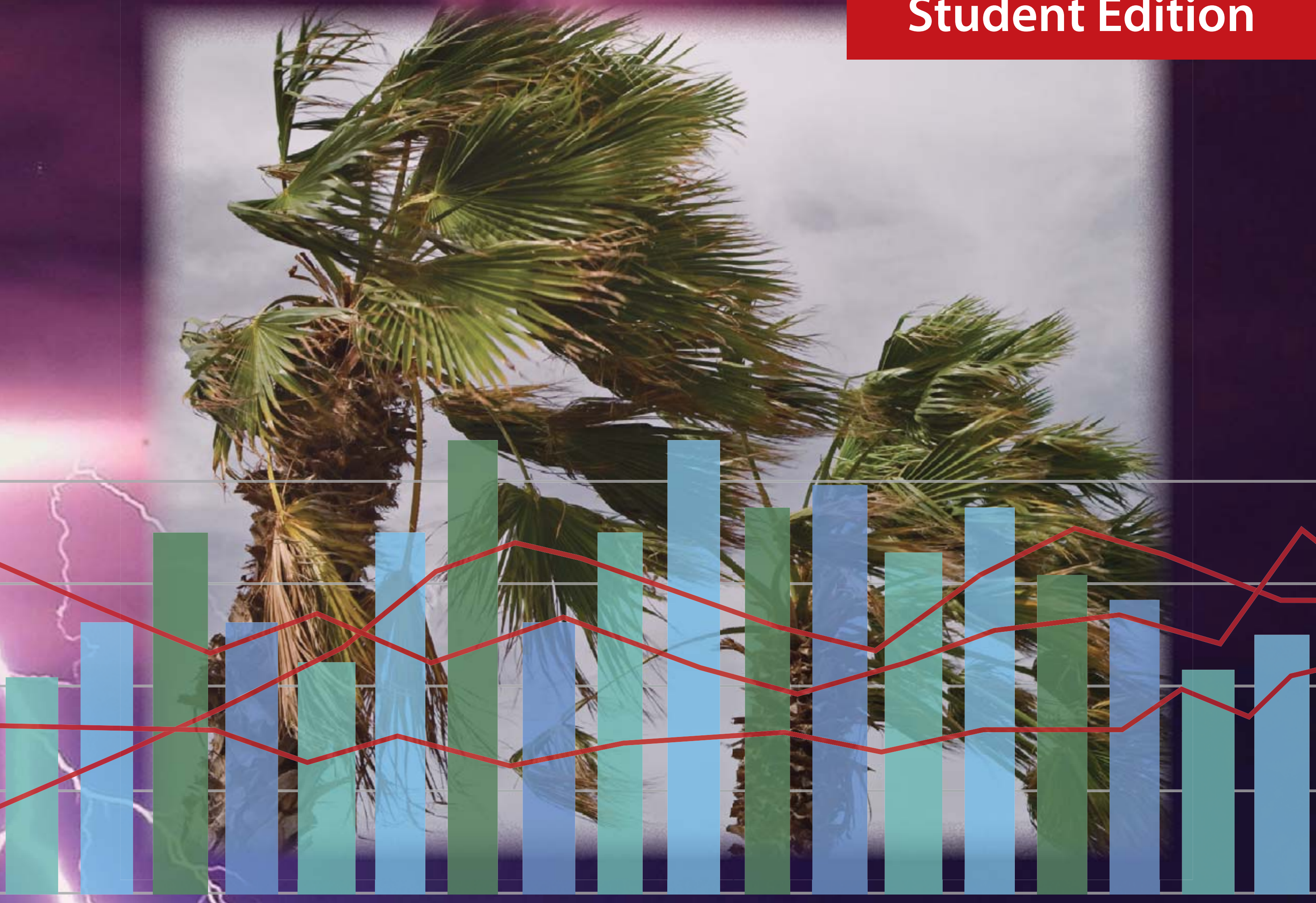
Probability & Statistics

Modular Learning Exercises

Student Edition



A Curriculum Resource for Accelerated Math Students

Think like an Actuary! Produce real world data from storm statistics.

THE ACTUARIAL FOUNDATION®

Preparing for tomorrow's possibilities® *through education*

About The Actuarial Foundation

The Actuarial Foundation, a 501(c)(3) nonprofit organization, develops, funds and executes education, scholarship and research programs that serve the public and the profession by harnessing the talents and resources of actuaries. Through an array of educational initiatives, the Foundation is *dedicated* to minimizing risk through maximizing education by providing the tools to help improve math and financial knowledge for all. We are proud to add *Probability & Statistics Modular Learning Exercises*, a curriculum resource for your accelerated math students to our library of math resources. Please visit the Foundation's web site at: www.actuarialfoundation.org for additional educational materials.

What is an Actuary? According to BeAnActuary.org, Actuaries are the leading professionals in finding ways to manage risk. It takes a combination of strong math and analytical skills, business knowledge and understanding of human behavior to design and manage programs that control risk. Careercast.com reported that "Actuary" is ranked as the 2nd best job out of 200 of the best and worst jobs. To learn more about the profession, go to: www.BeAnActuary.org.

Some of the activities in this book reference specific Web pages. While active at the time of publication, it is possible that some of these Online Resource links may be renamed or removed by their hosts. Note that these links were provided simply as a convenience; a quick search should reveal some of the many other online resources that can be used to complete these activities. Facts and opinions contained are the sole responsibility of the organizations expressing them and should not be attributed to The Actuarial Foundation and/or its sponsor(s).

Student Introduction - Background Information

Your class has been asked to help actuaries at an insurance company to assess the risk and potential loss due to hurricanes in the coastal town of Happy Shores. Happy Shores is a small oceanfront community of approximately 200 households. Some homes are located on the beach while others are further away, but all are within a few miles of the ocean. The actuaries want your help to determine the characteristics of the community and what losses might occur due to hurricanes.

Hurricanes

Hurricane Katrina devastated New Orleans, Louisiana, in the fall of 2005; it was one of the costliest natural disasters in the history of the United States. More than 1,800 people died in the actual hurricane and subsequent floods, and total property damage has been estimated at $81 billion by the National Hurricane Center.

Hurricanes are fierce storms with winds in excess of 72 miles per hour that form in the Atlantic Ocean. Every year the world experiences hurricane season when hundreds of storm systems spiral out from the tropical regions surrounding the equator. Almost half of these storms reach hurricane strength. In the Northern Hemisphere, hurricane season runs from June 1 to November 30.

Hurricanes can unleash incredible damage when they hit. With enough advance warning, however, cities and coastal areas can give residents the time they need to fortify the area and even evacuate. To better classify each hurricane and prepare those who would benefit by knowing the expected intensity of the storm, meteorologists rely on rating systems.

The Saffir-Simpson Hurricane Scale classifies hurricanes based on wind speed:

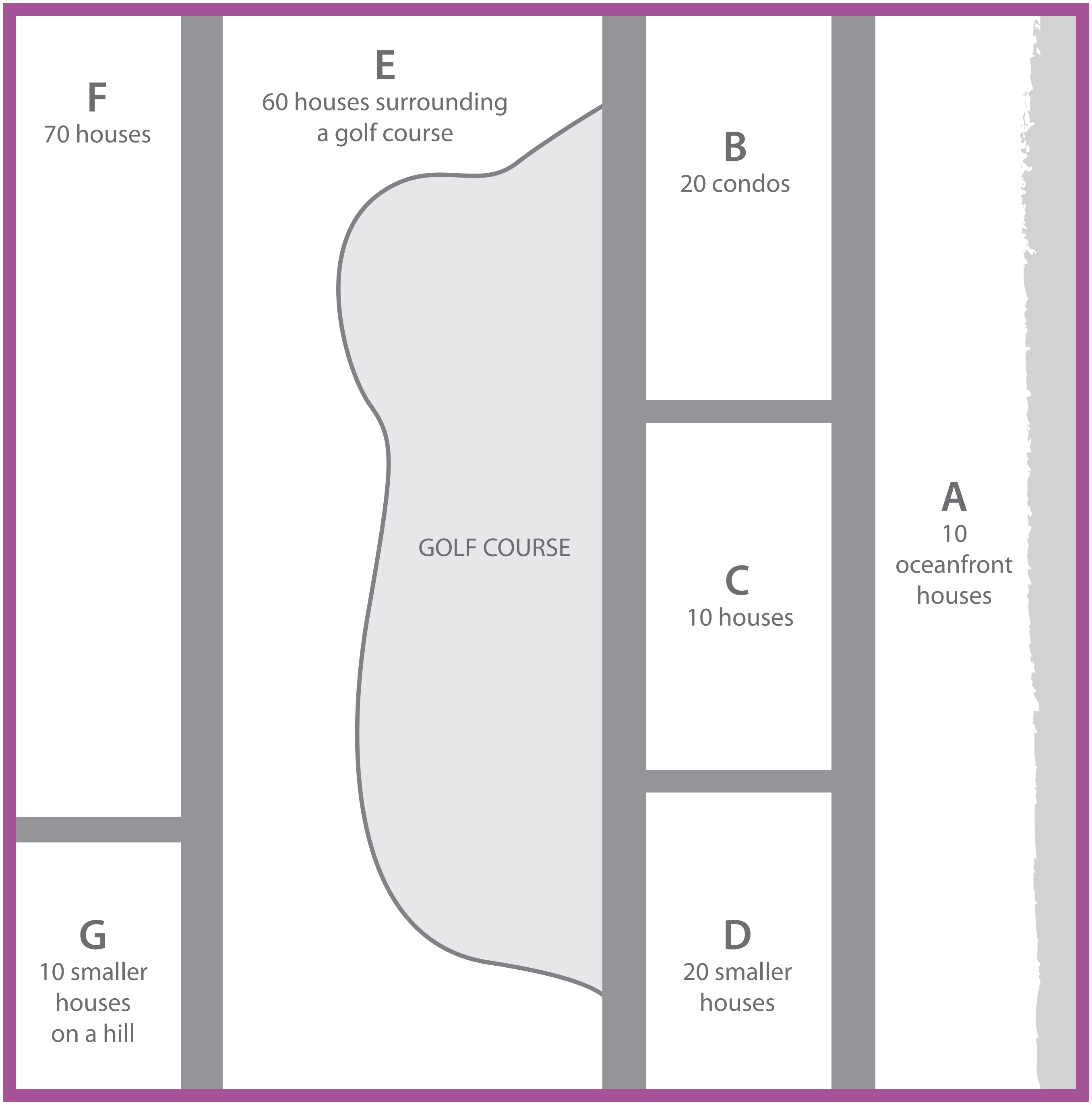
Category	Wind Speed (mph)	Storm Surge (feet)	Damages
1	74-95	4-5	Some flooding, little or no structural damage
2	96-110	6-8	Coastal roads flooded, trees down, roof damage (shingles blown off)
3	111-130	9-12	Severe flooding, structural damage in houses and mobile homes destroyed
4	131-155	13-18	Severe flooding inland, some roofs ripped off, major structural damage
5	> 155	> 18	Severe flooding farther inland, serious damage to most wooden structures

The extent of hurricane damage doesn't just depend on the strength of the storm, but also the way it makes contact with the land. In many cases, the storm merely grazes the coastline, sparing the shores its full power. Hurricane damage also greatly depends on whether the left or right side of a hurricane strikes a given area. The right side of a hurricane packs more punch because the wind speed and the hurricane's speed of motion complement one another there. On the left side, the hurricane's speed of motion subtracts from the wind speed.

This combination of winds, rain, and flooding can level a coastal town and cause significant damage to cities far from the coast. In 1996, Hurricane Fran swept 150 miles (241 km) inland to hit Raleigh, N.C. Tens of thousands of homes were damaged or destroyed, millions of trees fell, power was out for weeks in some areas, and the total damage was measured in the billions of dollars.

The Town of Happy Shores

Below is a map of the fictional town of Happy Shores. As you can see, the community is on the ocean.



The town has seven “neighborhoods” (labeled A-G). In each neighborhood, the homes are of similar value. For instance, the ten oceanfront homes are all worth around half a million dollars. In the modules, you will be analyzing data concerning Happy Shores.

Happy Shores has been affected by hurricanes before. Four years ago, a category 3 hurricane hit the town and caused extensive damage, especially to homes nearest the ocean. Happy Shores has been hit with a category 5 hurricane only once in the last 100 years; this was about 30 years ago. It caused widespread, severe damage. Over the years, other smaller storms have also affected the community. You will be analyzing data about damages caused by these storms in the modules.

Actuaries and Insurance

The company that wants help from you sells insurance. Insurance is a way to manage risk. As you go through life, there is always a chance that you will be in a car accident, you will get sick, or that your home will burn down or be damaged by a storm (such as a hurricane). The risk of these accidents is small, but if one of them were to happen, the results may be catastrophic. Without insurance, you would have to come up with the money on your own to repair your car, have needed surgery, or rebuild your home.

At insurance companies, actuaries build mathematical models to quantify risk, like the risk that your home will be damaged in a hurricane. They then help to determine how much the insurance company should charge for insurance to cover the likelihood and amount of possible claims. As Dan Tevet wrote in the March 2011 issue of Future Fellows (newsletter of the Casualty Actuary Society), “Actuaries use a combination of insurance knowledge, math, and historical data to predict future insurance events. It is sort of like a combination of being a math whiz and weather forecaster.”

There’s no way of knowing exactly who will be affected by events like hurricanes or when it might happen. With some data, actuaries can estimate how often these setbacks occur, who they are most likely to affect, and how much recovering from them will cost. For the town of Happy Shores, this means actuaries may be able to predict how often a catastrophic hurricane (or even a small hurricane) will occur, how much damage is likely, and which houses are most likely to be affected. Using this information, an insurance company can most equitably spread the risk among all its customers.

More on Insurance

Suppose an insurance company sells insurance to 100 different coastal communities, including Happy Shores. All these communities are approximately the same size. Every year during the past 25 years, one community has been hit with a hurricane or tropical storm that has resulted in $1 million in damage. Without insurance, the inhabitants of every community would have to save a million dollars to cope with the odds that their community would be the one damaged by a hurricane. At the end of the year, 99 communities would have to pay nothing, but one would have to pay the million dollars (and potentially be financially devastated).

With insurance, each community can join together to spread out the risk. If they create an insurance fund, all 100 communities will pay $10,000 at the start of the year (with the burden being shared by its residents). This $1 million total will then go to the community that is damaged by the hurricane.

Is it really fair to have each community pay the same amount into the insurance fund? Some communities may be more at risk because of location or elevation; some houses may be at more risk because of their location, size, or construction materials. With enough information, the insurance company can charge each community and household within a community a different rate depending on how likely it will be hit by a hurricane and other risk factors.

Insurance is typically good for the customer, in that it allows households and communities to spread risk. A formal request by a household to an insurance company asking for a payment is called an insurance claim. Insurance companies know that they will occasionally have to pay out claims. They'll also have to pay certain expenses, such as employees' salaries. And ultimately, insurance companies need to make a profit. So, they need to figure out what to charge each customer so that they can pay out claims, cover their expenses, and still make some money.

Your Role

The actuaries at an insurance company have asked your class to assess the risks involved with insuring homes in the Happy Shores Community where hurricanes may occur. They need to evaluate the potential damage of a hurricane and how it will be distributed amongst the 200 households in the town. They have some historical data about hurricanes hitting the town; however, since hurricanes causing major damage are (fortunately) somewhat infrequent, they also have to use historical data about hurricanes and damages sustained in similar towns in the coastal U.S. Ultimately, they need to decide how much to charge for insurance to each household, so the company can pay out claims when they occur and still make a profit. If they charge too little, they may not be able to pay out claims when they occur, and they will ultimately lose money. If they charge too much, then they may lose customers to less expensive companies.

The Math

In order to best understand what the data are telling us, we need to understand probability and statistics. Statistics is the mathematics of the collection, analysis, and interpretation of quantitative data in order to make better decisions, assess risk, and better understand the world. Probability is the mathematics of uncertainty and chance. These modules will introduce you to the world of probability and statistics. If you have studied calculus, you will find probability and statistics to be much different. Calculus is essentially the study of change focusing on limits, derivatives, integrals and infinite series. Statistics is much more focused on the interpretation of real data. It is very different than calculus, though some concepts in statistics ultimately depend on calculus and some concepts in statistics are applications of calculus.

In these modules you will learn about probability and statistics and apply them to better understand the hurricane risk and possible losses due to hurricanes in Happy Shores.

What you will learn

- In Module 1 you will learn about basic statistical concepts. You will learn how we can summarize data graphically and numerically with measures such as mean and standard deviation. Using these concepts, you will analyze the history of hurricanes by looking at how many storms occur each year. You will also gain an understanding of how a community like Happy Shores is damaged by hurricanes based on the category of storm that hits.
- In Module 2 you will learn about a specific model for distributions of data called the normal model. You will use this model to estimate probabilities of the insurance company receiving different value claims when a certain category of hurricane hits the town.
- In Module 3 you will learn about discrete probability distributions (model) and how to compute their expected values and standard deviations. You will use this to estimate potential claims that the insurance company would have to pay out based on the characteristics of the neighborhoods in Happy Shores and the category of the storm.
- In Module 4 you will learn how to create models for the relationships between two quantitative variables. You will use techniques such as correlation and linear regression. You will use these models to analyze the relationship between insurance claim amounts from hurricanes based on proximity to the ocean. You will also look at recent history of hurricanes and storms in the US to see if there are any trends.

Module 1: Basic Statistics Concepts

MODULE 1

In order to help the Actuaries, we will first look into the history of hurricanes in the U.S.

We will begin by looking at the distribution of the number of hurricanes and the number of tropical storms in the U.S. since 1932. A tropical storm is a storm whose sustained winds are at least 39 miles per hour, and a hurricane is a storm whose sustained winds are at least 74 miles per hour.

A **distribution** of a variable tells us what values a variable takes and how often it takes these values.

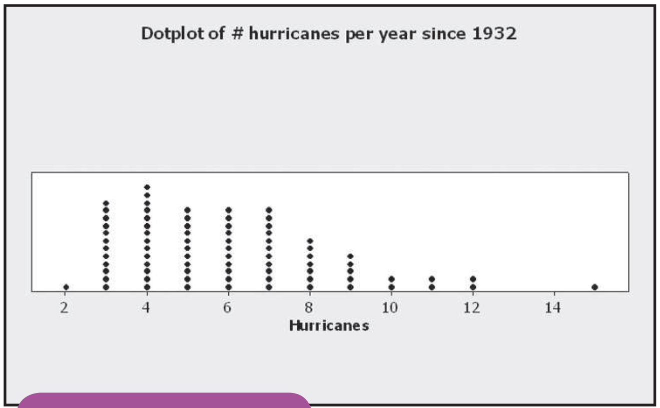
Instead of looking at a long list of numbers, making graphs summarizing data is often useful.

Two very common graphs for looking at the distribution of one quantitative variable are **dot plots** and **histograms**. In this module, we will focus on the interpretation of these graphs rather than the details of how to create them. We typically let computer software packages or calculators create these graphs.

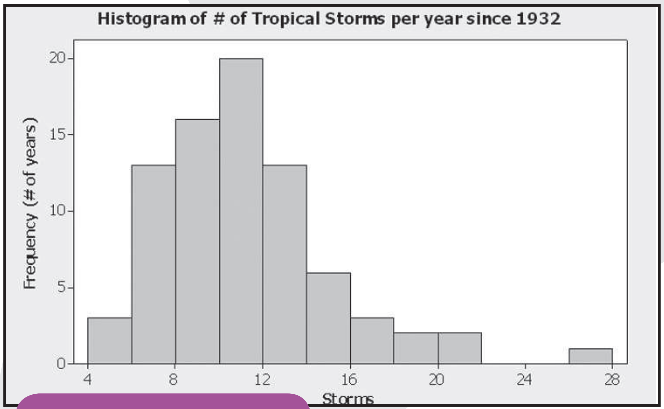
Examples of a dot plot (featuring number of hurricanes) and a histogram (featuring number of tropical storms) are shown below:

The dot plot plots every data value (in this case, the number of hurricanes in a particular year) as a dot above its value on the number line. A histogram is very similar except it may group values of variables together; for instance, the leftmost bar in the histogram below contains the number of years that had 4 or 5 tropical storms and the next bar contains the number of years that had 6 or 7 storms.

Making a graph is not an end in itself. The purpose of the graph is to help understand the data. After you make a graph, always ask, "What do I see?"



DOTPLOT



HISTOGRAM

Guidelines for Examining the Distribution of a Quantitative Variable

In any graph, look for the overall pattern and for any striking departures from that pattern. You can describe the pattern of a distribution by looking at:

- **Shape** – Is the distribution symmetrical or skewed? If it is skewed, is it skewed because most values are small and there are very few values that are high (we call this skewed right) or because most values are large and there a few that are very small (we call this skewed left). Is the data unimodal (around one "peak" to the graph) or bimodal?
- **Center** – What is the approximate value of the median (the value which divides the data in half)?
- **Spread** – The spread tells us how much variability there is in the data. One way to measure this is the range which is the largest value minus the smallest value.
- **Outliers** – Are there any values which deviate greatly from the overall pattern?

Discussion Questions

Q1: What do the histogram and dot plots tell us about the distribution of hurricanes and tropical storms since 1932?

Q2: What do the histogram and dot plot NOT show that might be important?

Numerical Measures Describing a Distribution

Two very common measures of center are **median** and **mean**.

The median (M) is the midpoint of a distribution, the number where half the observations are smaller and the other half are larger.

The mean ($\bar{x}$) is the numerical average of a distribution. It is given by the formula: $\bar{x} = \frac{\sum x_i}{n}$ where n is the number of observations.

A common measure of the spread of a distribution is the **standard deviation (s)**, which measures spread by looking at how far the observations are from the mean. Standard deviation is given by the following formula:

$$s = \sqrt{\frac{\sum (x_i - \bar{x})^2}{n-1}}$$

The summary data for the number of hurricanes and number of tropical storms is shown below:

Descriptive Statistics: Storms, Hurricanes

Variable	Total Count	Mean	StDev	Minimum	Median	Maximum
Storms	79	10.747	4.081	4.000	10.000	28.000
Hurricanes	79	6.000	2.557	2.000	6.000	15.000

Discussion Question

Q: What if there were one year that had 30 hurricanes? How would this affect the median, mean, and standard deviation of the data?

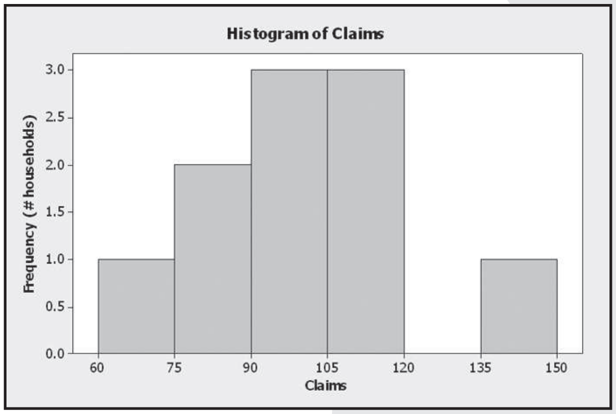
Data were obtained from http://www.wunderground.com/hurricane/hurrarchive.asp.

Let's Review...

Practice Exercise

The last time a major hurricane hit Happy Shores was 4 years ago when a Category 3 hurricane occurred. A Category 3 hurricane has sustained winds of 96-110 miles per hour. Following are insurance claims (in thousands of dollars) made by the ten households that are closest to the beach. We also have data concerning all of the 200 households in the area, but in order for you to gain an understanding of the main concepts in this module, we will focus on a small data set.

Claims ($000)
112
92
99
90
117
79
141
66
86
106



1. Describe the distribution.

2. Compute the median, mean, and standard deviation. The table below may help in computing the standard deviation:

Claims	Claim - Mean	(Claim – Mean)^2
112		
92		
99		
90		
117		
79		
141		
66		
86		
106		
Sum	xxxxxxxxxxxxxxxx	

- Compute the mean
- Compute the claim amount minus the mean for each row in the table
- Square these quantities
- Add up the squared quantities (column 3)
- Divide by (n – 1)
- Find the square root

The result is the standard deviation.

3. What percentage of claims in the data set is within the following:

   a. 1 standard deviation of the mean (that is, from the mean minus the standard deviation to the mean plus the standard deviation)?

   b. 2 standard deviations of the mean?

   c. 3 standard deviations of the mean?

Now consider claims from the ten homes that were farther from the beach:

Descriptive Statistics: Claims

Variable	Total Count	Mean	StDev	Minimum	Median	Maximum
Claims	10	1.100	1.729	0.000000000	0.000000000	5.000

Claims
5
3
2
1
0
0
0
0
0
0

4. Create a dot plot showing these claims.

5. Describe the distribution.

6. What percentage of claims in the data set is within the following:

   a. 1 standard deviation of the mean?

   b. 2 standard deviations of the mean?

   c. 3 standard deviations of the mean?

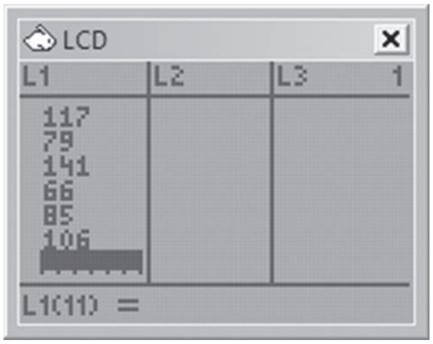
7. What are the main differences between the distributions of claims from homes farther from the beach and the one of claims from homes right on the beach?

Technology Connections

How to use TI-83/84 Calculator for Statistics

You can use your TI graphing calculator to enter data, create histograms, and compute summary statistics.

- Press STAT – EDIT
- Enter your data into a list:

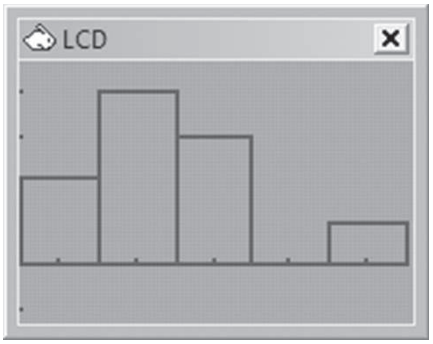


To create a histogram:

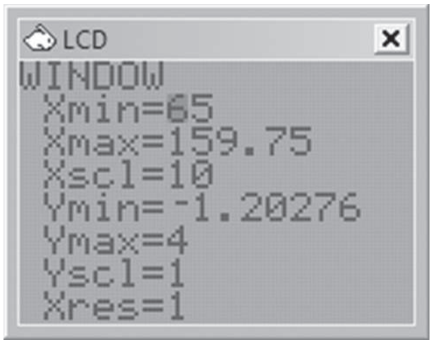
- Press 2^{nd} – Y=   (STATPLOT)
- Enter into Plot1 and set up the following:

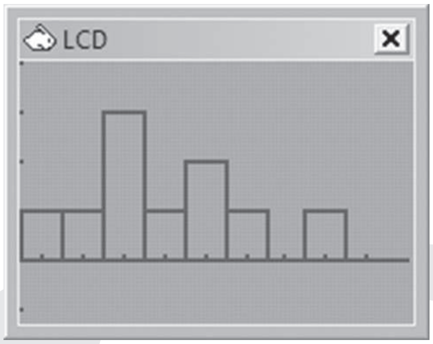
- Press ZOOM – 9   (ZoomStat)

- Press GRAPH



- You can change the histogram bin settings by going to WINDOW and changing XMIN and XSCL:

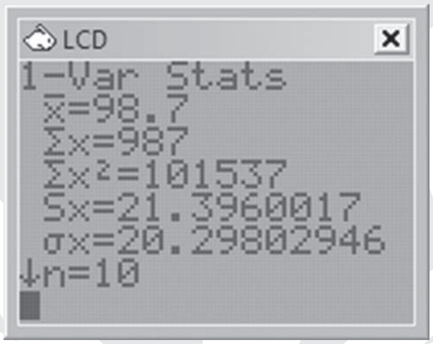




To compute summary statistics:

- Press STAT-CALC-1-Var Stats L1

- Press ENTER



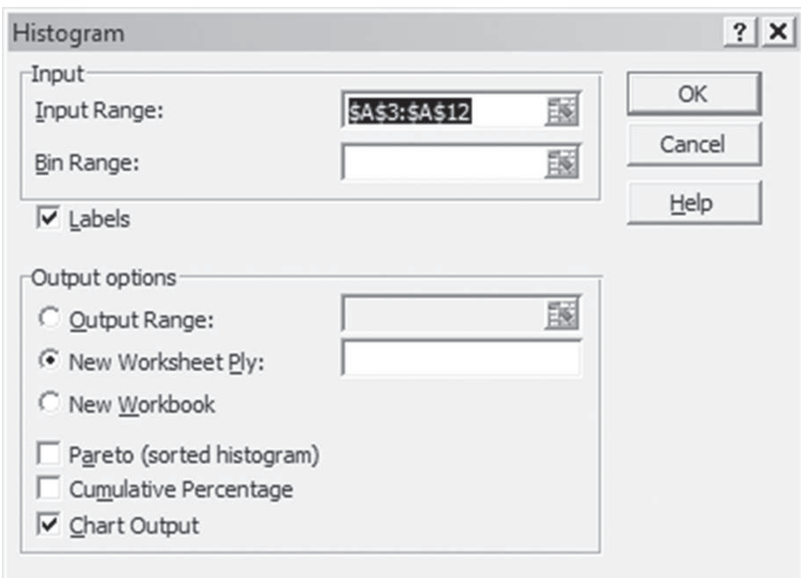
Instructions on How to Use Microsoft Excel for Statistics

The following functions are useful in Excel:

- AVERAGE(range)
- MEDIAN(range)
- STDEV(range)

In order to create histograms in Excel, you must add the “Data Analysis Add-In.” For details, type “Data Analysis Tool” in your help menu.

Once this has been done, you can create a histogram. Select *Data Analysis* from the *Data* tab. Set up the following dialog box:



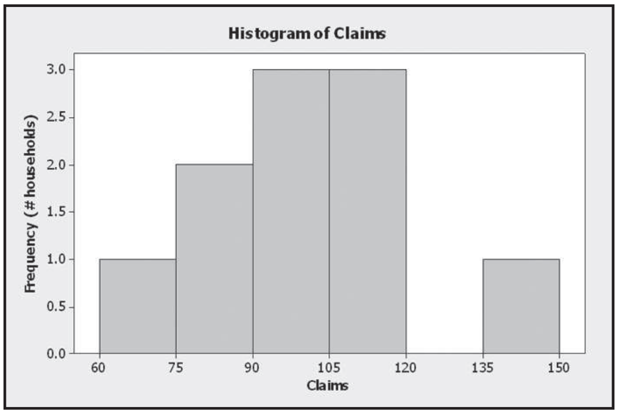
This will create a histogram. To experiment with the bin sizes and frequency, utilize Excel help.

Module 2: The Normal Model

MODULE 2

Let us return to the time when the last (Category 3) hurricane hit in order to review the distribution of claims from the houses in Happy Shores that are closest to the ocean.

Claims
112
92
99
90
117
79
141
66
86
106



For our sample, the mean is around 98.8, and the standard deviation is around 21.3. To make things a bit simpler, let's round the mean to 99 and the standard deviation to 21. For #3 on the practice exercise in Module 1, we found what percent of observations were within 1, 2 or 3 standard deviations of the mean. We could also look at each observation and compute the number of standard deviations from the mean.

For example, we can perform the following calculation to find out how many standard deviations from the mean the claim is for the house with a claim of 112,000:

$$\frac{(112-99)}{21} = \frac{13}{21} = 0.619$$

Discussion Question

Q: Compute the number of standard deviations from the mean for all the observations:

Claims	SStDev
112	0.619
92	
99	
90	
117	
79	
141	
66	
86	
106	

These values are often called **standardized values** because they allow us to compare values of one distribution to another by looking at the number of standard deviations from the mean. For example, comparing wind speed (measured in miles per hour) to storm surge (measured in feet) is difficult because they are in different units. How does a wind speed of 80 mph compare to a storm surge of 20 feet? It is difficult to tell. This is where standardizing becomes useful. We could say that a wind speed that is 1.5 standard deviations above average is more impressive than a storm surge that is 1.1 standard deviations above average.

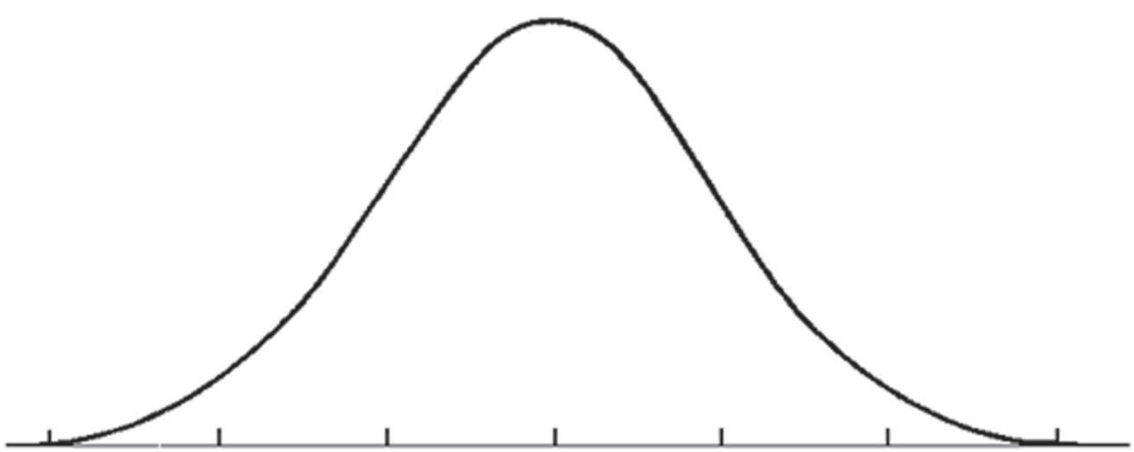
Standardized values are commonly called **z-scores**. As you discovered in computing z-scores for the claims for 10 houses near the beach, this is a formula to compute a z-score:

$$z = \frac{x - \bar{x}}{s}$$

A z-score gives us an indication of how unusual a value is because it tells us how far it is from the mean. If a data value is right at the mean then the z-score is 0. A z-score of 1 means that the value is 1 standard deviation greater than the mean. Note that z-scores can be negative as well. A z-score of -1 tells us that the value is one standard deviation below the mean. How far does a z-score have to be to be considered unusual? There is no universal standard, but the larger the z-score (negative or positive), the more unusual it is. It's not uncommon for over half the data to have z-scores between -1 and 1 (within 1 standard deviation of the mean). No matter what the shape of the distribution, a z-score of 3 (plus or minus) or more is considered rare.

To really understand how big we expect a z-score to be, we need a **model** to describe the distribution. A model describing a distribution is a mathematical curve that would approximately fit the histogram of the data. Models help our understanding in many ways even though they don't fit each data value exactly. All models in the real-world will be wrong—wrong in the sense that they can't match reality exactly. But models are very useful in that they are something we can look at and manipulate in order to learn more about the real world.

Creating a model to describe a distribution is often useful. Distributions that are symmetrical, bell-shaped and unimodal are often described by a **normal model**. A picture of a normal model is shown below.



The normal model with a mean of 0 and a standard deviation of 1 is called the standard normal model. Generally we define the mean of normal models as μ and standard deviation as σ. The mean and standard deviation don't come from the data. Rather, they are numbers (or **parameters**) which we specify to help describe the model.

This is the equation that describes a general normal model: $f(x) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$

This means that the standard normal model can be described by $f(x) = \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}x^2}$

Calculus Connection

Can you compute the area under the standard normal model?

HINT: Find: $\int_{\infty}^{-\infty} \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}x^2} dx$

You can try to find the anti-derivative of the function (but don't try too hard because it's impossible). Therefore you should use your calculator!

Using the same calculus techniques, find the following areas under the standard normal model:

- Area between -1 and 1
- Area between -2 and 2
- Area between -3 and 3

Because the entire area under the normal curve is 1, we can think of areas under the curve as proportions of observations or as probabilities. For instance, the probability that an observation occurs that is within 1 standard deviation of the mean for a normal model is around 0.68.

Let's Review…

Practice Exercise 1

Example: If SAT scores can be modeled with a normal distribution, and the mean score is 500 with a standard deviation of 100, then find the following probabilities:

1. The probability that someone scores between 400 and 600
2. The probability that someone scores over 600
3. The probability that someone scores over 650

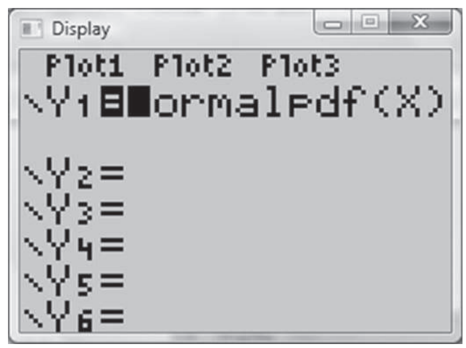
4. The probability that someone scores between 450 and 600

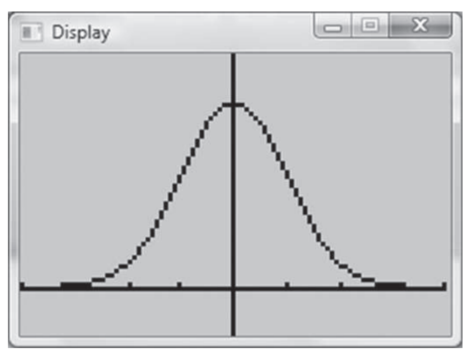
5. The probability that someone scores less than 420

Technology Connections

Your calculator knows the normal model. Have a look under 2nd-DISTR. There you will see the three "norm" functions

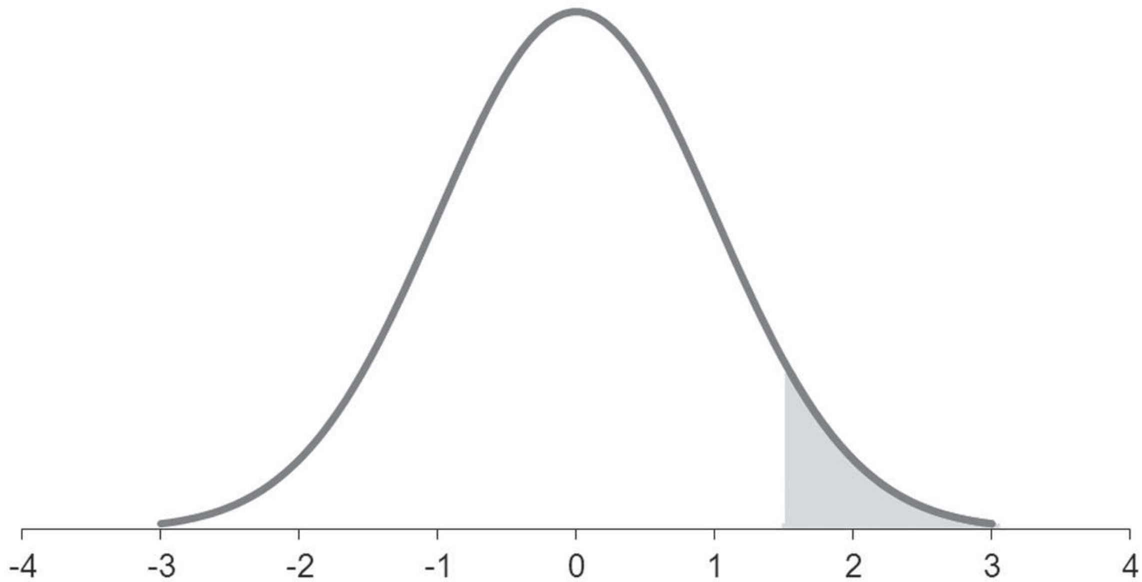
Normalpdf(calculates y-values for graphing a normal curve. You probably won't use this very often. If you want, graph Y1 = normalpdf(X)to try it:



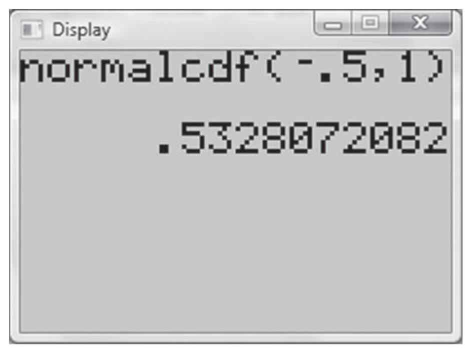


normalcdf finds the proportion of area under the curve between two z-score cutoff points, by specifying normalcdf(zLeft, zRight). You can use this function to find the integrals that you evaluated in the Calculus Connection section on page 17.

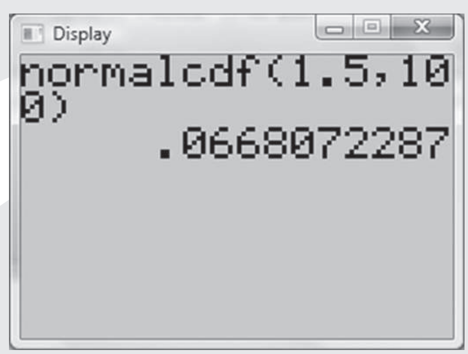
The normal model, shown below, shades the region between z = -0.5 and z = 1



To find the shaded area, you can do the following:



To find the area above z = 1.5, you could do the following:



Note that the zRight = 100 because if we are 100 standard deviations from the mean, essentially no area will be above this point.

The function invNorm finds the z-score that corresponds to a certain area below a value (this area below is called a **percentile**). For instance, the 75th percentile would represent the value such that 75% of the values are at or below this value. To find the z-score of the 75th percentile, you do the following:

Let's Review...

Practice Exercise 2

Suppose the data looking at insurance claims of oceanfront homes due to a category 3 hurricane is appropriately modeled by a normal curve with a mean of 99 (thousand) and a standard deviation of 21 (thousand).

1. If a category 3 hurricane hits, what is the probability that a particular household files a claim for more than $110,000?
2. What is the probability that a particular household files a claim for more than $150,000?
3. What is the probability that a particular household files a claim for less than $90,000?
4. What claim would represent the 90th percentile?
5. Approximately 5% of all claims would be below what amount?

Module 3: Discrete Probability Distributions

MODULE 3

Over the last 100 years, 16 hurricanes have hit Happy Shores. There have also been about 14 tropical storms (that didn't develop into hurricanes) in that time frame. The following table shows the frequency of these storms broken down by category. No more than one storm hit Happy Shores in any given year.

Intensity at Landfall	# Storms
1	1
2	1
3	1
4	5
5	8
Tropical Storm	14

Although, the occurrence, path and intensity of hurricanes depend on many things, at a very high level, we can consider them to be **random phenomena**. Is whether a hurricane hits Happy Shores completely unpredictable? When you think about it, we probably do expect some sort of regularity in the long-run. For instance, we might expect Happy Shores to be hit with a hurricane about once every 6.25 years.

In general, each occasion in which we observe a random phenomenon is called a **trial**. At each trial, we note the value of the random phenomenon, and call that the trial's **outcome**. In our context, we could consider each year to be a trial. In each year we can have the possible outcomes of no storms, Tropical Storm, Category 1 Hurricane, Category 2 Hurricane, etc. If we list all possible outcomes, then we call that the **sample space** of our random phenomenon. If we consider a large number of independent trials (independent means that one trial doesn't really affect the next), then we can estimate the **probability** of each outcome with the proportion of times the outcome occurs. Probability simply measures the likelihood or chance of a certain outcome occurring.

Because we have historical data, we can use it to estimate probabilities. Meteorologists may use other information to help estimate probabilities, but part of what they use is historical data. Based on our information, in 100 years, we have had one Category 5 hurricane. Therefore, we might estimate the probability of a category 1 hurricane making landfall in Happy Shores to be 1/100.

Discussion Question

Q: Fill in the following table based on the historical data:

Outcome	Probability
No Storms	
Tropical Storm	
Category 1 Hurricane	
Category 2 Hurricane	
Category 3 Hurricane	
Category 4 Hurricane	
Category 5 Hurricane	

This table of outcomes and probabilities is called a **probability distribution** (or **probability model**).

This is an example of a **discrete probability distribution** (or **model**) because the outcomes only take certain values. Notice that the probabilities all add up to one. This will always be true of a valid probability model.

We have seen other probability distributions (models) before. In module 2, we studied the normal probability model. This model is **continuous** because it can take on any value (theoretically). For instance, we looked at the number of insurance claims when a certain category of hurricane hit Happy Shores. We used the normal model to estimate the probability of a claim being in a certain range.

Expected Values (Means) of Discrete Probability Distributions

Consider the following simple example unrelated to hurricanes and Happy Shores but related to the concept of insurance. Suppose an insurance company offers a "death and disability" policy that pays $10,000 when you die and $5,000 if you are permanently disabled. It charges a premium of $50 per year for this plan. Is the company likely to make a profit selling such a plan?

To answer this question we will use historical data that tell us that the death rate in any one year is 1 out of every 1000 people, and that another 2 out of 1000 suffer some kind of disability.

Discussion Question

Q: What would the probability distribution for this insurance policy be (fill in the blanks):

Policyholder Outcome	Payout (x)	Probability P(X = x)
Death	10,000	
Disability	5,000	
Neither	0	

To see what the insurance company can expect, imagine that it insures exactly 1000 people. Also imagine that, in perfect accordance with the probabilities, 1 of the policyholders dies, 2 are disabled, and the remaining 997 survive the year without harm. The company will have to pay $10,000 to one client and $5,000 each to two clients. That's a total of 20000/1000=$20 per policy. Since it is charging $50 for the policy, the insurance company will have a profit of $30 per customer.

We can't predict what will happen in a given year, but we can say what we expect to happen. The **expected value** is the average amount of payout the company will make according to the model. It is the mean of the probability distribution. In this case it is $20 for the insurance company.

How did we come up with $20 as the expected value of the policy payout? Here is the computation:

$$E(X) = \frac{10,000(1) + 5,000(2) + 0(997)}{1000} = 10,000\left(\frac{1}{1000}\right) + 5000\left(\frac{2}{1000}\right) + 0\left(\frac{997}{1000}\right) = 20$$

As you should see, computing the expected value of a discrete random variable is easy—just multiply each possible outcome by its probability and add up these products. Here is the formula:

$$E(X) = \sum xP(x)$$

Let's Review…

Practice Exercise 1

Let us again consider the possible damages to the ten oceanfront homes in Happy Shores. In Modules 1 and 2, we saw that when a category 3 hurricane hit, we expected around $99,000 worth of damage to occur to a home right on the beach. This is approximately 20% of the home's value since these oceanfront homes are worth around $500,000 each.

Below are estimates (based on historical claims information) for the extent of damages to these ten oceanfront homes based on the category of hurricane.

- Category 5 Hurricane – Virtually wipes out 100% of the home ($500,000)
- Category 4 Hurricane – Wipes about 70% of the home ($350,000)
- Category 3 Hurricane – Wipes out about 20% of the home ($100,000)
- Category 2 Hurricane – Wipes out about 10% of the home ($50,000)
- Category 1 Hurricane – Wipes out about 5% of home ($25,000)
- Tropical Storm – Wipes out about 1% of home ($5,000)

1. Create a probability distribution for the possible claim amounts in a given year. Fill out the following table:

Results	Cat 5	Cat 4	Cat 3	Cat 2	Cat 1	TS	NONE
Claim Amt							
Probability							

2. Find the expected amount of the claim for these homes.

3. Based on these numbers, what do you think is a reasonable amount for the insurance company to charge as its premium for hurricane insurance for these homes? (Remember, the insurance company needs to make a profit!)

MODULE 3

Standard Deviation of a Discrete Probability Distribution

We now know that on average, the insurance company expects to pay out $14,700 in claims. Of course, the expected value is not what happens to a particular household in a particular year. No individual policy actually costs the company $14,700. In fact, 70% of the time, the company will not pay out any claims, and 1% of the time, it will pay out $500,000 to a household on the beach. Because the insurance company must anticipate this variability, it needs to know the **standard deviation** of the random variable.

Let us return to the "Death and Disability" Insurance:

Policyholder Outcome	Payout (x)	Probability P(X = x)
Death	10,000	1/1000
Disability	5,000	2/1000
Neither	0	997/1000

For data (in Module 1), we calculated the standard deviation by first computing the deviation from the mean and squaring it. We do that with discrete random variables as well. First we find the difference between the payout and the expected value ($20):

Policyholder Outcome	Payout (x)	Probability P(X = x)	X – E(x)
Death	10,000	1/1000	(10,000 – 20) = 9980
Disability	5,000	2/1000	(5,000 – 20) = 4980
Neither	0	997/1000	(0 – 20) = -20

Next we square each deviation. The **variance** is the expected value of those squared deviations, so we multiply the squared deviation by the appropriate probability and sum those products.

It looks like this:

$$Variance(X) = 9980^2(\frac{1}{1000}) + 4980^2(\frac{2}{1000}) + (-20)^2(\frac{997}{1000}) = 149,600$$

To get the standard deviation, we take the square root of the variance:

$$SD(X) = \sqrt{149,600} \approx \$386.78$$

So the insurance company can expect an average payout of $20 with a standard deviation of $386.78. The standard deviation is a measure of the risk of selling the policy. The standard deviation of $386.78 indicates that the risk is pretty big for an average profit of $30.

Here are the formulas for what we just did:

$$Var(X) = \sum (x - E(x))^2 P(x)$$

$$SD(X) = \sqrt{Var(X)}$$

Let's Review...

Practice Exercise 2

1. We found that the average claim for a home in Happy Shores on the beach in a given year is $14,700. Compute the standard deviation for the claim amount in a given year.

Additional Exercises

For each of the neighborhoods in Happy Shores, compute the mean and standard deviation of claim amounts based on the damage estimates and average home values below. Using a spreadsheet program to do these calculations may be helpful.

Extent of Damage (% of home value)

Neighborhood	Home Value ($ thousands)	Cat 5	Cat 4	Cat 3	Cat 2	Cat 1	TS
B	250	80%	50%	15%	10%	8%	5%
C	350	80%	50%	15%	10%	8%	5%
D	200	80%	50%	15%	10%	8%	5%
E	400	60%	30%	10%	5%	5%	3%
F	150	50%	20%	10%	5%	5%	3%
G	100	25%	10%	10%	5%	2%	1%

MODULE 4

Module 4: Correlation and Regression

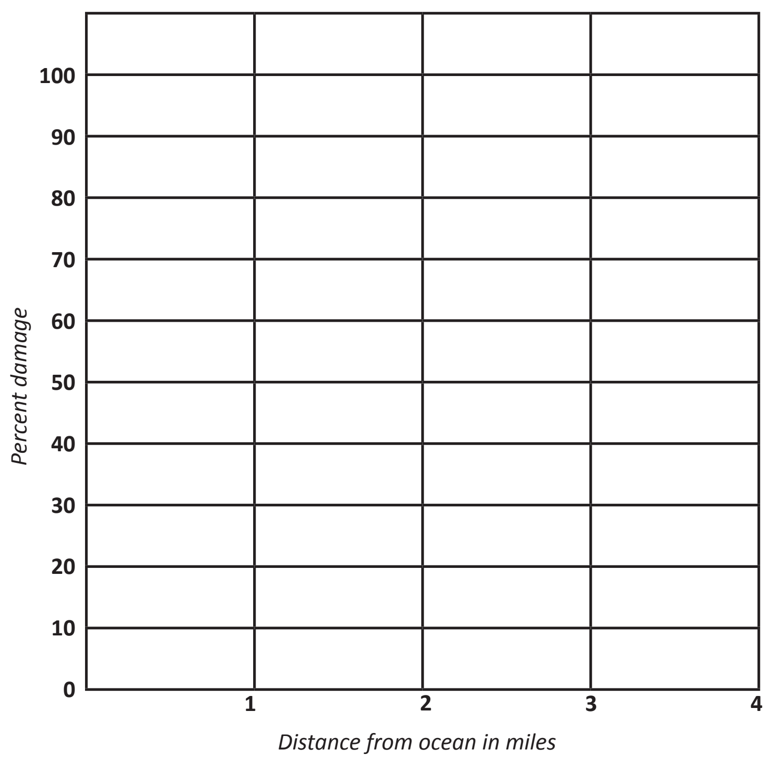
In Module 3 you computed the mean and standard deviation of the claim amounts. You should have noticed that there seems to be a relationship between the amount of damage and the distance a home is from the ocean. This seems obvious. The closer a home is to the ocean, the easier for it to be damaged by high winds or by a storm surge. For a category 5 hurricane, based on historical data, here are estimates concerning percent damage based on neighborhood.

Neighborhood	Avg Home Distance to Ocean	Cat 5 Damage % Estimate
A	0	100%
B	0.5	80%
C	0.6	80%
D	0.8	80%
E	2	60%
F	3	50%
G	3.5	25%

To show the relationship between the damage percent estimate and the distance from the ocean, it may be helpful to create a **scatterplot**.

Discussion Question

Q: Plot each of these points above on this graph:



A scatterplot is the most common graph for looking at the relationship between two quantitative variables. We call the y-axis the **response variable**, and the x-axis the **explanatory variable**. In this case, the distance from the ocean is the response variable and the percent of the home's value that is damaged is the explanatory variable.

The scatterplot clearly shows a negative association between the distance from the ocean and the percent damaged. This means that as the distance from the ocean increases, the damage the home receives tends to decrease.

When you look at a scatterplot, you should look for these three things:

- Direction of the relationship – Is it a positive or negative association, or has no association at all?
- Strength of the relationship – Is it strong or weak? A strong relationship shows very little "scatter"; the individual points cluster around an imagined line. A weak relationship shows the points dispersed across the graph with little relation to one another.
- Form of the relationship – Does it look like the relationship is linear?

These aspects may be somewhat subjective. What may look strong to you may look weak to your colleague. We need a quantitative measure for the strength and direction of the relationship.

The answer lies in what is called the **correlation coefficient**. Here is a formula for the correlation coefficient:

$$r = \frac{1}{n-1}\sum\left(\frac{x-\bar{x}}{s_x}\right)\left(\frac{y-\bar{y}}{s_y}\right)$$

Looks crazy, right? Well it is actually quite simple. You should recognize part of the formula

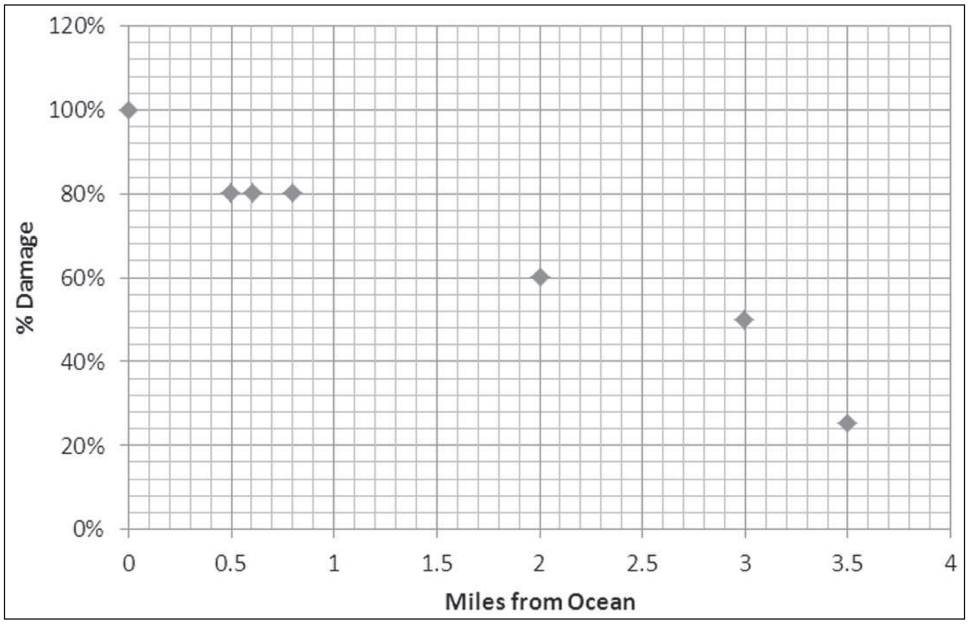
$\left(\frac{y-\bar{y}}{s_y}\right)$ and $\left(\frac{x-\bar{x}}{s_x}\right)$

as being formulas for z-scores (from Module 2). So the correlation coefficient adds up the products of the z-scores and divides that product by one less than the number of data points. This may not shed too much light on what the correlation coefficient measures. Let's not worry too much about the details. Here is what you need to know:

- The correlation coefficient measures the strength and direction of **linear** relationships.
- The correlation coefficient is between -1 and 1. R = -1 means a perfectly linear negative association and R=+1 means perfectly linear positive association.
- Correlation only measures the relationship between two quantitative variables.
- Correlation is a unitless measure.

What is the correlation coefficient for the relationship between the distance from the ocean and the percent damage?

R = -0.9714



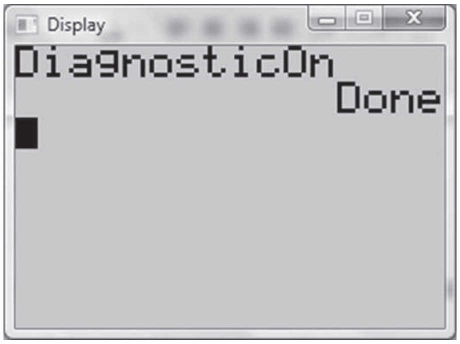
Since the correlation is close to negative 1, this shows a very strong negative association between the variables exists.

Technology Connection

This section has instructions on how to use TI-83/84 Calculator for finding correlation.

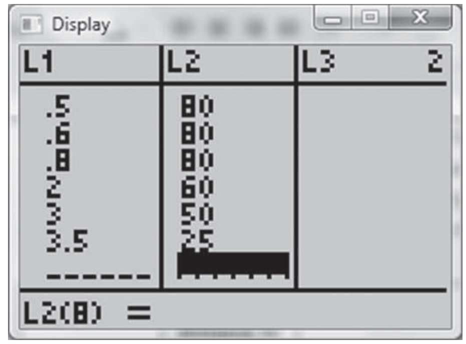
Unfortunately, the statistics package on your calculator does not automatically find correlation. You must change some settings. Here is how you do it.

Hit 2nd-Catalog (on the zero key). You now see a list of everything the calculator knows how to do. Scroll down until you find DiagnosticOn. Hit ENTER and it should say Done (shown below):

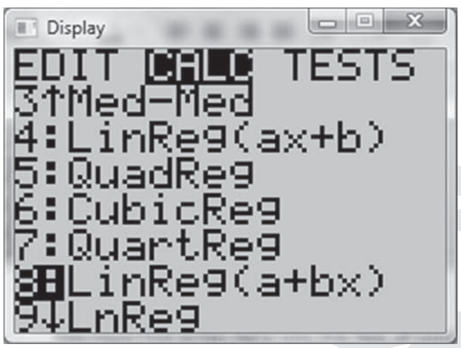


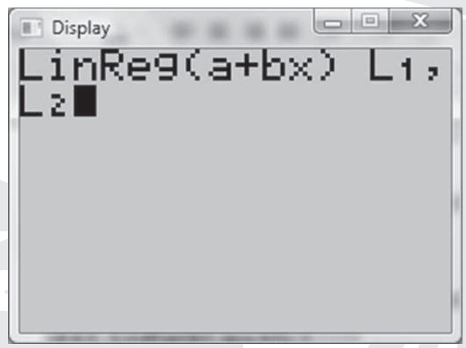
You must first enter data into the lists of your calculator. To do that, press STAT and then EDIT.

Enter the data as shown in the table:

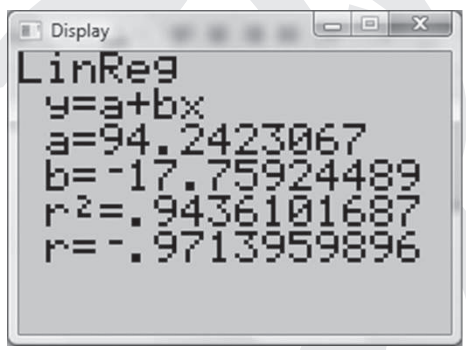


Press STAT CALC and select 8:LinReg(a+bx), and then put L1 (2nd – 1) and L2 (2nd – 2):





Press ENTER. You should see the following:



The correlation coefficient is listed. As you can see, for this example it is -.971.

But what do all those numbers mean?

When we computed the correlation, the calculator also performed **linear regression**. This technique (called "least squares regression") will create a line of best fit for the data. It does this by minimizing the sum of the squared deviations from each point to the line. What matters is that we now have a mathematical model which represents the relationship between the distance from shore and the extent of damage (in a category 5 hurricane).

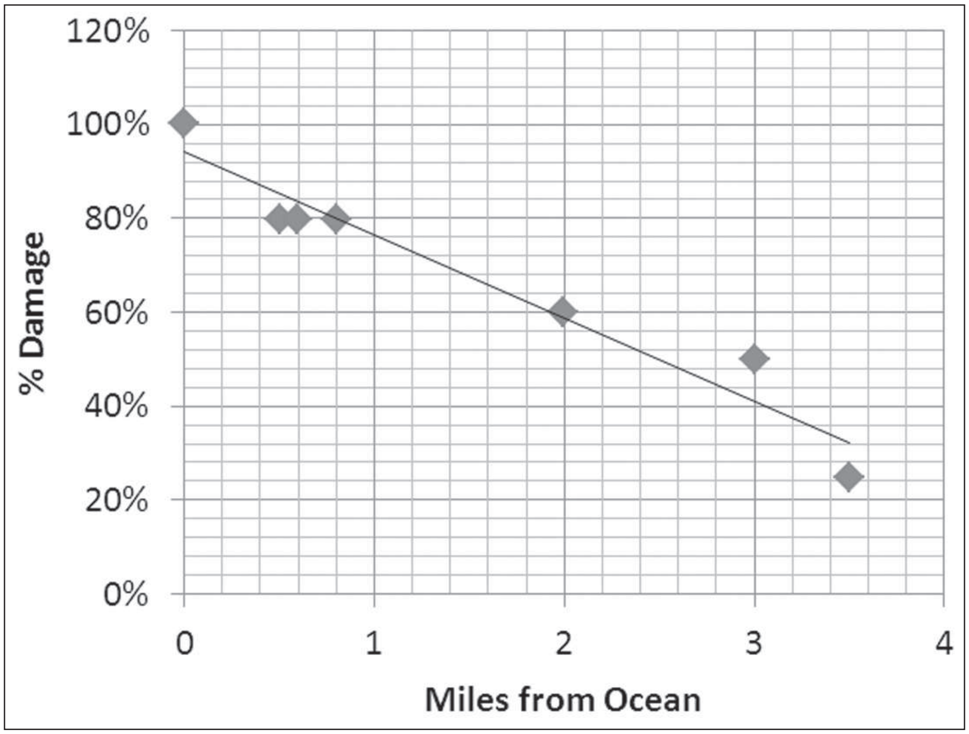
MODULE 4

Based on the calculator output, the equation relating the damage percent to the distance from shore is the following:

Percent damage = 94.24 – 17.76 (Distance from ocean)

Keep in mind that this is only a model to predict. It is not definitive. The percent damage numbers are also just estimates so we are doing a whole lot of estimating here!

The equation shown above is the one describing the trend line below:



Discussion Questions

Q1: What is the y-intercept in the model? What does it represent?

Q2: What is the slope in the model? What does it represent?

Q3: Use the model to predict the percent damage of a home that is 1.5 miles from the ocean.

Q4: Use the model to predict the percent damage of a home that is 5 miles from the ocean.

Using Microsoft Excel to Perform Linear Regression

Excel has useful functions that we can use to compute correlation and run linear regression.

First you must have your data in the spreadsheet:

	A	B
1	Dist	Cat 5
2	0	100%
3	0.5	80%
4	0.6	80%
5	0.8	80%
6	2	60%
7	3	50%
8	3.5	25%

- The function CORREL(list1, list2) will return the correlation coefficient. In the example above we would type =CORREL(A2:A8,B2:B8). The function would return the r-value of -0.9714.
- The function SLOPE(ylist, xlist) will return the slope of the regression equation. For instance, in the example above we would enter =SLOPE(B2:B8, A2:A8). The function would return the slope of -0.17759 (the percents were entered as decimals).
- The function INTERCEPT(ylist, xlist) will return the y-intercept of the regression equation. For instance, in the example above we would enter =INTERCEPT(B2:B8, A2:A8). The function would return the y-intercept of 0.942423.

Practice Exercise 1

We only looked at damages due to a Category 5 hurricane. What about other storm categories? Pick a storm category and create a regression model relating the percent damage to the miles from the ocean. Also report the correlation. Is the relationship stronger or weaker than the one we looked at for a category 5 hurricane?

Nbrhood	Distance	Cat 5	Cat 4	Cat 3	Cat 2	Cat 1	TS
A	0	100%	70%	20%	10%	5%	1%
B	0.5	80%	50%	15%	10%	8%	5%
C	0.6	80%	50%	15%	10%	8%	5%
D	0.8	80%	50%	15%	10%	8%	5%
E	2	60%	30%	10%	5%	5%	3%
F	3	50%	20%	10%	5%	5%	3%
G	3.5	25%	10%	10%	5%	2%	1%

Correlation Tables

Trying to find correlations between every pair of variables in a collection of variables and to arrange these correlations in a table is common in some fields. The rows and columns of the table name the variables, and the cells hold the correlations. Below is an example created from the data you worked with Practice Exercise 1.

	Dist	Cat 5	Cat 4	Cat 3	Cat 2	Cat 1	TS
Dist	**1**						
Cat 5	**-0.9714**	1					
Cat 4	**-0.97942**	0.98431	1				
Cat 3	**-0.89826**	0.895062	0.953233	1			
Cat 2	**-0.92895**	0.861827	0.898717	0.883883	1		
Cat 1	**-0.74275**	0.704623	0.635489	0.458333	0.766032	1	
TS	**-0.42122**	0.352001	0.266959	0.070014	0.495074	0.910182	1

Each row and column intersection shows the correlation between the variable in the corresponding row and column. For instance, we see that the correlation between the distance to shore and the damage associated with a category 5 hurricane is -0.9714 (what we found in the example).

We are most concerned, in this case, with the 1st column. We can see that the distance from the ocean matters most for category 4 and 5 hurricanes (they have the correlations closest to -1). For tropical storms, the distance the house is from the shore may matter less. Perhaps this is because the storm surge is less of an issue in lesser storms. Damage may be caused more by the wind than anything else and this may not vary that much as you move away from shore.

Practice Exercise 2

Consider the following correlation table for the variables about households in Happy Shores and the damage percentages caused by the Category 3 hurricane four years ago:

	% Damage	Distance to Ocean	Square Footage	Elevation	% of House Wood	# Inhabitants
% Damage	1					
Distance to Ocean	-0.8714	1				
Square Footage	0.3115	-0.1561	1			
Elevation	-0.5671	0.3125	-0.021	1		
% House Wood	0.9154	0.0531	-0.041	0.004	1	
# of Inhabitants	0.0233	0.0254	0.4521	-0.0141	0.051	1

- What seems to be correlated with % damage to the home? Explain each variable and the strength and direction of the correlation.
- What is NOT correlated strongly with % damage to the home?
- Describe any other patterns you may see.
- How could an insurance company use this information when trying to decide what to charge different households for hurricane insurance?

Practice Exercise 3

We are interested in the recent trends concerning hurricanes in the U.S. Consider the following information:

- In 2002, there were 12 total tropical storms, 4 of which were classified as hurricanes. The total damage to the U.S. was 2.6 billion dollars.
- In 2003, there were 16 total tropical storms, 7 of which were classified as hurricanes. The total damage to the U.S. was 4.4 billion dollars.
- In 2004, there were 15 total tropical storms, 9 of which were classified as hurricanes. The total damage to the U.S. was 50 billion dollars.
- In 2005, there were 28 total tropical storms, 15 of which were classified as hurricanes. The total damage to the U.S. was 130 billion dollars.
- In 2006, there were 10 total tropical storms, 5 of which were classified as hurricanes. The total damage to the U.S. was 0.5 billion dollars.
- In 2007, there were 15 total tropical storms, 6 of which were classified as hurricanes. The total damage to the U.S. was 3 billion dollars.
- In 2008, there were 16 total tropical storms, 8 of which were classified as hurricanes. The total damage to the U.S. was 47.5 billion dollars.
- In 2009, there were 9 total tropical storms, 3 of which were classified as hurricanes. The total damage to the U.S. was 0.1 billion dollars.
- In 2010, there were 16 total tropical storms, 9 of which were classified as hurricanes. The total damage to the U.S. was 8 billion dollars.

Create scatterplots, compute correlations, and create regression models for the following:

- Number of Hurricanes vs. Year
- Number of Total Storms vs. Year
- Damage vs. Number of Hurricanes

Definitions

Module 1

- **Statistics**—a branch of mathematics dealing with the collection, analysis, interpretation, and presentation of masses of numerical data
- **Data**—facts, statistics, or items of information
- **Distribution**—the values a variable takes and how often it takes those values
- **Histogram**—a type of bar graph that looks at the distribution of one quantitative variable; may group values of the variable together
- **Dot plot**—a graph that looks at the distribution of one quantitative variable by plotting every data value as a dot above its value on a number line
- **Median**—the midpoint of a distribution where half the observations are smaller and the other half are larger
- **Mean**—the numerical average of a distribution
- **Mode**—the value in a range of values that has the highest frequency
- **Unimodal**—a description of shape for a distribution with a single mode (either a single value or range of values)
- **Bimodal**—a description of the shape of a distribution with two modes (either a single value or range of values)
- **Standard deviation**—a measure of how spread out the observations are from the mean in a distribution
- **Variability**—the spread of a variable or distribution
- **Outlier**—a data point in a sample that is widely separated from the main cluster of data points in that sample

Module 2

- **Standardized values**—values that can be compared between distributions by looking at the number of standard deviations from the mean
- **Z-scores**—a common name for standardized values
- **Model**—the description of a distribution using a mathematical curve that approximately fits the histogram of the data
- **Normal model**—a distribution that is symmetrical, bell-shaped and unimodal

- **Parameters**—the mean and standard deviation of a model
- **Percentile**—the value in a distribution below which a certain percent of observations fall

Module 3

- **Random phenomena**—completely unpredictable outcomes in the short term
- **Trial**—each occasion in which a random phenomenon is observed
- **Outcome**—the value of the random phenomenon at each trial
- **Sample space**—all possible outcomes of the random phenomenon
- **Probability**—the likelihood or chance of a certain outcome occurring
- **Probability distribution (probability model)**—a table of outcomes and probabilities
- **Discrete probability model**—a distribution where the outcomes only take certain values
- **Continuous**—a distribution where the outcomes can take on any value in a given interval
- **Expected value**—the mean of the probability distribution
- **Standard deviation of a random variable**—a measure of the variation from the mean in a probability distribution

Module 4

- **Scatterplot**—the most common graph for looking at the relationship between two quantitative variables
- **Response variable**—the y-axis on a scatterplot
- **Explanatory variable**—the x-axis on a scatterplot
- **Correlation coefficient**—a measure of the strength and direction of the linear relationship between two quantitative variables
- **Linear regression**—a predictive model that creates a line of best fit for a set of data points
- **Correlation table**—a table showing the correlations between every pair of variables in a collection of variables

The Actuarial Foundation®

Preparing for tomorrow's possibilities® *through education*

475 North Martingale Road, Suite 600 • Schaumburg, IL 60173-2226

Ph 847-706-3535 • Fx 847-706-3599

www.actuarialfoundation.org