22.07	17.62

At first glance, this has not really helped. The table is in fact larger by one column. But by hiding the details of a grouped variable, we can get a much smaller table, letting us quickly compare the morning (Group 1) and noon (Group 2) rushes.

Average of Count		Location			
Time2	Time	Cranny	Hole	Nook	Grand Total
Group1		17.88	11.90	21.35	17.04
Group2		19.40	12.48	22.87	18.25
Grand Total		18.60	12.18	22.07	17.62

As a final way of looking at this data, consider using the count variable itself as a row variable. The counts in each section range from 0 to 65, leaving us with 66 rows in such a table. But by grouping them (which can easily be done with a numerical variable like "Count") we can quickly see how many 30 minute blocks of time (observations) in each area of the restaurant fell into each grouping of the number of patrons at our restaurant.

Count of Count	Location			
Count	Cranny	Hole	Nook	Grand Total
0-9	675	1030	517	2222
10-19	529	672	464	1665
20-29	498	311	516	1325
30-39	303	102	421	826
40-49	97	12	178	287
50-59	25	1	30	56
60-69	1		2	3
Grand Total	2128	2128	2128	6384

Essentially, this type of pivot table is a frequency table of the variable "Count". These are useful for creating histograms and other visual representations of one-variable data, which are discussed in chapter 5 (page 133).

3.2.3 Exploration 3B: Gender Discrimination Analysis with Pivot Tables

To help understand how pivot tables work and can help you analyze data to explore a problem context, we will consider a small private company called EnPact that produces environmental impact statements. (Basically, when a company wants to build in an area or manufacture a product, impact statements help predict the expected impact of this work on the local ecology.) Recently, the company has been sued by a group of female employees on the grounds that males have an unfair advantage in the salary process. By exploring this problem using pivot tables, you will learn a fundamental truth about data mining: the deeper you explore, the more you are forced to reconsider each and every piece of evidence you have.

The company salary data (and employee profiles) are in the file C03 EnPact.xls. Open this file. Using simple pivot tables (see the How To Guide for this section), answer the following questions.

1. How many male employees are there? How many female employees? What percentage of the employees is male? Female?

2. What is the average male salary? What is the average female salary?

3. Based on your answers to #1 and #2, write a sentence or two discussing the company's lawsuit.

4. Cross-section the data by both Gender and Education Level. Look at the average salaries of the employees and discuss the company's lawsuit.

5. Cross-section the data by both Gender and Job Level. What does the lawsuit look like now?

6. For a final look at just how complex this issue is, cross-section on three variables simultaneously. Set the pivot table up with Gender as the row variable, Education Level as the column variable, Job Level as the "page variable" (at the top of the pivot table) and average salary as the data.

7. Using the three-variable pivot table, pull down the "Job Level" menu and look at each job level separately in the pivot table. Are there any particular job levels where the male and female salaries, after accounting for education, are roughly the same? Are there any where the salaries are quite different?

8. Select one of the job levels that shows a large difference in salaries by gender. Go back to the original data. Can you account for these differences by looking at the numerical variables (Years of experience and Years Prior)?

3.2.4 How To Guide

Getting Started with Pivot Tables

1. Start by selecting a cell inside the range of the data. For example, the file C03 EnPact Data.xls has data in cells A1:G251 (with row 1 containing headings or variable names), so you might select cell A1 (or any other cell in the range of data).

2. Go to the "Insert" menu and select "Pivot Table". This will bring up the Pivot Table dialog box shown below in figure 3.4 (page 89).

3. Normally, you will not need to change anything else at this stage. Verify that the data range is correct, make sure it has "New worksheet" selected for where to create the table, and click "OK" and Excel will create a worksheet like the one shown in figure 3.5 (page 90).

4. What you see on this are (a) the pivot table area, (b) the pivot table ribbon (along the top) and (c) the pivot table field list (along the right-hand side; these fields should match the variable names/headings in the spreadsheet; remember: field = variable in a database).

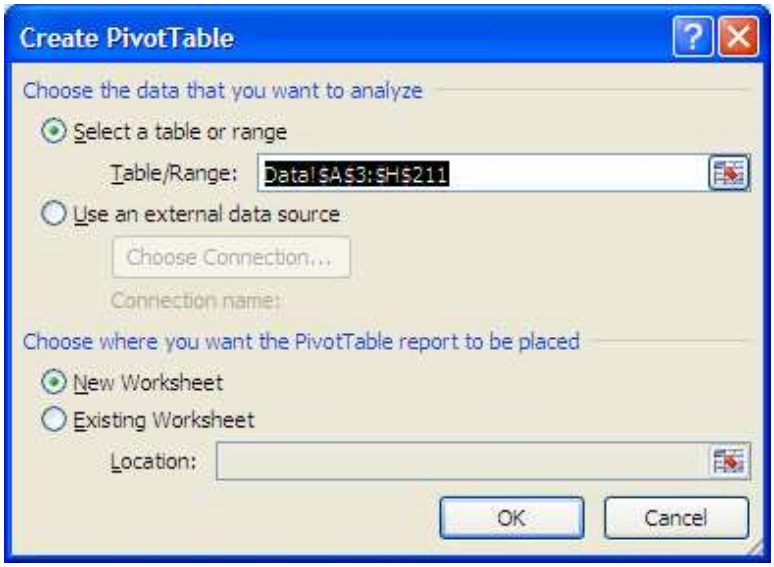

Figure 3.4: The basic Pivot Table dialog box.

Advice on Making Pivot Tables

While you can place any variable in any field, and in particular you can use either row or column variables interchangeably, if your variable has many values (for example, "Years Experience" in the C03 EnPact.xls file) you are best off making it a row variable, rather than a column variable in order to make it easier to read the resulting table, and avoid having to scroll horizontally to get information.

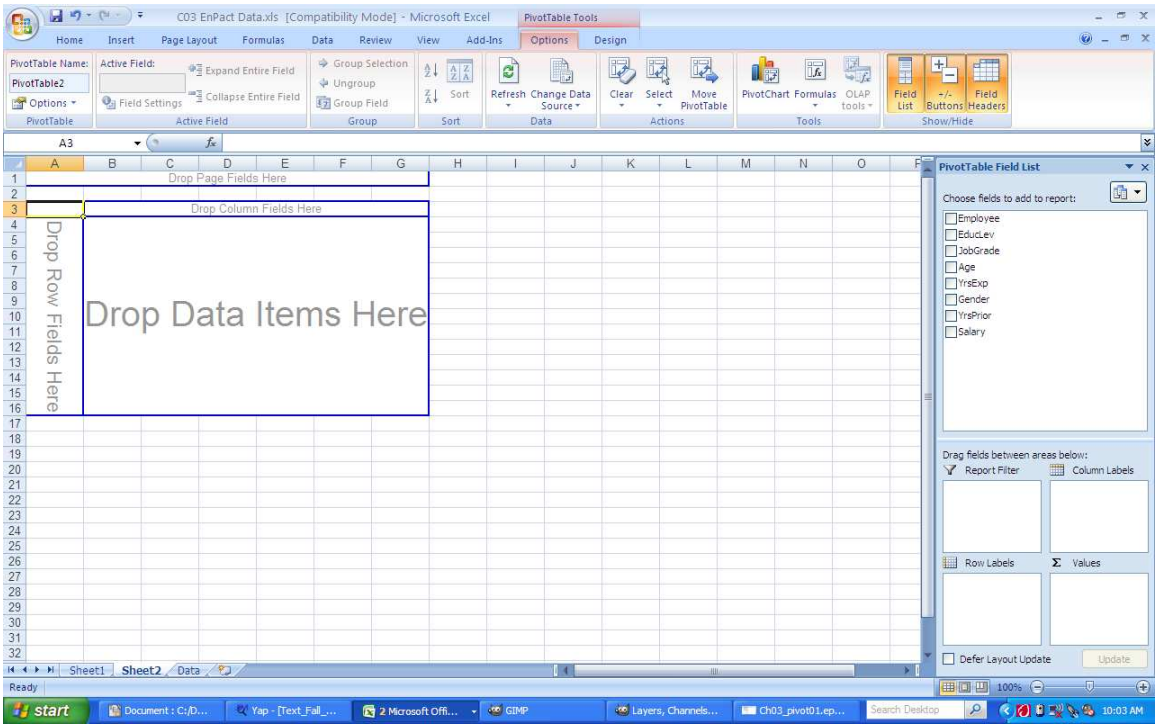

Figure 3.5: A blank pivot table.

Making a simple Pivot Table

Starting from the screen shown above, you now simply drag fields from the field list into the various regions of the pivot table. No data will be displayed until there is one field in the area marked "Drop Data Items Here". For example, to look at the average salaries of the employees, broken down by gender, complete the following.

1. Drag *Gender* to the area marked *Drop Row Fields Here* or drag it into the area in the lower right marked *row labels*.

2. Drag *Salary* to the area marked *Drop Data Items Here* or drag it into the area in the lower right marked *values*.

3. By default, the pivot table will either show the sum of the data variable (in this case, the total of all salaries for males and females) or the count of the data variable (the number of males and females). We would rather see the averages. To display the averages, double-click where it says *Sum of Salary*. You will see the dialog box shown at the right of figure 3.6 (page 91).

4. To summarize the data by averages, select *Average* from the list on the left. You can also format the data by clicking the *Number Format* button on the left side of the dialog box in figure 3.6 (page 91).

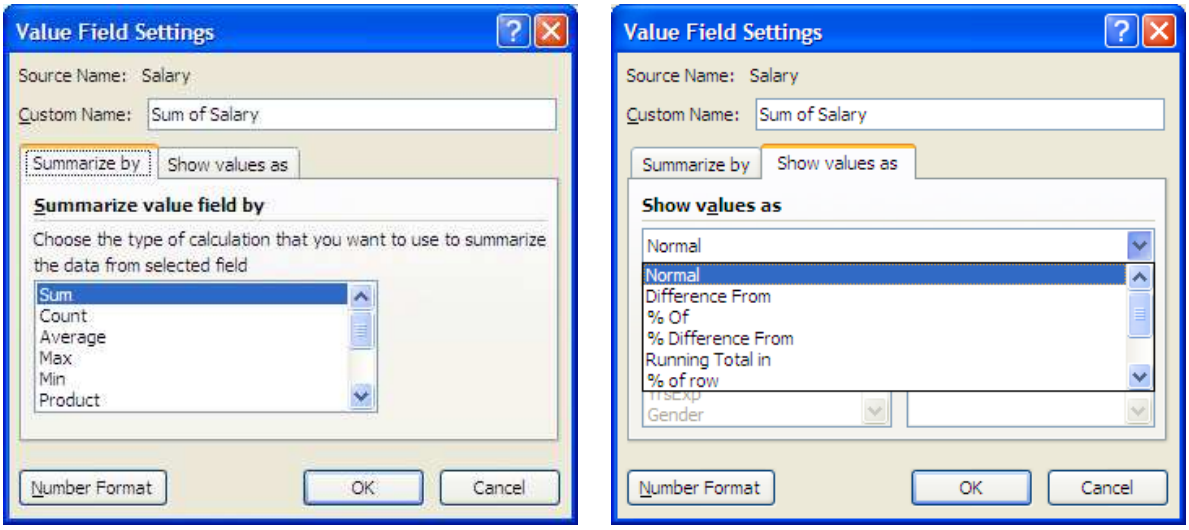

Figure 3.6: Pivot table field display options. These are accessed by double-clicking on the pivot table portion that explains what the data field is, in this case, *sum of salary*.

Figure 3.7: The pivot table ribbon.

Advanced Pivot Table Options

For more sophisticated tables, you can display the data as percentages, etc. To access this feature, simply click the tab marked *Show values as* in the pivot table field display dialog box shown in the right half of figure 3.6 (page 91). To select a different way of presenting the data, select one of the options from the pull down menu under *Show values as*. This gives you the various options for displaying the data. The most useful items from the list are probably *Normal*, *% of row* and *% of column*.

The Pivot Table Ribbon

The pivot table ribbon in figure 3.7 (page 91) provides a lot of flexibility for working with the pivot table. Two of the most important items on the ribbon are the *Refresh* button and the *Change Data Source* button. Refresh forces Excel to re-check the original data and re-build the current pivot table. This is useful if you change or add data to the original database. This makes it easy to update information, without having to create the pivot table again. If you have more data - that is, data outside the original range of the pivot table - you can use the *Change Data Source* button to modify the data range.

Figure 3.8: The pivot table grouping tool.

Grouping items in the table

This feature allows you to take a variable that has many values (like a numerical variable) and group it together in the pivot table. For example, one could easily use "Year Experience" as a row variable in the pivot table above, but the wide variety of values makes it hard to see any details or compare results. However, if you group some years experience together (like 0-9, 10-19, etc.) you can see more interesting results as illustrated below.

Average of Salary	Gender		
YrsExp	Female	Male	Grand Total
0-9	38.77	41.27	39.67
10-19	42.45	55.46	44.47
20-29	44.12	59.22	52.35
30-39	32.90	92.64	82.68
Grand Total	40.21	48.51	42.92

To group the salary categories in order to see the data more easily, select a range of years experience (like all the rows with fewer than ten (10) years experience) and click the *Group Selection* button on the ribbon. Repeat this with the other ranges of experience. Then you can collapse or expand the individual groups of experience to look at the data more easily.

You can also right click on the *YrsExp* field in the pivot table. Select *Group...* from the context-sensitive menu that appears. You should see a dialog box like the one in figure 3.8 (page 92). Here you can select the starting value, ending value and space between groupings. The settings below, for example, produce the groupings shown in the table above.

Making a more complex pivot table

Pivot tables can be used to cross-section the data on up to three variables at once. You can select one variable as the Row Field, one as the Column Field and one as the Page Field. For example, if we take the "StateEx_Deliveries.xls" file and create a pivot table like the one shown in example 7 (page 85), we can further explore the data by adding "Shift" as the page variable. Notice that next to each variable (field) there is a pull-down menu. This pull-down menu allows you to select specific values of the variable (field) to display in the table. For example, pulling down the menu next to the page variable "Shift" in this example allows

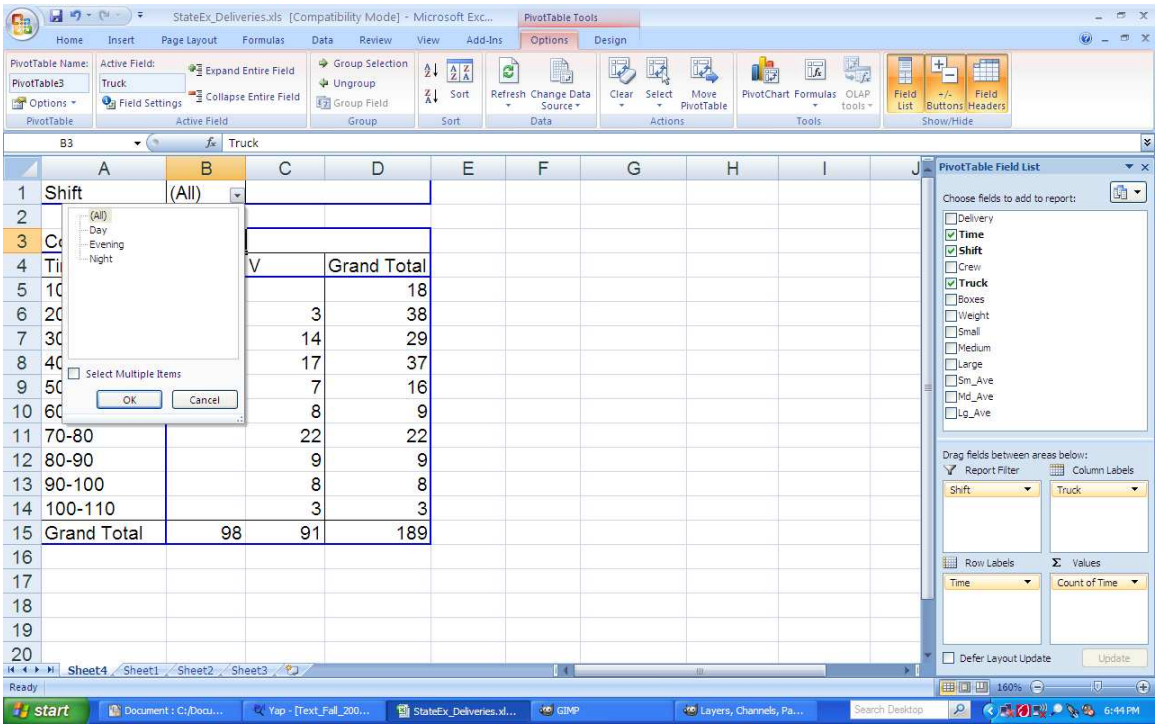

Figure 3.9: Selecting values of the "Page Field" variable for display in the pivot table.

you to display data from just one of the shifts (Day, Evening, or Night) as shown in figure 3.9 (page 93). One can then explore how each of the variables relates to the data.

One can also create more complex pivot tables by placing several variables in one location. For example, by dragging first the "Truck" and then the "Shift" variable into the Column Field area, we produce a pivot table like the one shown below. This pivot table breaks down the data even further than before, showing us how the "Truck" and "Shift" variables relate to the distribution of unloading times.

Count of Time	Truck	Shift							Grand
	Semi			Semi Total	Van			Van Total	Total
Time	Day	Evening	Night		Day	Evening	Night		
10-20		7	11	18					18
20-30	8	10	17	35			3	3	38
30-40	5	4	6	15	6	6	2	14	29
40-50	3	6	11	20	10	6	1	17	37
50-60		5	4	9	1	2	4	7	16
60-70		1		1	4	2	2	8	9
70-80					9	7	6	22	22
80-90					4	4	1	9	9
90-100					6	2		8	8
100-110					2	1		3	3
Grand Total	16	33	49	98	42	30	19	91	189

Using Autofilter to Sort and View Data

Often, our data contains so much information that it can be difficult to work with it and locate specific pieces of information. For example, in the pivot table above, we see that there is one semi load that took a little longer than the others. This occurred on the night shift. So it might be useful to arrange the data to focus on just the night shift. One tool for this is called "Autofilter". To activate the Autofilter, first click on any cell in the data, use the menu option "Data/ Filter/ Autofilter". (When you are finished, you select "Autofilter" again to turn it off; it works like a toggle switch.)

When Autofilter is active, you will see a little pull-down arrow next to each of the fields in your data. Excel assumes that you have arranged the data with columns as variables (fields) and rows as observations (records). Pulling down the arrow next to a field will allow you to select any of the values of that field; the table will then display only the records with that value. So, we could select "Evening" from the "Shift" field and "S" (for Semi) from the "Truck" field to display only the data for unloading times during the evening for a semi. While there is still a lot of data to explore, this simplifies the problem considerably. You could further reduce the data by filtering based on crew size or some other variable.

Notice that whatever fields have been filtered have a blue pull-down arrow; fields that have not been used to filter have a black pull-down arrow. You can unfilter on one variable by simply selecting "all" from its pull-down menu. The pull-down menus also allows you to "Sort ascending" or "Sort descending". Finally, if your variable has many values (like a numerical variable) you can select "Custom" from its menu and entering the information in dialog box. For the data in the example, if you display only semis unloaded on the evening shift and then select "Custom" for the "Time" field and use this to select only times greater than 48 minutes, we see that all the long unloading times were 2 or 3 person crews, giving us greater insight into the situation at StateEx.

3.3 Homework

3.3.1 Mechanics and Techniques Problems

3.1. Download the data file C03 Salaries.xls. This data represents salaries for employees at a small company.

1. Add in two new columns of computed data: The first column should contain the salaries of each employee after a flat $1000 raise. The second column should contain the salaries after a 5% raise.
2. What are the mean, median, and standard deviation of these three different salaries? (Be sure to copy and paste these statistics from Excel).
3. What happened to the mean after the $1000 increase? Why?
4. What happened to the mean after the 5% increase? Why?
5. What happened to the Standard Deviation after the $1000 increase? Why?
6. What happened to the Standard Deviation after the 5% increase? Why?

3.2. Data file C03 INCOMES.XLS contains a list of 1000 family incomes from each of four fictitious countries. All of the families are the same size (two parents and one child) and the families form a representative sample of such families in their country.

1. Place the four countries in order of increasing average income. Explain what this ordering tells you about these countries. Describe each country as either "Poor", "Average", or "Wealthy".
2. Place the four countries in order of increasing standard deviation. Explain what this ordering tells you about these countries. Describe each country as either "Shared Wealth" or "Disparate". Make sure you consider not just the standard deviation alone, but also how it compares with the mean income of the country.
3. Which country would you want to live in? Explain your reasoning.

3.3. These problems will involve the data set C03 AllTron.xls. The data shows information on the employees at AllTron, an electronics design company. Use pivot tables to answer the following questions.

1. How many of the employees in this sample are men (gender=0 for male, 1 for female)?
2. What percentage of the employees is female?
3. What is the average salary of the male employees?

4. What is the average salary of the male employees who have exactly four years of post-secondary education?

5. What is the total number of years of experience at AllTron for the male employees? For the female employees?

3.4. Using data file C02 HOMES.XLS, answer each question below by preparing a pivot table in Excel (Data/PivotTable). Put your answers into a Word document and be sure to include your answer in full sentences as well as the pivot table or other computations that support your answer.

1. How many of each STYLE of home do we have?

2. How many of each STYLE of home, broken down by the number of bedrooms?

3. What is the average VALUE of each STYLE of home, broken down by BED?

4. What is the Standard Deviation of VALUE, broken down by STYLE (row) and BED (column)?

5. What is the average SIZE of the houses, broken down by LOCATION (row), and BED (column)?

3.5. Using data file C03 BEEFNBUNS.XLS, answer each question below by preparing a pivot table in Excel (Data/PivotTable). Be sure to include your answer in full sentences as well as the pivot table or other computations that support your answer.

1. Use Statpro/SummaryStats to determine the mean and standard deviation for the variable WaitTime.

2. Now create a pivot table and calculate the AVERAGE of WaitTime.

3. Now break that statistic down with Venue for the Row Variable (C for Counter; D for Drive-thru.)

4. Can you break that down by Complexity (as the Column Variable)?

5. How about Venue (as the Row) and Size (as the Column Variable)?

6. How about Time (Time of Day) by Size?

3.3.2 Application and Reasoning Problems

3.6. Consider the data and work you did in problem 1.

1. Will any fixed amount of increase result in the same change to the mean? What about a salary decrease? Explain your reasoning.
2. Will any fixed amount of increase result in the same change to the standard deviation? What about a salary decrease? Explain your reasoning.
3. Will any percentage increase result in the same change to the mean? What about a salary decrease? Explain your reasoning.
4. Will any percentage increase result in the same change to the standard deviation? What about a salary decrease? Explain your reasoning.

3.7. Based on your work (and anything else you think is important to investigate) from problem 3, is it more important to have more experience or more education before working at Alltron? Support your claim.

3.8. Consider the quarterly sales figures at a national chain of pet supply stores. The stores are divided into four geographic regions (NE = Northeast, NW = Northwest, SE = Southeast, SW = Southwest).

Region	NE	NW	SE	SW
Mean Sales (thousands of dollars)	409	384	265	241
Standard Deviation in Sales (thousands of dollars)	112	77	73	120

1. Which region is performing better? Justify your answer.
2. Which region is performing the worst? Justify your answer.
3. Which typical sales figure is most trustworthy? Explain.
4. In which region do you expect to find the store with the highest sales? Explain.
5. In which region do you expect to find the store with the lowest sales? Explain.

3.3.3 Memo Problem

To: Carnivorous Crusie Lines Project Team
From: Director of Marketing
Date: May 14, 2008
Re: Preliminary analysis of venue attendance

As you know, we have won the contract on Carnivorous Cruise Lines and have begun data collection. The enclosed file contains attendance data for four venues over a one-week cruise. We want to analyze the data for this one cruise in order to determine the venues and days to focus our attention on for subsequent cruises. Since each venue has a different capacity, use a percent of capacity (as a decimal) to measure the attendance. I would like to see two separate analyses: One analysis that compares each venue and each night to the overall cruise data and one analysis which looks at each venue individually with respect to its own data. We need the two analyses to answer the following questions:

- Are there any venues performing so poorly (with respect to the entire data set for the cruise) that are candidates for elimination?
- Are there certain nights on which a particular venue does poorly with respect to its own performance and should be considered for possible closing on those nights?
- After deciding on which venues or nights might need to be eliminated what effects would such a change be likely to have on the other venues and nights?

Attachment: Data file "C03 Venues.XLS"