Chapter 9

Describing Data with Statistics

In chapter four, the descriptive statistics of average, minimum, maximum, and median were used to describe a group of numbers and explain how those statistics influence prices. In this chapter, more sophisticated statistical tools are used to describe groups of numbers and to infer the characteristics of large groups of data from samples.

Why Do I Need to Know This?

Some people say that you can prove anything with statistics so they do not believe any of them. Even though statistics is imperfect, it is still the best tool we have for understanding the behavior of groups that are too large to measure each component individually or where people are involved. Our news reports are constantly referring to "studies" that show "links". Like charts that do not start at zero for their vertical axis, assumptions can be made that exaggerate or misrepresent the facts. You can be an informed consumer of studies and news reports about them if you learn the capabilities, limitations, and assumptions used for the statistics on which these studies are based. This skill will be personally useful in a variety of ways from understanding a test that was graded on a "curve" to deciding if some invisible factor is really a threat to your family's health.

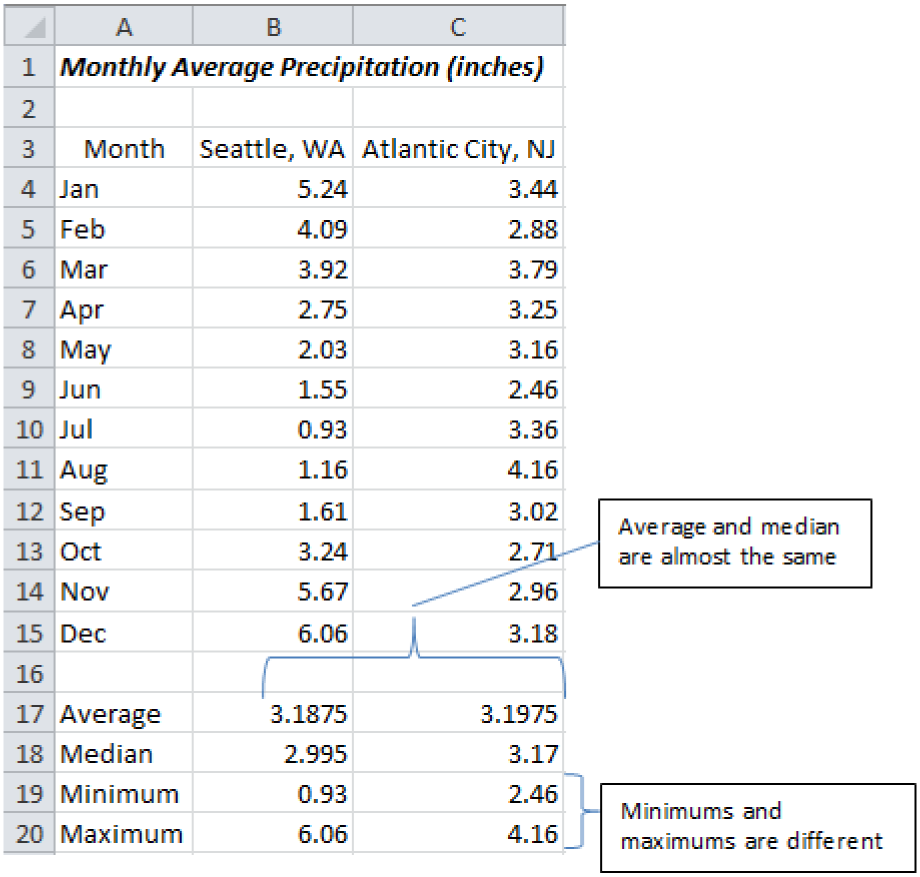
1 Describing a Group of Numbers

Learning Objectives

1. Define frequency distribution and identify its uses. [9.1.1]
2. Define Histogram and identify its uses. [9.1.2]
3. Describe the central tendency theorem. [9.1.3]
4. Identify normal and skewed distributions. [9.1.4]

Frequency Analysis

Descriptive statistics like average, median, minimum, and maximum provide valuable information about a group of numbers but they do not indicate how the values are distributed within the group. For example, consider the average monthly rainfall in two cities as shown in Figure 9.1.



	A	B	C
1	***Monthly Average Precipitation (inches)***		
2			
3	Month	Seattle, WA	Atlantic City, NJ
4	Jan	5.24	3.44
5	Feb	4.09	2.88
6	Mar	3.92	3.79
7	Apr	2.75	3.25
8	May	2.03	3.16
9	Jun	1.55	2.46
10	Jul	0.93	3.36
11	Aug	1.16	4.16
12	Sep	1.61	3.02
13	Oct	3.24	2.71
14	Nov	5.67	2.96
15	Dec	6.06	3.18
16			
17	Average	3.1875	3.1975
18	Median	2.995	3.17
19	Minimum	0.93	2.46
20	Maximum	6.06	4.16

Average and median are almost the same

Minimums and maximums are different

Figure 9.1. Comparison of precipitation in Seattle and Atlantic City.

By considering the minimums and maximums, one can observe that all of the values for monthly precipitation fall between zero and 7 inches. This range can be divided into equal intervals called bins and the number of values that occur in each interval can be counted. The number of times something occurs is its frequency and how the counts are distributed across the bins is the frequency distribution. The frequency distribution can be represented by a column chart. Frequency distribution charts for the precipitation in Seattle and Atlantic City are shown in Figure 9.2.

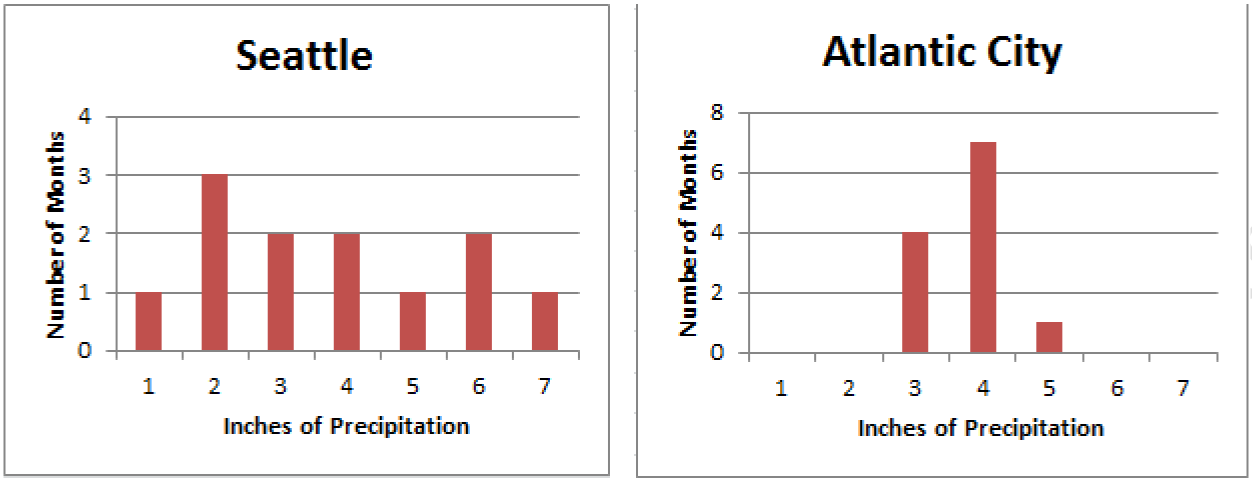


Figure 9.2. Frequency distribution of precipitation.

The sizes of the bins are 1 inch of precipitation and the height of the columns represents the count of the number of months that had precipitation in each range. Comparison of the two charts shows that the rain and snow that make up the precipitation in the two cities is distributed differently. In Atlantic City, all the values for precipitation fell within the bins for 2-3”, 3-4”, or 4-5” while in Seattle, the precipitation was more widely distributed. A special type of column chart where the column widths represent the size of the bins is called a histogram. The column charts can be modified so the width of the columns fills the bin ranges to make histograms, as shown in Figure 9.3.

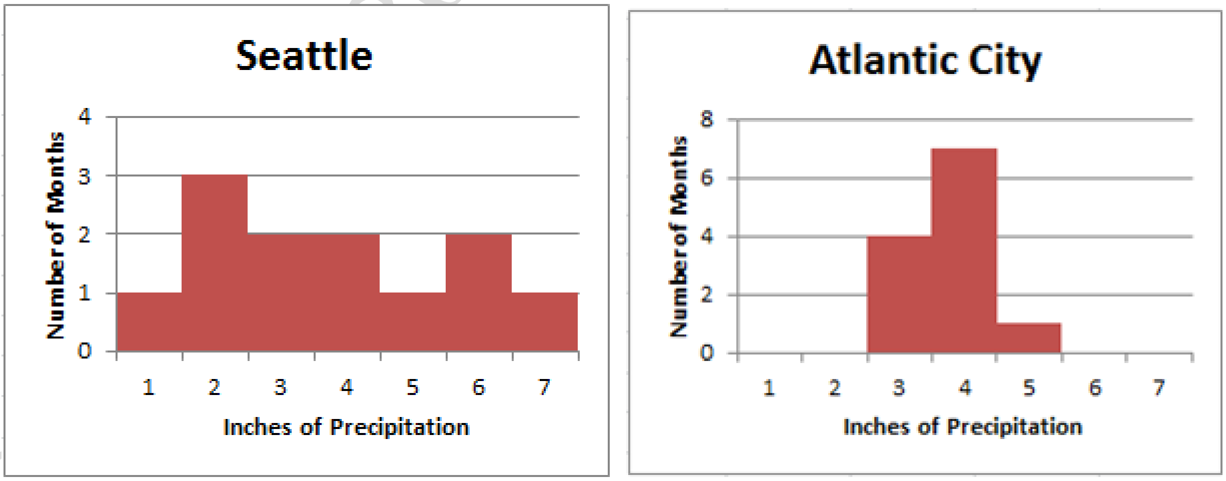


Figure 9.3. Histograms of precipitation in two cities.

Some educators use frequency distribution charts to analyze student performance on tests. For example, a test was given where students scored between 50 and 100. The instructor created a frequency distribution chart using bins that are 2 points wide and counted the test scores that fell within each bin. The chart of the distribution of scores is shown in Figure 9.4. Because the outline of this type of frequency distribution resembles the profile of a bell, it is often called a “bell curve”.

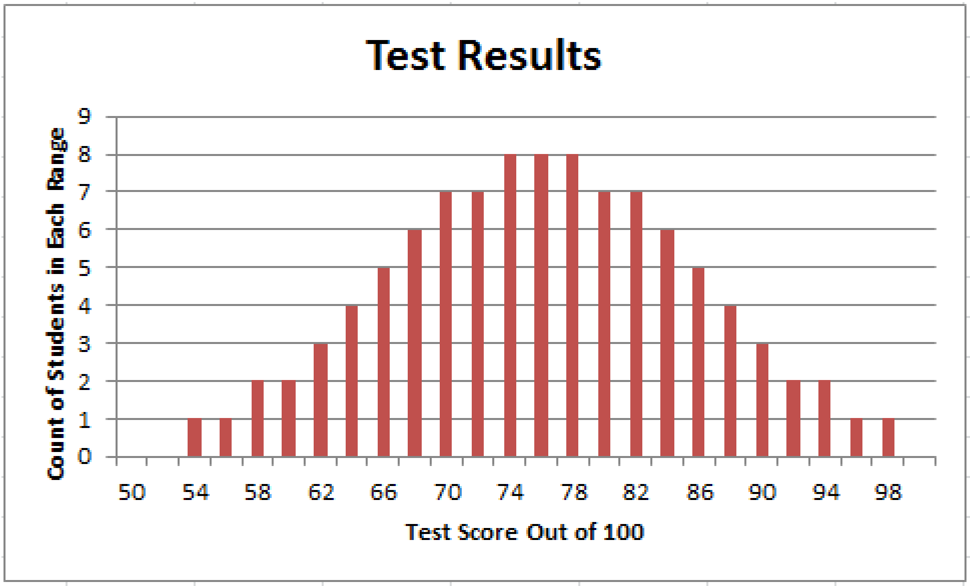


Figure 9.4. The shape of the distribution resembles the profile of a bell.

Normal Distribution

There are many different factors that influence the grades that students receive on a test. For some individuals in the group, the factors are all positive influences and they get the highest scores. For others they are all negative and they get the lowest scores. If the group of students is chosen without preference for any of these factors, the factors tend to cancel each other out for most people and the majority of the scores end up near the middle of the group close to the mean (average). This is called the central limit theorem. The bell-shaped distribution curve that is caused by competing random factors is called a normal distribution. If the frequency distribution is not symmetric on either side of the mean, it is skewed.

Key Takeaways

- If the range of values from the minimum to the maximum is divided into equal intervals called bins, the values in the group that fall into each interval can be counted. The count in each bin is the frequency distribution. [9.1.1]

- A histogram is a column chart of the frequency distribution where the width of the column represents the width of the bins and the height represents the count in each bin. [9.1.2]
- Random factors that cause members of the group to differ from the mean tend to cancel each other out most of the time, which results in most of the counts being close to the mean. [9.1.3]
- If the frequency distribution is symmetric and bell-shaped as a result of several competing factors it is a normal distribution. If the frequency distribution is not symmetrical it is skewed. [9.1.4]

2 Statistics and Quality

Learning Objectives

1. Define quality, grade, and statistics. [9.2.1]
2. Describe how the standard deviation is calculated. [9.2.2]
3. Define and explain statistical terms used in quality control. [9.2.3]
4. Estimate the likelihood of samples falling within one, two, or three standard deviations of the mean given a normal distribution caused by random factors. [9.2.4]
5. Relate tolerances to quality manufacturing programs such as Six Sigma. [9.2.5]
6. Recognize random patterns in run charts versus trends. [9.2.6]

Statistics are used in manufacturing to describe products. If a product is consistently measured to be what it is required to be, it is said to be of high quality. (International Organization for Standardization 2000) Similar products can have different requirements that separate them into grades. For example, gasoline has several different grades that are based on what its octane rating is required to be, as shown in Figure 9.5.

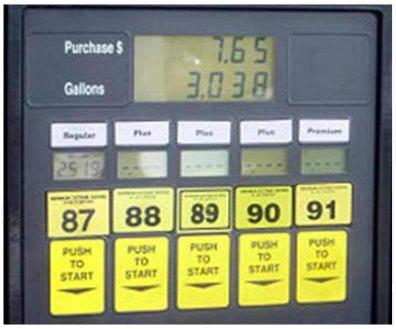
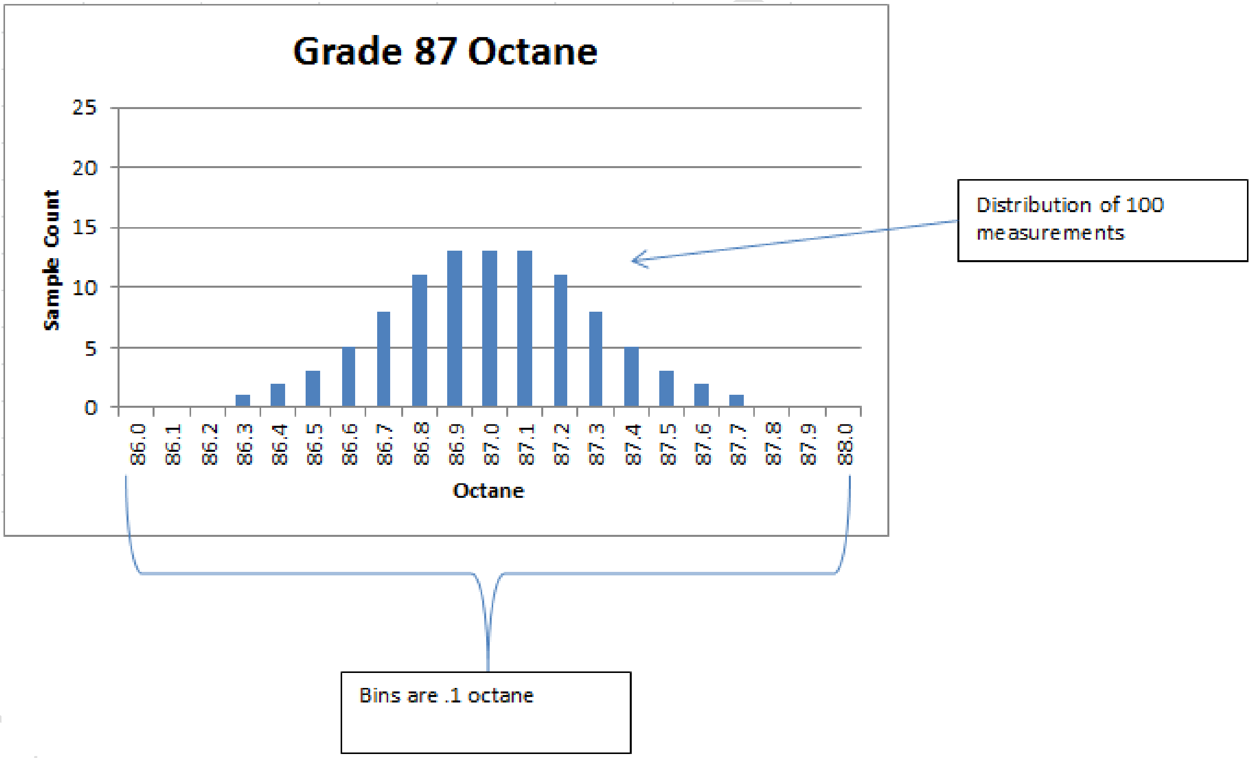


Figure 9.5. Grades of gasoline

The quality of each grade depends on how well it meets the requirements for that grade. At a refinery, the production manager might be producing 87 octane gasoline[1]. To check the quality, she takes 100 samples over a period of a day and measures the octane rating of each sample. A sample is a part of the group that has characteristics that represent the group. A frequency distribution of the samples is shown in Figure 9.6.



Distribution of 100 measurements

Bins are .1 octane

Figure 9.6. Frequency distribution of the octane of gasoline samples.

Tip: A common practice is to take samples and test them instead of attempting to test all

[1] This is a hypothetical example and any similarity to a particular brand of gasoline is coincidental.

of the members of a group. The characteristics of the unmeasured members of the group are inferred from the characteristics of the samples. Statistics that are based on samples but which are applied to the whole group are called inferential statistics. It is very important to select the samples carefully so they represent the group as well as possible.

After improvements are made to the equipment, more gasoline is made and another 100 samples are taken. This time there is less difference from the mean of 87, as shown in Figure 9.7.

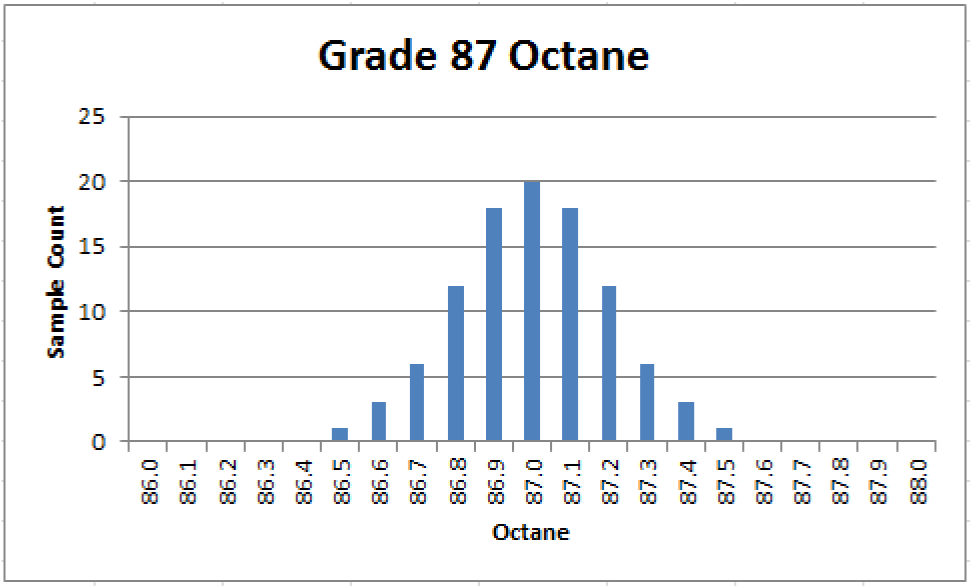


Figure 9.7. Higher quality production of 87 octane gasoline.

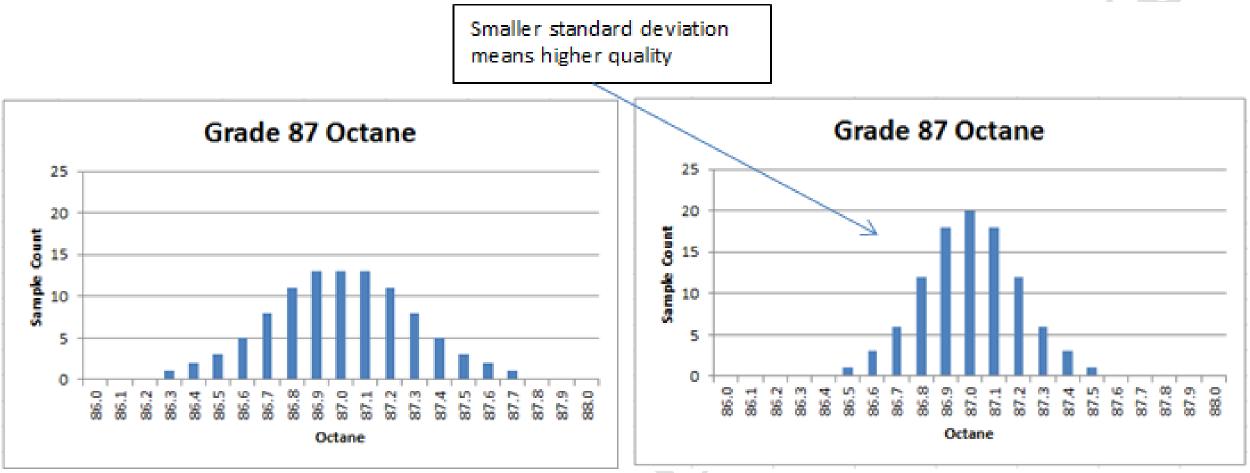
The second production run of 87 grade gasoline is of higher quality because the samples are closer to the requirement of its grade.

Both of these distributions are symmetric and are typical of a normal distribution. To describe the degree of spread of a normal distribution from the mean, a statistic called the standard deviation (STD) is used. The standard deviation is similar to the R^2 calculation for a trend line and uses the following steps:

1. The average (mean) of the measurements is found by summing them and dividing by the count.
2. The difference between each measurement and the mean is squared.
3. The average of the squared differences is found by summing them and dividing by the count.

4. The square root of the averaged squared difference is found.

Unlike the R^2 value, the result is not subtracted from zero. A small standard deviation means that most of the samples are close to the mean. For example, the standard deviation for the first run is .3 and the standard deviation for the second run is .2. A smaller standard deviation implies that the distribution curve is steeper and narrower, as indicated in Figure 9.8.



Smaller standard deviation means higher quality

Figure 9.8. Better manufacturing techniques result in smaller differences from the required value.

Instead of writing the term *standard deviation*, the Greek lowercase letter sigma, σ, is used. We would say the first run of gasoline had σ = .3 and the second was σ = .2.

The standard deviation is a very useful statistic. Because the normal distribution curves are all similar, we know from experience and mathematical calculations that about 2/3 of all the samples will be within one standard deviation on either side of the mean. Mathematicians use a theoretical model where there are many random factors and very small bins that create a smooth bell-shaped curve. If a real experiment is close to this theoretical model, 68% of the samples would be within one sigma on either side of the mean, 95% would be within two sigmas, and 99.7 would be within three sigmas. This is called the 68-95-99.7 rule. The data in our example does not exactly match this ideal distribution curve but it is close, as shown in Figure 9.9.

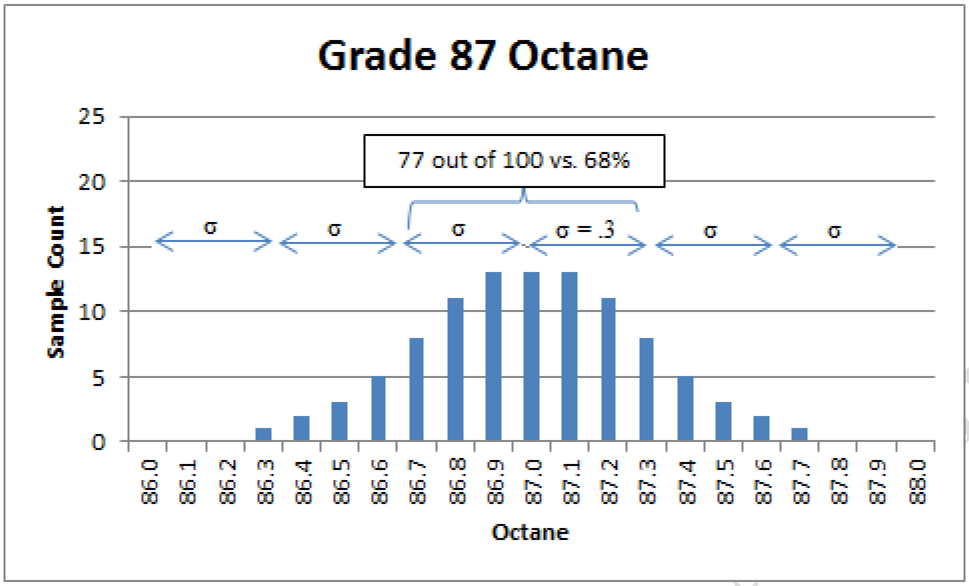


Figure 9.9. Normal distribution.

The manager of the refinery chooses to set the goal at producing 87 octane with σ =.2 octane. The manager also sets upper and lower limits that the process should rarely exceed called the control limits. The production manager samples the gasoline each day for twenty days and records the octane of the samples on a run chart as shown in Figure 9.10. Run charts show samples taken while the process is running.

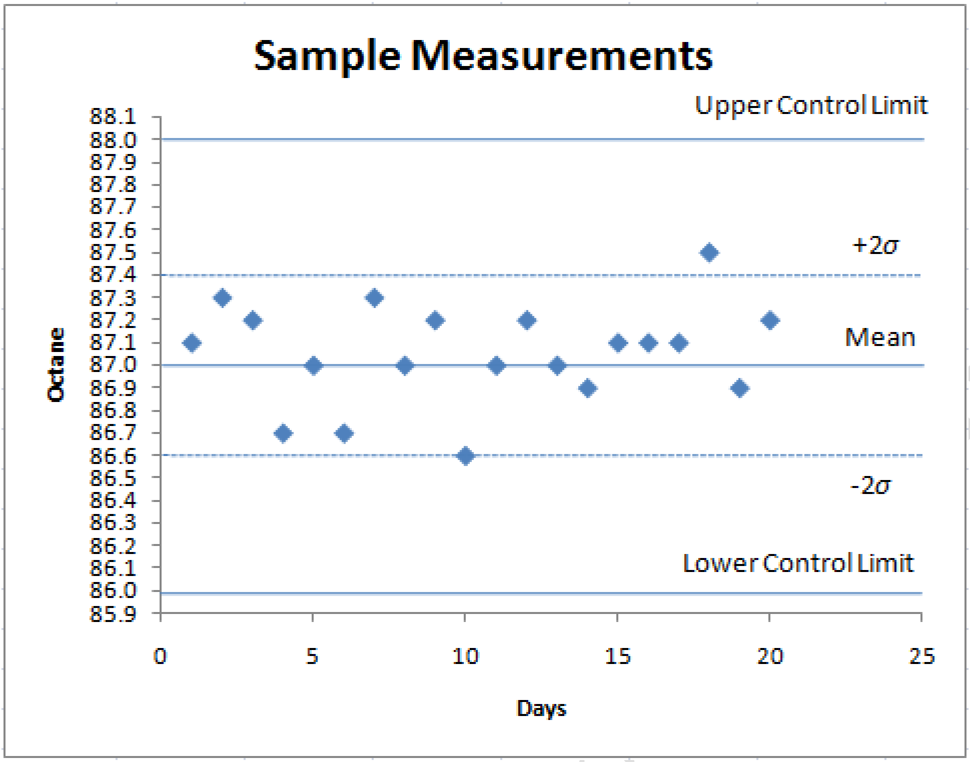
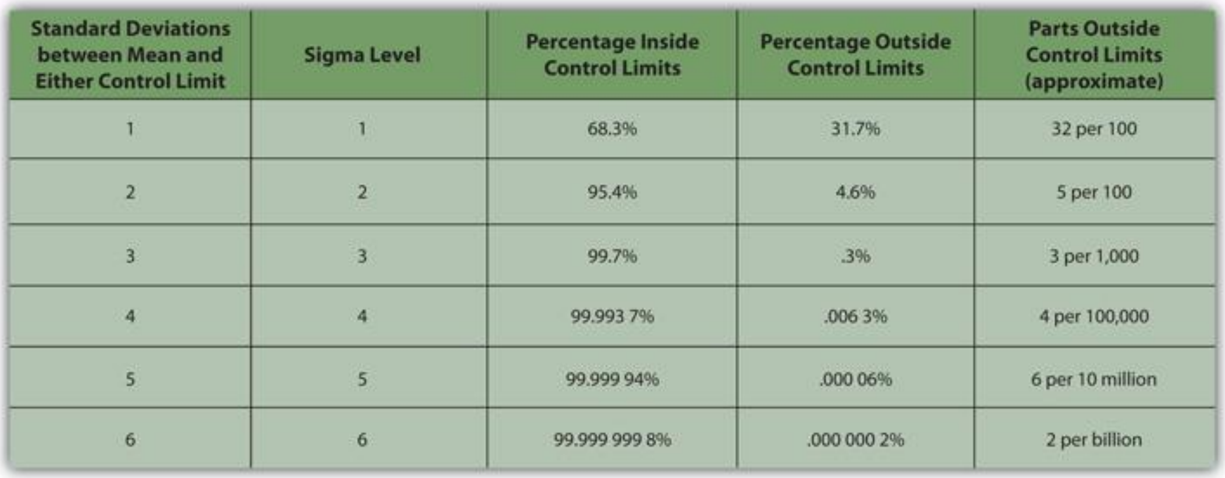


Figure 9.10. Run chart.

To choose a reasonable control limit, the manager refers to the table in Figure 9.11 that shows the likelihood of getting a sample outside of control limits.



Standard Deviations between Mean and Either Control Limit	Sigma Level	Percentage Inside Control Limits	Percentage Outside Control Limits	Parts Outside Control Limits (approximate)
1	1	68.3%	31.7%	32 per 100
2	2	95.4%	4.6%	5 per 100
3	3	99.7%	.3%	3 per 1,000
4	4	99.993 7%	.006 3%	4 per 100,000
5	5	99.999 94%	.000 06%	6 per 10 million
6	6	99.999 999 8%	.000 000 2%	2 per billion

Figure 9.11. Control limits table.

Observe from Figure 9.10 that the manager chose to set the control limits at 88 and 86 octane which would mark the border between 87 octane and the neighboring grades of 89 and 85. If the standard deviation is .2 octane, this represents five sigmas from the control limits (1 / .2 = 5). From the chart in Figure 9.11, the manager observes that only 6 measurements out of 10 million should be outside the control limits.

The range between the mean and the control limit is the tolerance. In this example the tolerance is 1 octane above or below the mean, which implies that the product will have an average (mean) value of 87 octane but rarely ever be lower than 86 or higher than 88. Tolerance is often indicated with the symbol, ±, which stands for *plus or minus*. The octane rating would be written as 87 ± 1 octane.

Tolerances are very important in manufacturing. For example if a rod is meant to fit into a hole, the largest rod should fit inside the smallest hole. If the standard for manufacturing the rod is 9 ± .6 millimeters and for drilling the hole is 10 ± .6 millimeters it would be acceptable to make a rod that is 9.6 mm and a part with a hole that is 9.4 mm in which case the rod would not fit. At the other extreme, the smallest rod could be 8.4 mm and if placed into the largest hole, 10.6 mm, the space between them might be so large that the rod would wobble and wear out too soon.

> Tip: When people talk about quality manufacturing, they are talking about making parts that have small tolerances and then creating processes with standard deviations that are small enough so that few products ever exceed the tolerances. In a complex product like an automobile that has thousands of parts, even high-quality processes will still produce products with a few defects. Ninety-five percent right might be an "A" in school but five bad parts per hundred (a 2σ process) is an "F" in manufacturing. Be ready for a much higher standard of accuracy in the working world.

One of the popular quality control methods is called Six Sigma, which implies that their goal is make products of such high quality that only two per billion are outside of control limits (see Figure 9.11).

Random variations do not repeat in predictable patterns but they have characteristics that can be recognized. As previously mentioned, if there are several different factors that vary randomly, they will tend to offset each other as stated in the central limit theorem. A production manager learns to recognize the difference between random variations in production runs and those that have an underlying trend. For example, the following run charts use four random variables that contribute to each measurement. Three runs using different random numbers produce the run charts shown in Figure 9.12.

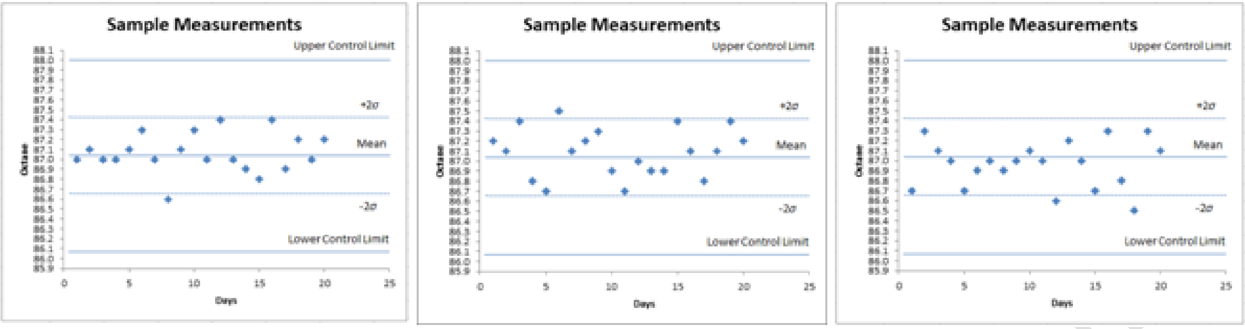


Figure 9.12. Production runs affected by random variables.

If there is an underlying trend that is not random, it can be masked by the random variables but might be recognized by looking at a run chart like the one shown in Figure 9.13 where there seems to be an upward trend.

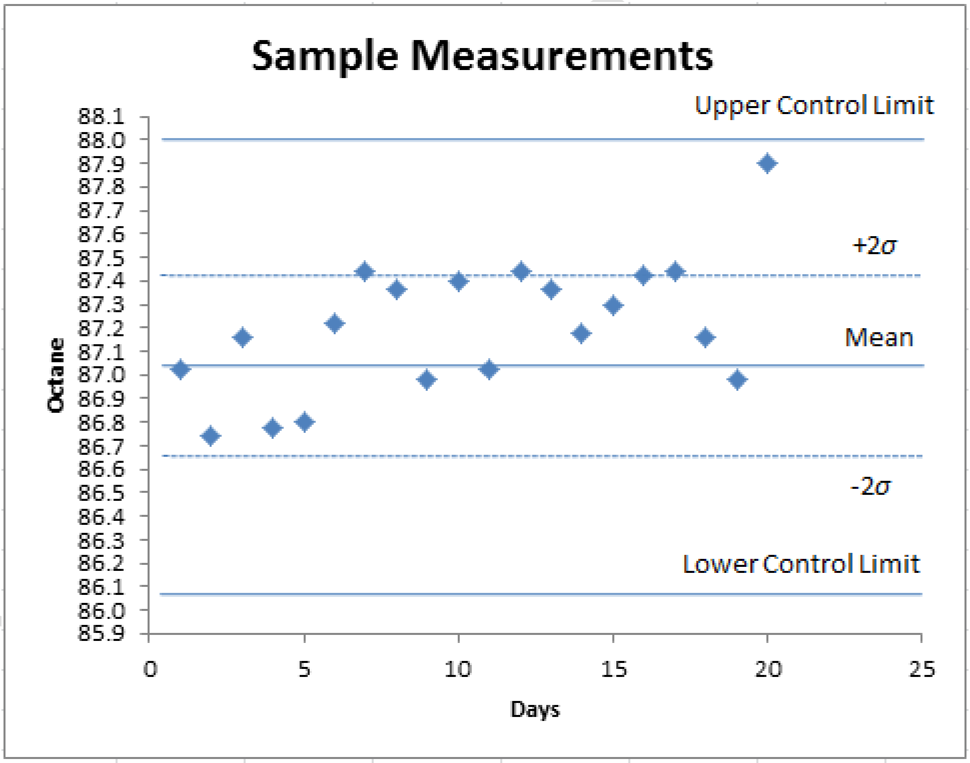


Figure 9.13. Non-random effect mixed with random effects.

The production manager would be alerted to look for an assignable cause behind this trend such as a valve that is wearing out.

Key Takeaways

- Grade is a set of requirements and quality is how closely the product matches the requirement for the grade. Statistics are numbers that describe a group of numbers. [9.2.1]
- Calculating the standard deviation of a normal distribution is accomplished by first finding the mean or average. Next each number in the group is subtracted from the mean and this difference is squared. The average of the squared differences is calculated and then its square root determined, which is the standard deviation. [9.2.2]
- The standard deviation is represented by the Greek letter sigma, σ. The quality is determined by the size of the standard deviation. Lower sigma values imply that a larger percentage of the samples are close to the mean. A tolerance is a range of differences from the mean that is acceptable. A run chart shows the value of sample measurements taken while the process is running. [9.2.3]
- The 68-95-99.7 rule describes the percentage of samples that are likely to fall within 1, 2, and 3 standard deviations from the mean. [9.2.4]
- A popular quality control method is called Six Sigma because it attempts to make the standard deviations so small that the control limits are six standard deviations (sigmas) from the mean. This implies that less than 2 parts per billion would be outside the control limits. [9.2.5]
- Some factors vary randomly while others have a cause called an assignable cause that can be identified and fixed. An assignable cause is determined when patterns in the data points on a run chart appear to have an overall trend. [9.2.6]

3 Studies of People

Learning Objectives

1. Define correlation and its relationship to causality. [9.3.1]
2. Describe the elements of the design of the study. [9.3.2]
3. Describe confounding variables and give an example. [9.3.3]
4. Describe the null hypothesis and interpretations. [9.3.4]
5. Identify tips for interpreting statistical studies. [9.3.5]

Measuring and describing variations of inanimate objects like the parts made in a manufacturing process is easier than measuring the effects of environmental factors on people. There is less certainty and there are more ways that mistakes can occur. Regardless of the potential for error, studies of people can be very useful for making decisions, like choosing new drugs or determining if something in the environment is harmful.

Correlation and Cause

Recall from the chapter on charting that a line chart shows the relationship between an independent and dependent variable and that the R^2 value indicated how closely the data points fit a line that represents a predicted behavior. It is possible to have two dependent variables that behave in a similar manner. For example, if we look at a chart that shows the number of people who took up smoking cigarettes in the U.S. and the number of lung cancers deaths, the two sets of data seem to depend on time in the same way with a 20-year lag between them, as shown in Figure 9.14.

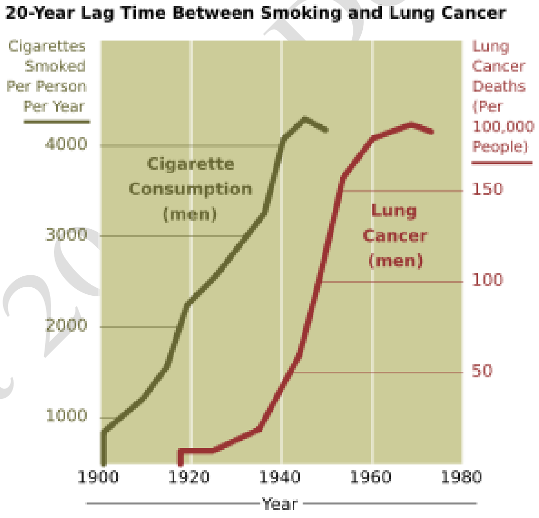


Figure 9.14. Cigarette smoking and lung cancer deaths.

The two factors, smoking and lung cancer deaths, are both dependent variables and the year is the independent variable. The similarity in the shapes of the lines indicates that the two factors might be related in some manner. If two variables change in the same way, we can say that they are co-related or correlated. If two factors are correlated, it raises the question of whether or not one of the factors causes the other. For example, if the position of the sun in the sky is charted versus the time of day and a similar chart is kept of the direction to which a sunflower points during the day, it would become apparent that the two factors are correlated. One might conclude that the position of the sun causes the sunflower to point in that direction or one could conclude that the sun moves where the sunflowers point. The presence of a correlation is an indicator that a causal relationship might exist but it is not proof and it does not indicate which factor is dependent on the other.

For example, many people are concerned about the correlation of the rise in global temperatures and the increase in carbon dioxide in the atmosphere. Both have risen in the last 150 years, as shown in Figure 9.15.

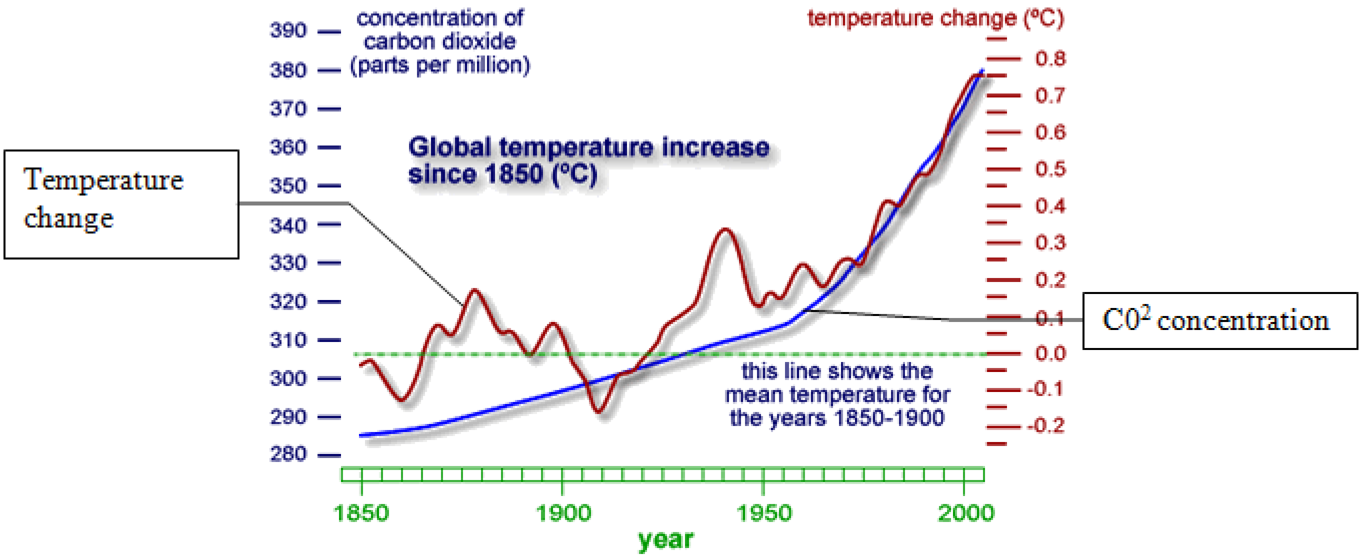


Figure 9.15. Possible correlation between global temperature and carbon dioxide.

If a correlation exists, the next step is to suggest a mechanism that would explain the correlation. Biologists can provide an explanation for the sunflower's turning toward the sun but no plausible explanation for how the sunflower could move the sun has been proposed. In the case of carbon dioxide and global temperature increase, the mechanism proposed for explaining how carbon dioxide increase causes temperature increase is the greenhouse effect. Others have argued that the ocean gives off carbon dioxide when it gets warmer which would be a mechanism that explains why carbon dioxide increases when temperatures rise.

Studies

To determine if one factor depends on the other, a study can be performed. The purpose of a study is to determine if changes in one variable cause changes in another.

Elements of a study:

- Choosing a sample that can be measured
- Identifying and blocking confounding variables
- Stating the null and alternative hypothesis
- Choosing a method of performing the study
- Performing the study
- Analyzing the data and reporting the results

Choosing a Sample

The sample must be selected or extracted from the larger group of data so that it is typical of the group. For example, air bubbles trapped in glaciers might be used as samples of the air from hundreds of years ago to measure the carbon dioxide concentration. The assumption is that the air that was trapped in the glacier in Greenland was typical of the rest of the atmosphere at the time.

Identifying and Blocking Confounding Variables

There might be other variables that explain the variation instead of the one that is being considered, which would make the study incorrect. For example, a study of women who took hormone replacements showed a correlation to lower incidence of heart disease and some people thought that taking the hormone was the cause of reduced heart problems. More studies were conducted to check this assumption of causality and it was found that the hormones actually increased the risk by a small amount. The confounding variable was the income level of the women who could afford the hormone treatments. Researchers had previously found that wealthier women usually had better health and less heart disease and the women who could afford the hormones were wealtheir than those who could not.

Typical confounding variables that most researchers check for are:

- Age
- Gender
- Social and economic status

Hypothesis Statement

The study is designed to address a hypothesis that must be carefully stated. There are usually two of them. The null hypothesis is a statement that is stated negatively such as *"Living near a cell phone tower does not increase the risk of brain cancer."* The alternative hypothesis is usually the positive version such as; *"Living near a cell phone tower increases the risk of cancer."* Another example is used in criminal trials where the hypothesis is; *"The defendant did not commit the crime."* and the altertenative hypothosis is; *"The defendant did commit the crime."*

There are two types of studies, experimental and observational.

Experimental Studies

Experimental studies are done by applying a treatment to a group and then measuring its effect. Studies of the effectiveness of a new drug begin by choosing a symptom that can be measured to indicate the effectiveness of the drug. Researchers recognize that random factors will affect their measurements and their selection of a sample of people but they hope that the central limit theorem will apply and that they will get a normal frequency distribution. Next, they divide the volunteers into two groups using an arbitrary method such as assigning a code number to them sequentially as they volunteer, and then placing every other person in group A and the others in Group B. The researchers recognize that even though the members of the two groups were selected arbitrarily, the means and frequency profiles of the two groups will not be exactly the same.

To illustrate this effect, two groups of numbers can be generated using a computer. Each of the two groups has 1,000 randomly generated numbers between zero and 200 with a mean of about 100 and a normal distribution, as shown in Figure 9.16.

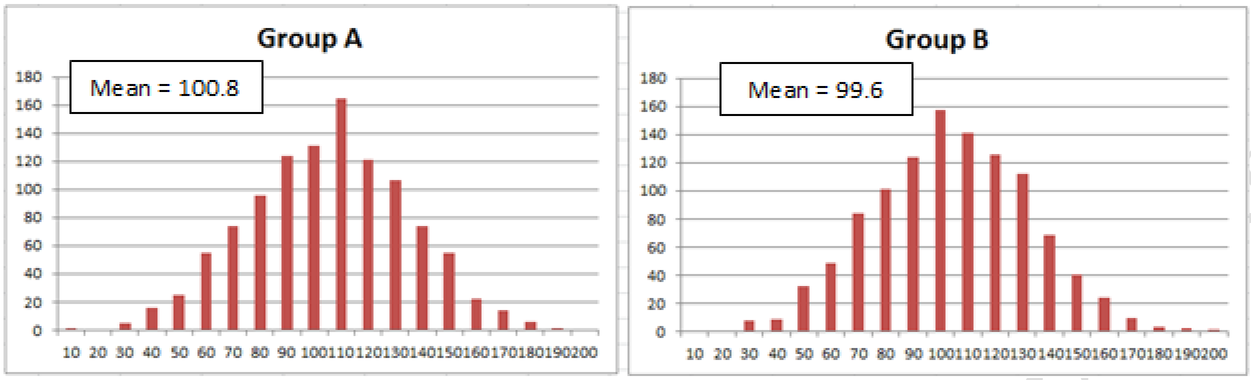
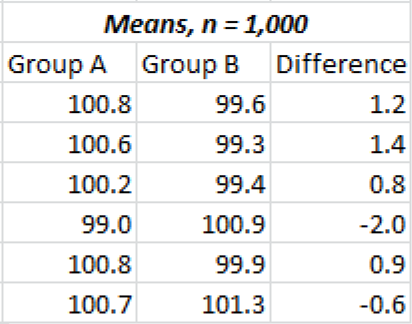


Figure 9.16. Means and distributions are not exactly the same.

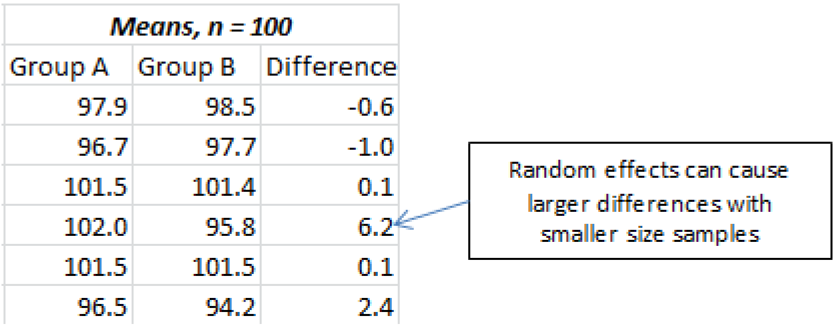
The number of measurements or members of a group is called the n value. If the two groups of numbers are generated several more times with the computer, we can see how the means of the two groups can vary from the random effects, as shown in the table in Figure 9.17.



Means, n = 1,000		
Group A	Group B	Difference
100.8	99.6	1.2
100.6	99.3	1.4
100.2	99.4	0.8
99.0	100.9	-2.0
100.8	99.9	0.9
100.7	101.3	-0.6

Figure 9.17. Differences in the means of the two groups.

Because there are 1,000 numbers in each group, the random effects are more likely to offset each other and the means will not differ by much. If the groups are reduced to only 100, n=100, the random effects are more likely to cause larger differences in the means. If the numbers are generated six times by the computer, with only 100 members in each group, a table of means in Figure 9.18 shows that some of the differences can be much larger for smaller values of n.



Means, n = 100		
Group A	Group B	Difference
97.9	98.5	-0.6
96.7	97.7	-1.0
101.5	101.4	0.1
102.0	95.8	6.2
101.5	101.5	0.1
96.5	94.2	2.4

Random effects can cause larger differences with smaller size samples

Figure 9.18. More variation with smaller sample size.

The next step in the experimental study is to give the members of one of the groups the drug and give the members of the other group a placebo. A placebo is a pill that looks exactly like the drug. To assure that the people who are handing out the drugs and placebos do not give anything away, they are not told which pills are placebos and which contain the drug. When the people who take the pills and the people who are handing them out do not know the difference between the pills with the placebo or the drug, the procedure is called a double-blind study. The group that gets the placebo is called the control group.

Observational Studies

Studying factors that are expected to be hazardous to people requires a different approach. It is not ethical to expose people to harmful conditions intentionally to test a treatment, although history records several examples of this practice that are widely condemned. Instead of an experimental study, researchers work with existing data that requires different techniques for selecting samples and screening for confounding variables. These studies are called observational studies.

Analysis of the data

Before the groups are measured, researchers use the rules that apply to statistical studies of each type to determine a threshold value for how much the means of the two groups should differ before they can conclude that the drug or environmental factor made a difference. The typical practice is to choose a threshold value that will account for 95% of the random factors that might cause the samples to differ from the population they represent. Because there is more variability with smaller samples, the threshold value for the difference is higher for small samples than it is for large samples.

If the means of the two groups are different by more than the threshold value, the difference is called a significant difference which means that there is a 95% likelihood that the difference was not caused by random variations in the samples.

Recall that the null hypothesis is a statement that the drug or treatment does not make a difference. The conclusion of a study is also carefully worded. If the means of the two groups are different but they do not meet the threshold value, the researcher would say; *"We fail to reject the null hypothesis."* Using a double negative is confusing but it is more accurate than saying that we know the drug makes no difference. It would be like a jury returning a verdict that said; *"The prosecution did not prove the case against the defendant beyond a reasonable doubt."* which is not the same as saying the defendant is innocent. If the means are different by more than the threshold value, the conclusion would be; *"We reject the null hypothesis."* This implies that a causal relationship is likely.

Tips for Interpreting Studies

There are many studies conducted on a wide variety of subjects and the results are the basis for making important decisions. A student who grasps the concepts described in this chapter, can be a much better consumer of the reports that come from these studies if a few precautions are followed:

- If the news agency does not provide the name of the study or a way to look at the actual report, the interpretation of the reporter should not be accepted without verification. If a hyperlink is provided, read the summary statement of the actual study to see if the reporter read it correctly. If no hyperlink is provided, attempt to search online for the study to which the reporter refers. The study might be available through a college library if it is not available to the general public.

- Watch for the use of the word *link*. It implies causality when the study might only confirm a correlation. Remember that a correlation does not prove cause.

- Determine the size of the sample. Small samples have higher chance of random factors accounting for observed differences.

- Do not assume that a significant difference means a large difference. It just means that there is a 95% confidence that the difference was not due to random factors in the choice of a sample.

- There is a 5% chance that a positive result is due to random variations in the sample. This is 1 in 20 which means that if there are hundreds of studies being done each year several of them will come up with an incorrect result. If a single study has a positive result, check to see if other studies have confirmed it.

Key Takeaways

- Correlation means that two variables seem to change in the same way to one another or proportionally to each other. It does not imply that one causes the other: [9.3.1]

- The design of a study has the following elements: [9.3.2]

  Choosing the sample that can be measured
  Identifying and blocking confounding variables
  Stating the null and alternative hypothesis
  Choosing a method of performing the study
  Performing the study
  Analyzing the data and reporting the results

- A confounding variable is a hidden factor that might explain the difference instead of the factors that are being considered. Examples are age, gender, and socio-economic status. [9.3.3]

- The null hypothesis is the negative statement of the expected outcomes of a study, such as *the drug does not reduce pain*. If the study does not show a significant difference, the conclusion is that we fail to reject the null hypothesis. If the difference is significant, we reject the null hypothesis [9.3.4]
- Tips for interpreting studies are: [9.3.5]

  Confirm that the source of the study is given and read the summary statement yourself.
  The term "link" is misleading. Determine if they mean correlation or significant difference.
  Check the sample size—small samples have more random variability.
  Do not assume that a significant result means a large difference. Studies with large samples can have small differences that are deemed significant.
  Recognize that one study has a 5% chance of having a sample with random factors that produce an incorrect result. Look for other studies that confirm the study.

4 Using the Frequency and Standard Deviation Functions

Learning Objectives

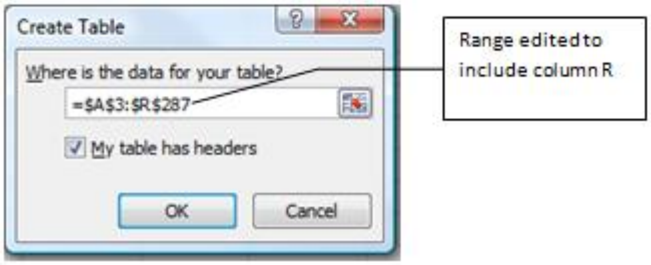
1. Use descriptive statistics functions SUM MIN, MAX, AVERAGE, in an Excel table. [9.4.1]
2. Use the FREQUENCY function to create a distribution of events and chart the results. [9.4.2]
3. Generate random numbers; create a FREQUENCY distribution and related charts to illustrate the effect of the central limit theorem. [9.4.3]
4. Choose the appropriate standard deviation function to use for a group of data. Create a grading curve based on the standard deviation and determine the rank order of the grades. [9.4.4]

Statistical tools help describe data and create information. In this chapter you will use the standard descriptive statistics to describe precipitation, create a frequency distribution of precipitation in two cities and chart the results. Then you will generate random numbers and observe the effect of a larger sample on the distribution of the results. Finally you will determine the standard deviation for a group of test scores, determine the grade distribution, and then chart the results.

Describing Weather Data Using Statistical Functions

Functions such as SUM, AVERAGE, MEDIAN, MIN, and MAX can provide descriptions of data that can be used for a variety of comparisons. In the first exercise you will work with a table of weather data of precipitation for over 280 cities in the United States and Puerto Rico. You will format the data as a table and then use functions to determine the total, average, median, minimum, and maximum precipitation. These basic statistical functions use one argument, which is the range of the group of numbers for which you want to find the statistic.

1. Start Excel. Navigate to the location of files for this class, and then open Ch09Statistics.xlsx and then save the file as Ch09StatisticsStudentName using your own name in place of StudentName.
2. If a Security Warning displays at the top of the file, click the Options button in the Warning bar and then click the *Enable this content* option button. Click OK.
3. Click the Precipitation sheet tab. Right-click the row 2 heading and from the short cut menu click Insert to insert a blank row between the title and the data. You will format this data as a table and the title needs to be separated from the data so that Excel will recognize the data and the labels for each column in the top row of the data.
4. Click in a cell in the data. On the Insert tab, in the Tables group, click the Table button.
5. In the displayed Create Table dialog box, edit the data range so that the last cell is $R$287 as shown in Figure 9.19, and then click OK. In the Information box that displays to advise you that converting the selection to a table will remove all external connections, click Yes to continue. This extends the table definition to column R so that it will include the statistical calculations that you will add to those columns.



Where is the data for your table?

=$A$3:$R$287

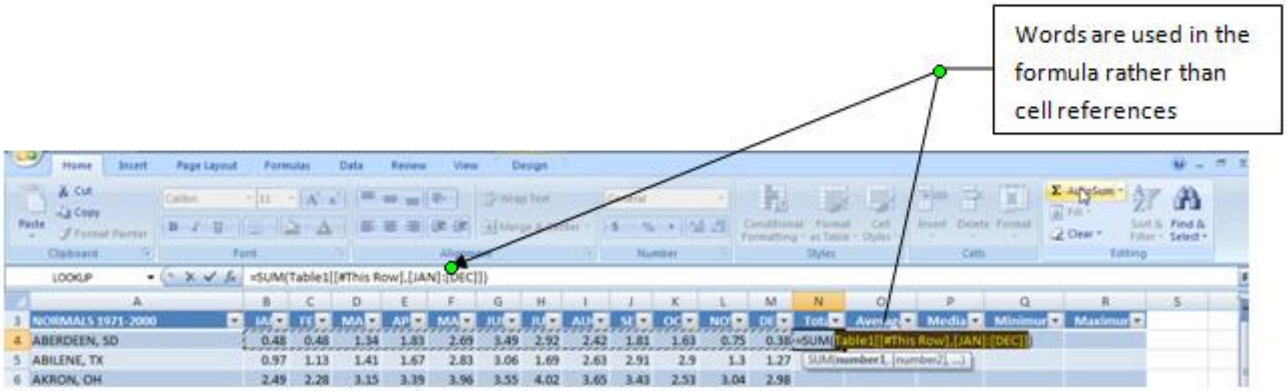
☑ My table has headers

Range edited to include column R

Figure 9.19. Data range for table.

6. Click cell N4 and on the Home tab, in the Editing group, click the AutoSum function to sum the data in row 4, cells B4:M4. Notice that the program uses words rather than cell references in the formula. It

refers to the *Table* and the column labels *Jan:Dec* rather than the cell references B4:M4. Because the data is defined as a table, Excel recognizes this as a table and uses this structured syntax in the formula.



Words are used in the formula rather than cell references

=SUM(Table1[[#This Row],[JAN]:[DEC]])

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R
3	NORMALS 1971-2000	JA	FE	MA	AP	MA	JU	JU	AU	SE	OC	NO	DE	Tot	Average	Media	Minimum	Maximum
4	ABERDEEN, SD	0.48	0.48	1.34	1.83	2.69	3.49	2.92	2.42	1.81	1.63	0.75	0.38	=SUM(Table1[[#This Row],[JAN]:[DEC]])				
5	ABILENE, TX	0.97	1.13	1.41	1.67	2.83	3.06	1.69	2.63	2.91	2.9	1.3	1.27	SUM(number1, [number2], ...)				
6	AKRON, OH	2.49	2.28	3.15	3.39	3.96	3.55	4.02	3.65	3.43	2.53	3.04	2.98					

Figure 9.20. Words are used in the formula.

7. Click the AutoSum button a second time and observe that the formula is filled down column N. It is not necessary to use the fill handle to fill the formula to the rest of the rows in the table. This Auto Calculate is a feature of Excel tables.
8. Click cell O4. On the Home tab, in the Editing group, click the AutoSum list arrow as shown in Figure 9.21. Some of the basic statistical functions that are used frequently can be accessed from the AutoSum button.

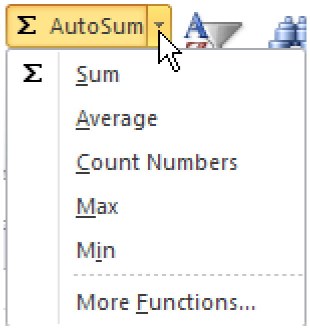
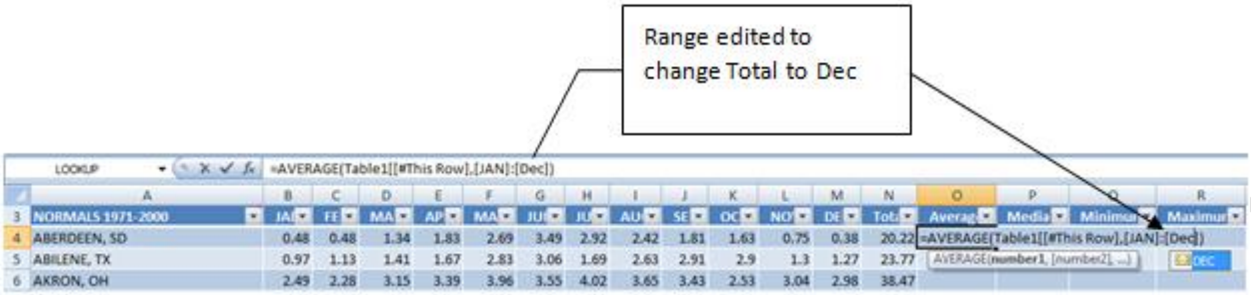


Figure 9.21. Frequently used functions available on the AutoSum button.

9. Click Average and notice that the formula now includes the Total cell. The Total should not be included in the average.
10. In the formula that displays in the cell or on the Formula bar, select TOTAL and type Dec as shown in

Figure 9.22.



Range edited to change Total to Dec

=AVERAGE(Table1[[#This Row],[JAN]:[Dec])

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R
3	NORMALS 1971-2000	JA	FE	MA	AP	MA	JU	JU	AU	SE	OC	NO	DE	Tot	Averag	Media	Minimu	Maximu
4	ABERDEEN, SD	0.48	0.48	1.34	1.83	2.69	3.49	2.92	2.42	1.81	1.63	0.75	0.38	20.22	=AVERAGE(Table1[[#This Row],[JAN]:[Dec])			
5	ABILENE, TX	0.97	1.13	1.41	1.67	2.83	3.06	1.69	2.63	2.91	2.9	1.3	1.27	23.77	AVERAGE(number1, [number2], ...)			DEC
6	AKRON, OH	2.49	2.28	3.15	3.39	3.96	3.55	4.02	3.65	3.43	2.53	3.04	2.98	38.47				

Figure 9.22. Range for Average is from Jan through Dec.

11. Press Enter. Again the formula is filled down the column.

12. Click cell P4 and type =MEDIAN(

13. Drag to select the range B4:M4, confirm that the formula displays =MEDIAN(Table1[@[JAN]:[DEC]]) and then press Enter.

Tip: In Excel 2007 the formula will display =MEDIAN(Table1[#ThisRow],[JAN]:[DEC]]).

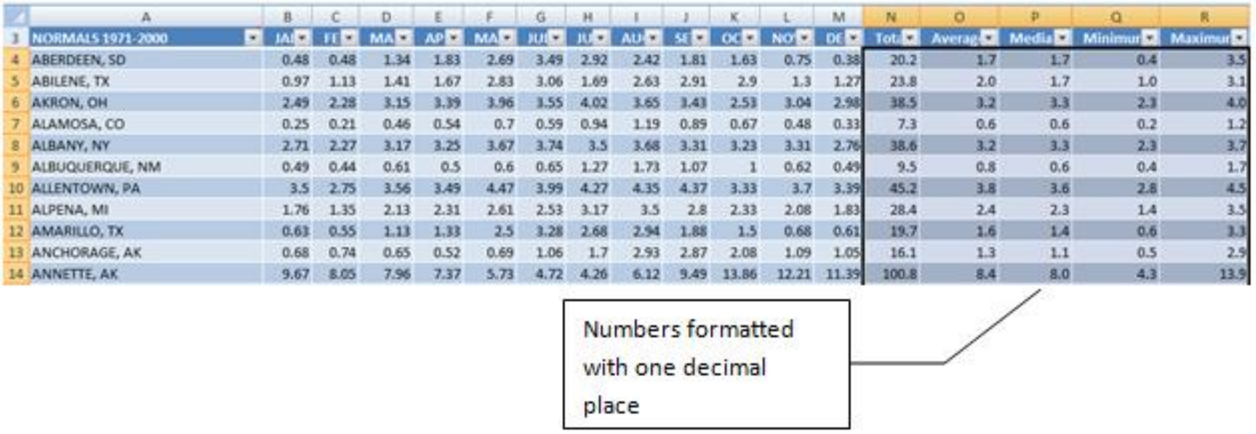
14. In Cell Q4 use the skills you just practiced to use the MIN function to calculate the minimum for row 4. Be sure that the range in the formula is for JAN through DEC. You can type the formula, or use the AutoSum button and select MIN from the list.

15. In cell R4 use the MAX function to calculate the maximum in row 4, ensuring that the range of cells is JAN through DEC.

16. Click in a cell in columns O, P, Q, and R and verify that the range for the formulas is from Jan:Dec.

17. Select cells N4:R4. On the Home tab, in the Number group, click the Decrease Decimal button as needed until all the cells display the numbers with one decimal.

18. With cells N4:R4 still selected, point to the fill handle in cell R4 and double-click the fill handle. This will copy the format of one decimal down the columns to all of the other rows. Compare your result with Figure 9.23.



	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R
3	NORMALS 1971-2000	JA	FE	MA	AP	MA	JU	JU	AU	SE	OC	NO	DE	Tot	Averag	Media	Minimur	Maximur
4	ABERDEEN, SD	0.48	0.48	1.34	1.83	2.69	3.49	2.92	2.42	1.81	1.63	0.75	0.38	20.2	1.7	1.7	0.4	3.5
5	ABILENE, TX	0.97	1.13	1.41	1.67	2.83	3.06	1.69	2.63	2.91	2.9	1.3	1.27	23.8	2.0	1.7	1.0	3.1
6	AKRON, OH	2.49	2.28	3.15	3.39	3.96	3.55	4.02	3.65	3.43	2.53	3.04	2.98	38.5	3.2	3.3	2.3	4.0
7	ALAMOSA, CO	0.25	0.21	0.46	0.54	0.7	0.59	0.94	1.19	0.89	0.67	0.48	0.33	7.3	0.6	0.6	0.2	1.2
8	ALBANY, NY	2.71	2.27	3.17	3.25	3.67	3.74	3.5	3.68	3.31	3.23	3.31	2.76	38.6	3.2	3.3	2.3	3.7
9	ALBUQUERQUE, NM	0.49	0.44	0.61	0.5	0.6	0.65	1.27	1.73	1.07	1	0.62	0.49	9.5	0.8	0.6	0.4	1.7
10	ALLENTOWN, PA	3.5	2.75	3.56	3.49	4.47	3.99	4.27	4.35	4.37	3.33	3.7	3.39	45.2	3.8	3.6	2.8	4.5
11	ALPENA, MI	1.76	1.35	2.13	2.31	2.61	2.53	3.17	3.5	2.8	2.33	2.08	1.83	28.4	2.4	2.3	1.4	3.5
12	AMARILLO, TX	0.63	0.55	1.13	1.33	2.5	3.28	2.68	2.94	1.88	1.5	0.68	0.61	19.7	1.6	1.4	0.6	3.3
13	ANCHORAGE, AK	0.68	0.74	0.65	0.52	0.69	1.06	1.7	2.93	2.87	2.08	1.09	1.05	16.1	1.3	1.1	0.5	2.9
14	ANNETTE, AK	9.67	8.05	7.96	7.37	5.73	4.72	4.26	6.12	9.49	13.86	12.21	11.39	100.8	8.4	8.0	4.3	13.9

Numbers formatted with one decimal place

Figure 9.23. Formulas filled and numbers formatted with one decimal place.

19. Select cells A1:R1. On the Home tab, in the Alignment group, click the Merge & Center button. In the Font group, click the Bold button **B** and the click the Increase Font Size button A˄ until the Font Size box displays 14.

20. In cell N3—*Total*—click the list arrow and then click Sort Smallest to Largest. The entire table is sorted by the Total column with the smallest total of 3.0 for Yuma, AZ displayed at the top of the list.

21. Click in cell A1 and drag down to cell A3 to select rows 1:3. Right-click on the selected rows and click Copy. Click the Distribution sheet tab. On the Distribution sheet right-click cell A1 and then click Paste. Alternatively, use the keyboard shortcuts or the copy and paste buttons on the Home tab in the Clipboard group to copy and paste the first three rows of the Precipitation worksheet to the Distribution worksheet.

22. On the Precipitation worksheet review the data and pick two cities in different states with similar totals but where the other statistics are different. Look for data where the minimum and maximum are widely different. Copy the data from the first city into row four of the Distribution sheet.

23. On the Distribution sheet the Paste Options button displays at the end of row 4 after you paste the data. Click the Paste Options button list arrow to display the options as shown in Figure 9.24.

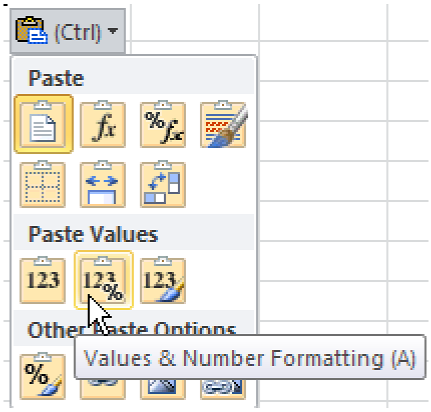


Figure 9.24. Data can be pasted with and without formatting

24. On the displayed list, click the Values and Number Formatting option. The table formatting is removed, but the one decimal number format is retained.

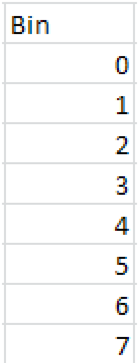
25. Repeat this procedure to copy the data for the second city into row 5. Save your work.

Creating a Frequency Distribution Table and Histogram Chart

A frequency distribution can be created to show how often something occurs in a group of data. This is another tool that is used to describe a group of data. To do this you examine the data and select upper and lower limits—the minimum and maximum—and then create a list of numbers that are at equal intervals. This creates bins. The Frequency function is used to count how many records fall into each range or bin. In this example, a bin could be .2 inches of rainfall and the program could count how many months fall into each range.

1. On the Distribution worksheet, in cell D9 type Bin. In cell E9 type =A4 and in cell F9 type =A5 to transfer the names of the cities and states from cells A4 and A5. Adjust the column width of columns A, E and F to fully display the city and state information.
2. In cell Q7 write a formula to determine the minimum value in cells Q4:Q5. In cell R7 write a formula to determine the maximum value in cells R4:R5. These numbers determine the upper and lower limit for the bins. The goal is to create at least eight bins.

3. Examine the minimum and maximum and determine an interval that you can use to create a list of eight to ten bin numbers that will create equal ranges between the lowest and highest numbers. For example if the minimum is 1 and the maximum is 8, use 1 as the interval so that the bins would be 1, 2, 3, 4 ….up to 8. Your bin should look similar to the one in Figure 9.25. It might have different numbers but the intervals must be equal.



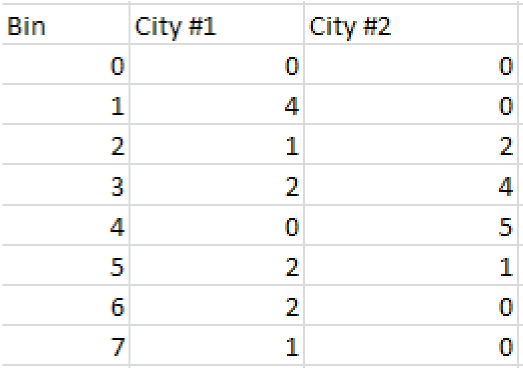
Bin
0
1
2
3
4
5
6
7

Figure 9.25. Sample Bin for the range 0.1 to 6.5 using 1 as the interval

Tip: The FREQUENCY function is a special type called an array function. It fills in a range of answers so you begin by selecting the range of output cells. To enter the function, you use CTRL+SHIFT+ENTER instead of just Enter. Do the following steps carefully. If it does not work, try again and read each step.

4. In column E, select the cells next to the bins, from E10 to E18 or E19 so that it is the same number of cells as the Bin range.
5. In E10, type =FREQUENCY(
6. Drag to select the range of cells in B4:M4, and then type a comma.
7. Drag to select the range of bin values in column D, and then type a closing parenthesis. Do not press Enter.
8. Press and hold CTRL and SHIFT, and then press Enter. The Frequency function is filled down column E. This is the distribution of the precipitation for the first city in your selected data.
9. In column F, repeat this process to enter a Frequency function for the data in cells B5:M5. Select the range in column F where the results will go, type =FREQUENCY(B5:M5,D10:D19) and press and hold

CTRL and SHIFT, and then press Enter. Your results should look similar to Figure 9.26 but with different numbers in each column depending on how the precipitation is distributed for the two cities that you selected and the bin numbers that you used. Your table should display the city and state for data you selected.



Bin	City #1	City #2
0	0	0
1	4	0
2	1	2
3	2	4
4	0	5
5	2	1
6	2	0
7	1	0

Figure 9.26. Frequency distribution of precipitation for two cities.

10. Verify that the numbers in each column add up to 12—the number of measurements in each range. If it does not total 12, if you have more than twelve, you might have included the statistics in columns N through R. If you have less than twelve, the bin ranges might not include all the possible values.
11. Select all the data in your bin table. Include the headings in row 9, the bins in column D and the frequencies in columns E and F. On the Insert tab, in the Charts group, click the Column Chart button and then select the 2-D Clustered Column Chart. The bin values will display on the chart as if it is data that should be charted, so this needs to be removed.
12. On the Chart Tools Design tab, in the Data group, click the Select Data button. In the Select Data Source dialog box, under Legend Entries, click Bin, and then click the Remove button.
13. On the Horizontal (Category) side of the dialog box, click the Edit button and then drag the bin values in column D. Click OK two times.
14. Move the chart so that it is to the right of the Bin table, with the upper left corner in cell H9.
15. In the Chart Layout group, click the Layout 1 button. Click the Chart Title and type Precipitation

Comparison and then press Enter.

16. Click the Chart Tools Layout tab. In the Labels group click the Axis titles and add a horizontal title that reads Inches of Precipitation. Add a rotated vertical title that reads Number of Months.
17. If the vertical axis does not display as whole numbers, right-click the numbers in the vertical axis and then click Format Axis. In the Format Axis dialog box, change the Major unit to Fixed and type 1.0 in the box for this option. Close the dialog box.
18. Right-click one of the columns and from the shortcut menu click Format Data Series. In the dialog box change the Gap Width to 0%. You can do this by dragging the slider to No Gap or by typing 0% in the Gap Width box. By changing the gap between the columns to zero, you have changed the column chart to a histogram.
19. Save your work.

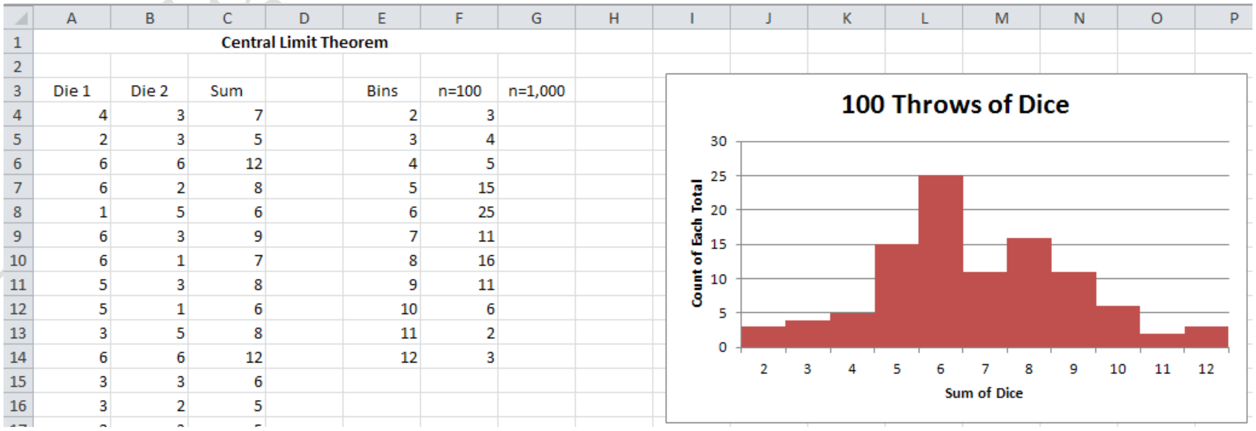
Creating Random Numbers to Demonstrate the Central Limit Theorem

Random numbers are used to simulate real situations that are effected by random events. Random numbers can be generated by the computer and then used to create models that behave similarly to real systems.

The two most commonly used functions are RAND and RANDBETWEEN. The RAND function generates fifteen digit numbers between zero and 1 such as 0.124577437776544. The RAND function does not need arguments so it is written with nothing between the parentheses, =RAND(). The RAND function is typically used in combination with other numbers and functions. For example if random numbers between zero and ten were required, the formula would be written =RAND() * 10. The RANDBETWEEN function returns integers—whole numbers— between the two numbers specified as arguments. For example to simulate the roll of a single die it would be =RANDBETWEEN(1,6). When two dice are rolled the sum of the dice will range from 2 to 12. Because the two random factors tend to offset each other, the sums will tend to be near the mean, which is seven ((2+12)/2=7). This is an example of the central limit theorem.

On the next worksheet you will practice using the RANDBETWEEN function to simulate the throwing of two dice.

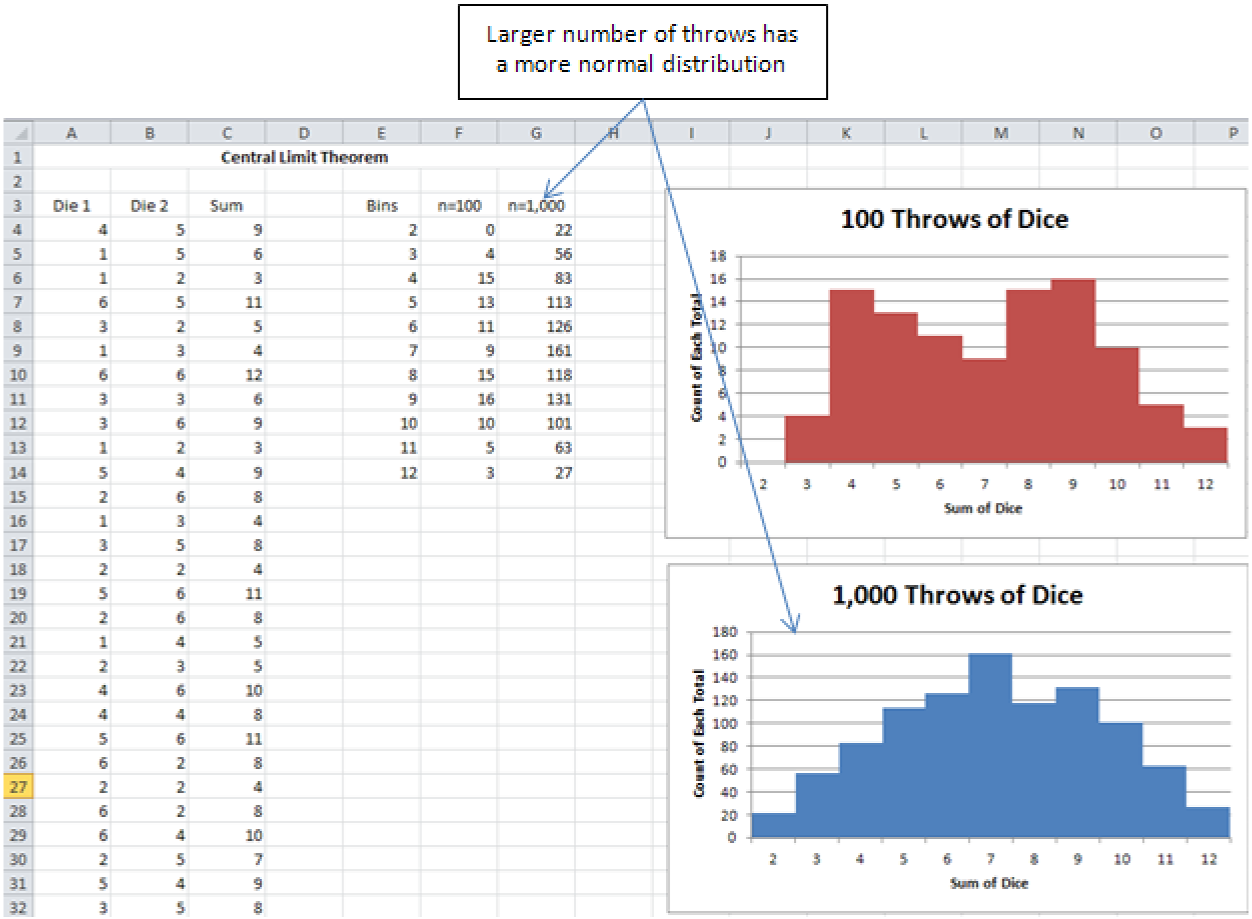
1. Click the Random sheet tab. Click cell A4, type =RANDBETWEEN(1,6) and press Enter, to simulate throwing a single dice. Repeat this process in cell B4.
2. In cell C4, use the SUM function to sum the values of the two RANDBETWEEN functions to show the sum of the two die.
3. Select cells A4:C4 and use the fill handle in cell C4 to fill the functions down 100 rows to row 104. You may notice that the numbers in each cell changes each time you add another function or fill the columns.
4. Notice that bins from 2 to 12 are entered in column E. Select cells F4:F14. Type =FREQUENCY(C4:C104,E4:E14) and then press and hold CTRL plus SHIFT and press ENTER.
5. Select cells E3:F14 and create a 2-D Clustered Column chart. Use the Select Data button to remove the Bin numbers from the Legend Entries. In the Select Data Source dialog box, under Horizontal (Category) Axis Labels, click the Edit button, and then drag to select cells E4:E14.
6. Delete the legend. Select the chart title and type 100 Throws of Dice as the chart title. Add a horizontal axis title: Sum of Dice and a vertical axis title: Count of Each Total. Format columns to reduce the gap to 0%. Move the chart to the right of the Bin table, beginning in column I. Compare your results with Figure 9.27. Your numbers and the shape of your chart will be different because of the random numbers generated by the RANDBETWEEN function.



	A	B	C	D	E	F	G
1			Central Limit Theorem				
2							
3	Die 1	Die 2	Sum		Bins	n=100	n=1,000
4	4	3	7		2	3	
5	2	3	5		3	4	
6	6	6	12		4	5	
7	6	2	8		5	15	
8	1	5	6		6	25	
9	6	3	9		7	11	
10	6	1	7		8	16	
11	5	3	8		9	11	
12	5	1	6		10	6	
13	3	5	8		11	2	
14	6	6	12		12	3	
15	3	3	6				
16	3	2	5				

Figure 9.27. Frequency distribution of 100 throws of dice.

7. Press the F9 key to generate new random numbers and observe the effect on the chart
8. Select cells A104:C104 and use the fill handle to extend the pairs to row 1004.
9. In cells G4:G14 enter the FREQUENCY function to create a second distribution analysis of the sum of the1000 numbers generated.
10. Select cells G3:G14 and create a 2-D Clustered Column chart. Click the Select Data button and on the Horizontal (Category) side of the dialog box click the edit button and then drag E4:E14 to select the range for the horizontal axis. Click OK.
11. Add a chart title—1,000 Throws of Dice—and the same horizontal axis title and vertical axis title as you did in the previous chart. Change the gap width to 5%. Move the chart so that it is below the previous chart. Compare your results with Figure 9.28.



Larger number of throws has a more normal distribution

	A	B	C	D	E	F	G
1			Central Limit Theorem				
2							
3	Die 1	Die 2	Sum		Bins	n=100	n=1,000
4	4	5	9		2	0	22
5	1	5	6		3	4	56
6	1	2	3		4	15	83
7	6	5	11		5	13	113
8	3	2	5		6	11	126
9	1	3	4		7	9	161
10	6	6	12		8	15	118
11	3	3	6		9	16	131
12	3	6	9		10	10	101
13	1	2	3		11	5	63
14	5	4	9		12	3	27
15	2	6	8				
16	1	3	4				
17	3	5	8				
18	2	2	4				
19	5	6	11				
20	2	6	8				
21	1	4	5				
22	2	3	5				
23	4	6	10				
24	4	4	8				
25	5	6	11				
26	6	2	8				
27	2	2	4				
28	6	2	8				
29	6	4	10				
30	2	5	7				
31	5	4	9				
32	3	5	8				

Figure 9.28. The Central Limit Theorem is demonstrated with the larger sample size

12. Press the F9 key to see how the distribution still changes more with the lower sample size compared to the larger sample size. Save your changes

Using Standard Deviation to Create a Grading Curve

At some schools where hundreds of students take the same classes and the same tests, instructors use standard deviations to determine the range of letter grades to assign to the scores. This is known as *grading on a curve* where the curve refers to the normal distribution. The test is made difficult enough so that no one gets 100% so all of the scores fit on the scale. A letter grade of C is assigned to scores that are within one standard deviation of the mean, which will account for approximately 67% of the students. The scores between one and two standard deviations above the mean get Bs and those between one and two standard deviations below the mean get Ds. According to the 68-95-99.7 rule, this will account for approximately 95% of the students. Students who score higher than two standard deviations get As and those below two standard deviations get Es, which accounts for the remaining 5%. On the Grade sheet you will create a grading curve based on the mean and the standard deviation of the test scores.

1. Click the Grade sheet tab. In cell D3 use the AVERAGE function to calculate the average of the scores in cells A4:A504. The results should display as 50.7
2. Click cell D4 and type =STDEVP(A4:A504) to calculate the standard deviation of the text results. Format the results to display one decimal place.

> Tip: The STDEVP function is used if the scores or measurements include all the values, which is called the population. If the data is a sample of a larger group then the STDEV function is used because the entire population is not included in the measurement. STDEV results in a slightly larger number because there is more chance for error when dealing with samples instead of all of the data.

3. Notice bin numbers have been entered in cells C8 to C28 from 0 to 100 with a 5 point interval. Select cells D8:D28 and type =FREQUENCY(A4:A504,C8:C28) and then press Ctrl+Shift+Enter.
4. Select cells D8:D28 and create a 2-D Cluster Column chart. Using the techniques you have practiced

previously, display the Select Data Source dialog box and set the horizontal axis to cells C8:C28.

5. Move the chart so that the upper left corner of the chart begins in cell F10. If necessary, use a corner sizing handle to expand the chart slightly so that the numbers along the horizontal axis are aligned horizontally. Delete the legend and add the chart title, horizontal axis title and vertical axis title as shown in Figure 9.29.

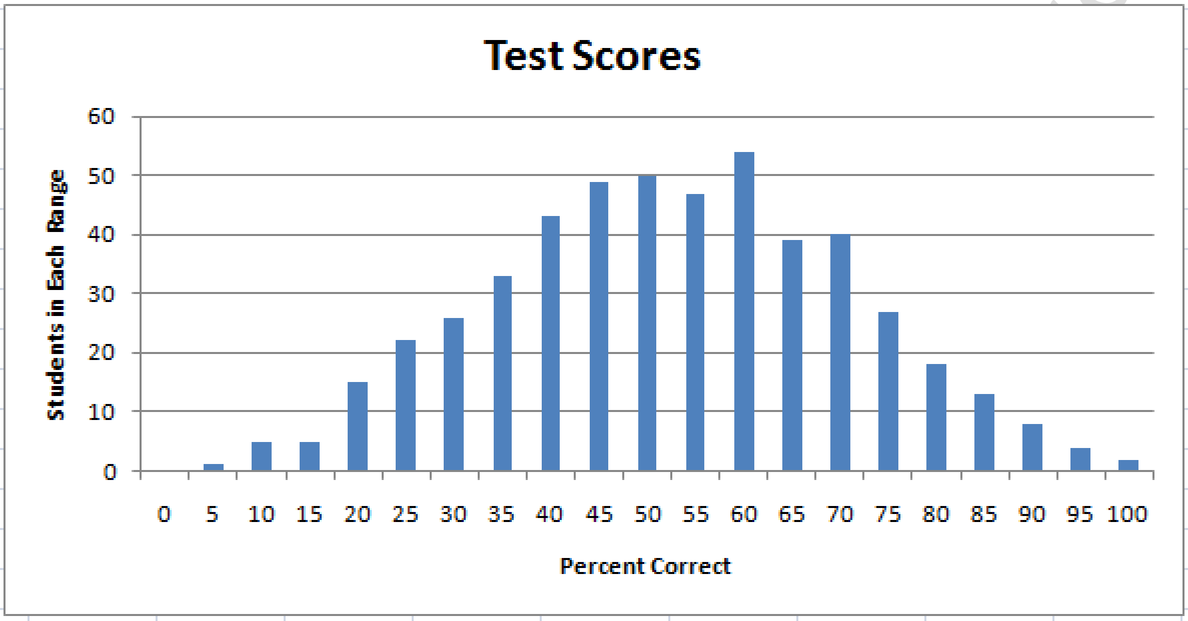


Figure 9.29. Distribution of test scores

6. On the Insert tab, in the Illustrations group click the Shapes button. Click the Text box button as shown in Figure 9.30.

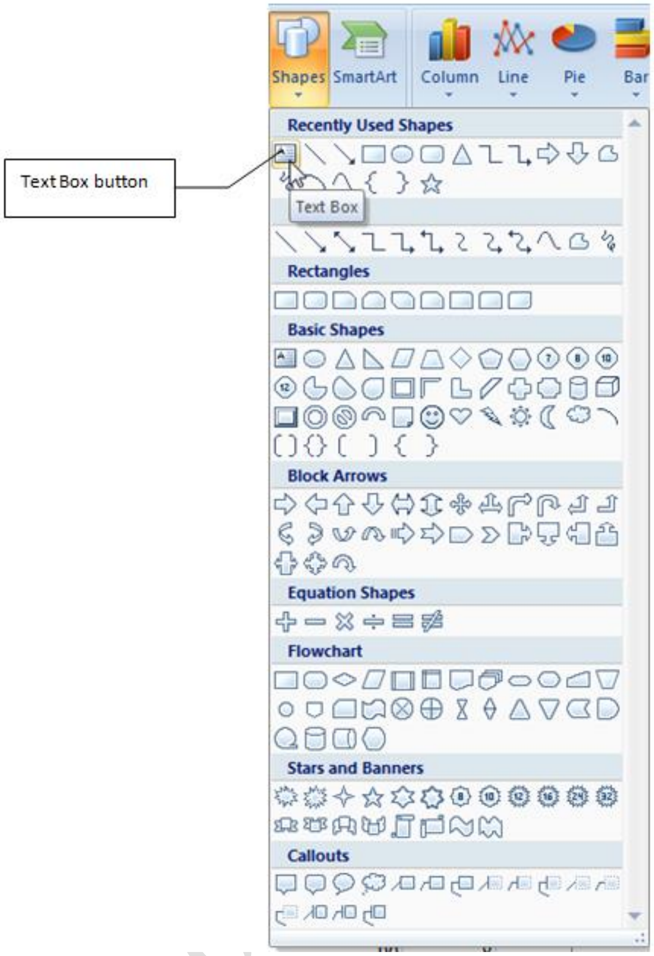
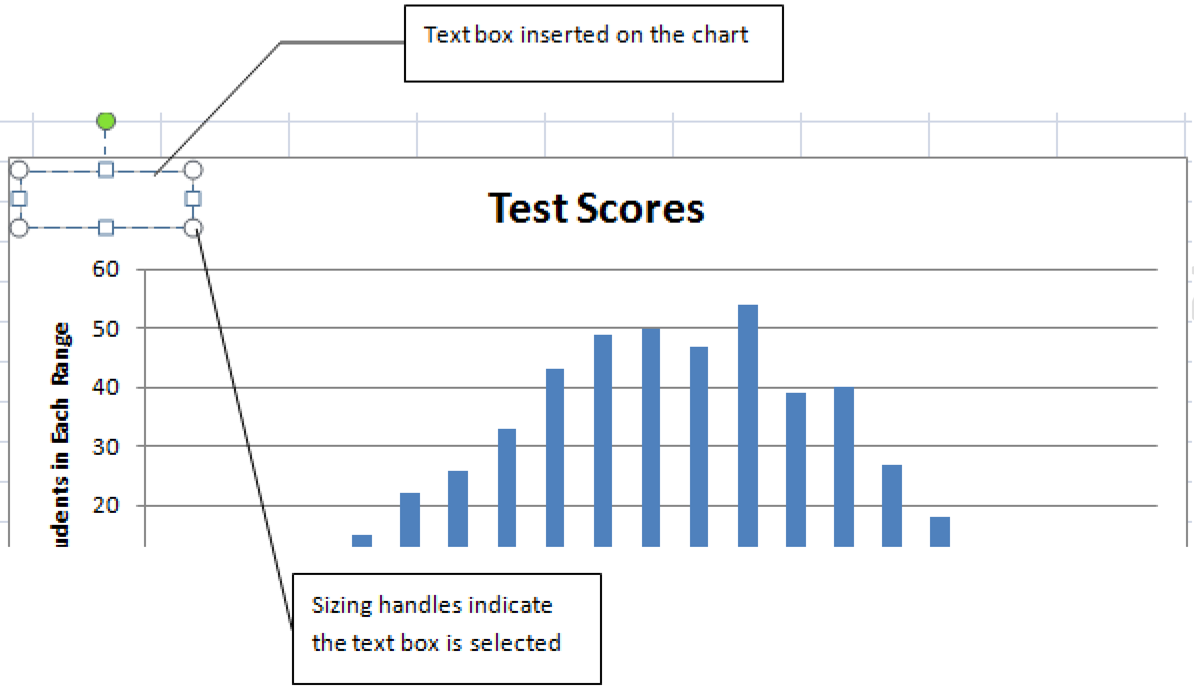


Figure 9.30. Text Box selected from the Shapes list.

7. In the upper-left corner of the chart drag a small rectangle to create the text box as shown in Figure 9.31.

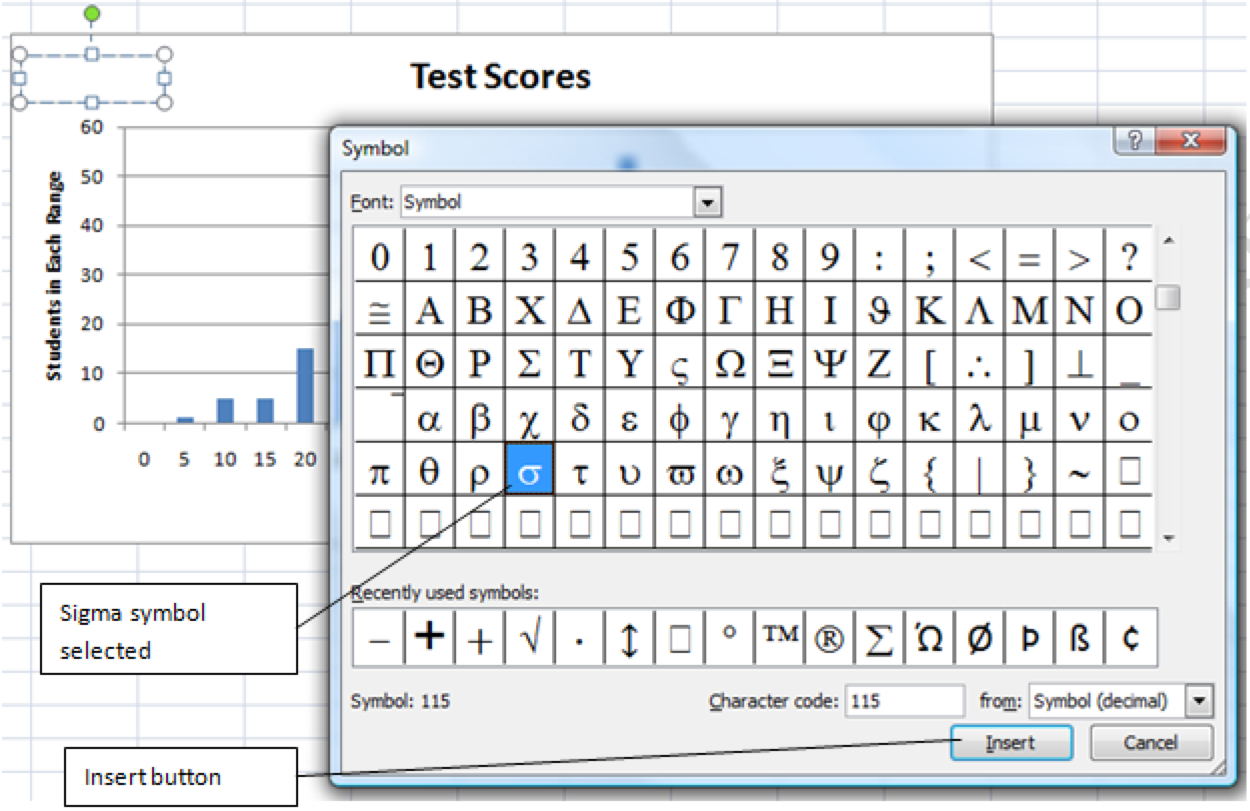


Text box inserted on the chart

Sizing handles indicate the text box is selected

Figure 9.31. Text box inserted in upper left corner of the chart

8. With the text box selected, on the Insert tab, in the Text group, click the Symbol button. If necessary, use the list arrow in the Font box to change the display to Symbol. In the Symbol box locate and click the lower case sigma symbol as shown in Figure 9.32, and then click the Insert button.



Font: Symbol

Symbol: 115

Character code: 115

from: Symbol (decimal)

Sigma symbol selected

Insert button

Figure 9.32. Sigma symbol selected to insert in the text box on the chart.

9. Close the Symbol dialog box. In the text box, press the space bar and then type =18.3.

10. In cell G4 type 100, in cell H4 type =D3+2*D4. This sets the first range for the grades, an A begins at 100 and goes to two standard deviations above the mean—2 *D4—the standard deviation, +D3—the mean.

11. In cell G5 type =H4-.1. Fill G5 down to G8. The results display as -.01 because the reference cells in column H has not been completed yet.

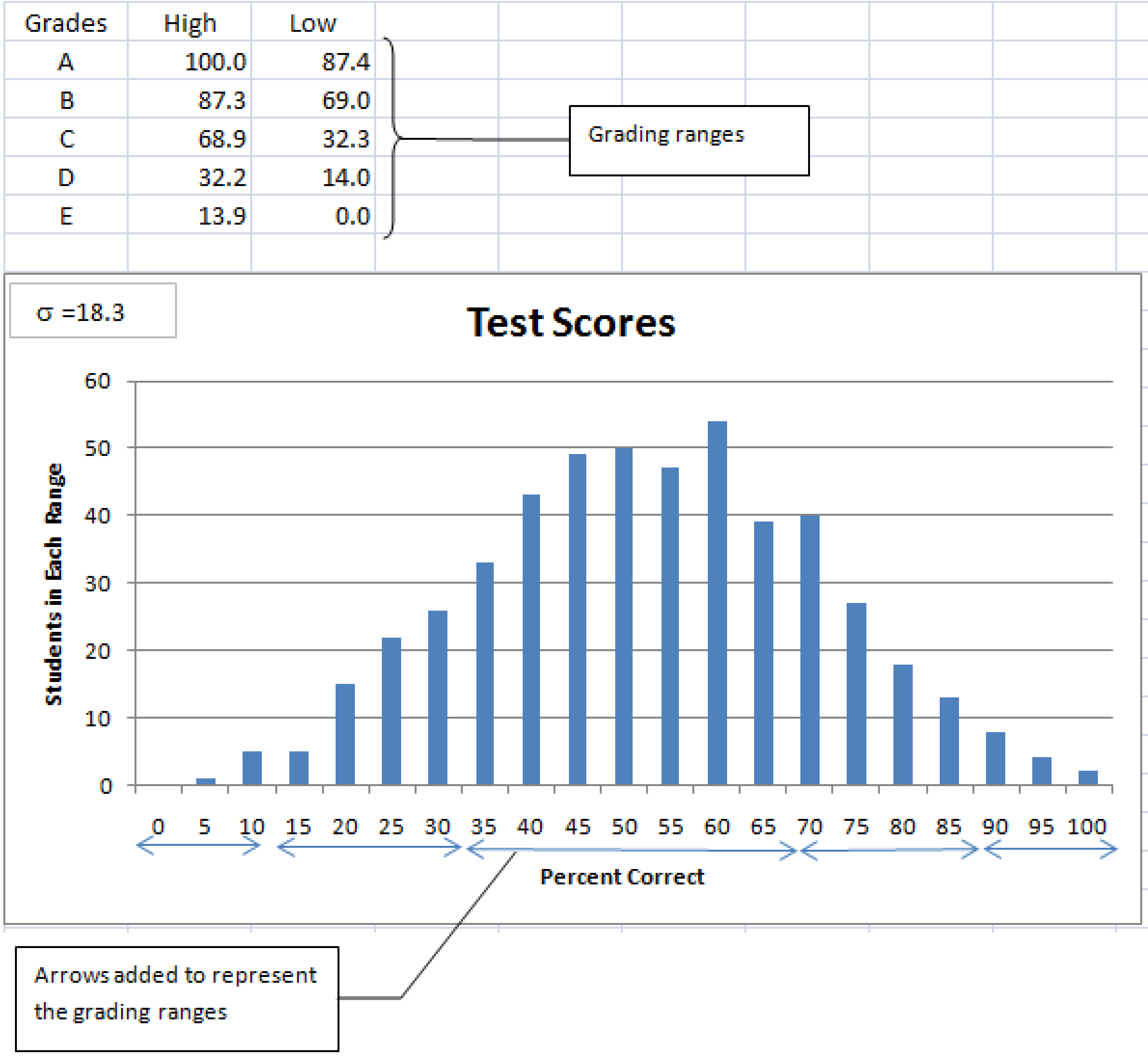
12. In cell H5 type =D3+D4 to create the range for the Bs which is from one tenth of a point below an A to one standard deviation above the mean—D3+D4.

13. In cell H6 type = D3-D4 to create the range for the Cs, which begins one-tenth of a point below a B, and goes to one standard deviation below the mean—D3-D4.

14. In cell H7 type =D3-2*D4 to create the range for the Ds, which begins one-tenth of a point below a C,

and goes to two standard deviations below the mean—D3-2*D4.

15. In cell H8 type 0. Select cells G4:H8 and format the cells to one decimal.
16. On the Insert tab, in the Illustrations group, click the Shapes button and then click the Double Arrow. Drag to draw an arrow on the chart under the numbers between 0 and approximately 13.9 to represent the lowest grade range. After you have inserted a shape the Drawing Tools Format tab is active and an abbreviated Insert Shapes list displays at the left end of this tab. The most recent shapes display.
17. Click the Double Arrow again and draw it on the chart below the numbers for the next grade range from 14 to 32. Continue this process to create arrows for each of the grading ranges. Compare your results with Figure 9.33.



Grades	High	Low
A	100.0	87.4
B	87.3	69.0
C	68.9	32.3
D	32.2	14.0
E	13.9	0.0

Grading ranges

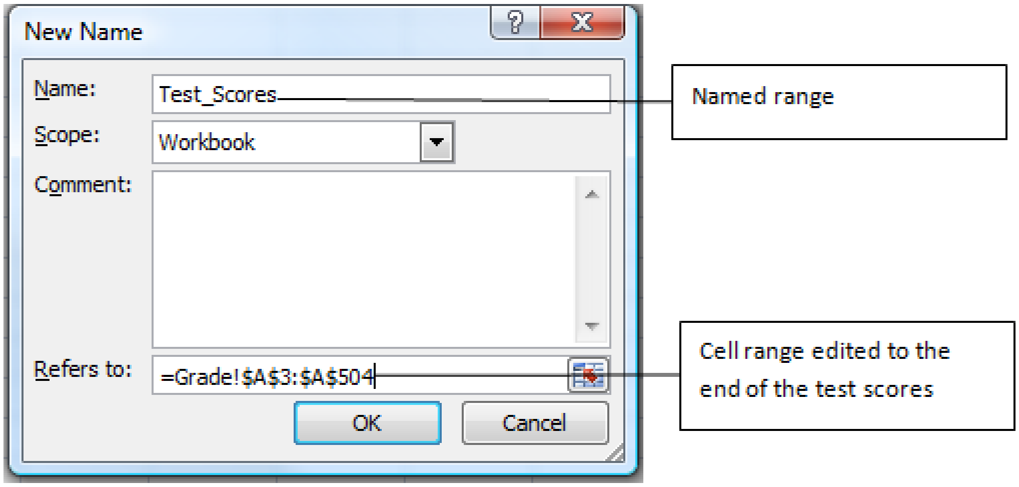
σ =18.3

Arrows added to represent the grading ranges

Figure 9.33. Arrows added to represent grading ranges.

18. In cell B3, type Rank and press enter. The rank is a sequential number that shows the sequence in which the data could be arranged by a certain criterion. In this case, the rank is determined by the test score.

19. Click cell A3. On the Formulas tab, in the Defined Names group, click the Define Name button. In the New Name dialog box, confirm that Test_Scores displays in the Name box. In the Refers to box edit the range to add the last cell in the list. Click at the end of the cell reference and type :A504. Compare your results to Figure 9.34, and then click OK.



New Name

Name: Test_Scores

Scope: Workbook

Comment:

Refers to: =Grade!$A$3:$A$504

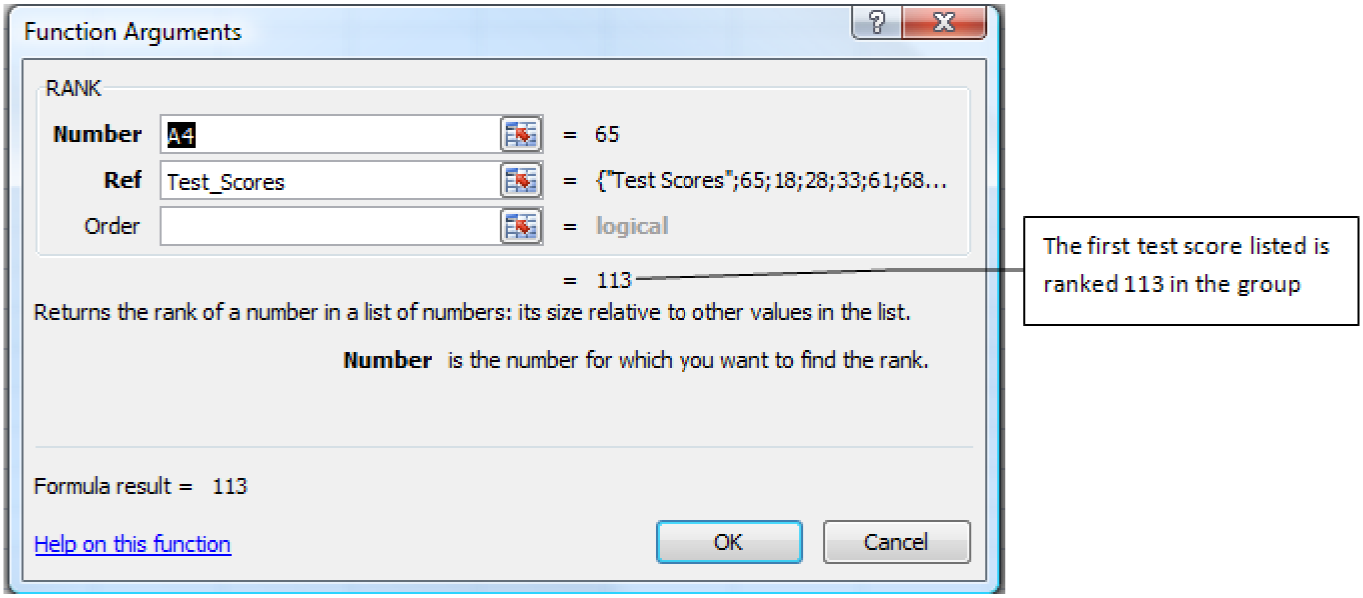
Named range

Cell range edited to the end of the test scores

Figure 9.34. Range of test scores is named.

20. Click in cell B4. On the Formulas tab, in the Function Library group, click the More Functions button, and then click Statistical. There are numerous statistical functions that can be used in Excel as shown by this list. The functions are arranged alphabetically.

21. Scroll the list until you locate Rank.EQ. and then click it. In the Function Arguments dialog box, in the Number box type A4—this is the number for which you want to find the ranking in the list of text scores. In the Ref box type Test_Scores. Compare your results with Figure 9.35.



Function Arguments

RANK

Number A4 = 65

Ref Test_Scores = {"Test Scores";65;18;28;33;61;68…

Order = logical

= 113

Returns the rank of a number in a list of numbers: its size relative to other values in the list.

Number is the number for which you want to find the rank.

Formula result = 113

The first test score listed is ranked 113 in the group

Figure 9.35. Rank function can be used to rank scores or measurements.

22. Click OK. In cell B4, point to the fill handle and then double-click the fill handle to fill the rank function down column B.

Tip: Rank.EQ is a function that returns the rank of a number in a list of numbers. If two numbers are the same, both are given the same rank, and the next rank is skipped. The next number in the list is given the next ranking after the one that was skipped. RANK.AVG is similar except when more than one item has the same rank, they are all given the average rank instead of the top rank.

23. Click in the Name box and type A3:B504 and then press enter. This selects the test scores and the ranking columns and is an easy way to select a large range of data that may be off your screen.

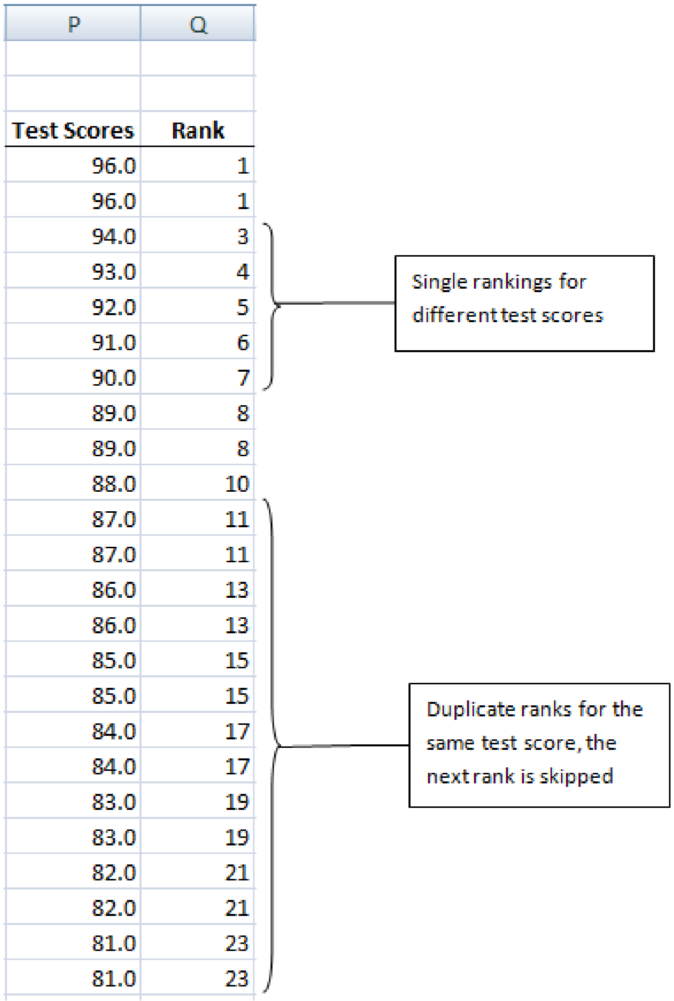
24. Right-click on the selected range and from the shortcut menu click Copy. Right-click cell P3 and click Paste to copy the test scores and rankings to columns P and Q.

25. Click in a cell in the data in column Q. On the Data tab, in the Sort & Filter group, click the Sort button.

26. In the Sort dialog box, confirm that the *My data has header*s check box is selected. In the Sort by box use the list arrow if necessary and select Rank. In the Order column, use the list arrow to select Smallest to Largest. The sort dialog box can be used to sort data on more than one field at a time.

27. Click OK. The data is resorted in rank order. Notice that when the test scores are the same, the rank is duplicated and the next rank is skipped. This is characteristic of the RANK.EQ function.

28. Select cells P3:Q3. Right-click and on the Mini toolbar click the Bold button. Click the Border button and the click Bottom Border. Apply the same formatting styles to cells A3:B3. Compare your results with Figure 9.36.



P	Q
Test Scores	Rank
96.0	1
96.0	1
94.0	3
93.0	4
92.0	5
91.0	6
90.0	7
89.0	8
89.0	8
88.0	10
87.0	11
87.0	11
86.0	13
86.0	13
85.0	15
85.0	15
84.0	17
84.0	17
83.0	19
83.0	19
82.0	21
82.0	21
81.0	23
81.0	23

Single rankings for different test scores

Duplicate ranks for the same test score, the next rank is skipped

Figure 9.36. Test scores sorted in rank order

29. Save your work.

Key Takeaways

- If you add a column to an Excel table and then enter a function like SUM, AVERAGE, MIN, or MAX, the function is filled into the rest of the column. [9.4.1]
- The FREQUENCY function is an array function that counts the number of measurements that are within each interval called a bin. The range of bins should include the minimum and maximum values and be at equal intervals. To use the FREQUENCY function, select a range of cells the same size as the range of bins and type the function and its arguments. The arguments are the range of data, and the range of bins. To complete the function, press and hold both the CTRL and SHIFT keys, and then press Enter. This creates a

frequency distribution that can be charted as a column chart, using the bins as the horizontal axis values. Use the Format Data Series dialog box to remove the gaps between the columns in the chart to create a histogram of the data.[9.4.2]

- Use the RANDBETWEEN function to create a list of random whole numbers. The arguments are the smallest number and the largest number that you want to include in the group. Two frequency distribution charts based on a small sample and a large sample demonstrate the central limit theorem by showing that the larger sample has a distribution that is closer to the theoretical normal distribution. A new set of random numbers are regenerated by pressing the F9 key, which alters the chart to difference in variability between small and large samples.[9.4.3]

- Use the standard deviation (STDEV) when using a sample of a larger group, otherwise use the standard deviation of the population (STDEVP) when the entire group is included. The argument of either function is the group of numbers or measurements in the sample or population. To create a grading scale that resembles a normal distribution, ranges that correspond to letter grades can be chosen based on the standard deviation where C grades are assigned to scores within one standard deviation on either side of the mean. Ds and Bs are in the range between one and two standard deviations, and the Es and As are scores beyond two standard deviations in either direction. This is an application of the 68-95-99.7 rule whereby approximately 95% of the grades will fall between B and D. The RANK function is used to find the placement or rank order of a measurement within a group. If more than one measurement is the same, the same rank is applied to the identical numbers and the next rank is skipped. The arguments of the rank function are the number or cell reference that you want to rank, and the range of cells within which the number is found.[9.4.4]

Key Terms

68-95-99.7 Rule

In a normal distribution, 68% of the measurements will be within one standard deviation on either side of the mean, 95% will be within two, and 99.7% will be within three.

Alternative hypothesis

A statement of a study’s expected result, stated as a positive.

Assignable cause

Reason for difference from the mean that is not random.

Bin

Interval used with frequency distributions.

Central limit theorem

Multiple random factors tend to offset each other resulting in most measurements being near the mean.

Control group

Group that gets the placebo or the treatment with a known effect.

Control limits

Values above and below the desired value that indicate the acceptable difference.

Correlation

Related to each other so they vary in similar ways.

Double-blind

Neither the participants nor the administrators know who is getting the treatment and who is getting the placebo.

Experimental studies

Type of study where there are two groups and a treatment is administered to one of the groups to see if it makes a significant difference.

Frequency

Count of the number of members of a group that fall within each interval or bin.

Frequency distribution

The pattern of counts of each numbers in each interval.

Grade

Set of requirements for a product.

Histogram

Column chart with no space between columns that indicates frequency distribution.

Inferential statistics

Describe the characteristics of samples. These characteristics are assumed to be the characteristics of the unmeasured members of the group.

n

Number of items in the sample.

Normal

Frequency distribution caused by competing random factors that form a bell-shaped curve.

Null hypothesis

A statement of a study's expected result stated as a negative.

Observational study

Study that uses existing data instead of actively treating a group.

Placebo

Non-active substitute.

Population

All of the items or people in a group.

Quality

The degree to which the product matches the requirements for its grade.

Rank

Sequential number that shows the sequence in which the data could be arranged by a certain criterion.

Run chart

Data chart that displays sample values during the manufacturing process.

Sample

Members of a group that represent the group.

Significant difference

Greater difference than would be due to random effects in the sample 95% of the time.

Six-Sigma

Quality program that has a goal of making products within six standard deviations of the mean within tolerance.

Skewed

Frequency distribution that is not symmetric around the mean.

Standard deviation (STD)

Statistic that describes the spread of a normal frequency distribution.

Tolerance

Range between the mean and the high and low acceptable limits.

Bibliography

International Organization for Standardization. *Quality Management Systems--Fundamentals and Vocabulary.* Geneva: ISO Press, 2000.

Microsoft Corporation. *Introducing array formulas in Excel.* 2010. http://office.microsoft.com/en-au/excel-help/introducing-array-formulas-in-excel-HA001087290.aspx (accessed November 7, 2010).

Last edited on January 30, 2011