

Unit 5: Sampling Distributions

Introduction:

Until this point, we have been analyzing raw data from samples and populations. Since we often don't have a way of doing a census and even well-designed studies can run into confounding variables, replication of results is crucial. We can take repeated samples from populations and look at the trends of their results to help us get an idea of what the actual population mean is when we have no way of getting a true census. Sampling distributions are different in part because these are not normal distributions of data but of *means* of collected samples from that data. Assuming that our sampling methods are sound (see chapter 3), repeated sampling will allow us to approximate μ and σ

The importance of replication: Remember that when we cannot actually measure a whole population which is most of the time, replication means that much more. Among our assumptions (such as the distribution is normal) we are also assuming that the sampling method and design in each case is sound. When we have all of that, replication is the final piece to making the best possible inference.

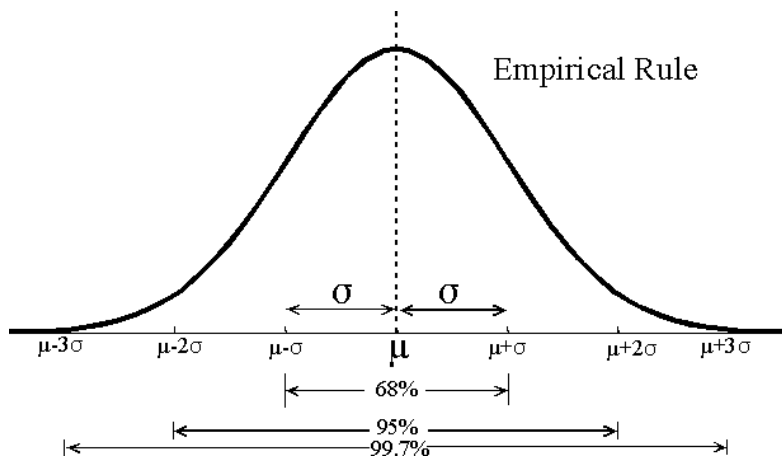
5.1 The Normal Distribution Revisited

Objectives:

- Review the Empirical Rule and Percentiles on a Normal Curve.
- Using the Standard Normal Curve to calculate z-scores and percentiles.
- Determine the interval associated with a given area in a normal distribution
- Using the Standard Normal Curve to calculate the probability that a particular value lies within a given interval of a normal distribution

Review of The Empirical Rule (the 68 – 95 – 99.7 Rule)

If the histogram of values in a data set can be reasonably well approximated by a normal curve, then...



- Approximately 68% of the observations are within 1 standard deviation of the mean.
- Approximately 95% of the observations are within 2 standard deviation of the mean.
- Approximately 99.7% of the observations are within 3 standard deviation of the mean.

z-scores (a quick review)

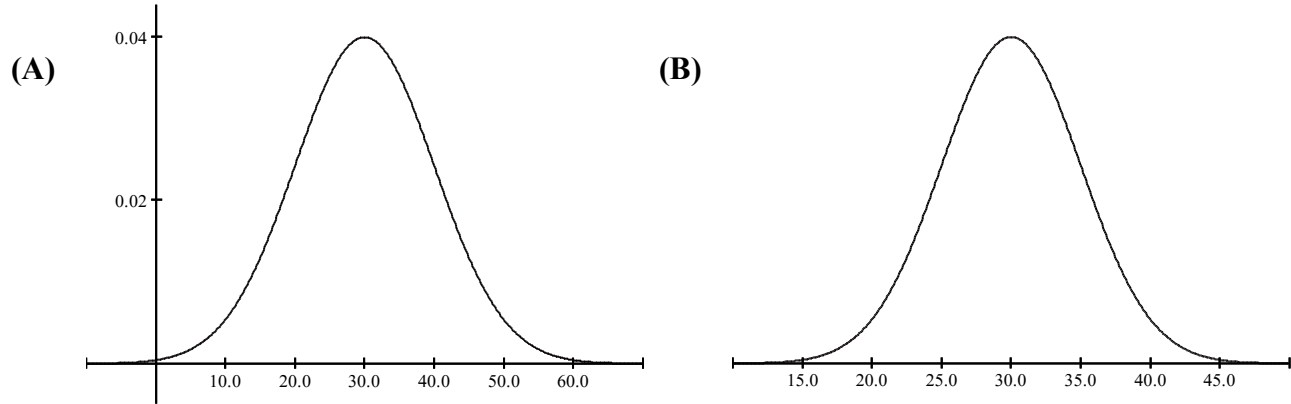
- It is positive or negative depending on whether the observed value is above or below the mean. The **z-score** associated with a particular value is given by the following:

$$z\text{-score} = \frac{\text{value} - \text{mean}}{\text{standard deviation}} \quad \text{or.} \quad z_i = \frac{x_i - \mu}{\sigma}$$

- The **z-score** tells us how many standard deviations **an observed value** is from the mean.
- It is positive or negative according to whether the value lies above or below the mean.
- A more positive or more negative z-score is further from the mean.

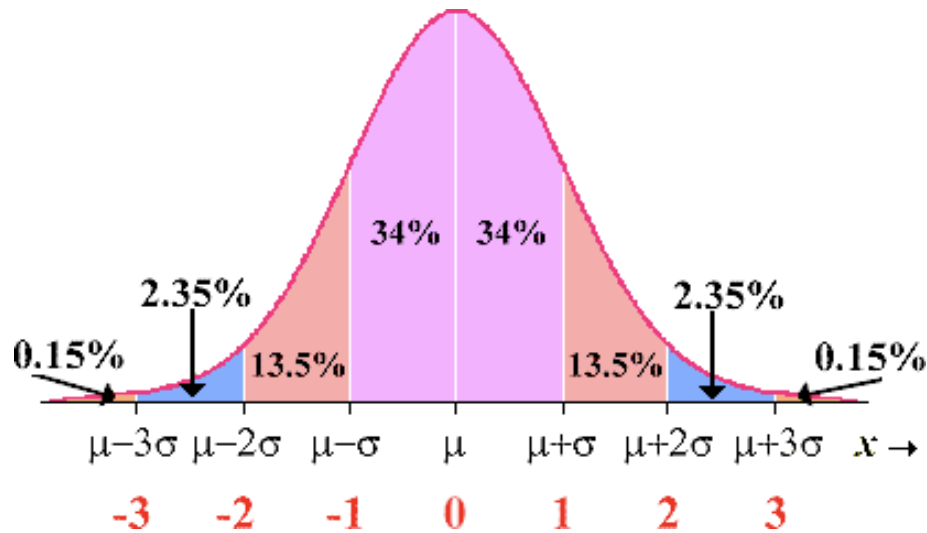
- Anything more than a z-score of 2 (or less than -2) is actually quite rare in a population (only 2.5% of a normal population falls above 2, and the same amount falls below -2)

Example 1 Which of the following distributions has a mean of 30 and a standard deviation of 5?



Answer: B The x -axis shows that three standard deviations on either side of the mean are contained by almost all of the graph per the Empirical Rule

Another graphical view of this rule looks at the percentages on each side of the mean ...

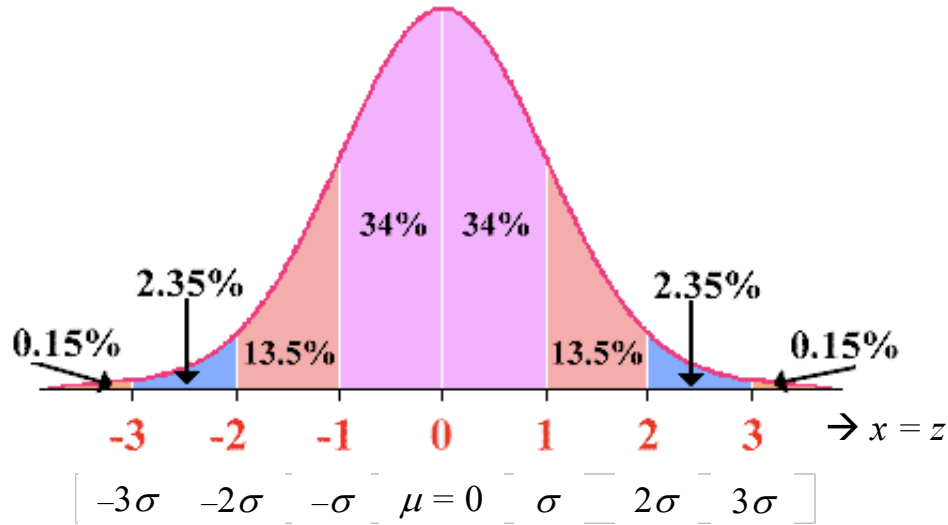


You can see the 34% on each side one standard deviation from the mean totals to 68%, two standard deviations from the mean total to 95%, etc.

You can also see the z-scores across the bottom of the curve. This is an important juxtaposition for what we will discuss next.

The Standard Normal Curve

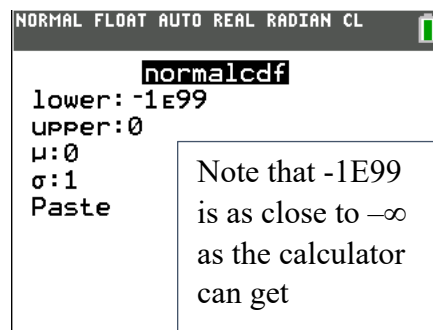
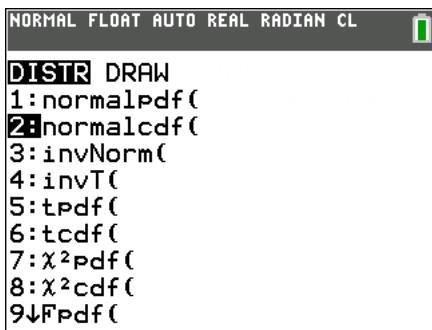
When comparing normal distributions from different sets of data (presumably with different means and standard deviations, we need a standard from which to compare z scores and percentiles. The standard normal distribution provides this standard. The graph below of the standard normal distribution shows how the x value is equal to the z score



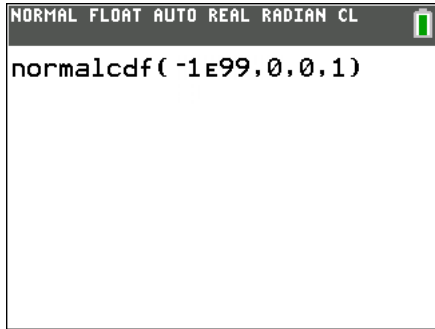
We can easily do this on our calculators to find the left half of the standard normal curve:

1) Calculator: 2nd → Vars → 2

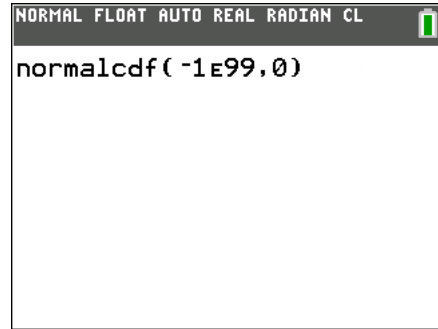
2) Here we're going to find the left half of the standard normal curve. If you are using a TI-83, see step 3



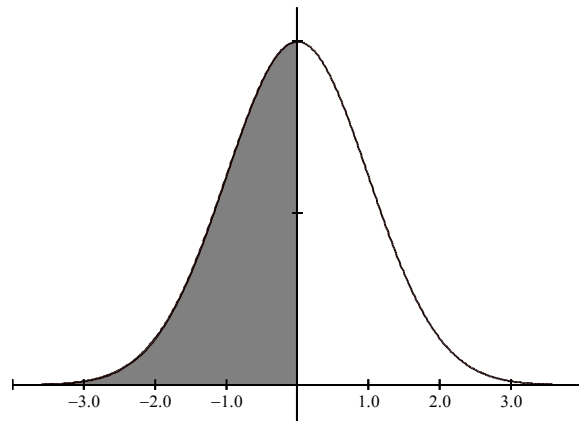
3) This is how it will appear on your display. If you are using TI-83, you will have to enter your parameters in this display



4) When using the Standard Normal Distribution, you need not enter the mean or standard deviation on the calculator because it defaults to 0 and 1



5) Your calculator should have given you an answer of 0.5 since we just calculated half the area under the standard normal curve (see graph to the right)



Definition:

- **Standard Normal Distribution:** A normal distribution in which the mean is 0, the standard deviation is 1, and $x = z$ (the x value is equal to the z score).

Example 1 In this year's county mathematics competition, a student scored 40; in last year's competition, the student scored 35. The average score this year was 38 with a standard deviation of 2. Last year's average score was 34 with a standard deviation of 1. In which year did the student score better?

- The student scored better on this year's exam.
- The student scored better on last year's exam.
- The student scored equally well on both exams.
- Without knowing the number of test items, it is impossible to determine the better score.

- e) Without knowing the number of students taking the exam in the county, it is impossible to determine the better score.

Each score is 1 standard deviation above the mean so both z scores are equal to 1

Answer: c)

Example 2 The weights of women are approximately normally distributed. This week, the z scores of weight for a member of a weight-watching group is 1.25. Which of the following is a correct interpretation of this z -score?

- a) This week the member weighs 1.25 lbs. more than last week.
- b) This week the member weighs 1.25 lbs. less than last week.
- c) This week the member weighs 1.25 lbs. more than the average woman.
- d) This week the member weighs 1.25 standard deviations more than she did last week.
- e) This week the member weighs 1.25 standard deviations more than the average woman.

Answer: e) which directly matches the definition of a z score

Example 3 Calculate the following probabilities with your calculator.

- (a) $P(z < -1.76) =$
- (b) $P(z \leq 0.58) =$
- (c) $P(-2 < z < 2) =$
- (d) $P(z > 1.18) =$
- (e) $P(z > 1.96) =$
- (f) $P(-1.76 < z < 0.58) =$

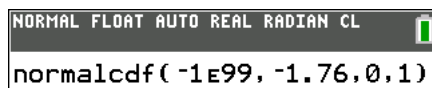
We can easily find these using the *normalcdf* feature on the calculator.

Calculator:  →  → 

Display: *normalcdf*(lower bound, upper bound, mean, s.d)

Note: On the calculator, $\infty = 1E99$ and $-\infty = -1E99$

(a) $P(z < -1.76) = 0.0392$



(b) $P(z \leq 0.58) = 0.71904$

```
NORMAL FLOAT AUTO REAL RADIAN CL
normalcdf(-1E99,0.58)
```

Note that we do not need to enter 0 and 1 if we are using a standard normal distribution. We also do not need to distinguish between $<$ and \leq

(c) Answer: 0.9545

```
NORMAL FLOAT AUTO REAL RADIAN CL
normalcdf(-2,2)
```

(d) Answer: 0.119

```
NORMAL FLOAT AUTO REAL RADIAN CL
normalcdf(1.18,1E99)
```

(e) Answer: 0.025

```
NORMAL FLOAT AUTO REAL RADIAN CL
normalcdf(1.96,1E99)
```

(f) Answer: 0.6798

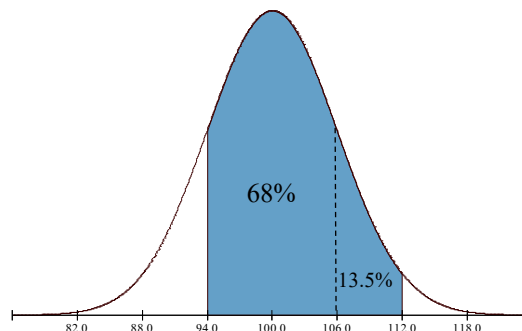
```
NORMAL FLOAT AUTO REAL RADIAN CL
normalcdf(-1.76,0.58)
```

Finding Probabilities for any Normal Distribution

Example 4 The growth of children can be an important indicator of general levels of nutrition and health. Data suggest that a reasonable model for the probability distribution of the continuous numerical variable $x =$ height of a randomly selected 5-year old child is a normal distribution with mean $\mu = 100$ cm and standard deviation $\sigma = 6$ cm. What proportion of the heights is between 94 and 112 cm?

$$P(94 < x < 112) = normalcdf(94, 112, 100, 6) = 0.819$$

Alternative answer: we can also notice that 94 is one standard deviation to the left and 112 is two standard deviations to the right of the mean. Using the Empirical Rule, we would have $0.68 + 0.135 = 0.8185$



What is the probability that a randomly chosen child will be taller than 110 cm?

$$P(x > 110) = normalcdf(110, 1E99, 100, 6) = 0.048$$

Example 5 Although there is some controversy regarding the appropriateness of IQ scores as a measure of intelligence, IQ scores are commonly used for a variety of purposes. One commonly

used IQ scale has a mean of 100 and a standard deviation of 15, and IQ scores are approximately normally distributed. If we define the random variable $x =$ IQ score of a randomly selected individual, then x has approximately a normal distribution with $\mu = 100$ and $\sigma = 15$.

One way to become eligible for Mensa, an organization purportedly for those of high intelligence, is to have an IQ score above 130.

(a) What proportion of the population would qualify for Mensa membership?

$$P(x > 130) = \text{normalcdf}(130, 1E99, 100, 15) = 0.023$$

(b) What proportion of the population has IQ scores below 80?

$$P(x < 80) = \text{normalcdf}(-1E99, 80, 100, 15) = 0.091$$

(c) What proportion of the population has IQ scores between 75 and 125?

$$P(75 < x < 125) = \text{normalcdf}(75, 125, 100, 15) = 0.904$$

Example 6 An electronic product takes an average of 3.4 hours to move through an assembly line. If the standard deviation is 0.5 hours, what is the probability that an item will take between 3 and 4 hours? Assume a normal distribution.

- a) 0.2119
- b) 0.2295
- c) 0.3270
- d) 0.3811
- e) 0.6730

$$P(3 < x < 4) = \text{normalcdf}(3, 4, 3.4, 0.5) = 0.673$$

Answer: e)

The Backwards “Situation”

Sometimes you will be given a probability and asked for an x value or a z score.

In these cases, we use InvNorm on the calculator or the formula to solve the problem.

Calculator

InvNorm : 2nd \rightarrow Vars \rightarrow 3 \rightarrow percentile \rightarrow mean \rightarrow s.d.

Display: $invNorm(\text{percentile}, \text{mean}, \text{s.d.})$

Note that $invNorm(\text{percentile})$ presumes a standard normal distribution

Formula

To convert a z score back to an x value, use

$$x = \mu + z\sigma$$

Example 7 Data on the length of time required to complete registration for classes using a telephone registration system suggest that the distribution of the variable $x = \text{time to register}$ for students at a particular university can be well approximated by a normal distribution with mean $\mu = 12$ min and standard deviation $\sigma = 2$ min.

Because some students do not sign off properly, the university would like to disconnect students automatically after some amount of time has elapsed. It is decided to choose this time such that only 1% of the students are disconnected while they are still attempting to register.

What is the cutoff time for disconnection?

$$Invnorm(0.99, 12, 2) = 16.7 \text{ minutes}$$

Formula method

$$x = \mu + z\sigma = 12 + z(2)$$

$$invNorm(0.99) = 2.326$$

$$x = 12 + 2.326(2) = 16.653 \text{ minutes}$$

Example 8 A sales person ranks in the top 5% of all sales people in a large company. If the annual mean sales amount is \$750,000 and the standard deviation is \$150,000, how much does the person sell each year? Assume a normal distribution.

- (a) \$757,500
- (b) \$996,750
- (c) \$1,044,000
- (d) \$1,697,100
- (e) \$2,650,500

Formula method

$$x = \mu + z\sigma = 750000 + z(150000)$$

$$invNorm(0.95) = 1.644\dots$$

(I recommend using the
store function on your
calculator for this)

$$x = 750000 + (1.644\dots)(150000) = \$996,728$$

$$invNorm(0.95, 750000, 150000) = \$996,728 \text{ Answer: (b)}$$

Summary:

- **The Empirical Rule:** In a normal distribution
 - 68% of the data lie within one standard deviation of the mean
 - 95% of the data lie within two standard deviations of the mean
 - 97.5% of the data lie within three standard deviations of the mean
- **The Standard Normal Curve:** A specific normal distribution in which the mean is 0 and the standard deviation is 1 to calculate z-scores and percentiles.
- Use *normalcdf*(a, b, μ, σ) to find the probability that a random variable will fall between a and b
- Use *normalcdf*(a, b) to find the probability that a z score will fall between a and b
- Use *invNorm*(p, μ, σ) to find a percentile location in a normal distribution given percentile value p

Checkpoint, Section 5-1:

Multiple Choice

1. A factory dumps an average of 2.43 tons of pollutants into a river every week. If the standard deviation is 0.88 tons, what is the probability that in a week more than three tons are dumped? Assume a normal distribution.

- (a) 0.2578
- (b) 0.2843
- (c) 0.6500
- (d) 0.7157
- (e) 0.7422

2. Runners competed in a local road race. The mean finishing time for the race was 43.5 minutes with a standard deviation of 16.2 minutes. The sponsors wanted to have a special race for those who were in the fastest 10%. Assuming the times were normally distributed, which of the following is the cutoff time?

- (a) 22.8 minutes
- (b) 25.7 minutes
- (c) 39.2 minutes
- (d) 42.2 minutes
- (e) 64.3 minutes

3. Which of the following are true?

I. The area under a normal curve is always equal to 1, no matter what the mean and standard deviation are.

II. The smaller the standard deviation of a normal curve, the higher and narrower the graph.

III. Normal curves with different means are centered around different numbers.

(a) I and II (b) I and III (c) II and III (d) I, II, and III (e) None of the above

4. A fire department in a rural county reports the mean response is 22 minutes. A home owner was told the response time of 30 minutes to his neighborhood was at the 3rd quartile. What standard deviation did the report use if the times were normally distributed?

(a) 2.7 minutes

(b) 3.4 minutes

(c) 7.1 minutes

(d) 11.9 minutes

(e) 16.0 minutes

5.1 Homework

1) Mr. Maychrowitz' tracks his average time solving a four by four Rubik's cube and finds that he has a mean of 7 minutes and a standard deviation of 2 minutes. From his database of all his times, a random sample is taken. Given that Mr. Maychrowitz' times are normally distributed, find the probability that the selected time is less than 4 minutes.

2) Mr. Murphy is happy just to be able to solve a three by three cube but while his times are also normally distributed, they are a paltry 12 minutes with a standard deviation of 3 minutes. He does however boast that 25% of the time, he can do it in under 10 minutes. Is this statistically the case?

3) A certain type and size of anchor rod is used to support a walk way on a bridge. For this particular function, the rods need to have a length of 1.5 feet. Suppose that quality control at the manufacturer requires the length to be within 0.003 inches of the expected length in order to be accepted for use. The length of these anchor rods is normally distributed with a mean of 18 inches and a standard deviation of 0.0012 inches. Of 1000 rods that come off the assembly line, how many of these would we expect to be rejected?

Based on studies over the past five years, US men average $\mu_m = 69$ inches with $\sigma_m = 2.5$ while women average $\mu_w = 64.5$ and $\sigma_w = 2.2$. Use these statistics to answer the remaining questions.

- 4) When selecting one male and one female from the population,
 - (a) Find the probability that the male will be taller than 6 foot 2 inches (74 inches)
 - (b) Find the probability that the female will be between 5 feet 4 inches and 5 feet 10 inches (64 and 70)

- 5) When randomly selecting 100 females from the population, approximately how many can we expect to be 5 foot 8 (68) or taller?

5-2 Sample Means & The Central Limit Theorem

Objectives:

- Explain the concepts of sampling variability and sampling distribution.
- Determine the sampling distribution for \bar{x} .
- Calculate probabilities based on the distribution of \bar{x} .
- Explain the Central Limit Theorem.

Example 1 Consider a small population consisting of the 20 students enrolled in an upper division class. The students are numbered 1 to 20, and the amount of money (in dollars) each student spent on textbooks for the current semester is shown in the following table:

Calculate μ and σ . Describe the SOCS.

Student	Amount Spent on Books	Student	Amount Spent on Books
1	267	11	319
2	258	12	263
3	342	13	265
4	261	14	262
5	275	15	333
6	295	16	184
7	222	17	231
8	270	18	159
9	278	19	230
10	168	20	323

Here we chose a window setting to give us an interval size of 30 for each histogram rectangle and that can fit the range of values of x

NORMAL FLOAT AUTO REAL RADIAN CL

1-Var Stats

$\bar{x}=260.25$
 $\Sigma x=5205$
 $\Sigma x^2=1403695$
 $Sx=50.83189529$
 $\sigma x=49.54480296$
 $n=20$
 $\text{min}X=159$
 $\downarrow Q_1=230.5$

$$\mu = \$260.25 \quad \sigma = \$49.54$$

S – Roughly symmetric

O – no outliers

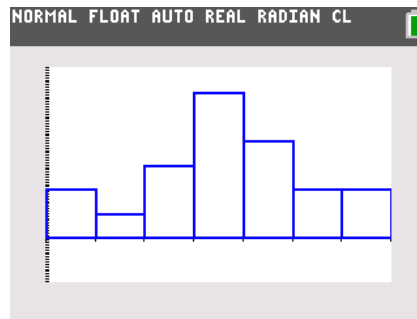
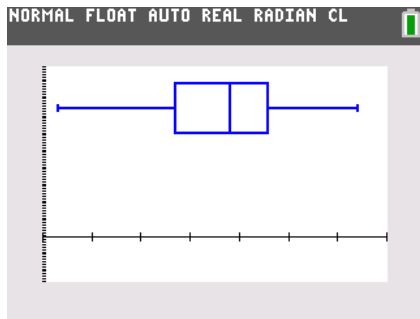
C - $\mu = \$260.25$ Median = 264

S – spread is 159 to 342 range = 183

NORMAL FLOAT AUTO REAL RADIAN CL

WINDOW

$X_{\text{min}}=150$
 $X_{\text{max}}=360$
 $X_{\text{sc1}}=30$
 $Y_{\text{min}}=-1.80414$
 $Y_{\text{max}}=7.02$
 $Y_{\text{sc1}}=.1$
 $X_{\text{res}}=1$
 $\Delta X=.795454545455$
 $\text{TraceStep}=1.5909090909091$



Notice that the distribution of both the box plot and the histogram are approximately normal.

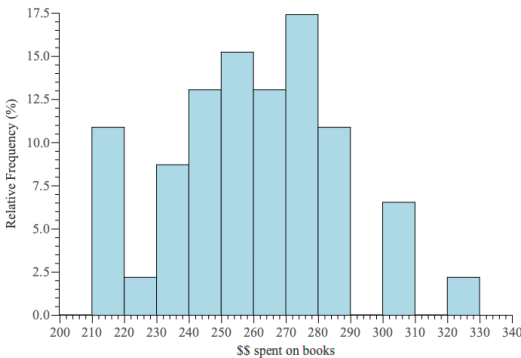
Now take a random sample of 5 students and calculate \bar{x} .

Did we all get the same \bar{x} ?

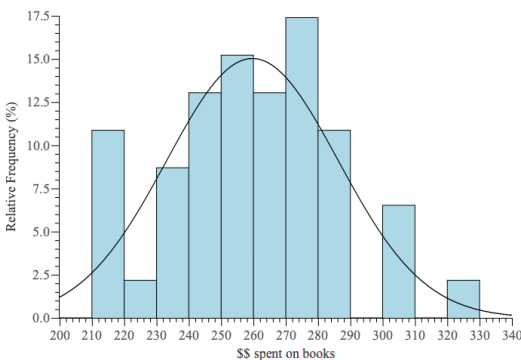
I took 45 samples, each of size 5, and calculated each \bar{x} . Here are the sample means:

Sample #	\bar{x}	Sample #	\bar{x}	Sample #	\bar{x}
1	274.6	16	255.0	31	253.4
2	256.6	17	307.2	32	279.6
3	263.8	18	280.0	33	252.6
4	254.8	19	277.4	34	242.2
5	238.2	20	241.2	35	262.6
6	275.6	21	216.0	36	215.4
7	279.2	22	270.0	37	266.2
8	241.6	23	301.0	38	261.4
9	255.4	24	247.0	39	275.0
10	263.8	25	273.8	40	301.4
11	239.6	26	282.8	41	237.0
12	248.2	27	220.4	42	287.4
13	330.8	28	213.6	43	249.2
14	288.8	29	287.6	44	236.8
15	252.8	30	214.8	45	264.6

Here is a histogram of all the \bar{x} values. What shape does the **sampling distribution** have? What can be said about the spread of this distribution as compared to the spread of the original data?



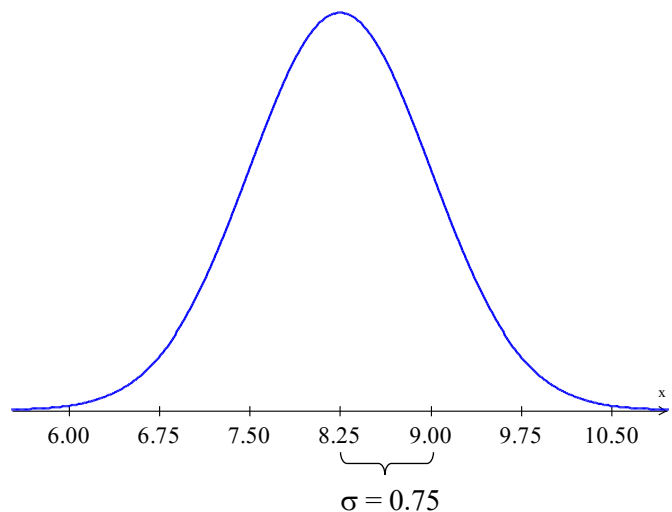
This graph clearly shows a trend towards a normal distribution. In fact, the graph below shows the projected normal curve.

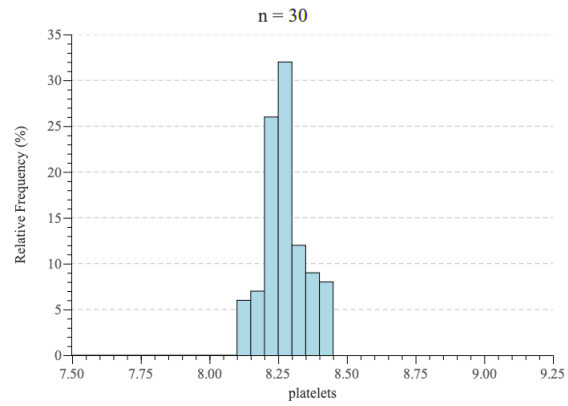
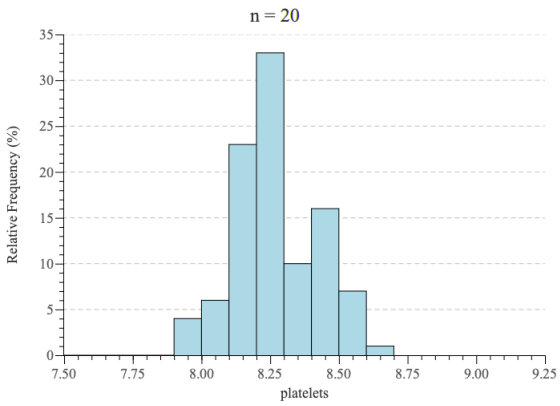
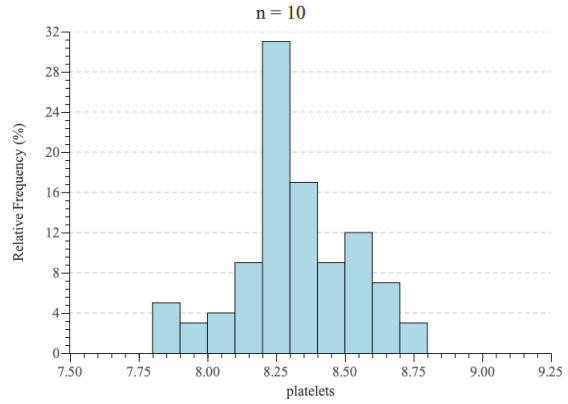
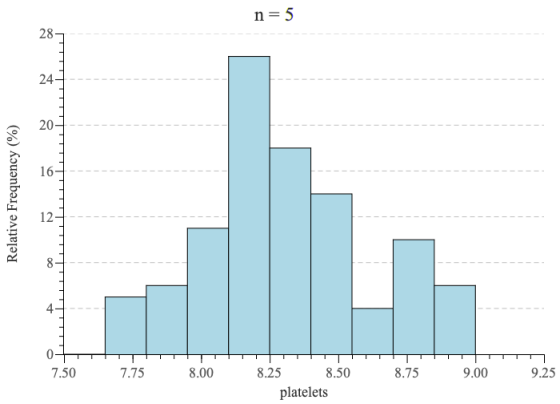


In fact, this second graph has the projected normal curve overlap with $\mu_{\bar{x}} = \$259.69$

Example 2 Data suggested that the distribution of platelet size for patients with non-cardiac chest pain is approximately normal with mean $\mu = 8.25$ and standard deviation $\sigma = 0.75$. The figure below shows the corresponding normal curve.

MINITAB was used to select 500 random samples from this normal distribution, with each sample consisting of $n = 5$ observations. The process was repeated for sample of size $n = 10$, $n = 20$, and $n = 30$. The resulting 500 \bar{x} values appear in the following histograms.

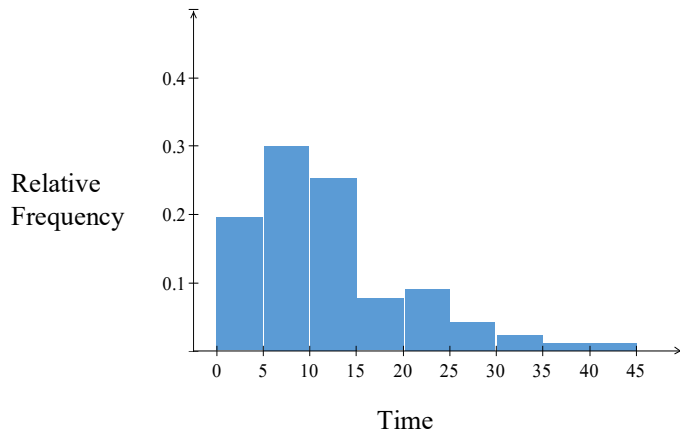




What can be said about the histograms? In terms of SOCS?

Example 3 Now consider the properties of the \bar{x} distribution when the population is quite skewed (and thus very unlike a normal distribution). The Winter 1995 issue of *Chance* magazine gave data on the length of overtime period for all 251 National Hockey League play-off games between 1970 and 1993 that went into overtime. In hockey, the overtime period ends as soon as one of the teams scores a goal. This histogram of the data is below.

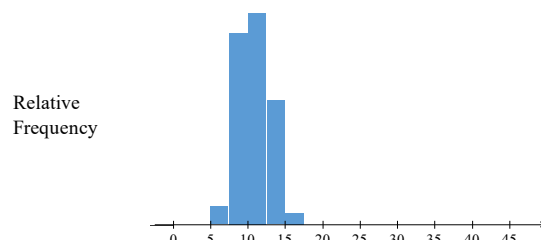
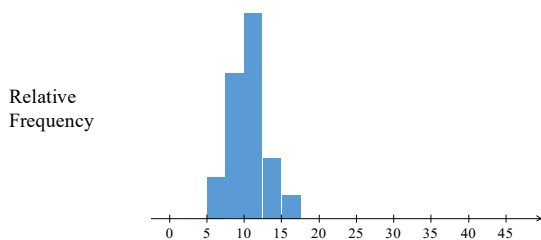
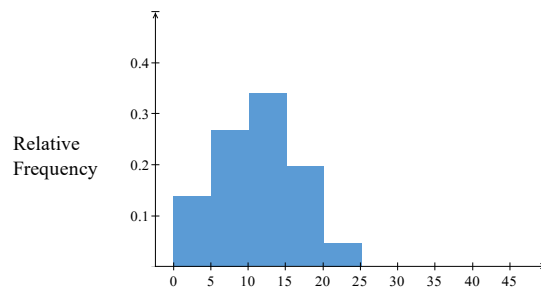
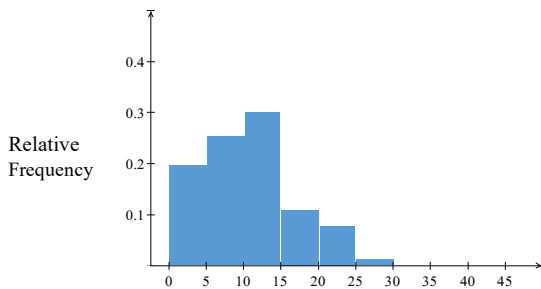
Describe this histogram. (SOCS)



Let's say we found that $\mu = 9.841$ so that is the balance point for the population histogram. The median is 8. What does that tell us about our histogram?

For each of the sample sizes $n = 5, 10, 20,$ and 30 , we selected 500 random samples of size n . We then constructed the following histograms of the 500 \bar{x} values.

What do we notice about the histograms? In terms of SOCS.



Notice also that despite the fact that the population graph was skewed right, as the sample sizes increase, the sample mean distribution becomes approximately normal and the spread of the data

shrinks. This pattern with all of these examples can be explained by the following properties of sampling distributions:

General Properties of the Sampling Distribution of \bar{x}

Let \bar{x} denote the mean of the observations in a random sample of size n from a population having mean μ and standard deviation σ . Denote the mean value of the \bar{x} distribution by $\mu_{\bar{x}}$ and the standard deviation of the \bar{x} distribution by $\sigma_{\bar{x}}$. Then the following rules hold:

$$\text{Rule \#1: } \mu_{\bar{x}} = \mu$$

$$\text{Rule \#2: } \sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}}$$

Assumption #1: When the population distribution is normal, the sampling distribution of \bar{x} is also normal for any sample size n .

Assumption #2 (Central Limit Theorem): When n is sufficiently large, the sampling distribution of \bar{x} is well approximated by a normal curve, even when the population distribution is not itself normal.

• One of the above assumptions must be met in order to continue.

• If n is large or if the population distribution is normal, then the standardized variable

$$z = \frac{\bar{x} - \mu_{\bar{x}}}{\sigma_{\bar{x}}} = \frac{\bar{x} - \mu}{\sigma / \sqrt{n}}$$

has (at least approximately) a standard normal (z) distribution.

• The Central Limit Theorem can be safely applied if n exceeds 30.

Ex4 What are the mean and the standard deviation of a sampling distribution consisting of samples of size 16? These samples were drawn from a population whose mean is 25 and whose standard deviation is 5.

- (a) 25, 1.25
- (b) 5, 5
- (c) 25, 5
- (d) 5, 1.25
- (e) 25, $\sqrt{5}$

Can we consider this sampling distribution normal?

Answer: No because we do not have enough info about the population distribution.

Example 5 The mean TOEFL score of international students at a certain university is normally distributed with a mean of 490 and a standard deviation of 80. Suppose that groups of 30 students are studied. The mean and the standard deviation for the distribution of sample means, respectively, will be

- (a) 490 and $8/3$
- (b) 16.33 and 80
- (c) 490 and 14.61
- (d) 490 and 213.33

Can we consider this sampling distribution normal?

Answer: Yes because it meets both assumptions about the CLT; normal population distribution and sample size of at least 30

Ex6 A certain brand of lightbulb has a mean lifetime of 1500 hours with a standard deviation of 100 hours. If the bulbs are sold in boxes of 25, the parameters of the distribution of sample means are

- (a) 1500 and 100
- (b) 1500 and 4
- (c) 1500 and 2
- (d) 1500 and 20

Can we consider this sampling distribution normal?

Answer: No because it meets neither assumption of the CLT

Ex7 A soft-drink bottler claims that, on average, cans contain 12 oz. of soda. Let x denote the actual volume of soda in a randomly selected can. Suppose that x is normally distributed with $\sigma = 0.16$ oz. Sixteen cans are to be selected, and the soda volume will be determined for each one. Let \bar{x} denote the resulting sample mean soda volume.

What is the distribution of \bar{x} ?

Calculate $P(11.96 \leq \bar{x} \leq 12.08)$.

Calculate $P(\bar{x} \leq 12.08)$.

Example 8 Samples of size 49 are drawn from a distribution that's highly skewed to the right with a mean of 70 and a standard deviation of 14. What's the probability of getting a sample mean between 71 and 73?

- (a) 0.0563
- (b) 0.00023
- (c) 0.2417
- (d) 0
- (e) We can't answer this question because the distribution is highly skewed.

Summary:

- The observed value of a statistic depends on the particular sample selected from the population; typically, it varies from sample to sample. This variability is called **sampling variability**.
- The distribution of the values of a statistic is called its **sampling distribution**.
- The **Central Limit Theorem** states that as long as the population distribution is approximately normal or the sample size $n \geq 30$ then the sampling distribution will be approximately normal even if the population distribution is not.
- The mean of the sampling distribution is equal to the mean of the population distribution

$$\mu_{\bar{x}} = \mu$$

- The standard deviation of the sampling distribution is equal to the standard deviation of the population distribution divided by the square root of the sample size

$$\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}}$$

5.2 Homework

- 1) A few SI grads with engineering degrees develop a new brand of lithium laptop battery, called the Maychro Battery boasting that it lasts a revolutionary average life between charges of 11.09 hours with a standard deviation of 0.5 hours. Some current SI students decide to run their own tests to verify. They select samples of 20 batteries and tabulate their means in a spreadsheet.
 - (a) If the sampling distribution of this turns out to be normal when the students analyze their data, does that imply that the distribution of the original developer's data is also normal? Explain your answer.
 - (b) If the developer's data is normal, what will be the mean and standard deviation of the sampling distribution of the Maychro Battery life?

- 2) As discussed in Unit 5-1, studies over the past five years have left us with this overall pop data: US Men average $\mu_m = 69$ inches with $\sigma_m = 2.5$ while women average $\mu_w = 64.5$ and $\sigma_w = 2.2$. If random samples of 100 men and 100 women are drawn, find the sampling distribution of each sample mean for both men and women.

Problems 3-9 refer to the collection of data on freshmen classes from fall 2016 through 2021 shared with you in class.

- 3) Describe the student population distribution (SOCS). Is this distribution approximately normal? Why or why not?
- 4) You were emailed one or more small samples of which to calculate the mean and to submit via google form. When the distribution of those sample means was displayed, how did their distributions compare to the population?
- 5) Now look at the results when samples of 50 and 100 were taken from the population and a sampling distribution generated, the sampling distribution was approximately normal.
- 6) Find the standard deviation of these two sampling distributions.
- 7) Describe each population distribution when the data is separated by gender.
- 8) Explain why when sampling by gender, samples of 5 and 10 could still produce sampling distributions that were approximately normal.
- 9) What are the standard deviations of the sampling distributions for men and women?

5-3 Biased and Unbiased Point Estimates

Objectives:

- Identify point estimates in sampling distributions
- Distinguish between *biased* and *unbiased* statistics

Any sample statistic used to approximate the population parameter is called a *point estimator*. The value of that statistic is called the *point estimate*.

We consider any point estimate to be *unbiased* if, on average, the estimator is equal to the population parameter. In other words, if a particular measurement of our sample matches the same measurement of the population, then we consider that measurement to be *unbiased*. If it is not then we consider it to be *biased*.

Keep in mind here that we are not using the term *bias* in the usual colloquial way with which you are likely familiar. This bias is not personal. Here the term is being used to describe an estimate from a sample that does not describe the population precisely enough.

Example: The mean score on the 2022 AP Statistics Exam was 2.89. We took 10 different random samples of 50 AP Stats students and calculated the mean for each sample. Below is the distribution of the sample means. Is the mean of this distribution a biased or unbiased estimator of the actual mean score?

\bar{x}_1	\bar{x}_2	\bar{x}_3	\bar{x}_4	\bar{x}_5	\bar{x}_6	\bar{x}_7	\bar{x}_8	\bar{x}_9	\bar{x}_{10}
3.08	2.66	3.1	2.72	2.82	2.66	3.02	3.06	2.9	2.88

Since the mean of these sample means is 2.89 then the mean of this distribution is considered *unbiased*

Each one of these sample means is called a *point estimate*

Now let's try another 10 samples of only 10 scores per sample. The table below contains our results.

Samples:	1	2	3	4	5	6	7	8	9	10
	1	1	4	5	4	4	3	5	3	5
	2	4	2	4	1	1	5	5	3	3
	1	5	2	3	2	4	3	3	4	1
	3	2	1	1	2	4	1	2	1	4
	4	1	3	1	3	4	1	3	2	4
	3	5	3	3	3	2	2	4	1	2
	2	5	3	3	3	5	4	3	4	2
	2	1	4	5	2	4	3	4	1	3
	2	4	5	4	1	1	5	4	4	2
	1	3	1	2	4	4	2	4	3	2
Average	2.1	3.1	2.8	3.1	2.5	3.3	2.9	3.7	2.6	2.8

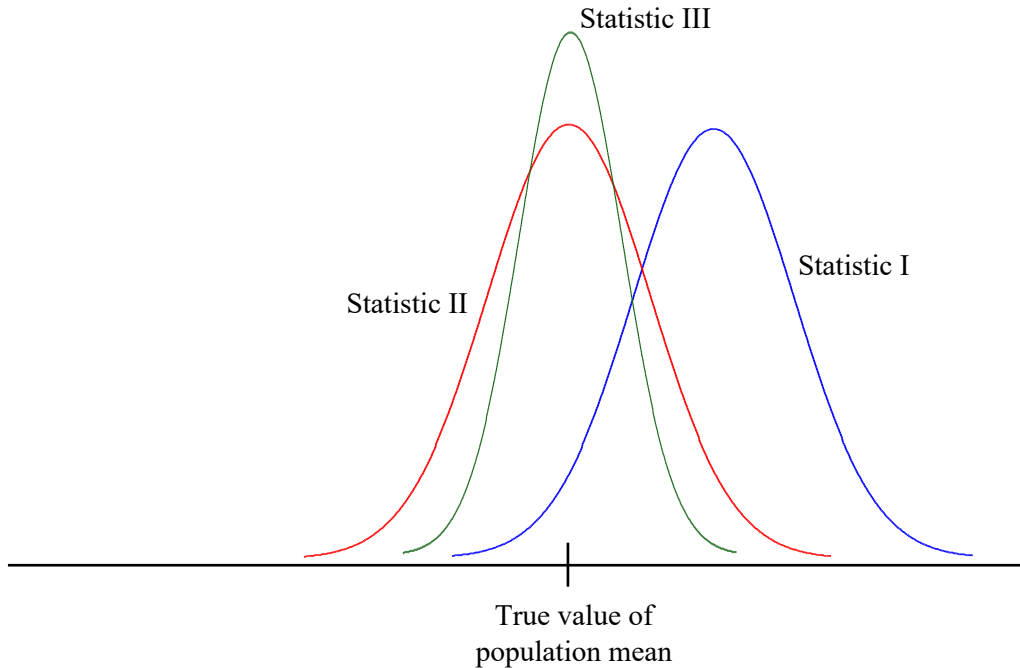
Mean of Sample Means
→ 2.89

Samples 5 and 9 are missing a 5 while sample 8 is missing a 1 making the range of those three samples equal to 3. When we average the range for all ten samples we get 3.7 when the true range is $5-1 = 4$.

In fact, we can expect that in about 20% of our samples of size 10, a score of 5 will not be drawn. This alone means that using the range of any of these samples is problematic in estimating the range of the population. This makes range an example of a *biased* estimate since we would expect that the mean of all the ranges would be 3 but they won't be if we know that 20% won't contain a 5 and about 7% won't contain a 1. If we repeated the samples and took the mean of the ranges collected we will always get a number less than 5 because those times when the sample range is smaller will drag the average range down. This is why range is a good example of a *biased* estimator.

5.3 Homework

- 1) Three different statistics are being considered for estimating a population characteristic. The sampling distributions of the three statistics are in the illustration shown below. Rate in order from best to least which statistic is the best estimator of the true value of the population mean. Explain your answers.



- 2) A random sample of 10 houses heated with natural gas in a particular area, is selected, and the amount of gas (in therms) used during the month of January is determined for each house. The resulting observations are as follows:

103 156 118 89 125 147 122 109 138 99

- Let μ_j denote the average gas usage during January by all houses in this area. Compute a point estimate of μ_j .
- Suppose that 10,000 houses in this area use natural gas for heating. Let x denote the total amount of gas used by all of these houses during January. Estimate x using the given data. What statistic did you use in computing your estimate?
- Use the data in Part (a) to estimate p , the proportion of all houses that used at least 100 therms.
- Give a point estimate of the population median usage based on the sample of Part (a). Which statistic did you use?

5-4 Sampling Distributions for Differences in Sample Means

Objectives:

- Find the **difference in means** and **standard deviations** between two **independent samples**
- Find the **difference in sample means** and their **standard deviations** between two **independent samples**

In this section, we consider using sample data to **compare** two population means or two treatment means.

If two random variables X and Y are independent and have distributions that are approximately normal then we can say that $\mu_{X \pm Y} = \mu_X \pm \mu_Y$ and their sampling distributions would have this relationship:

$$\mu_{\bar{X} \pm \bar{Y}} = \mu_{\bar{X}} \pm \mu_{\bar{Y}}$$

The standard deviations are not as simple however.

For any two ***independent*** random samples: $\sigma_{X \pm Y} = \sqrt{\sigma_X^2 + \sigma_Y^2}$ Note that we ***always*** add variances.

For the difference in sample means of any two ***independent*** random samples we have

$$\sigma_{\bar{X} \pm \bar{Y}} = \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}$$

Note the Pythagorean similarities in how the standard deviations are combined. While beyond the scope of this class, a recommended explanation of this can be found at

<https://apcentral.collegeboard.org/courses/ap-statistics/classroom-resources/why-variances-add-and-why-it-matters>

Example 1 The mean and standard deviation of scores on the 2022 AP Calculus exam were 2.91 and 1.44 while the mean and standard deviation of AP Stats exams were 2.89 and 1.38.

Using the data above and assuming the scores on each exam are independent, find the mean and standard deviation of the difference between the AB Calculus scores and the AP Stats scores.

Let C = a randomly selected AB Calc Test score

Let S = a randomly selected AP Stats score

$$\bar{C} = 2.91 \quad \sigma_C = 1.44 \quad \bar{S} = 2.89 \quad \sigma_S = 1.38$$

The mean of the difference is $\mu_{C-S} = 2.91 - 2.89 = 0.02$

Given the formula for standard deviation of the difference of two variables, we will use the variance of each variable:

$$\sigma_C^2 = 1.44^2 \quad \sigma_S^2 = 1.38^2 \quad \sigma_{C-S} = \sqrt{\sigma_C^2 + \sigma_S^2} = \sqrt{2.0736 + 1.9044} = 1.99$$

Example 2 The mean number of visits per day to Dr. Quattrin's website is 41.2 with a standard deviation of 8.1 while the mean number of visits per day to Mr. Murphy's website is 29.1 with a standard deviation of 10.1. A few students decide to track the web traffic to each site by taking samples in intervals of seven days to find the total number of visits to both sites. They repeat this sample over the course of a school year only during weeks when school is in session

- (a) Find the mean and standard deviation of the sampling distribution.
- (b) Find the probability that on a randomly selected week the mean number of total visits to both sites will be at least 80

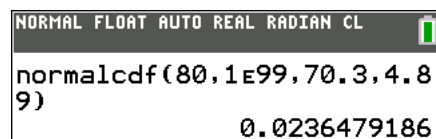
(a) Let Q = visits to Dr. Q's site, M = Visits to Mr. Murphy's site, and $T = Q + M$

$$\mu_Q = 41.2 \quad \mu_M = 29.1 \quad \sigma_Q = 8.1 \quad \sigma_M = 10.1$$

For this sampling distribution the average total visits would be $\mu_{\bar{T}} = 41.2 + 29.1 = 70.3$

The standard deviation is $\sigma_{\bar{T}} = \sqrt{\frac{\sigma_Q^2}{n_1} + \frac{\sigma_M^2}{n_2}} = \sqrt{\frac{8.1^2}{7} + \frac{10.1^2}{7}} = 4.89$

(b) For the probability, we just use *normalcdf*:



NORMAL FLOAT AUTO REAL RADIAN CL

normalcdf(80, 1E99, 70.3, 4.89)

0.0236479186

So the probability of that the mean number of visits would be greater than or equal to 80 is 0.024

Example 3 Mr. Murphy, after much pain and suffering, is finally able to solve a four by four Rubik's cube. Over time, he finds that his solving times settle into a normal distribution with mean 22 minutes and a standard deviation of 8 minutes. Comparing this to Mr Maychrowitz' normal distribution of scores with a mean of 7 minutes and standard deviation of 2 minutes, find the mean and standard deviation of the difference between Mr. Maychrowitz' time and Mr. Murphy's.

Let G = Mr Maychrowitz' time solving a rubik's cube on any given attempt

L = Mr. Murphy's time solving a rubik's cube on any given attempt

$$\mu_{L-G} = 22 - 7 = 15 \text{ minutes} \quad \sigma_{L-G} = \sqrt{8^2 + 2^2} = 8.246$$

Example 4 A few less trusting students want to see the data for themselves but have work to do for other classes so they divide the work among themselves and each randomly selects 10 of Mr Maychrowitz' times and 7 of Mr. Murphy's times (Because of his slower times, Mr. Murphy's data base is slightly smaller than Mr. Maychrowitz'). What should we expect $\mu_{\bar{L}-\bar{G}}$ and $\sigma_{\bar{L}-\bar{G}}$ to be for this sampling distribution?

$$\mu_{\bar{L}-\bar{G}} = 22 - 7 = 15 \text{ minutes}$$

$$\sigma_{\bar{L}-\bar{G}} = \sqrt{\frac{\sigma_L^2}{n_L} + \frac{\sigma_G^2}{n_G}} = \sqrt{\frac{2^2}{10} + \frac{8^2}{7}} = 3.089 \text{ minutes}$$

Example 5 Given any one of the random selections described in Example 4, would it be unusual to find a mean difference of 10 minutes or less?

$$\text{normalcdf}(-1E99, 10, 15, 3.089) = 0.0528$$

Since we expect about 5.28% of the sample means to be within 10 minutes, we would find Mr. Murphy being within 10 minutes of Mr. Maychrowitz to be unusual

Note: In future units what qualifies as 'unusual' will be discussed in more depth.

Checkpoint

Multiple Choice

1. A random sample of 32 games is chosen for a professional basketball team, team A, and their results are recorded. The team averages 88 points per game with a standard deviation of 8. The same is done for a second team, team B, with this team averaging 90 points per game with a standard deviation of 6. One game result is randomly selected from each team. Assuming that the two distributions are approximately normal, the probability that two randomly drawn results will differ by more than 7 points is

- (a) 0.184
- (b) 0.368
- (c) 0.507
- (d) 0.260
- (e) 0.520

2. Two professors, A and B, got into an argument about who grades tougher. Professor A insisted that his grades were lower than those for Professor B. In order to test this theory, each professor took a random sample of 25 student grades and compared results. The graphical displays showed each grade distribution was approximately normal. The results are recorded below.

	Professor A	Professor B
Count:	25	25
Mean:	79	82
Std dev:	6	4

The sampling distribution of the difference in sample means of the two professors' scores (Prof A – Prof B) is

- (a) negative
- (b) not determinable because the sample sizes are too small
- (c) skewed right
- (d) skewed left
- (e) approximately normal

3. Assuming that the population parameters match the data from the sample above, the standard deviation of the sampling distribution from #3 is

- (a) 6
- (b) 4
- (c) $2\sqrt{13}$
- (d) $\frac{2\sqrt{13}}{5}$
- (e) $\sqrt{52}$

5.4 Homework

- 1) Mr. Murphy and Mr. Maychrowitz decide to go into business selling all sorts of brain twister puzzles, opening a Murph & Mac's Puzzle Palace in SF and one in Pacifica. Their expected profit per item at the SF store is \$7 with a standard deviation of \$2 while their expected profit per item at the Pacifica store is \$5 with a standard deviation of \$1. Find the mean and standard deviation of the difference in profits between the SF and Pacifica locations.
- 2) Some inquiring students decide to do their own study in hopes of tracking certain specific trends in sales at Murph & Mac's Puzzle Palace. They conduct several samples of 50 items

sold at the SF store and 40 at the Pacifica store and measure the sample mean of profit per item for each store.

- (a) Find the mean and standard deviation of this sampling distribution
- (b) Explain why it is not necessary for either distribution in problem #1 to be normal for the sampling distribution to be normal.

- 3) To compete with the Maychro Battery (See problem #1 in Unit 5-2), a few other SI grads turned engineers develop the QuattriPower Battery that boasts an average life of 12.1 hours with standard deviation 0.8 hours. Assuming both battery charge life distributions are normally distributed, find the mean and standard deviation of the difference in battery charge life between the two batteries.
- 4) The same inquiring SI students get word of this and decide to do some more testing, this time aiming to compare the performance of the two batteries. They are able to obtain samples of 15 Maychro Batteries and 10 QuattriPower Batteries. Find the probability that their sample mean difference in battery charge life will be less than 1.5 hours.

Given the data from the class height study, use random variable notation to find the mean and standard deviation of the difference in height between boys and girls in the population.

- 3) Find the standard deviation of the difference in heights between boys and girls in the population distribution
- 4) Given a sample of size 100 from the boys and 110 from the girls (who slightly outnumbered the boys in the collected population).

5-5 Sampling Distributions for Sample Proportions

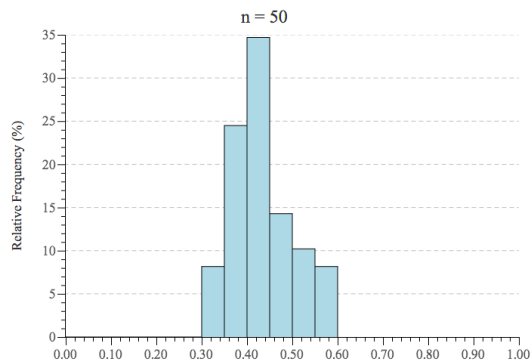
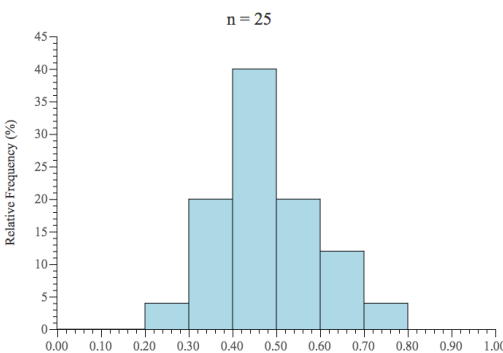
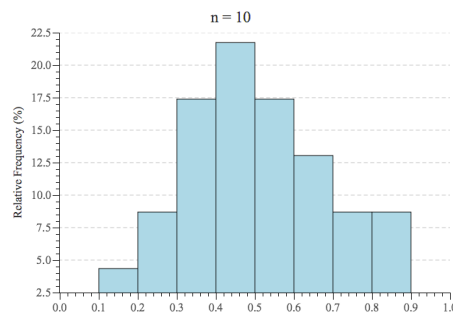
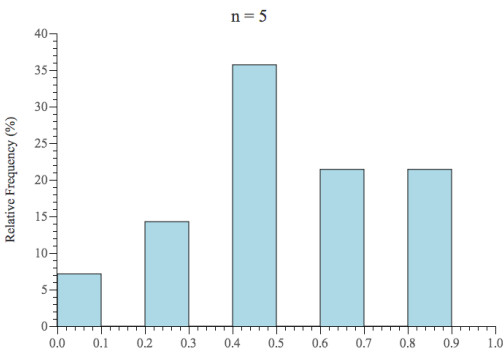
Objectives:

- Determine the sampling distribution for a sample proportion \hat{p} .
- Calculate probabilities based on the distribution of a sample proportion \hat{p}

Ex1 In the fall of 1999, there were 16,250 students enrolled at Cal Poly, SLO. Of these students, 6825 (42%) were female. With F denoting a female student and M denoting a male student, the proportion of F's in the population is $p = 0.42$.

Let's say we selected 500 samples of size $n = 5$ and recorded \hat{p} . Then we took 500 samples of size $n = 10$ and recorded \hat{p} , then 500 samples of size $n = 25$ and recorded \hat{p} , and finally 500 samples of size $n = 50$ and recorded \hat{p} .

Below are the corresponding histograms.

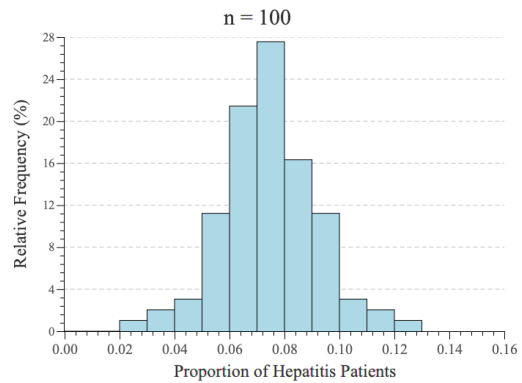
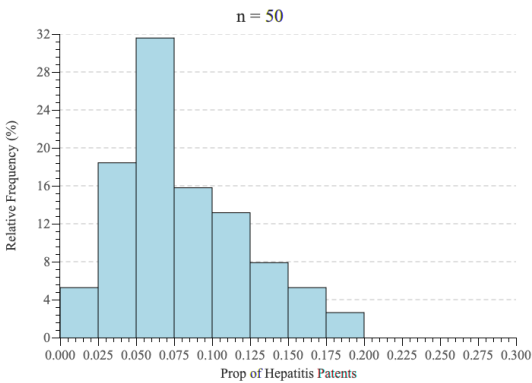
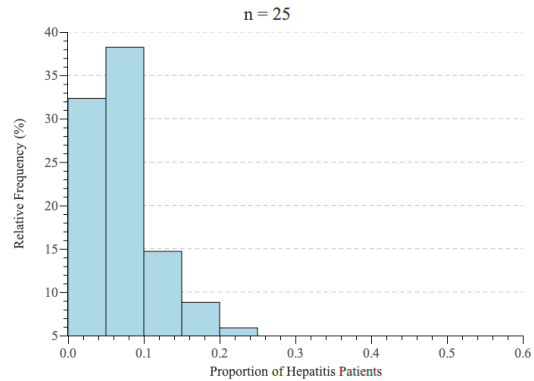
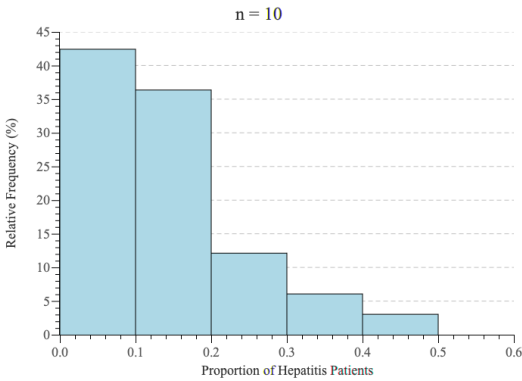


What do you notice about the histograms in terms of SOCS?

Example 2 The development of viral hepatitis subsequent to a blood transfusion can cause serious complications for a patient. Suppose it is known that hepatitis occurs in 7% of patients who receive multiple blood transfusions during heart surgery.

Let S denote a recipient who contracts hepatitis. Then $p = 0.07$.

The following displays the histograms of 500 values of \hat{p} for the four sample sizes $n = 10, 25, 50,$ and 100 .



What do you notice about the histograms in terms of SOCS?

General Properties of the Sampling Distribution of a sample proportion \hat{p}

Let \hat{p} be the proportion of successes in a random sample of size n from a population whose proportion of successes is p . Denote the mean value by $\mu_{\hat{p}}$ and the standard deviation by $\sigma_{\hat{p}}$. Then the following rules hold:

Rule 1: $\mu_{\hat{p}} = p$

Rule 2: $\sigma_{\hat{p}} = \sqrt{\frac{p(1-p)}{n}}$

Assumption #1: When n is large and p is not too near 0 or 1, the sampling distribution of \hat{p} is approximately normal. A conservative rule of thumb is that if both

$np \geq 10$ and $n(1-p) \geq 10$, then it is safe to use a normal approximation.

Assumption #2: The sample size is less than 10% of the population size.

Ex3 You have an SRS of 300 students selected from over 100,000 college students. Of your sample, 35% said they had fallen asleep in their English class at least once during the previous semester. The mean and standard deviation of this statistic are:

(a) $\mu_{\hat{p}} = 0.35, \sigma_{\hat{p}} = 0.000758$

(b) $\mu_{\hat{p}} = 0.105, \sigma_{\hat{p}} = 8.26$

(c) $\mu_{\hat{p}} = 0.105, \sigma_{\hat{p}} = 0.028$

(d) $\mu_{\hat{p}} = 0.35, \sigma_{\hat{p}} = 0.028$

(e) This question cannot be answered because the sample size is not large enough relative to the population size.

Example 4 The article “Should Pregnant Women Move? Linking Risks for Birth defects with Proximity to Toxic Waste Sites” reported that in a large study carried out in the state of New York, approximately 30% of the study subjects lived within 1 mile of a hazardous waste site. Let p denote the proportion of all New York residents who live within 1 mile of such a site, and suppose that $p = 0.3$.

(a) What are the mean value and standard deviation of \hat{p} based on a random sample of size 10?

(b) Would \hat{p} based on a random sample of only 10 residents have an approximately normal distribution? Explain why or why not.

(c) When $n = 10$, what is $P(0.25 \leq \hat{p} \leq 0.35)$?

(d) What are the mean value and standard deviation of \hat{p} based on a random sample of size 400?

(e) Would \hat{p} based on a random sample of only 400 residents have an approximately normal distribution? Explain why or why not.

(f) When $n = 400$, what is $P(0.25 \leq \hat{p} \leq 0.35)$?

Example 5 Given that 58% of all gold dealers believe next year will be a good one to speculate in South African gold coins, in a simple random sample of 150 dealers, what is the probability that between 55% and 60% believe that it will be a good year to speculate?

- (a) 0.0500 (b) 0.1192
- (c) 0.3099 (d) 0.4619
- (e) 0.9215

Example 6 According to the manufacturer, the average proportion of red candies in a package is 20%. An 8 oz package contains about 250 candies. What is the probability that a randomly selected 8 oz bag contains less than 45 red candies?

- (a) 0.788 (b) 0.317
- (c) 0.215 (d) 0.155
- (e) None of these

Checkpoint:
Multiple Choice

1. According to the US Census, the proportion of adults in a certain county who owned their own home was 0.71. An SRS of 100 adults in a certain section of the county found 65 owned their homes. What is the probability of obtaining a sample of 100 adults in which 65 or fewer own their own homes, assuming this section of the county has the same overall proportion of adults who own their homes as does the entire county?

- (a) 0
- (b) 0.093
- (c) 0.106
- (d) 0.186
- (e) 0.907

2. Historically for the Central Florida Blood Bank, 13% of first-time donors return to make a second donation within three months. In an effort to determine if 13% was out of date, in the fall of 2007 they tracked 6000 first-time donors and found that 891 donated a second time within 90 days. What is the probability of seeing a sample proportion this high or higher if, in fact, the 13% is the true long term rate?

- (a) Less than 0.001
- (b) 0.007
- (c) 0.014
- (d) 0.492
- (e) 0.983

3. About 25% of all dogs live more than 10 years. Out of a random sample of 80 dogs, what's the probability that between 15 and 20 dogs will live more than 10 years?

- (a) 0.40
- (b) 0.09646
- (c) 0.02
- (d) 0.50
- (e) You cannot do a problem like this unless the population is at least 10 times the sample size.

4. Suppose voters from a simple random sample of 500 ($N > 1,000,000$) are interviewed and asked which presidential candidate they're going to vote for. Of these, 35% say they'll vote for the Statistics Party. You want to know the probability, assuming this proportion is correct for the population, that more than 40% of a random sample of 500 people will vote for the Statistics Party. Which of these shows two of the ways you could find your answer.

- (a) Proportion, using a normal approximation: $1 - \mathit{binomcdf}(500, 0.35, 200)$
Exact binomial count: $\mathit{normalcdf}(0.40, 1E99, 0.35, 0.0213)$
- (b) Proportion, using a normal approximation: $\mathit{normalcdf}(200, 1E99, 175, 10.66)$
Exact binomial count: $\mathit{normalcdf}(0.40, 1E99, 0.35, 0.0213)$
- (c) Proportion, using a normal approximation: $\mathit{normalcdf}(0.40, 1E99, 0.35, 0.0213)$
Exact binomial count: $1 - \mathit{binomcdf}(500, 0.35, 200)$
- (d) Using Central Limit Theorem: $1 - \mathit{binomcdf}(500, 0.35, 200)$
As a sampling distribution: $\mathit{normalcdf}(200, 1E99, 175, 10.66)$

(e) Using Central Limit Theorem: $1 - \text{binomcdf}(500, 0.35, 200)$

As a sampling distribution: $1 - \text{binomcdf}(500, 0.40, 200)$

5.5 Homework

- 1) Student heights were collected at the time of freshmen registration at SI from 2016 to 2021. The proportion of girls who are 5 foot 6 or taller is 0.32 while the proportion of boys who are 5 foot 6 or taller is 0.66. If samples of 20 boys and 20 girls are collected, will the sampling distributions of each sample proportion be normal? Explain.
- 2) What is the smallest value of the sample size n for which the sampling distribution can be considered to be approximately normal?
- 3) Determine the mean and standard deviation of each sampling distribution for the n value in #2.
- 4) Early in the year this class collected and sorted M & M's by color. From the resulting spreadsheet, the proportion of M & M's that are orange was determined to be $p_o = 0.230$. To test whether this proportion is an unbiased estimate of the proportion of all orange M & M's made, what criteria would have to be met? What is the minimum amount of M & M's that would have to be in any randomly drawn bag?
- 5) A standard vending machine bag of M & M's contains approximately 50 M & M's. Based on the value of p in #4, several bags of 50 are randomly opened and counted. What is the probability of finding at least 15 orange M & M's?

5-6 Sampling Distributions for Differences in Sample Proportions

Objectives:

- Compare the sampling distributions for two sample proportions
- Determine the mean of the difference between two sample proportions
- Determine the variance and standard deviation of the difference between two sample proportions.
- Calculate probabilities based on the distribution of a sample proportion \hat{p}

\hat{p} = sample proportion

Difference in population proportions $p_1 - p_2$

Difference in sample proportions $\hat{p}_1 - \hat{p}_2$

Mean for a distribution of differences in sample proportions: $\mu_{\hat{p}_1 - \hat{p}_2} = p_1 - p_2$

Standard Deviation for a distribution of differences in sample proportions:

$$\sigma_{\hat{p}_1 - \hat{p}_2} = \sqrt{\frac{p_1(1-p_1)}{n_1} + \frac{p_2(1-p_2)}{n_2}}$$

Assumption #1: When both sample sizes n_1 and n_2 are large and both p_1 are not too near 0 or 1, the sampling distribution of \hat{p} is approximately normal. A conservative rule of thumb is that if both

$np \geq 10$ and $n(1-p) \geq 10$, then it is safe to use a normal approximation.

Assumption #2: The sample size is less than 10% of the population size.

Example 1: According to a 2014 study by the American Academy of Ophthalmology, 45% of Americans have brown eyes while 27 % have blue eyes. If we were to use a sampling distribution to demonstrate that the difference in proportions of brown vs blue eyes is 0.18, what is the smallest number for n to make $\hat{p}_{br} - \hat{p}_{bl}$ an unbiased estimate? Using this value for n , find the probability that a random sample will contain at least 20 left handed people

Example 1: Mr Maychrowitz can solve a three by three Rubik's cube in under a minute 65% of the time while Mr. Murphy can do the same 25% of the time. Some students decide to record

each one's times over a sample of attempts each at solving a cube. They collect 50 for Mr. Maychrowitz and 40 (because they just couldn't wait that long) for Mr. Murphy. Find the probability that the difference between their percentages will be at most 20%

5-6 Checkpoint

Multiple Choice

1. In a random sample of 200 University of Manitoba graduate students, it was found that 66% of them had previously attended some other college or university. In a random sample of 100 University of Waterloo graduate students, it was found that 35% of them had previously attended some other college or university. The standard deviation of the difference in proportions between the two colleges is:

- (a)
- (b)
- (c)
- (d)
- (e)

The next two questions refer to the following situation:

One criticism of reforestation efforts after timber harvesting is that too few of the seedling survive. An experiment was conducted to assess if mulching the slash (limbs, roots, small branches, etc.) and leaving the mulch on the ground improves the survival rate compared to just leaving the slash on the ground. It is believed that mulching will cause the material to break down sooner and release the nutrients to the seedlings. A total of 500 seedlings were randomly assigned to the two treatments and the two year survival rate was measured. Of the 250 seedling receiving the "mulching" treatment, 75 survived; of the 250 seedlings receiving the "control" treatment, 55 survived.

2. The difference between the sample proportions of surviving plants treated with mulch and those treated with the control ($m - c$) is

- (a) 25
- (b) 0.3
- (c) 0.08
- (d) 20
- (e) 0.52

3. The variance of the difference between the sample proportions of surviving plants treated with mulch and those treated with the control ($m - c$) is

- (a)
- (b)
- (c)
- (d)
- (e)

Unit 5-6 Homework

- 1) From the data on student height that we collected, the proportion of girls who are 5 foot 6 or taller is 0.32 while the proportion of boys who are 5 foot 6 or taller is 0.66. If samples of 50 boys and 60 girls are taken
 - (a) Explain how we can assume normality with both sampling distributions?
 - (b) Find the mean and standard deviation of the difference between the sampling distributions between the boys and the girls.
 - (c) Find the probability that a random sample of girls will have a higher proportion than a random sample of boys.

- 2) Early in the year this class collected and sorted M & M's by color. From the resulting spreadsheet, the proportion of M & M's that are orange $p_o = 0.230$ and the proportion that are yellow is $p_y = 0.141$
 - (a) What is the smallest sample size needed for a sampling distribution of the sample proportion of yellow M & M's to be assumed to be normal?
 - (b) Find the mean and standard deviation of the difference between the sampling distributions between the orange and yellow M & M's using the value of n from part (a).
 - (c) Find the probability that a random bag of M & M's will have more yellow than orange.

Unit 5 Practice Test: