Unit 2

Probability, Random Variables, and Sampling Distributions

Unit Objectives

Upon completion of this unit, you should be able to:

1. demonstrate a basic understanding of probability;

2. find the probability distribution for a discrete random variable;

3. recognize the normal distribution and how to compute areas (or probabilities) under any normal curve;

4. explain the sampling distribution of the mean; and

5. discuss the Central Limit Theorem and the situations in which it may be applied

Instructor's Notes

Probability and Random Variables

Thus far, we have dealt only with **descriptive** statistics – methods for organizing and summarizing data. In general, it would be useful if we could take the results from information obtained from a sample of the population and make generalizations about the whole population. Making generalizations about a population based on sample data is called **inferential** statistics. Because such conclusions are made based on partial information about a population, we can never be sure that our conclusions are correct. Uncertainty will be present in all inferential statements about a population. Probability is the language of inferential statistics. For this reason, we will need to acquaint ourselves with the language of probability.

We frequently encounter statements of probability in everyday life. The local weatherman may tell us that there is a 40% chance of rain or scattered showers. In a football game tossing a coin determines who gets to kickoff. Each team knows they have a probability of winning equal to 1/2, or a "50-50 chance." Suppose Bobby calls directory assistance for ten different numbers and it turns out that only 7 of them are the correct number. Then Bobby would argue that directory assistance is useful only 70% of the time, or the probability of getting the correct number is 0.70. These are a sampling of the different ways that probability may be expressed.

In the weatherman example, the probability is most likely computed using intuition or some type subjective approach based on previous experience and knowledge of weather patterns. The probability computed in the coin toss example is based on the idea that

events of tossing a head or a tail are considered to be equally likely and therefore share the same probability of occurring. In the example of directory assistance, the probability is determined based upon the results of previous "experiments."

Before we get into the theory of probability it is necessary to introduce a few new vocabulary terms. An **outcome** is the result of an experiment or procedure. An **event** is an outcome or a collection of outcomes. The **sample space** is the set of all possible outcomes for an experiment. The **frequency** refers to the number of times a favorable outcome occurs (or the number of favorable outcomes). The **relative frequency** refers to the number of favorable outcomes divided by the number of possible outcomes, i.e., relative frequency = (# favorable outcomes)/(# possible outcomes).

The two primary methods that we will use for determining the probability of an event are the classical approach of equally likely outcomes, and the relative frequency approach (your book calls this the **frequentist interpretation of probability**).

In the classical approach, which can only be applied if each outcome may be considered equally likely to occur, we will simply define probability as follows:

$$\text{probability of an event} = \frac{\#\text{ favorable outcomes}}{\#\text{ possible outcomes}}$$

You may notice that our definition for probability is precisely the same as our definition for relative frequency. In other words, the relative frequency tables of the text may be thought of as probability tables.

Example: Consider a standard deck of 52 cards, not including jokers. The probability of selecting the ace of spades is equally likely as selecting any other card. This can be said of all the cards in the deck, in other words, each card has a probability of 1/52 of being selected. Therefore, the

$$\text{probability of selecting a heart} = \frac{\#\text{ favorable outcomes}}{\#\text{ possible outcomes}} = \frac{13}{52}$$

Using the notation of your book the above statement could be written

$$\boldsymbol{P}(\text{heart}) = \frac{\boldsymbol{f}}{\boldsymbol{N}} = \frac{13}{52}$$

where $\boldsymbol{f}$ = the number of favorable outcomes and $\boldsymbol{N}$ = the number of possible outcomes.

It is often useful to use letters such as *A, B, C,* etc. to represent events so that we may further condense the notation.

The frequentist interpretation of probability would define the probability to be the proportion of times a favorable outcome occurs in a large number of repetitions of an experiment. Technically, calling directory assistance 10 times is not really considered to be a large number of repetitions. However, computing the probability of getting a correct

number based upon the proportion of times directory assistance has been correct in the past certainly makes use of the frequentist approach to probability.

Since probability is defined to be the proportion of times that something occurs it should seem reasonable that the probability for every event should always be a number between 0 and 1, inclusive. Furthermore, if we are interested in the probability of an impossible event occurring, then we should expect that there would be no favorable outcomes. Hence, the probability of an event that cannot occur is 0. Conversely, the probability of an event that is certain to occur will be 1. This should provide us with some intuition concerning the computation of probabilities. If we suspect that an event has a great likelihood of occurring then we can expect that its probability will be close to 1. If an event is unlikely to occur then we can expect that its probability will be close to 0.

The examples above are very simplistic and not entirely realistic. In general, we are interested in compound events. Section 5.2 deals with situations with two or more events happening simultaneously. Section 5.3 introduces the more formal notation and formulas used when computing probabilities of such events.

Two events are considered to be **mutually exclusive** if they have no common outcomes. Determining if events are mutually exclusive can be useful in reducing the amount of work needed to compute probabilities. The complement of event A is defined to be the event consisting of all outcomes in which event A does not occur. Your book denotes the complement of A by (**not *A***).

The sum of all probabilities for all outcomes of an experiment will equal 1. At times, computing the probability of an event A may be very time consuming or difficult. For such situations we will consider the complement of A, if its probability is reasonably easy to compute we will find it and subtract it from 1 in order to find the probability of A. This is the **complementation rule** for probability: $\boldsymbol{P(A) = 1 - P(not\ A)}$.

A **random variable** is a quantitative variable whose value is determined by chance. For example, we may define the random variable *x* to be the number times a head appears when we toss a coin 50 times. Similarly, we may define the random variable *y* to be the lifetime of a Duracell battery. Random variables may be classified as either discrete or continuous.

In chapter 2 the relative frequency distribution for a variable provided information regarding the possible values of the variable and the proportion of times each value occurs. We can extend this idea to random variables using probability. A **probability distribution** is a graph, table, or formula that provides information regarding the possible values of the **random** variable and the proportion of times each value occurs. Similarly, we may construct **probability histograms** to provide the graphical display of a probability distribution.

Example: Suppose a particular contest sells "lottery" tickets for $1 each. The prizes will be distributed according to the following: 1 ticket wins $500, 18 tickets win $200, 120 tickets win $25, and 270 tickets win $20.

Let the random variable *x* be defined as the amount of money that a ticket-holder can win. Then the possible values of *x* are: $0, $20, $25, $200, and $500. The probability distribution for *x* is given below:

x	Frequency f	Probability $P(X = x)$
$0	99591	0.99591
$20	270	0.0027
$25	120	0.0012
$200	18	0.00018
$500	1	0.00001

The table provides a great deal of information. For example, the probability that a ticket-holder will win the jackpot of $500 is $1/100000 = 0.00001$.

Suppose we define event A to consist of the outcome that the ticket-holder wins money. To compute the probability of event A we could add up the bottom four numbers of the probability column, since the individual outcomes are mutually exclusive. However, we might notice that in this instance it might be quicker to apply the complementation rule. The complement of A, (**not** A), is the event that the ticket-holder wins no money. From the table we can see that $P(\textit{not } A) = 0.99591$, therefore $P(A) = 1 - .99591 = 0.00409$.

The mean and standard deviation of a probability distribution may be computed using the following formulas:

$$\mu = \sum \left(x \cdot P(X = x) \right)$$

$$\sigma = \sqrt{\left(\sum x^2 \cdot P(X = x) \right) - \mu^2}$$

These formulas look intimidating, but we can modify our table for the probability distribution in order to make it easier to compute the mean and standard deviation.

x	Probability $P(X = x)$	$x \cdot P(X = x)$	$x^2 \cdot P(X = x)$
\$0	0.99591	0	0
\$20	0.0027	0.054	1.08
\$25	0.0012	0.03	0.75
\$200	0.00018	0.036	7.2
\$500	0.00001	0.005	2.5
SUM	1	0.125	11.53

The mean is the sum of the third column $\mu = 0.125$. We may interpret the mean to say that if we play the lottery we can **expect** to win 12.5 cents. To find the standard deviation we use the sums of the last two columns: $\sigma = \sqrt{11.53 - (0.125)^2} = 3.39$.

Often in probability and statistics we are interested in the results of the repetition of an experiment with only two possible outcomes. For example, suppose we are trying to demonstrate that a particular weight-loss program is effective. We might collect a sample of 25 people. Each subject would be required to implement the weight-loss program for the same amount of time. Then at the end of the allotted time we would determine whether or not the subject lost weight. We might classify losing weight as a "success" and **not** losing weight as a "failure."

Each time the experiment is repeated we call it a **trial**. For each trial we are interested in the outcome as being either a **success** or a **failure**. The probability of success is denoted by p and the probability of failure is given by $q = 1 - p$. The **binomial distribution** is the probability distribution for x number of successes in a sequence of n trials.

Example Nomar Garciaparra plays for the Boston Red Sox. During the 2001 season, his batting average was .289. In other words, for every time that he was up to bat there was a 28.9% chance that he would get a hit. Suppose that in a certain game Nomar goes up to bat 4 times ($n = 4$). We can construct a binomial probability distribution to describe the possible outcomes of Nomar's 4 at bats.

Let the random variable X denote the number of hits in 4 at bats. The possible values of X are 0, 1, 2, 3, and 4. We must keep in mind that there is more than one way for most of these events to occur. To see this it is helpful to list all possible outcomes of his 4 at bats where H = hit and M = no hit (or miss):

HHHH	MHHH	MHMM	HMMH
HHHM	MMHH	MMHM	MHHM
HHMH	MMMH	MHMH	HHMM
HMHH	MMMM	HMHM	MMHH

If we think of this example in terms of a binomial distribution, we can classify a hit as a **success** and a miss as a **failure**. Furthermore, the probability of success $p = 0.289$, while the probability of failure $q = 0.711$. The probability that Nomar gets one hit followed by 3 misses (HMMM) is $(0.289)\cdot(0.711)^3 = 0.10387$. If we are only interested in the probability that he gets one hit (and three misses with no restriction on order) then we need to count the number of times it occurs in the list above and multiply by the individual probability: $4\cdot(0.10387) = 0.4155$. We can do this for all values of the random variable and create a table for the probability distribution of *X*.

Number of Hits x	Frequency f	Probability $P(X = x)$
0	1	$1\cdot(0.711)^4 = 0.2556$
1	4	$4\cdot(0.289)(0.711)^3 = 0.4155$
2	6	$6\cdot(0.289)^2(0.711)^2 = 0.2533$
3	4	$4\cdot(0.289)^3(0.711) = 0.0686$
4	1	$1\cdot(0.289)^4 = 0.0070$
Total		1.0000

We can look at the table and see that the probability of Nomar getting 3 hits (and one "no hit") is 0.0686. Likewise, the probability that Nomar gets at least one hit is 0.7444 (the sum of the last four numbers: 0.4155, 0.2533, 0.0686, and 0.0070).

Clearly, this whole process would be more complicated if we increased the number of trials. The above probability distribution may be classified as a binomial distribution with $n = 4$ and probability of success $p = 0.289$.

Your textbook provides the **binomial probability formula**, which greatly reduces the amount of work to be done. Be sure to familiarize yourself with the formula and its new notation.

The mean and standard deviation for a binomial distribution with *n* trials and probability of success *p* are given by $\mu = np$ and $\sigma = \sqrt{npq}$, respectively, with $q = 1 - p$. For our example, the mean is $\mu = 4\cdot(0.289) = 1.156$ and the standard deviation is $\sigma = \sqrt{npq} = \sqrt{4\cdot(0.289)\cdot(0.711)} = 0.907$.

Written Assignment

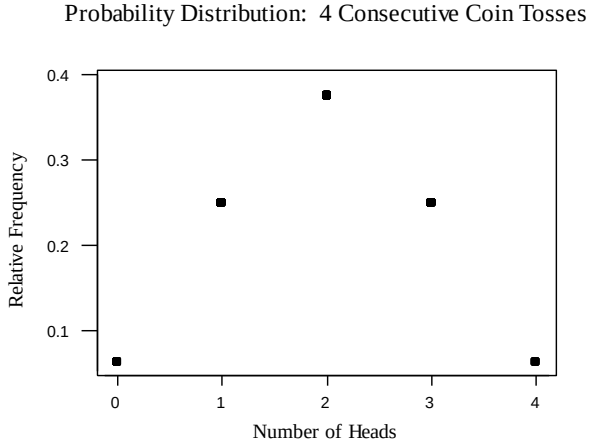
Reminder: these written assignments are for your benefit and are **NOT** to be turned in for a grade.

Do problems 5.8, 5.13, 5.19, 5.37, 5.47, 5.45, 5.65, 5.69, 5.79, 5.93, 5.103, 5.112, 5.188, 5.121

The Normal Distribution

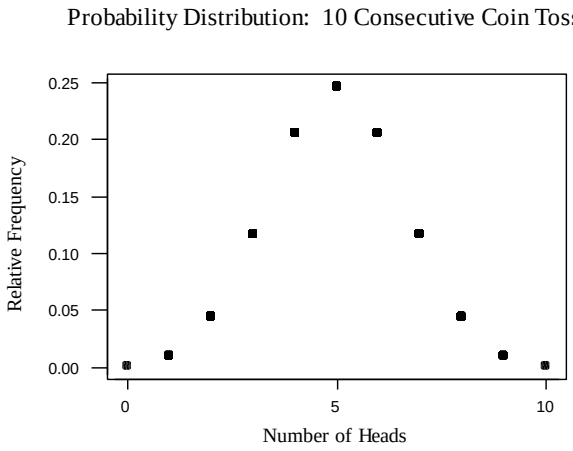
Suppose we want to construct a probability distribution for an experiment consisting of counting the number of heads that appear in 4 consecutive tosses of a fair coin. The probability distribution is derived using the methods outlined in your text. We can list the sample space and calculate the associated probability for each value of the random variable (like we did in the Nomar Garciaparra example) and then construct a table or graphical display. A probability distribution table and associated graph are given below.

Number of Heads x	Probability $P(X = x)$
0	0.0625
1	0.2500
2	0.3750
3	0.2500
4	0.0625
Total	1.0000

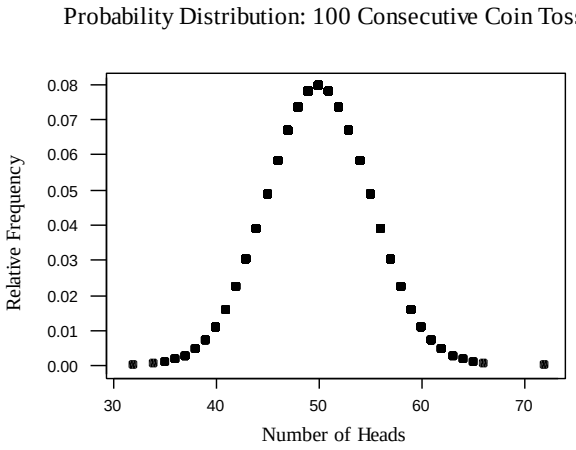


Both the table and the graph show the **theoretical** probability distribution for the number of heads that appear when flipping a coin 4 times. Note: since this is a **discrete** random variable the height of the dot above a number on the horizontal axis represents the probability of the random variable taking on that variable.

What happens if we modify our experiment to consist of tossing a coin 10 times and recording the number of heads? The sample space now consists of $2^{10} = 1024$ possible outcomes. That is too many outcomes for us to list by hand. Instead we can let Minitab do the work for us. The graph of the probability distribution is given below.



If we flip a coin 100 consecutive times we get the following distribution.

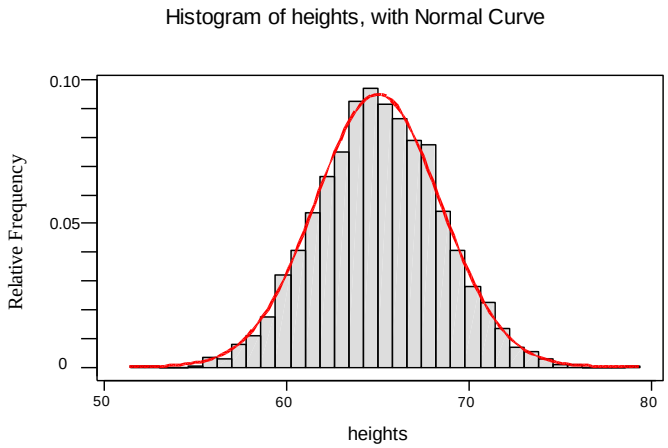


Here there are no dots corresponding to numbers less than 30 and greater than 75. This is due to limitations of the graphical display. The probabilities associated with such values are so small (e.g., $\boldsymbol{P(X} = 10 \text{ heads}) = 1.4 \times 10^{-17} = 0.0000000000\ 00000014$) that it cannot be seen given the scale of the graph.

After comparing the three graphs a few properties seem to be shared by all. In particular, the graphs are bell-shaped and symmetric about the high point in the middle.

A continuous random variable that follows a symmetric, bell-shaped distribution (similar in shape to the distributions of the discrete random variables above) is called a **normally distributed variable**. Such variables are said to follow a **normal distribution**. It turns out that the normal distribution occurs frequently in nature.

Example: The random variable for women's height is normally distributed. The frequency histogram for a random sample of 2500 women's heights is given below. The distribution is bell-shaped and symmetric about its mean of 65 inches.



We compute probabilities associated with continuous random variables by calculating the area under a curve bound between two values. In general, the idea is to obtain the percentage of observations that lie within a specific interval of numbers. In the example above, we can approximate the percent of observations falling within a certain range by finding the area of the rectangle, for the corresponding class, in the above histogram.

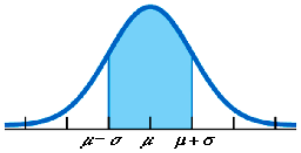
To find areas under a normal curve requires the use of a table. However, there are infinitely many normal curves. We distinguish between them by identifying their mean and standard deviation. It is not reasonable to think that we could have a table for each normal curve. Instead we will make use of a process which transforms every normal curve into the **standard normal curve** (or *z*-curve). The standard normal curve is a normal curve with mean $\mu = 0$ and standard deviation $\sigma = 1$. After converting to the standard normal curve we will refer to the **standard normal table** (*z*-table) to determine probabilities. We **standardize** a variable by computing its *z*-score, $z = \frac{x - \mu}{\sigma}$.

At this point you should familiarize yourself with the techniques of section 6.2 regarding the use of the *z*-table. It is important to keep in mind that the total area under the standard normal curve is equal to 1. Furthermore, since the distribution is symmetric about the mean, located at 0, precisely half of the area (equal to 0.5) will lie to both the left and right of zero. Finally, due to the way the *z*-table is constructed, it will often be necessary to implement the complementation rule in order to compute the area of the desired region.

Example: Suppose that the population mean $\mu = 65$ inches and $\sigma = 3.25$ inches for women's height. If a woman is selected at random, what is the probability that she will be within one standard deviation of the population mean?

In other words, we are interested in finding the probability that the woman is between 65 – 3.25 = 61.75 and 65 + 3.25 = 68.25 inches tall. In probability notation, we want to find $P(61.75 \le x \le 68.25)$. Since we need to find the area bound under the normal curve between these two values it will be necessary to standardize our values of x.

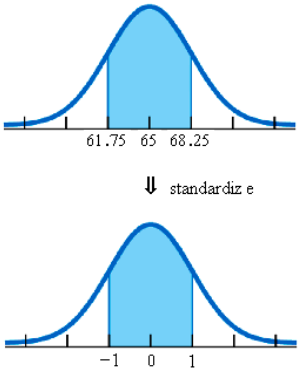
within one standard deviation of the mean:



converting to z-scores:

$$z_1 = \frac{61.75 - 65}{3.25} = -1$$

$$z_2 = \frac{68.25 - 65}{3.25} = 1$$



Thus, we want the area bounded between –1 and 1 under the standard normal curve. Using the z-table, we find the area to be 0.6826. Hence, our randomly selected woman has a probability of 0.6826 of being within one standard deviation of the mean. Similarly, we may interpret our problem to say that we can expect 68.26% of the population of women to have a height within one standard deviation of the mean.

Written Assignment

Do problems 6.14, 6.15, 6.37-6.45, 6.51-6.61, 6.71, 6.69, 6.73

The Sampling Distribution of the Sample Mean

Typically we do not have access to all measurements of an entire population because of constraints due to time, money, and effort, etc. As a result, the only practical way to gather information about a population is through sampling. In particular, we will use a statistic to make inferences (or generalizations) about corresponding population parameters. For example, we will use the sample mean to make generalizations about the population mean.

The inherent problem associated with using sample data to make inferences is that we cannot be certain that information collected from a sample will accurately describe characteristics about the population. Therefore, we should expect a certain amount of error to exist as a result of sampling.

Most generalizations about population parameters may be categorized as either estimating a value of a population parameter or formulating a decision about a population parameter. In order to determine the reliability of such inferences, it will be necessary to

know the probability distribution associated with a particular statistic. Such a probability distribution is called a **sampling distribution**.

We will focus on the **sampling distribution of the sample mean** (or the sampling distribution of $\bar{x}$). This basically means we will be taking several samples of a given size, computing the mean for each individual sample, and then looking at the distribution of those means. What can we say about this distribution? Will it look symmetric? Skewed? Bell-shaped? Will it follow a binomial distribution? Or perhaps a normal distribution? The **Central Limit Theorem** will provide the answers to such questions.

Example: The Mars Company uses a special "recipe" when making a batch of M&M's, i.e. the percent of each color in a batch is pre-determined. For example, the percent of yellow found in a bag of Peanut M&M's is different than the percent of yellow found in a bag of Crispy M&M's. It turns out that the target percent of red in a bag of Plain M&M's is 20%. If we define a variable x to be the percent of red in a bag of Plain M&M's, then the population mean percent of red M&M's is $\mu = 20$.

You have recently been hired as the floor manager at the local M&M factory where part of your job involves Quality Assurance. In particular, you must determine if the colors of M&M's in a Plain 1.69 oz. bag meet the recipe standards. Suppose you walk up to the conveyor belt and select two 1.69 oz. bags and discover the first contains only 13% red while the second contains 32% red. Neither of these selections is close to the satisfying the requirement of 20% red. Does this mean the whole batch should be thrown out? What exactly is the percent of red M&M's that goes into a 1.69 oz. bag?

Clearly you cannot pull every bag off the conveyor belt, open it, and survey its contents. In order to answer the question, you decide to sample the population and record some data. Every day for 40 days you collect a random sample of ten 1.69 oz. bags and determine the percent of each color in each bag. Below is a table of values for the 40 samples and their associated sample means. By looking at the far right column we can get a sense of the **average** percent of red M&M's.

Table 1: Percent of Red M&M's in a 1.69 oz. Bag of Plain M&M's

sample	bag 1	bag 2	bag 3	bag 4	bag 5	bag 6	bag 7	bag 8	bag 9	bag 10	mean %
1	17	16	21	24	14	11	18	11	18	17	17
2	13	16	7	24	25	14	10	19	12	18	16
3	25	24	19	14	24	16	13	22	17	25	20
4	16	33	20	15	18	14	19	14	15	24	19
5	20	26	15	15	22	30	32	23	25	27	24
6	13	5	17	19	19	20	15	15	20	14	16
7	20	26	21	23	19	18	9	16	29	20	20
8	23	22	11	5	25	21	23	31	18	20	20
9	25	17	23	22	29	18	9	20	23	23	21
10	30	21	25	24	22	16	22	14	27	33	23
11	28	11	19	21	21	22	15	9	13	11	17
12	25	32	27	25	24	28	31	24	30	22	27
13	19	11	26	18	16	12	15	23	15	19	17
14	18	27	20	23	11	22	19	23	18	11	19
15	9	21	9	11	15	5	9	18	13	12	12
16	26	24	15	19	16	19	22	15	18	21	20
17	16	22	25	22	22	23	26	30	33	22	24
18	32	26	26	16	28	22	18	19	22	24	23
19	18	20	26	13	32	16	16	20	23	28	21
20	20	22	21	15	27	16	16	25	23	21	21
21	26	16	19	20	20	21	22	18	28	23	21
22	20	20	24	22	33	15	22	28	24	17	23
23	29	23	32	28	24	26	30	36	27	26	28
24	31	9	13	31	36	25	14	29	31	30	25
25	14	17	24	17	19	13	22	14	21	16	18
26	32	13	30	22	27	16	19	30	23	34	25
27	18	16	19	19	18	19	18	19	19	23	19
28	23	26	32	22	20	18	30	23	16	31	24
29	16	7	14	17	16	7	9	12	14	13	12
30	14	14	19	18	18	30	15	27	29	23	21
31	22	19	20	13	16	13	10	22	24	21	18
32	17	22	19	14	15	11	23	12	22	18	17
33	10	14	11	15	23	16	27	23	18	25	18
34	14	16	17	19	13	20	15	21	11	21	17
35	16	8	27	19	11	18	8	11	23	16	16
36	21	22	18	28	23	26	16	19	20	20	21
37	23	32	16	23	19	18	27	26	21	10	21
38	17	20	25	14	33	15	18	20	24	29	21
39	15	12	14	22	28	32	30	29	23	25	23
40	16	23	22	29	26	31	27	25	17	20	24

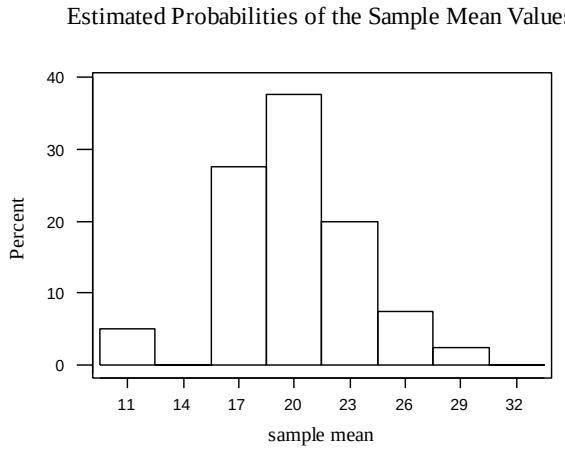
Looking at the 40 values of the sample mean does not tell us very much. Sometimes the mean is the target value of 20% while other times it is as low as 12% or as high as 28%. However, we may compute the mean of the sample means (i.e., the average of the values in the far right column) and find it equals 20%. Furthermore, we might notice that, in general, there appears to be less variation in the values of the far right column than for most individual samples. Let us organize our sample mean values into a frequency table.

Table 2: Frequency Table for 40 Values of $\bar{x}$

% Red	Class Midpt	Frequency	Relative Frequency
10 - 12 %	11	2	0.050
13 - 15 %	14	0	0.000
16 - 18 %	17	11	0.275
19 - 21 %	20	15	0.375
22 - 24 %	23	8	0.200
25 - 27 %	26	3	0.075
28 - 30 %	29	1	0.025
31 - 33 %	32	0	0.000
Total		40	1.000

The far right column contains relative frequencies, which may be thought of as probabilities. Therefore, we may view the table as describing a probability distribution. Since $\bar{x}$ represents the mean percent of red M&M's, we can estimate the probability of $\bar{x}$ falling into each class by using the relative frequencies. For example, approximately 37.5% of the time we can expect that a randomly selected bag will contain 19-21% red M&M's.

The graph below represents a **probability sampling distribution** for the sample mean $\bar{x}$ of percent of red M&M's based on random samples of size 10.



The distribution is roughly bell-shaped. The irregularities are due to the small number of samples used (only 40 sample means) and the rather small sample size (10 bags per sample). These irregularities would be less pronounced and even disappear if the number samples increased, if we used a larger number of classes, and if we used more bags per sample. Specifically, the possible sample means should cluster more closely around the population mean $\mu = 20\%$ as the sample size increases. Therefore, a larger sample size should yield a smaller amount of sampling error when using samples to estimate the population mean. Therefore, if we have a large sample size then we should expect the mean of the sample means $\mu_{\bar{x}}$ to be very close to the population mean μ. Similarly, as

sample size increases we should expect the **standard error** $\sigma_{\bar{x}}$ (or the standard deviation of $\bar{x}$) to get smaller and smaller.

Essentially, the **Central Limit Theorem** tells us that for a given variable ***x*** (with mean μ and standard deviation σ), if we take all possible samples of size ***n*** (where $n \geq 30$) then the distribution of the sample mean $\bar{x}$ will be approximately normally distributed with mean $\mu_{\bar{x}} = \mu$ and standard deviation $\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}}$. If the original variable ***x*** is normally distributed then the distribution of $\bar{x}$ will be normal regardless of sample size. If we know nothing of the distribution of ***x*** then we will require a sample size of 30 or larger to invoke the Central Limit Theorem.

It is, indeed, surprising that regardless of the initial distribution if we allow sample size to get larger and larger then the distribution of $\bar{x}$ will approach a **normal** distribution. Only now can we begin to appreciate the significance of the normal distribution.

Example: Assume that the population of human body temperatures is normally distributed with mean μ = 98.6°F and standard deviation σ = 0.62. If a sample of size *n*=15 is randomly selected, find the probability of getting a mean body temperature between 98.4°F and 98.8°F.

To answer this question we need to recognize that unlike questions from the text, we are now asked to find a probability associated with the average temperature of a collection of values instead of an individual value. Therefore, we recognize that since the original distribution is normal then the Central Limit Theorem applies despite the small sample size of 15. This means that the distribution of $\bar{x}$ is normal and has a mean of $\mu_{\bar{x}} = \mu = 98.6^{\circ}\text{F}$ and standard deviation of $\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}} = \frac{0.62}{\sqrt{15}} = 0.16$. Therefore, converting to *z*-scores and using the *z*-table yields

$$P(98.4 \leq \bar{x} \leq 98.8) = P\left(\frac{98.4 - 98.6}{0.16} \leq z \leq \frac{98.8 - 98.6}{0.16}\right) = P(-1.25 \leq z \leq 1.25) = 0.7888$$

In other words, there is about a 79% chance that the average body temperature will be in the given range.

Written Assignment

Do problems 7.8, 7.25, 7.31, 7.47, 7.49, 7.53, 7.59

You are now ready to take the Unit 2 test. Please send in the test, with the attached assignment cover, as your assignment for Unit 2.

Unit 2 Test

Use only hand calculators on this exam; Minitab or any other computer software is **not** allowed. The textbook (but no other notes) may be used during the exam. Be sure to submit all your work in order to receive partial credit.

1. The United States National Center for Education Statistics compiles enrollment data on American public schools and reports the information in Digest of Education Statistics. The following table displays a frequency distribution for the enrollment by grade level in public secondary schools for a given year. Frequencies are in thousands.

Grade ***x***	**Frequency** ***f***
9	3604
10	3131
11	2749
12	2488
Total	

   a. Determine the probability distribution of the random variable ***x***. (Hint: complete the table by finding relative frequencies.)
   b. Construct a probability histogram for the random variable ***x***.
   c. What is the probability that a randomly selected student is in either 9th or 10th grade?
   e. What is the probability that a randomly selected student is in at most 11th grade?
   f. Find the mean and standard deviation of the random variable ***x***.

2. Shaquille O'Neal is a popular basketball player, but is not recognized as a great free throw shooter. His career free throw average is 0.532 (in other words he has made approximately 53.2% of his free throws). Suppose in a particular game that Shaq takes 5 free throws.

   a. What type of probability distribution is this? Explain.
   b. What are the mean and standard deviation for the number of made free throws?
   c. What is the probability that he will miss 4 of the next 5 free throws he takes?

3. Consider two normally distributed random variables, ***X*** and ***Y***, such that

$$\mu_X = \mathbf{2} \quad \textbf{and} \quad \sigma_X = \mathbf{5}$$

$$\mu_Y = \mathbf{2} \quad \textbf{and} \quad \sigma_Y = \mathbf{1}$$

In other words, both variables have the same mean but different standard deviations. Draw rough sketches of the two normal curves on the same graph. **Be sure to label your curves.**

4. Determine the area under the standard normal curve corresponding to the following regions. **Be sure to draw a picture showing the specified region**.

   a. the region to the left of $z = \mathbf{1.3}$
   b. the region between $z = \mathbf{-1.42}$ and $z = \mathbf{.07}$
   c. the region to the right of $z = \mathbf{2.45}$
   d. the region between $z = \mathbf{-1.96}$ and $z = \mathbf{1.96}$

5. The average length of stay in a chronic disease hospital for a certain type of patient is 60 days with a standard deviation of 15. Suppose it is reasonable to assume an approximately normal distribution of lengths of stay.

   a. What percentage of patients stay less than 50 days?
   b. Fill in the blanks. The probability is 0.9534 that a patient will stay between

      _____ and _____ days.

   (Hint: to start, find two *z*-scores equidistant from the center such that the area between them is 0.9534.)

6. In the study of fingerprints, an important quantitative characteristic is the total ridge count for the 10 fingers of the individual. Suppose that the total ridge counts of individuals in a certain population are approximately normally distributed with a mean of 140 and a standard deviation of 50.

   a. Find the probability that an individual picked at random from this population will have a ridge count of 200 or more.
   b. In a population of 10,000 people how many would you expect to have a ridge count of 200 or more?

7. The National Health and Nutrition Examination Survey of 1976-80 found that the mean serum cholesterol level for U.S. males aged 20-74 years was 211. The standard deviation was approximately 90. Consider the **sampling distribution of the sample mean** based on samples of size 100 drawn from this population of males.

   a. What are the mean and standard deviation of the sampling distribution?
   b. Is it necessary for the original distribution to be normal? Explain.
   c. Suppose a random sample of 100 is taken, what is the probability that the mean serum cholesterol level will be between 198 and 220?