Lesson 5: Probability Distributions

Introduction

Learning objectives for this lesson

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

- distinguish between discrete and continuous random variables
- explain the difference between population, parameter, sample, and statistic
- determine if a given value represents a population parameter or sample statistic
- find probabilities associated with a discrete probability distribution
- compute the mean and variance of a discrete probability distribution
- find probabilities associated with a binomial distribution
- find probabilities associated with a normal probability distribution using the standard normal table
- determine the standard error for the sample proportion and sample mean
- apply the Central Limit Theorem properly to a set of continuous data

Random Variables

A **random variable** is numerical characteristic of each event in a sample space, or equivalently, each individual in a population.

Examples:

- The number of heads in four flips of a coin (a numerical property of each different sequence of flips).
- Heights of individuals in a large population.

Random variables are classified into two broad types

- A **discrete random variable** has a countable set of distinct possible values.
- A **continuous random variable** is such that any value (to any number of decimal places) within some interval is a possible value.

Examples of discrete random variable:

- Number of heads in 4 flips of a coin (possible outcomes are 0, 1, 2, 3, 4).
- Number of classes missed last week (possible outcomes are 0, 1, 2, 3, ..., up to some maximum number)
- Amount won or lost when betting $1 on the Pennsylvania Daily number lottery

Examples of continuous random variables:

- Heights of individuals
- Time to finish a test

- Hours spent exercising last week.

Note : In practice, we don't measure accurately enough to truly see all possible values of a continuous random variable. For instance, in reality somebody may have exercised 4.2341567 hours last week but they probably would round off to 4. Nevertheless, hours of exercise last week is inherently a continuous random variable.

Probability Distributions: Discrete Random Variables

For a discrete random variable, its **probability distribution** (also called the probability distribution function) is any table, graph, or formula that gives each possible value and the probability of that value. **Note** : The total of all probabilities across the distribution must be 1, and each individual probability must be between 0 and 1, inclusive.

Examples:

(1) Probability Distribution for Number of Heads in 4 Flips of a coin

Heads	0	1	2	3	4
Probability	1/16	4/16	6/16	4/16	1/16

This could be found be listing all 16 possible sequences of heads and tails for four flips, and then counting how many sequences there are for each possible number of heads.

(2) Probability Distribution for number of tattoos each student has in a population of students

Tattoos	0	1	2	3	4
Probability	0.850	0.120	0.015	0.010	0.005

This could be found be doing a census of a large student population.

Cumulative Probabilities

Often, we wish to know the probability that a variable is less than or equal to some value. This is called the **cumulative probability** because to find the answer, we simply add probabilities for all values qualifying as "less than or equal" to the specified value.

Example: Suppose we want to know the probability that the number of heads in four flips is 1 or less. The qualifying values are 0 and 1, so we add probabilities for those two possibilities.

P(number of heads = 2) = P(number of heads = 0) + P (number of heads = 1) = (1/16)+(4/16) = 5/16

The **cumulative distribution** is a listing of all possible values along with *the cumulative probability* for each value

Examples:

(1) Probability Distribution and Cumulative Distribution for Number of Heads in 4 Flips

Heads	0	1	2	3	4
Probability	1/16	4/16	6/16	4/16	1/16
Cumulative Probability	1/16	5/16	11/16	15/16	1

Each cumulative probability was found by adding probabilities (in second row) up to the particular column of the table. As an example, for 2 heads, we add probabilities for 0, 1, and 2 heads to get 11/16. This is the probability the number of heads is two or less.

(2) Probability Distribution and Cumulative Distribution for number of tattoos each student has in a population of students

Tattoos	0	1	2	3	4
Probability	0.850	0.120	0.015	0.010	0.005
Cumulative Probability	0.850	0.970	0.985	0.995	1

As an example, probability a randomly selected student has 2 or fewer tattoos = =0.985 (calculated as 0.850+0.120+0.015).

Mean, also called Expected Value, of a Discrete Variable

The phrase **expected value** is a synonym for **mean** value in the long run (meaning for many repeats or a large sample size). For a discrete random variable, the calculation is Sum of (value× probability) where we sum over all values (after separately calculating value× probability for each value), expressed as:

E(X) = $\sum x_i p_i$, meaning we take each observed X value and multiply it by its respective probability. We then add these products to reach our expected value labeled E(X). [NOTE: the letter X is a common symbol used to represent a random variable. Any letter can be used.]

Example : A fair six-sided die is tossed. You win $2 if the result is a “1”, you win $1 if the result is a “6” but otherwise you lose $1.

The probability distribution for X = amount won or lost is

X	+2	+1	-1
Probability	1/6	1/6	4/6

Expected Value = $(2 \times \frac{1}{6}) + (1 \times \frac{1}{6}) + (-1 \times \frac{4}{6})$ = -1/6 = -$0.17.

The interpretation is that if you play many times, the average outcome is losing 17 cents per play.

Example : Using the probability distribution for number of tattoos given above (not the cumulative!),

The mean number of tattoos per student is

Expected Value = (0 ×0.85) + (1 ×0.12) + (2×0.015) + (3×0.010) + (4×0.005) = 0.20.

Standard Deviation of a Discrete Variable

Knowing the expected value is not the only important characteristic one may want to know about a set of discrete numbers: one may also need to know the spread, or variability, of these data. For instance, you may "expect" to win $20 when playing a particular game (which appears good!), but the spread for this might be from losing $20 to winning $60. Knowing such information can influence you decision on whether to play.

To calculate the standard deviation we first must calculate the variance. From the variance, we take the square root and this provides us the standard deviation. Your book provides the following formula for calculating the variance:

$$\sigma^2 = \sum (x_i - u)^2 p_i$$ and the standard deviation is: $$\sigma = \sqrt{\sum (x_i - u)^2 p_i}$$

In this expression we substitute our result for E(X) into *u* , and *u* is simply the symbol used to represent the mean of some population .

However, an **easier** formula to use and remember for calculating the standard deviation is the following:

$$\sigma^2 = \sum x_i^2 p_i - u^2$$ and again we substitute E(X) for μ.

The standard deviation is then found by taking the square root of the variance. Notice in the summation part of this equation that we only square each observed X value and **not** the respective probability.

Example : Going back to the first example used above for expectation involving the die, we would calculate the standard deviation for this discrete distribution by first calculating the variance:

$$\sigma^2 = \sum x_i^2 p_i - u^2 = (2^2 \mathbf{x} \frac{1}{6}) + (1^2 \mathbf{x} \frac{1}{6}) + (-1^2 \mathbf{x} \frac{4}{6}) - (-\frac{1}{6})^2 = \frac{4}{6} + \frac{1}{6} + \frac{4}{6} - \frac{1}{36} = \frac{53}{36} = 1.472$$

So the standard deviation would be the square root of 1.472, or 1.213

Binomial Random Variable

This is a specific type of discrete random variable. A binomial random variable counts how often a particular event occurs in a fixed number or tries. For a variable to be a binomial random variable, these conditions must be met:

- There are a fixed number of trials (a fixed sample size).
- On each trial, the even of interest either occurs or does not.
- The probability of occurrence (or not) is the same on each trial.
- Trials are independent of one another.

Examples of binomial random variables:

- Number of correct guesses at 30 true-false questions when you randomly guess all answers

- Number of winning lottery tickets when you buy 10 tickets of the same kind
- Number of left-handers in a randomly selected sample of 100 unrelated people

Notation

n = number of trials (sample size)

p = probability event of interest occurs on any one trial

Example : For the guessing at true questions example above, n = 30 and p = .5 (chance of getting any one question right).

Probabilities for binomial random variables

The conditions for being a binomial variable lead to a somewhat complicated formula for finding the probability any specific value occurs (such as the probability you get 20 right when you guess as 20 True-False questions.)

We'll use Minitab to find probabilities for binomial random variables. Don't worry about the “by hand” formula. However, for those of you who are curious, the by hand formula for the probability of getting a specific outcome in a binomial experiment is:

$$P(x) = \frac{n!}{x!(n-x)!} p^x (1-p)^{n-x}$$

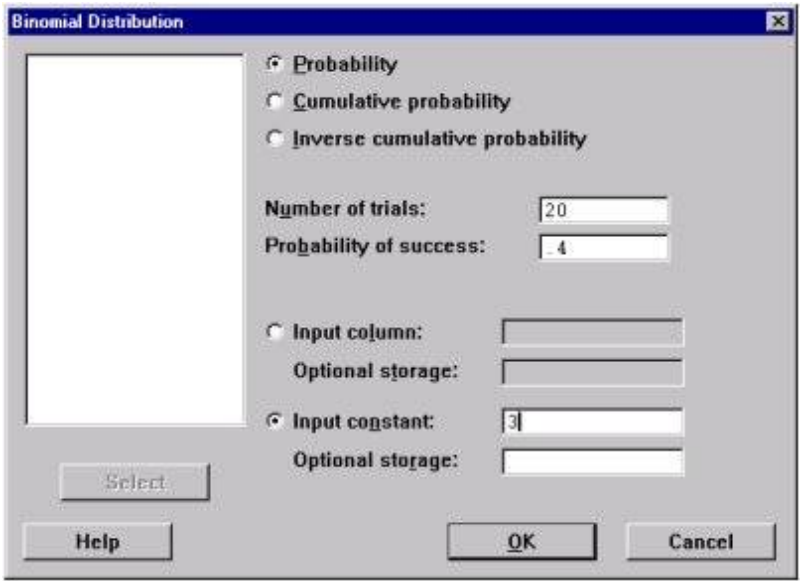
Evaluating the Binomial Distribution

One can use the formula to find the probability or alternatively, use Minitab to find the probability. In the homework, you may use the one that you are more comfortable with unless specified otherwise.

Example Minitab: Using Minitab, find $P(x)$ for n = 20, x =3, and □ = 0.4.

Calc > Probability Distributions > Binomial

Choose Probability since we want to find the probability x = 3. Choose input constant and type in 3 since that is the value you want to evaluate the probability at. {NOTE: The following graphic is from Minitab Version 14. If using Version 15, Probability of Success has been edited to Event Probability.



Minitab output:

Probability Density Function

Binomial with $n = 20$ and $p = 0.4$

x	$P(X = x)$
3.00	0.0123

In the following example, we illustrate how to use the formula to compute binomial probabilities. If you don't like to use the formula, you can also just use Minitab to find the probabilities.

Example by hand: Cross-fertilizing a red and a white flower produces red flowers 25% of the time. Now we cross-fertilize five pairs of red and white flowers and produce five offspring.

Find the probability that:

a. There will be no red flowered plants in the five offspring.

X = # of red flowered plants in the five offspring. Here, the number of red flowered plants has a binomial distribution with n = 5, p = 0.25.

$$P(X = 0) = \frac{5!}{0!(5-0)!} p^0 (1-p)^5 = 1\,(0.25)^0\,(0.75)^5 = 0.237$$

b. **Cumulative Probability** There will less than two red flowered plants.

Answer:

$P(X$ is 1 or less) = $P(X = 0) + P(X = 1)$ =

$$\frac{5!}{0!(5-0)!}0.25^0(1-0.25)^5+\frac{5!}{1!(5-1)!}0.25^1(1-0.25)^4$$

=0.237 + 0.395 = 0.632

In the previous example, part a was finding the P(X = x) and part b was finding P(X <= x). This latter expression is called finding a **cumulative probability** because you are finding the probability that has accumulated from the minimum to some point, i.e. from 0 to 1 in this example

To use Minitab to solve a cumulative probability binomial problem, return to Calc > Probability Distributions > Binomial as shown above. Now however, select the radio button for Cumulative Probability and then enter the respective Number of Trials (i.e. 5), Event Probability (i.e. 0.25), and click the radio button for Input Constant and enter the x-value (i.e. 1).

Expected Value and Standard Deviation for Binomial random variable

The formula given earlier for discrete random variables could be used, but the good news is that for binomial random variables a shortcut formula for expected value (the mean) and standard deviation are:

Expected Value = *np* Standard Deviation = $\sqrt{np(1-p)}$

After you use this formula a couple of times, you'll realize this formula matches your intuition. For instance, the “expected” number of correct (random) guesses at 30 True-False questions is *np* = (30)(.5) = 15 (half of the questions). For a fair six-sided die rolled 60 times, the expected value of the number of times a “1” is tossed is *np* = (60)(1/6) = 10. The standard deviation for both of these would be, for the True-False test $\sqrt{(30)(0.5)(1-0.5)}=\sqrt{7.5}=2.74$ and for the die $\sqrt{(60)(\frac{1}{6})(1-\frac{1}{6})}=\sqrt{\frac{50}{6}}=2.89$

Probability Distributions: Continuous Random Variable

Density Curves

Previously we discussed discrete random variables, and now we consider the contuous type. A **continuous random variable** is such that all values (to any number of decimal places) within some interval are possible outcomes. A continuous random variable has an infinite number of possible values so we can't assign probabilities to each specific value. If we did, the total probability would be infinite, rather than 1, as it is supposed to be

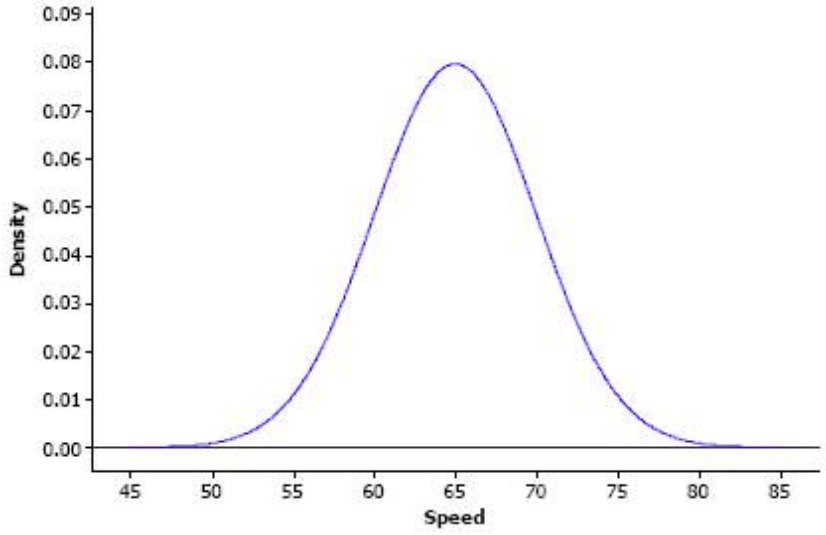
To describe probabilities for a continuous random variable, we use a *probability density function.* A **probability density function** is a curve such that the area under the curve within any interval of values along the horizontal gives the probability for that interval.

Normal Random Variables

The most commonly encountered type of continuous random variable is a **normal random variable** ,

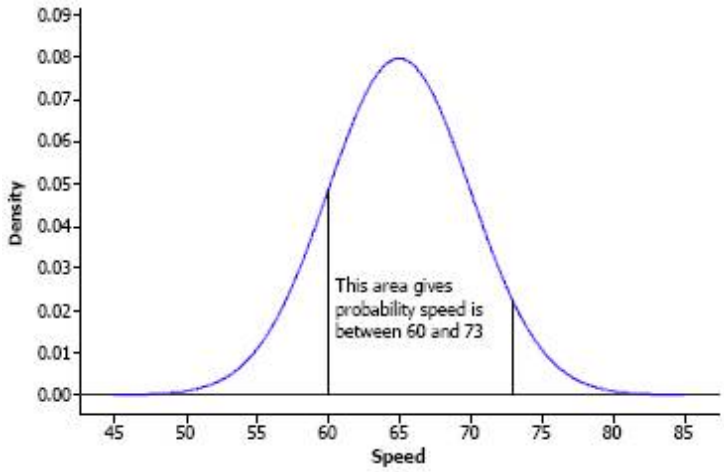
which has a symmetric bell-shaped density function. The center point of the distribution is the mean value, denoted by μ (pronounced "mew"). The spread of the distribution is determined by the variance, denoted by σ^2 (pronounced "sigma squared") or by the square root of the variance called standard deviation, denoted by σ (pronounced "sigma").

Example : Suppose vehicle speeds at a highway location have a normal distribution with mean $\mu = 65$ mph and standard deviation $s = 5$ mph. The probability density function is shown below. Notice that the horizontal axis shows speeds and the bell is centered at the mean (65 mph).



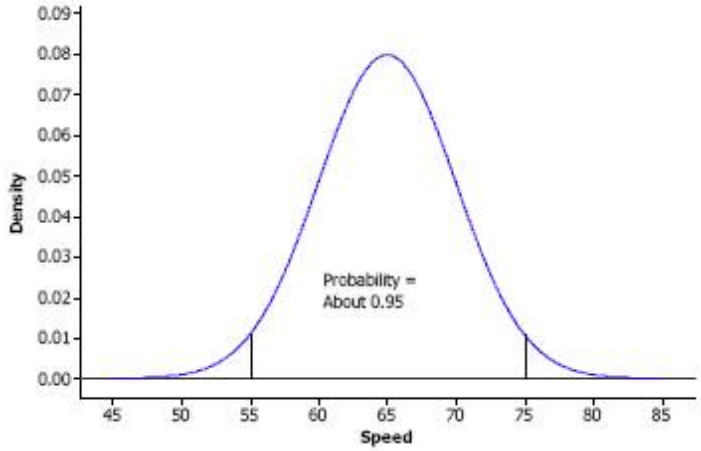
Probability for an Interval = Area under the density curve in that interval

The next figure shows the probability that the speed of a randomly selected vehicle will be between 60 and 73 mile per hour, with this probability equal to the area under the curve between 60 and 73.



Empirical Rule Review

Recall that our first lesson we learned that for bell-shaped data, about 95% of the data values will be in the interval *mean* ± (2 × *std. dev*) . In our example, this is 65 ± (2 × 5), or 55 to 75. The next figure shows that the probability is about 0.95 (about 95%) that a randomly selected vehicle speed is between 55 and 75.



The Empirical Rule also stated that about 99.7% (nearly all) of a bell-shaped dataset will be in the interval *mean* ± (3 × *std. dev*) . This is 65 ± (3 × 5), or 50 to 80 for example. Notice that this interval roughly gives the complete range of the density curve shown above.

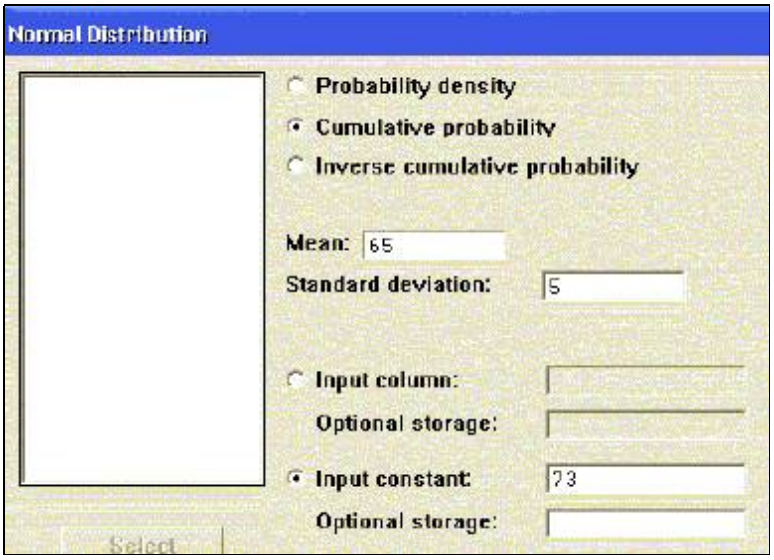
Finding Probabilities for a Normal Random Variable

Remember that the **cumulative probability** for a value is the probability less than or equal to that value. Minitab, Excel, and the TI-83 series of calculators will give the *cumulative probability* for any value of interest in a specific normal curve.

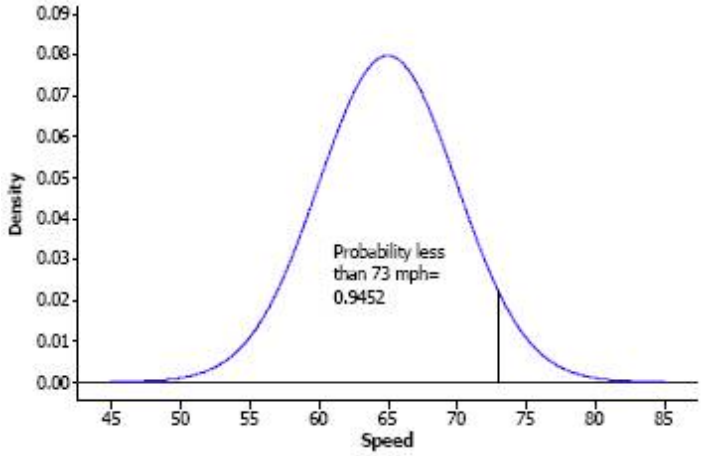
For our example of vehicle speeds, here is Minitab output showing that the probability = 0.9542 that the speed of a randomly selected vehicle is less than or equal to 73 mph.

```
Normal with mean = 65 and standard deviation = 5
 x       P( X <= x )
73         0.9452
```

To find this probability, use **Calc>Probability Distribution> Normal**, specify the mean and standard deviation and enter the value of interest as "Input Constant." Here's what it looks like for our example.



Here is a figure that illustrates the cumulative probability we found using this procedure.



"Greater than" Probabilities

Sometimes we want to know the probability that a variable has a value **greater than** some value. For instance, we might want to know the probability that a randomly selected vehicle speed is greater than 73 mph, written P(X > 73).

For our example, probability speed is greater than 73 = 1 - 0.9452 = 0.0548.

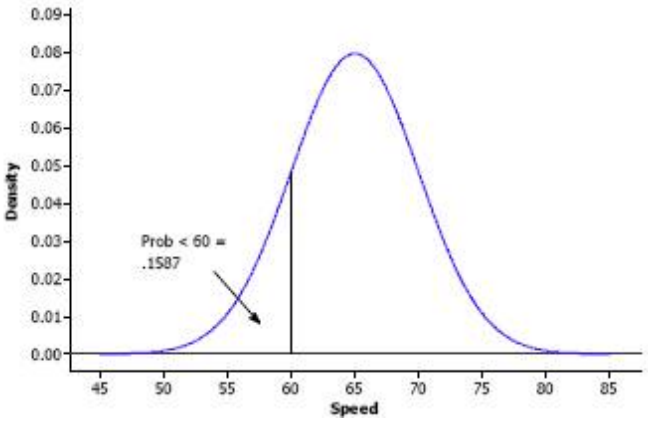
- The general rule for a "greater than" situation is

    P (greater than a value) = 1 - P(less than or equal to the value)

Example : Using Minitab we can find that the probability = 0.1587 that a speed is less than or equal to 60 mph. Thus the probability a speed is greater than 60 mph = 1 - 0.1587 = 0.8413.

The relevant Minitab output and a figure showing the cumulative probability for 60 mph follows:

```
Normal with mean = 65 and standard deviation = 5
x     P( X <= x )
60      0.1587
```



"In between" Probabilities

Suppose we want to know the probability a normal random variable is **within** a specified interval. For instance, suppose we want to know the probability a randomly selected speed is between 60 and 73 mph. The simplest approach is to subtract the cumulative probability for 60 mph from the cumulative probability for 73. The answer is

Probability speed is between 60 and 73 = 0.9452 − 0.1587 = 0.7875.

This can be written as P(60 < X < 73) = 0.7875, where X is speed.

- The general rule for an "in between" probability is

P(between *a* and *b*) = cumulative probability for value *b* − cumulative probability for value *a*

Finding Cumulative Probabilities

Using the Standard Normal Table in the appendix of textbook or see a copy at Standard Normal Table

Table A.1 in the textbook gives normal curve cumulative probabilities for standardized scores.

- A **standardized score** (also called *z*-score) is $z = \frac{\text{value} - \text{mean}}{\text{s.d.}} = \frac{x - \mu}{\sigma}$.
- Row labels of Table A.1 give possible *z*-scores up to one decimal place. The column labels give the second decimal place of the *z*-score.

The cumulative probability for a value equals the cumulative probability for that value's *z*-score. Here, probability speed less than or equal 73 mph = probability z-score less than or equal 1.60. How did we arrive at this z-score?

Example

In our vehicle speed example, the standardized scores for 73 mph is

$$z = \frac{73 - 65}{5} = 1.60.$$

We look in the ".00" column of the "1.6" row (1.6 **plus** .00 equals 1.60) to find that the cumulative probability for z = 1.60 is 0.9452, the same value we got earlier as the cumulative probability for speed = 73 mph.

z	*.00*	*.01*	*.02*	*.03*	*.04*	*.05*	*.06*	*.07*	*.08*	*.09*
1.6	.9452	.9463	.9474	.9484	.9495	.9505	.9515	.9525	.9535	.9545

Example

For speed = 60 the z-score is

$$z = \frac{60 - 65}{5} = -1.00.$$

Table A.1 gives this information:

z	*.00*	*.01*	*.02*	*.03*	*.04*	*.05*	*.06*	*.07*	*.08*	*.09*
−1.0	.1587	.1562	.1539	.1515	.1492	.1469	.1446	.1423	.1401	.1379

The cumulative probability is .1587 for *z* = -1.00 and this is also the cumulative probability for a speed of 60 mph.

Example

Suppose pulse rates of adult females have a normal curve distribution with mean μ =75 and standard deviation s = 8. What is the probability that a randomly selected female has a pulse rate **greater than 85** ? *Be careful* ! Notice we want a "greater than" and the interval we want is entirely above average, so we know the answer must be less than 0.5.

If we use Table A.1, the first step is to calculate a z-score of 85.

$$z = \frac{85 - 75}{8} = 1.25$$

Information from Table A.1 is

z	*.00*	*.01*	*.02*	*.03*	*.04*	*.05*	*.06*	*.07*	*.08*	*.09*
1.2	.8849	.8869	.8888	.8907	.8925	.8944	.8962	.8980	.8997	.9015

Use the "05" column to find that the cumulative probability for z = 1.25 is 0.8944.

This is not yet the answer. This is the probability the pulse is less than or equal to 85. We want a greater than probability so the answer is

P(greater than 85) = 1 - P(less than or equal 85) = 1 − 0.8944 = **0.1056.**

Finding Percentiles

We may wish to know the value of a variable that is a specified percentile of the values.

- We might ask what speed is the 99.99 th percentile of speeds at the highway location in our earlier example.
- We might want to know what pulse rate is the 25 th percentile of pulse rates.

In Minitab, we can find percentiles using the Calc>Probability Distributions> Normal but we have to make two changes to what we did before. (1) Click on the "Inverse Cumulative Probability" radio button (rather than cumulative probability) and (2) enter the percentile ranking as a decimal fraction in the "Input Constant" box.

- The 99.99 th percentile of speeds (when mean = 65 and standard deviation = 5) is about 83.6 mph. Output from Minitab follows. Notice that now the specified cumulative probability is given first, and then the corresponding speed.

```
Normal with mean = 65 and standard deviation = 5
  P( X <= x )         x
     0.9999     83.5951
```

- The 25 th percentile of pulse rates (when μ = 75 and s= 8) is about 69.6. Relevant Minitab output is

```
Normal with mean = 75 and standard deviation = 8
    P( X <= x )         x
       0.25          69.6041
```

Normal Approximation to the Binomial

Remember binomial random variables from last week's discussion? A binomial random variable can also be approximated by using normal random variable methods discussed above. This approximation can take place as long as:

1. The population size must be **at least** 10 times the sample size.
2. $np = 10$ **and** $n(1-p) = 10$. [These constraints take care of population shapes that are unbalanced because p is too close to 0 or to 1.]

The mean of a binomial random variable is easy to grasp intuitively: Say the probability of success for each observation is 0.2 and we make 10 observations. Then on the average we should have 10 * 0.2 = 2 successes. The spread of a binomial distribution is not so intuitive, so we will not justify our formula for standard deviation.

If sample count X of successes is a binomial random variable for n fixed observations with probability of success p for each observation, then X has a mean and standard deviation as discussed in section 8.4 of:

Mean = np and standard deviation = $\sqrt{np(1-p)}$

And as long as the above 2 requirements are for n and p are satisfied, we can approximate X with a normal random variable having the same mean and standard deviation and use the normal calculations discussed previously in these notes to solve for probabilities for X.

Review of Finding Probabilities

Click on the Inspect icon for an audio/visual example for each situation described. When reviewing any of these examples keep in that they apply when:

1. The variable in question follows a normal, or bell-shaped, distribution

2. If the variable is not in standardized, then you need to standardized the value first by $z = \frac{\text{value} - \text{mean}}{\text{s.d.}} = \frac{x - \mu}{\sigma}$.

Finding "Less Than" Probability

Finding "Greater Than" Probability

Finding "Between" Probability

Inspect! Finding "Either / Or" Probability

Population Parameters and Sample Statistics

S1 - A survey is carried out at a university to estimate the proportion of undergraduates living at home during the current term. **Population:** undergraduates at the university **Parameter:** the true proportion of undergraduates that live at home **Sample:** the undergraduates surveyed **Statistic:** the proportion of the sampled students who live at home - used to estimate the true proportion

S2 - A study is conducted to find the average hours college students spend partying on the weekend. **Population:** all college students **Parameter:** the true mean number of hours college students spend partying on the weekend **Sample:** the students sampled for the study **Statistic:** the mean hours of weekend partying calculated from the sample

S1 is concerned about estimating a proportion *p* where *p* represents the true (typically unknown) parameter and $\hat{p}$ [pronounced "p-hat"] represents the statistic calculated from the sample

S2 is concerned about estimating a mean *u* where *u* [pronounced "mew"] represents the true (typically unknown) parameter and $\bar{x}$ [pronounce "x-bar"] represents the statistic calculated from the sample.

In either case the statistic is used to estimate the parameter. The statistic can vary from sample to sample, but the parameter is understood to be fixed.

The statistic, then, can take on various values depending on the result of repeated random sampling. The distribution of these possible values is known as the **sampling distribution.**

Overview of symbols

The following table of symbols provides some of the common notation that we will see through the remaing sections.

Parameter Names and Description	Symbol for the Population Parameter	Symbol for the Sample Statistic
For Categorical Variables:		
One population proportion (or probability)	p	$\hat{p}$
Difference in two population proportions	$p_1 - p_2$	$\hat{p}_1 - \hat{p}_2$
For Quantitative Variables:		
One population mean	μ	$\bar{x}$
Population mean of paired differences (dependent or paired)	μ_d	$\bar{d}$
Difference in two population means (independent)	$\mu_1 - \mu_2$	$\bar{x}_1 - \bar{x}_2$

The difference between "paired" samples and "independent" samples can be most easily explained by the situation where the observations are taken on the same individual (e.g. measure a person's stress level before and after an exam) where independent would consist of taking observations from two distinct groups

(e.g. measure the stress levels of men and women before an exam and compare these stress levels). An exception to this a situation that involves analyzing spouses. In such cases, spousal data is often linked as paired data.

Sampling Distributions of Sample Statistics

Two common statistics are the sample proportion, $\hat{p}$, (read as "pi-hat") and sample mean, $\overline{x}$, (read as "x-bar"). Sample statistics are random variables and therefore vary from sample to sample. For instance, consider taking two random samples, each sample consisting of 5 students, from a class and calculating the mean height of the students in each sample. Would you expect both sample means to be exactly the same? As a result, sample statistics also have a distribution called the **sampling distribution**. These sampling distributions, similar to distributions discussed previuosly, have a mean and standard deviation. However, we refer to the standard deviation of a sampling distribution as the **standard error**. Thus, the standard error is simply the standard deviation of a sampling distribituion. Often times people will interchange these two terms. This is okay as long as you understand the distinction between the two: standard error refers to *sampling* distributions and standard deviation refes to *probability* distributions.

Sampling Distributions for Sample Proportion, $\hat{p}$

If numerous repetitions of samples are taken, the distribution of $\hat{p}$ is said to approximate a normal curve distribution. Alternatively, this can be assumed if BOTH *n**p and *n* *(1 - p) are **at least** 10. [**SPECIAL NOTE:** Some textbooks use 15 instead of 10 believing that 10 is to liberal. We will use 10 for our discussions.] Using this, we can estimate the true population proportion, p, by $\hat{p}$ and the true standard deviation of p by s.e.($\hat{p}$) = $\sqrt{\frac{p(1-p)}{n}}$, where s.e.($\hat{p}$) is interpreted as the **standard error of** $\hat{p}$

Probabilities about the number X of successes in a binomial situation are the same as probabilities about corresponding proportions.

In general, if *np* >= 10 and *n*(1- *p*) >= 10, the sampling distribution of $\hat{p}$ is about normal with mean of p and standard error SE($\hat{p}$) = $\sqrt{\frac{p(1-p)}{n}}$.

Example. Suppose the proportion of all college students who have used marijuana in the past 6 months is *p* = .40. For a class of size *N* = 200, representative of all college students on use of marijuana, what is the chance that the proportion of students who have used mj in the past 6 months is less than .32 (or 32%)?

Solution. The mean of the sample proportion $\hat{p}$ is p and the standard error of $\hat{p}$ is SE($\hat{p}$) = $\sqrt{\frac{p(1-p)}{n}}$.

For this marijuana example, we are given that p = .4. We then determine **SE($\hat{p}$)** = $\sqrt{\frac{p(1-p)}{n}}$ = $\sqrt{\frac{p(1-p)}{n}}$

= $\sqrt{\frac{(.4)(1-.4)}{200}}$ = 0.0346

So, the sample proportion $\hat{p}$ is about normal with mean p = .40 and SE($\hat{p}$) = 0.0346.

The z-score for .32 is z = (.32 - .40) / 0.0346 = -2.31. Then using Standard Normal Table

Prob($\hat{p}$ < .32) = Prob(Z <. -2.31) = 0.0104.

Question to ponder: If you observed a sample proportion of .32 would you believe a claim that 40% of college students used mj in the past 6 months? Or would you think the proportion is less than .40?

Sampling Distribution of the Sample Mean $\bar{x}$

The **central limit theorem** states that if a large enough sample is taken (typically $n > 30$) then the sampling distribution of $\bar{x}$ is approximately a normal distribution with a mean of μ and a standard deviation of $\frac{\sigma}{\sqrt{n}}$.

Since in practice we usually do not know μ or σ we estimate these by $\bar{x}$ and $\frac{s}{\sqrt{n}}$ respectively. In this case s is the estimate of σ and is the standard deviation of the sample. The expression $\frac{s}{\sqrt{n}}$ is known as the standard error of the mean, labeled **s.e.($\bar{x}$)**

Simulation: Generate 500 samples of size heights of 4 men. Assume the distribution of male heights is normal with mean m = 70" and standard deviation s = 3.0". Then find the mean of each of 500 samples of size 4.

Here are the first 10 sample means:

70.4 72.0 72.3 69.9 70.5 70.0 70.5 68.1 69.2 71.8

Descriptive Statistics: xbars for Samples of Size 4:

Variable	N	Mean	Median	TrMean	StDev	SE Mean
xbars	500	**69.957**	69.968	69.946	**1.496**	0.067

Variable	Minimum	Maximum	Q1	Q3
xbars	**65.838**	**74.410**	68.958	70.927

Theory says that the mean of ($\bar{x}$) = μ = **70** which is also the Population Mean and SE($\bar{x}$) = $\frac{\sigma}{\sqrt{n}} = \frac{3}{\sqrt{4}} =$ **1.50.**

Simulation shows: Average (500 $\bar{x}$'s) = **69.957** and SE(of 500 $\bar{x}$'s) = **1.496**

Change the sample size from n = 4 to n = 25 and get descriptive statistics:

Descriptive Statistics: xbars for Samples of size 25:

Variable	N	Mean	Median	TrMean	StDev	SE Mean
xbars25	500	69.983	69.971	69.987	0.592	0.026

Variable	Minimum	Maximum	Q1	Q3
xbars25	67.938	71.817	69.559	70.426

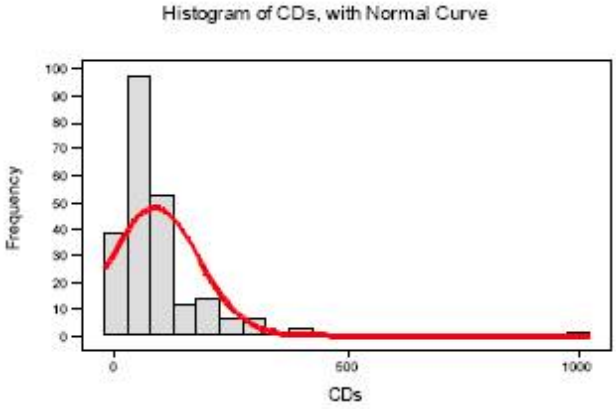
Theory says that the mean of ($\bar{x}$) = μ = **70** which is also the Population Mean and SE($\bar{x}$) = $\frac{\sigma}{\sqrt{n}} = \frac{3}{\sqrt{25}} =$

0.60.

Simulation shows: Average (500 $\bar{x}$'s) = **69.983** and SE(of 500 $\bar{x}$'s) = **0.592**

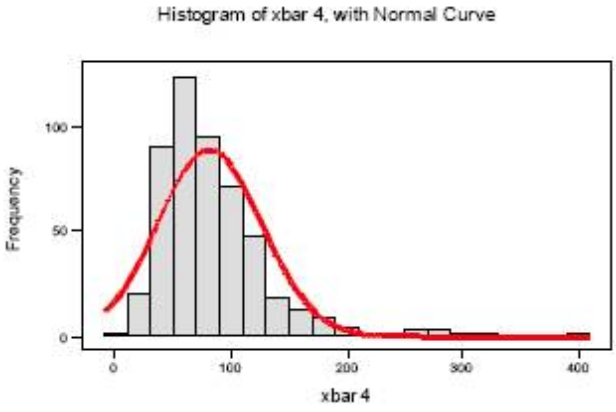
Sampling Distribution of Sample Mean $\bar{x}$ from a Non-Normal Population

Simulation: Below is a Histogram of Number of Cds Owned by PSU Students. The distribution is strongly skewed to the right.



Assume the Population Mean Number of CDs owned is μ = 84 and s = 96

Let's obtain 500 samples of size 4 from this population and look at the distribution of the 500 x-bars:

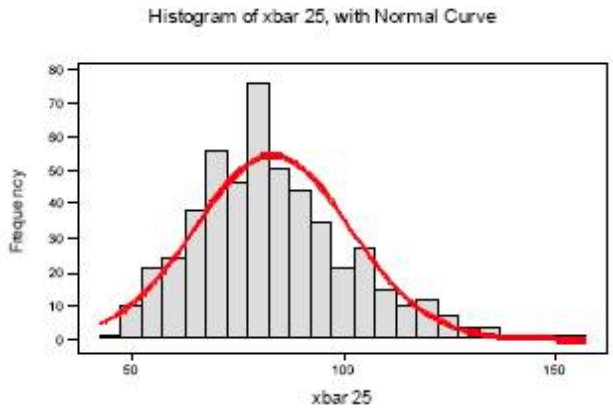


Variable	N	Mean	StDev
xbar 4	500	81.11	45.12

Theory says that the mean of ($\bar{x}$) = μ = **84** which is also the Population Mean the SE($\bar{x}$ **) = 48 =** $\frac{84}{\sqrt{4}}$

Simulation shows Average(500 $\bar{x}$'s) = **81.11** and SE(500 $\bar{x}$'s for samples of size 4) = **45.1**

Change the sample size from *n* = 4 to *n* = 25 and get descriptive statistics and curve:



Variable	N	Mean	StDev
xbar 25	500	83.281	18.268

Theory says that the mean of ($\bar{x}$) = μ = **84** which is also the Population Mean and the SE($\bar{x}$ **)** = $\frac{96}{\sqrt{100}}$ = **19.2** Simulation shows Average(500 $\bar{x}$'s) = 83.281 and SE(500 $\bar{x}$'s for samples of size 25) = 18.268. A histogram of the 500 $\bar{x}$'s computed from samples of size 25 is beginning to look a lot like a normal curve.

i. The Law of Large Numbers says that as the sample size increases the sample mean will approach the population mean.

ii. The Central Limit Theorem says that as the sample size increases the sampling distribution of $\bar{X}$ (read x-bar) approaches the normal distribution. We see this effect here for *n* = 25. Generally, we assume that a sample size of *n* = 30 is sufficient to get an approximate normal distribution for the distribution of the sample mean.

iii. The Central Limit Theorem is important because it enables us to calculate probabilities about sample means.

Example. Find the approximate probability that the average number of CDs owned when 100 students are asked is between 70 and 90.

Solution. Since the sample size is greater than 30, we assume the sampling distribution of $\bar{x}$ is about normal with mean *m* = 84 and SE($\bar{x}$) = $\frac{\sigma}{\sqrt{n}} = \frac{96}{\sqrt{100}}$ = 9.6. We are asked to find Prob(70 < $\bar{X}$ < 90). The z-scores for the two values are

for 90: z = (90 - 84)/ 9.6 = 0.625 and for 70: z = (70-84)/9.6 = -1.46. From tables of the normal distribution we get P(-1.46 < Z < 0.625) = .734 - .072 = .662.

Suppose the sample size was 1600 instead of 100. Then the distribution of $\bar{x}$ would be about normal with mean 84 and standard deviation $\frac{\sigma}{\sqrt{n}} = \frac{96}{\sqrt{1600}}$ = 96 / 40 = 2.4. From the empirical rule we know that almost all x-bars for samples of size 1600 will be in the interval

84 ± (3)(2.4) or in the interval 84 ± 7.2 or between 76.8 and 91.2. The Law of Large Numbers says that as we increase the sample size the probability that the sample mean approaches the population mean is 1.00!

APPLET

Here is an applet developed by the folks at Rice University that simulates "sampling distribution". The object here is to give you a chance to explore various aspects of sampling distributions. When the applet begins, a histogram of a normal distribution is displayed at the topic of the screen.

The distribution portrayed at the top of the screen is the population from which samples are taken. The mean of the distribution is indicated by a small blue line and the median is indicated by a small purple line. Since the mean and median are the same, the two lines overlap. The red line extends from the mean one standard deviation in each direction. Note the correspondence between the colors used on the histogram and the statistics displayed to the left of the histogram.

The second histogram displays the sample data. This histogram is initially blank. The third and fourth histograms show the distribution of statistics computed from the sample data. The number of samples (replications) that the third and fourth histograms are based on is indicated by the label "Reps=."

Basic Operation

The simulation is set to initially sample five numbers from the population, compute the mean of the five numbers, and plot the mean. Click the "Animated sample" button and you will see the five numbers appear in the histogram. The mean of the five numbers will be computed and the mean will be plotted in the third histogram. Do this several times to see the distribution of means begin to be formed. Once you see how this works, you can speed things up by taking 5, 1,000, or 10,000 samples at a time.

Notice that as you increase the sample size, regardless of the shape you create, the distribution (i.e. look at the histogram) becomes more bell-shaped. This is the theoretical meaning behind the central limit theorem: as sample size increases, then despite that the population from which the sample originated is not normal (e.g. uniform or chi-square), the **sample mean** will approximate a normal distribution

---

Review of Sampling Distributions

In later part of the last lesson we discussed finding the probability for a continuous random variable that followed a normal distribution. We did so by converting the observed score to a standardized z-score and then applying Standard Normal Table. For example:

IQ scores are normally distributed with mean, μ, of 110 and standard deviation, σ, equal to 25. Let the random variable *X* be a randomly chosen score. Find the probability of a randomly chosen score exceeding a 100. That is, find P(X > 100). To solve,

$$\mathbf{P}(\mathbf{X} > 100) = \mathbf{P}(\frac{\mathbf{X} - 110}{25} > \frac{100 - 110}{25}) = \mathbf{P}(\mathbf{Z} > -0.40) = 1 - \mathbf{P}(\mathbf{Z} < 0.40) = 0.6554$$

But what about situations when we have more than one sample, that is the sample size is greater than 1? In practice, usually just one random sample is taken from a population of quantitative or qualitative values and the statistic $\overline{x}$ the sample mean or $\hat{p}$ the sample proportion, respectively, is measured - one time only. For

instance, if we wanted to estimate what proportion of PSU students agreed with the President's explanation to the rising tuition costs we would only take one random sample, of some size, and use this sample to make an estimate. We would not continue to take samples and make estimates as this would be costly and inefficient. For samples taken at random, sample mean {or sample proportion} is a **random variable**. To get an idea of how such a random variable behaves we consider this variable's **sampling distribution** which we discussed previously in this lesson.

Consider the population of possible rolls *X* for a single six-side die has a mean, μ, equal to 3.5 and a standard deviation, σ, equal to 1.7. [If you do not believe this recall our discussion of probabilities for discrete random variables. For the six-side die you have six possible outcomes each with the same 1/6 probability of being rolled. Applying your recent knowledge, calculate the mean and standard deviation and see what you get!] If we rolled the die twice, the sample mean, $\bar{x}$ of these two rolls can take on various values based on what numbers come up. Since these results are subject to the laws of chance they can be defined as a random variable. From the beginning of the semester we can apply what we learned to summarize distributions by its **center, spread, and shape**.

1. Sometimes the mean roll of 2 dice will be less than 3.5, other times greater than 3.5. It should be just as likely to get a lower than average mean that it is to get a higher than average mean, but the sampling distribution of the sample mean should be **centered** at 3.5.
2. For the roll of 2 dice, the sample mean could be **spread** all the way from 1 to 6 - think if two "1s" or two "6s" are tossed.
3. The most likely mean roll from the two dice is 3.5 - all combinations where the sum is 7. The lower and higher the mean rolls, the less likely they are to occur. So the **shape** of the distribution of the sample means from two rolls would take the form of a triangle.

If we increase the sample size, i.e. the number of rolls, to say 10, then this sample mean is also a random variable.

1. Sometimes the mean roll of 10 dice will be less than 3.5 and sometimes greater than 3.5. Similar to when we rolled the dice 2 times, the sample distribution of $\bar{x}$ for 10 rolls should be **centered** at 3.5.
2. For 10 rolls, the distribution of the sample mean would not be as **spread** as that for 2 rolls. Getting a "1" or a "6" on all 10 rolls will almost never occur.
3. The most likely mean roll is still 3.5 with lower or higher mean rolls getting progressively less likely. But now there is a much better chance of the for the sample mean of the 10 rolls to be close to 3.5, and a much worse chance for this sample mean to be near 1 or 6. Therefore, the **shape** of the sampling distribution for 10 rolls bulges at 3.5 and tapers off at either end - ta da! The shape looks bell-shaped or normal!

This die example illustrates the general result of the central limit theorem: regardless of the population distribution (the distribution for the die is called a *uniform* distribution because each outcome is equally likely) the distribution of the sample mean will approach normal as sample size increases and the sample mean, $\bar{x}$ has the following characteristics:

1. The distribution of $\bar{x}$ is centered at μ

2. The spread of $\bar{x}$ can be measured by its standard deviation, σ, equal to $\frac{\sigma}{\sqrt{n}}$.

Example

Assume women's heights are normally distributed with μ = 64.5 inches and σ = 2.5 inches. Pick *one* women at random. According to the Empirical Rule, the probability is:

68% that her height *X* is between 62 inches and 68 inches

95% that her height *X* is between 59.5 inches and 69.5 inches

99.7% that her height *X* is between 57 inches and 72 inches

Now pick a random sample of size 25 women. The sample mean height, $\bar{x}$ is normal with expected value (i.e. mean) of 64.5 inches and standard deviation, $\frac{\sigma}{\sqrt{n}}$, equal to 0.5. The probability is:

68% that their sample mean height $\bar{x}$ is between 64 inches and 65 inches

95% that their sample mean height $\bar{x}$ is between 63.5 inches and 65.5 inches

99.7% that their sample mean height $\bar{x}$ is between 63 inches and 66 inches

Using Standard Normal Table for more exact probabilities instead of the Empirical Rule, what is the probability that the sample mean height of 25 women is **less than** 63.75 inches?

$$\mathbf{P}(\bar{\mathbf{x}} < 63.75) = \mathbf{P}\left(\frac{\bar{\mathbf{x}} - 64.5}{\frac{2.5}{\sqrt{25}}} < \frac{63.75 - 64.5}{0.5}\right) = \mathbf{P}(\mathbf{Z} < -1.5) = 0.0668$$

Proportions

Similar laws apply for proportions. The differences are:

1. For the Central Limit Theorem to apply, we require that **both** nρ >= 10 and n(1 - ρ) >= 10, where ρ is the true population proportion. If ρ is unknown then we can substitute the sample proportion, $\hat{p}$.
2. The distribution of the sample proportion, $\hat{p}$, will have a mean equal to ρ and standard deviation of $\sqrt{\frac{p(1-p)}{n}}$.

To find probabilities associated with some $\hat{p}$ we follow similar calculations as that for sample means:

$$\mathbf{P}(\mathbf{p} < \hat{\mathbf{p}}) = \mathbf{P}\left(\mathbf{Z} < \frac{\hat{\mathbf{p}} - \mathbf{p}}{\sqrt{\frac{\mathbf{p}(1-\mathbf{p})}{\mathbf{n}}}}\right) = \mathbf{P}(\mathbf{Z} < \mathbf{z}) \textbf{ and use Table A1}$$