Chapter 2

Sampling Distributions

This ActiveStats document contains a set of activities for Introduction to Statistics, MA 207 at Carroll College. This is a non-calculus based statistics class which serves many majors on campus. This document is intended for the classroom teacher to support students in active engagement with statistics on a daily basis. This document is not designed to be given to students as is. Rather, it is a teacher resource.

The activity set is designed to work alongside the OpenIntro *Introductory Statistics with Randomization and Simulation* textbook by Diez, Barr, and Cetinkaya-Rundel. The chapters in ActiveStats are numbered to align with OpenIntro, though the subsections may differ. OpenIntro is an open source curriculum with accompanying data sets. OpenIntro is the textbook resource to direct students to for out-of-class reading assignments and review. We also use the Cartoon Guide to Statistics as a supplement for assigned reading.

Data sets for ActiveStats can be found at mathquest.carroll.edu/activestats/data/ or on the class Moodle page.

2.1 Introduction to sampling distributions

Teacher Note: The goal of this lesson is to begin to explore sampling distributions through simulations in TinkerPlots. At this point, sampling distributions are used to quantify weirdness and begin to reinvent the idea of p-values in an intuitive way.

Reading Assignment 2.1. Null and alternative hypotheses

- In the Cartoon Guide to Statistics, read pages 137-142. Focus on the language of hypothesis testing.
- In your notes, record definitions for: null hypothesis, alternative hypothesis, test statistic, and significance level.

▲

Preview Activity 2.2. Coin Flipping:

- If you flip a coin 20 times, how many heads do you anticipate? Rather than providing a one number answer, provide an interval of what you think is reasonable, i.e. between (lower number) and (upper number).

- Conduct an experiment. Flip a coin 20 times and record the number of heads. Record the number of heads on the class Google spreadsheet linked on Moodle.
- Did your count of heads surprise you? Was it outside the range of what you expected?
- In the coin flipping experiment, what is your null hypothesis?

Activity 2.3. Variability in coin flips

In the preview you were asked to state an interval for how many heads you might expect if you flip a coin 20 times. You may have relied on your intuition for this. One tool statisticians use is simulations. Statisticians use software to try an experiment over and over and determine what types of outcomes are common. This allows a statistician to recognize when something unusual or unexpected happens.

Your instructor will guide you through how to build a fair coin simulation in TinkerPlots to model flipping a coin 20 times and repeating that experiment many many times. Be sure to click along and create the same simulation on your own computer.

Once your simulator is built and run, you will have a graph displaying the sampling distribution. Use your sampling distribution to answer the following:

1. How likely is it to get exactly 10 heads (out of 20)?
2. How likely is it to get 8 heads or less?
3. How likely is it to get 5 heads or less?
4. How likely is it to get between 40% and 60% (inclusive) heads?
5. Fill in the blank: The middle 95% of experiments had between ____ and ____ heads.
6. If you flip a coin 20 times and get 18 heads, would you think the coin was unfair/weighted? Why or why not?

▲

The fair coin example is a classic example because it uses a concrete object and repeated sampling to build a sampling distribution. However, statisticians do not spend a great deal of their time wondering about fair coins. They do spend time thinking about whether a random sample from a popultion is typical or highly unusual. The same tools that are used to predict fair and unfair coins can also be used to examine claims about populations.

Activity 2.4. Building and making sense of a sampling distribution
Underage drinking scenario

A few college students were discussing the prevalence of underage drinking on campus. Based on anecdotes from their peers, the students believe that roughly 50% of underage college students drink. They later find out that a survey was given to 76 randomly selected underage college students on their campus. Only 32 of the 76 reported regularly participating in underage drinking.

- Identify the research question.
- Is this a controlled experiment? If so, what are the treatment(s)?
- Null Hypothesis:
  Alternative Hypothesis:
- We can use a simulator to determine the chance of getting a random sample as low as 32 out of 76 (42.1%). If the null hypothesis is true, the probability of getting data this unusual or more unusual is called the **p-value.**
- Estimate the p-value. What conclusion can we make?
- Can we draw a cause / effect relationship out of this study?

Teacher Note: Build a simulation (e.g. TinkerPlots) to go along with the drinking study. Talk about the impact of the underlying assumption of 50%. If you thought 60% of students drank, how would that change your mathematics? Where would it change in the simulator? Encourage the students to build the simulators along with you, so they are more prepared for the next lab. Also keep in mind that 32 out of 76 is about 42% and choose counts or percents purposefully.

A sampling distribution shows us how spread out we expect randomly selected samples to be. A sampling distribution gives us the power to recognize whether one sample (of size 76 in this case) is weird or typical. If a sample is particularly weird, we should suspect that

- It didn't come from the population we thought it did ... [interesting], OR
- Our null hypothesis might be wrong ... [interesting], OR
- Our participants are lying ... [interesting], OR
- The sample was unusual because variability is part of statistics [This is possible, but it has a low probability of being the explanation]

If our sample is not very weird, its variability could be just from random chance. Our job is to quantify "weird" in a mathematical / statistical way

▲

Activity 2.5. Underage drinking scenario - Modified

At a larger state school, a study from 2005 indicated that 65% of underage students drank alcohol. The school engaged is a sustained campaign to lower the rates of drinking across campus. This year, a survey was given to 150 randomly selected underage students at the school. Only 83 of the students reported drinking (55.3%).

- What is the null hypothesis and how do you use it in your simulator?
- What is the alternative hypothesis and why does it not appear in your simulator?

- If the drinking rates have not changed since 2005, whats the chance of getting a sample as low (or lower) than the one in this study? i.e. What portion of your randomly generate samples were this low?
- What is the p-value and what does it mean?
- Was this a controlled experiment?
- Can we draw a cause / effect relationship out of this study?

▲

Activity 2.6. Building and making sense of a sampling distribution
At a typical 4 year college, about 25% of students are in each class. Build a sampler that has 25% Freshmen, 25% Sophomores, 25% Juniors, and 25% Seniors. The number of students who access health services on campus should be the same across the 4 classes, but there is a concern that freshmen are less likely to seek help. Last week, out of 120 students who visited health services, only 20% were freshmen. Is this 20% in the range of normal variability, or does it represent a statistically extreme situation?

- What is the null hypothesis and how do you use it in your simulator?
- What is the alternative hypothesis?
- Build a simulation to learn how random samples of size 120 behave. One might anticipate that there would be 30 students from each class, but the simulation will show what level of variability is typical.
- Create a sampling distribution for the proportion of freshmen and determine where the middle 95% of samples are.
- How unusual is a sample with 20% or fewer freshmen? Find the p-value and state your results as a complete sentence related to the context.

Teacher Note: This could be used as a χ^2 problem if we offer data about all 4 student types. However, this can also be done as a simple one-proportion problem if we classify as freshmen vs. non-freshmen. When building the simulator for this task, you can use a spinner with 4 sections for the 4 classes. This problem could also be modified to reflect different proportions. For example, the freshmen class may be larger than 25% at your school. ▲

Activity 2.7. Where are you from?
A local reporter recently mentioned that 30% of Carroll students are from Montana and the other 70% are from other states. We are curious if that claim is accurate.

- State the null and alternate hypotheses.
- Create a simulation. When chosing the sample size for this experiment, let's consider a convenient source of data. Your class (and possibly other sections) have recently completed a demographic survey, which includes home state. The teacher will provide the results from this survey, including the sample size.

- Compare the results from the survey (the proportion $\hat{p}$ of students from Montana) to determine whether the reporters claim is reasonable or not.
- The data we used from the demographic survey wasn't really a random sample. Why not? And what bias might this cause? How could we get a better random sample?

Teacher Note: If home state isn't particularly interesting, choose any of the other qualitative attributes from the class demographic survey. Sample sizes between 20 and 100 are appropriate for this task, so long as there are more than 5 people in each category of interest. Later we set our minimum at 10 items per category, when we want to approximate the sampling distribution with a normal distribution. ▲

Definition 2.8. A statistical *simulation* is a way to model a statistical situation, generally using a computer to generate repeated samples of a statistical situation. TinkerPlots is a tool for creating statistical simulations.

Definition 2.9. A *sampling distribution* is a collection of all the possible values a statistic (e.g. $\bar{x}$ or $\hat{p}$) can be, along with the likelihood of each result. When working with the TinkerPlots simulation for proportions, the sampling distribution is the set of collected statistics, usually displayed as a plot which is roughly bell curve in shape.

While the sampling distribution for sample means and sample proportions is roughly bell shaped, other sampling distributions can take on different shapes, e.g. F or χ^2.

2.2 Lab Day - Lab 2: Simulations for proportions

Reading Assignment 2.10. • Task #1 Read section 2.1 in the OpenIntro online textbook (p. 61-65). This is a case study about gender discrimination.

▲

Preview Activity 2.11. Follow-up based on the reading

1. Why is it reasonable that the sampling distribution centers near 0 for this scenario? Circle all that apply.
   (a) Because men and women are treated equally in the work force. Under the Equal Rights Amendment, this is required by law.
   (b) Because if gender does not affect promotion, about half of the time women will be promoted at a higher rate and about half of the time men will be promoted at a higher rate.
   (c) Because the simulation uses the assumption that gender is not affecting promotion rates. i.e. promotion and gender are independent.
   (d) Because this experiment was conducted at 100 different companies and this data reflects what really happened with 100 experiments.

2. According to the sampling distribution from the simulation discussed in this section, how rare is it to find a difference in promotion rates of 15% or more. Include both the cases when men outnumber women and when women outnumber men.

   Solution: 18% chance... code this with a tolerance of 1% for counting errors

3. Did this case study report strong evidence of gender discrimination? If so, how strong was the evidence?

   Solution: The study reports "we determined that there was only a 2% probability of obtaining a sample where 29.2% more males than females get promoted by chance alone, so we conclude the data provide strong evidence of gender discrimination against women by the supervisors" (Diez et al., p.65)

Activity 2.12. Lab 2 includes simulations for 3 scenarios:

- Tire company
- Weight loss drug
- Spicy salsa

All scenarios in Lab 2 are one-proportion problems that can be simulated with a spinner in TinkerPlots.

Teacher Note: There is a video link to learn how to create a simulation in TinkerPlots. I share this with students only after we do this in class. I prefer to demo this in class, together, and use the video as a back-up / review. The video is the *same* task as the lab. Video link - youtu.be/tX90rXswax4

▲

2.3 Probability and Intro to Pivot Tables

Reading Assignment 2.13. Probability

- Read the intro probability section in the Cartoon Guide to Statistics pages 27 to 39.
- Work with the videoGamesXBoxPS3 data set (from Moodle) to find probabilities.

▲

Preview Activity 2.14. 1. Now that you're done reading the section on probability from the Cartoon Guide to Statistics please answer these questions:

(a) The largest a probability can be is:

(b) The smallest a probability can be is:

(c) Consider rolling a black 6-sided die and a white 6-sided die. Let A be the event that the white die is a 2. Let B be the event that the black die is even. What is the probability: P(A and B)?

(d) Consider rolling a black 6-sided die and a white 6-sided die. Let A be the event that the white die is a 2. Let B be the event that the black die is even. What is the probability: P(A or B)?

(e) Consider rolling a black 6-sided die and a white 6-sided die. Let A be the event that the white die is a 2. What is the probability: P(not A)?

2. Open the *videoGamesXBoxPS3* data set from Moodle. This data set is a randomly selected subset of video games, taken from *kaggle.com/datasets*. The data set is limited to games on the XBox and PS3 platforms which contained reviews from critics and users. A larger data set is available on *kaggle.com/datasets* if you would like to explore further.
Use your Excel sorting and counting skills to find counts for each of the following. **Teacher Note:** This is a preview before the formal introduction of pivot tables in class. This preview motivates finding a more efficient strategy for counting items in Excel.

(a) How many XBox games are listed in the data set?

(b) How many PS3 games are listed in the data set?

(c) How many of the games listed are rated M for mature audiences?

(d) (optional) Of those games rated M, how many are for the XBox?

(e) (optional) Of those games rated M, how many are for the PS3?

In Lab 2, you built simulators for a tire company and a pharmaceutical company. In both cases, you were told about prior information in order to set up your simulators. For the tire company, you knew that 75% of the tires last at least 30,000 miles. If you had been told that 80% of tires lasted at least 30,000 miles, it would have altered the probabilities in the simulator. Therefore the 75% or 80% are conditions which affect the probability. We are going to spend a few days learning about probability, especially conditional probability.

Definition 2.15. Probability

The **probability** of an outcome is the proportion of times the outcome would occur if we observed the random process an infinite number of times. - OpenIntro

Another definition: The **probability** of an event A is the ratio of the cases that are part of event A out of all possible cases in the sample space.

ex) With a deck of 52 cards, Probability(picking a facecard) = 12 / 52.

ex) With a fair 6-sided die, Probability(rolling a 2) = 1/6

Definition 2.16. Independent

When the probability of an event does not change just because another event occurs, then that event is **independent** of the other. If events A and B are independent then $P(A\ and\ B) = P(A) \times P(B)$

Activity 2.17. Pivot table introduction

Open the *videoGamesXBoxPS3* data set from Moodle. Suppose we want to know how many of the games for each console type have a rating of M for Mature, T for teen, E for everyone, or E+10. One option is to sort the data in Excel and count different subsets. A more efficient option is to use pivot tables in Excel. Your instructor will demonstrate how to set up a basic pivot table to find counts.

Fill in the following chart with counts from Excel.

	M for mature	T for teen	E+10	E for everyone	Total
PS3					
Xbox					
Total					

1. If a game is selected at random, what's the chance it is rated E?
2. If a game is selected at random, what's the chance it is rated E and it's an XBox game?
3. If a game is selected at random, what's the chance it is rated E and it's a PS3 game?
4. If a game is selected at random, what's the chance it is rated M?
5. What proportion of M rated games are for XBox? (This time, limit your focus to only M-rated games)

▲

Activity 2.18. A second pivot table task

Open the NCBabySmoke data set from Moodle. This data set is from the OpenIntro textbook. It represents a random sample of 1000 mothers and their newborns in North Carolina. A codebook for the data set is provided below.

column name	description and units
fage	father's age
mage	mother's age
mature	under 35 vs. 35 or older
weeks	length of pregnancy
premie	premie or full term
visits	number of doctor visits
marital	married or not married
gained	weight gained by mom (lbs)
weight	weight of baby (lbs)
lowbirthweight	low is ≤ 5.5 lbs
gender	baby's gender
habit	smoking habit of mom
whitemom	white or not white

Use pivot tables to answer the following probability questions.

1. If a baby is randomly selected from this data set, what's the chance it is a premie?
2. If a baby is randomly selected from this data set, what's the chance it is male?
3. If a baby is randomly selected from this data set, what's the chance it is a male premie?
4. What's the chance a randomly selected baby weighs less than 5.5 lbs?
5. If a mom is randomly selected from this data set, what's the chance she smokes?
6. If a mom is randomly selected from this data set, what's the chance she is 35 or older?
7. If a mom is randomly selected from this data set, what's the chance she is married?
8. What's the chance of randomly selecting a mom who is 28 years old?

▲

Teacher Note: The following set of activities are optional, depending on the level of emphasis on probability in your course.

Activity 2.19. Multiple choice probability questions:

A consumer organization estimates that over a 1-year period 17% of cars will need to be repaired once, 7% will need repairs twice, and 4% will require three or more repairs. What is the probability that a car chosen at random will need exactly 1 repair?

1. What is the probability that a car chosen at random will need exactly 1 repair?
   (A) 0.17 (B) 0.07 (C) 0.89 (D) 0.28 (E) 0.72

2. What is the probability that a car chosen at random will need no repairs?
   (A) 0.17 (B) 0.07 (C) 0.89 (D) 0.28 (E) 0.72

3. What is the probability that a car chosen at random will need at least 1 repair?
(A) 0.17 (B) 0.07 (C) 0.89 (D) 0.28 (E) 0.72

4. What is the probability that a car chosen at random will need no more than 1 repair?
(A) 0.17 (B) 0.07 (C) 0.89 (D) 0.28 (E) 0.72

5. If you own two cars, what is the probability that both will need exactly one repair?
(A) 0.0289 (B) 0.0049 (C) 0.7921 (D) 0.0784 (E) 0.5184

6. If you own two cars, what is the probability that neither care will need a repair?
(A) 0.0289 (B) 0.0049 (C) 0.7921 (D) 0.0784 (E) 0.5184

7. What other probability questions could you write for this scenario?

▲

Activity 2.20. The following table provides information on housing units in some part of the U.S. The top row indicates how many rooms the housing unit has. The second row indicates how many thousands of units correspond with that number.

Rooms	1	2	3	4	5	6	7	8+
Thousands of units	47	140	1170	2350	2450	2130	1370	1560

Find the probability of each of the events A through E and then answer the follow-up questions.
A = the unit has at most 4 rooms
B = the unit has at least 2 rooms
C = the unit has between 5 and 7 rooms (inclusive)
D = the unit has more than 7 rooms
E = the unit has less than 3 rooms

1. Find: P(A or D)

2. Find: P(A and D)

3. Find: P(A or C)

4. Find: P(B or E)

5. Find: P(not B)

▲

2.4 Conditional Probability

Reading Assignment 2.21. Watch the 20 minute TED talk by Peter Donnelly link here

▲

Preview Activity 2.22. Follow-up to the TED talk

1. Suppose that we have a test for a rare disease that gets it right 99% of the time. If you take a person from the general populous at random and they test positive for the disease, what is the chance that they have the disease?
   (a) much less than 99%
   (b) a bit less than 99%
   (c) 99%
   (d) more than 99%
2. Why is the correct answer what it is in the previous question?
   (a) there will be many false positives
   (b) there will be many false negatives
   (c) there will be 1 person in 100 who tests positive
   (d) the true accuracy of the test cannot be known
3. There are two unlikely events that occur in this scenario: Either (A) a person has the disease (unlikely) and tests positive (likely), or (B) a person does not have the disease (likely) and tests positive (unlikely). In a rare disease, if 1,000,000 people are in the population and only 100 people have the disease, then which is more likely:
   (a) A
   (b) B
   (c) They are about the same.

Activity 2.23. Risk Assessment

In the book *Gut Feelings*, the author describes a study where he asked 24 doctors to estimate the following probability. Only 8 of the doctors were anywhere close! (G. Gigerenzer Gut Feelings: The Intelligence of the Unconscious, Penguin Books 2008)

"In your clinic the probability that one woman has breast cancer is 0.8 percent. If a woman has breast cancer, the probability is 90 percent that she will have a positive mammogram. If a woman does not have breast cancer, the probability is 7 percent that she will still have a positive mammogram. Imagine a woman who has a positive mammogram. What is the probability that she actually has breast cancer?" Make your best guess for the probability:

(a) Less than 25% (b) Between 25% and 50% (c) Between 50 and 75% (d) More than 75% Justify your thinking.

Using the same breast cancer information as above, let's create a strategy for understanding the problem.

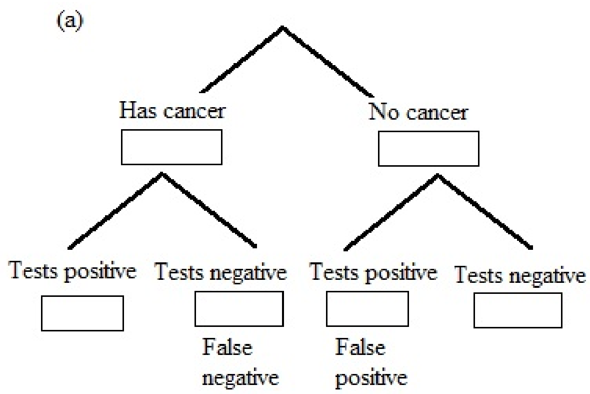
- Assume that 100,000 women come through the clinic. Fill in the contingency table. Hint: In this example it makes sense to fill in the bottom row first.

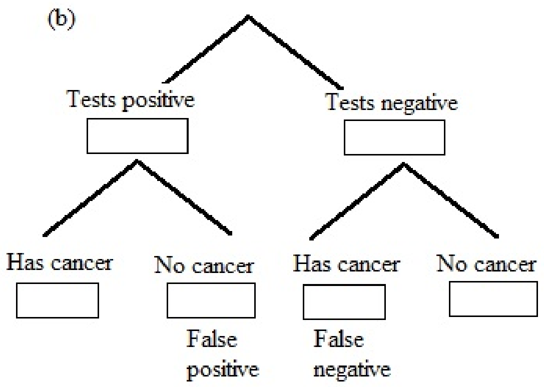
	Has cancer	No cancer	Total
Tests positive			
Tests negative			
Total			

Solution:

	Has cancer	No cancer	Total
Tests positive	720	6944	7664
Tests negative	80	92256	92336
Total	800	99200	100000

- A flowchart or tree can also be used to organize your thinking. For the problem described above, why is flowchart (a) a better idea than flowchart (b) for this particular problem?





Solution: Because the information provided tells us how likely cancer is, we can use that as our starting point. The probability of positive and negative results then relies on whether the patient has cancer or not. If we use the tree on the right, we do not start with enough info to know the probabilities to put under tests-positive and tests-negative.

- Fill in the numbers on flowchart (a) with proportions. The sum of all four bottom leaves should be approximatley 1. If you choose to use counts instead of proportions, the total would be 100,000.

- Now to answer some conditional probability questions

1. What proportion of women with cancer get a positive result?
   $P(positive|cancer) =$ **Solution:** 720/800
2. What proportion of women without cancer get a positive result?
   $P(positive|cancer) =$ **Solution:** 6944/99200
3. If a woman gets a positive result, is it more likely that she is from the cancer subset of the population or the non-cancer subset? **Solution:** 720 vs. 6944... she's probably in the non-cancer group
4. If a woman gets a positive test result, what is the likelihood she has cancer?
   $P(cancer|positive) =$ **Solution:** 720/7664
5. If a woman gets a positive test result, what is the chance she does not have cancer?
   $P(no\ cancer|positive) =$ **Solution:** 6944/7664
6. What portion of tests results are correct?
   $P(true\ negatives + true\ positives) =$ **Solution:** (720+92256)/100000
7. If a woman gets a negative result, what is the chance she does not have cancer?
   $P(no\ cancer|negative) =$ **Solution:** 92256/92236
8. According to the contingency table, only 80 women with cancer get a negative test result. Why would doctors want to keep this false negative number as low as possible? i.e. Why do we want it to be so much lower than the other boxes?

▲

Activity 2.24. Let's return to the *videoGamesXBoxPS3* data set from Moodle. In the previous section we used a pivot table in Excel to fill in the chart below. Recreate your pivot table in Excel and answer the conditional probability questions below.

	M for mature	T for teen	E+10	E for everyone	Total
PS3					
Xbox					
Total					

For each of the following, be careful in your selection of the divisor. Is the divisor the full count of video games, or a specific subset?

1. If a game is selected at random, what's the chance it is rated M?
2. If a game is rated M, what's the chance it's for PS3?
3. If a game is rated M, what's the chance it's for XBox?
4. If a game is for Xbox, what's the likelihood it is rated E or E+10?
5. If a game is for PS3, what's the likelihood it is rated E or E+10?
6. Problems (2) and (3) should sum to 1. Why?
7. Problems (4) and (5) likely do not sum to 1. Why?

▲

Activity 2.25. A certain disease has a prevalence of 2% in a population. Let's assume that a test for the disease is developed and it is estimated that the test is 99% accurate. In this case, "accurate" means that if you have the disease the test is positive and if you do not have the disease the test is negative.

1. Use your intuition to estimate the probability that a person with a positive test actually has the disease.

2. Assume there are 10,000 people in the population. Determine the probability via the contingency table.

	Has Disease	No Disease	Total
Tests positive			
Tests negative			
Total			

3. Then use a flowchart to organize the probability.

4. Find the following probabilities

   (a) What is the probability of a true positive test? A true negative test?
   Solution: true pos=0.0198, true neg=0.9702

   (b) What is the probability of a false positive test? A false negative test?
   Solution: false pos=0.0098, false neg =0.0002

   (c) What is the probability of a negative test given that you have the disease?
   Solution: P(neg result||have disease)=0.01

   (d) What is the probability of having the disease if you test positive?
   Solution: P(having disease||pos test)=0.669

   (e) What is the probability of having the disease if you test negative?
   Solution: P(having disease||neg test)=0.000206

▲

Activity 2.26. Binge drinking and car accidents

Teacher Note: Use as a worksheet, have students work with a partner.

For men, binge drinking is defined as having five or more drinks in a row, and for women as having four or more drinks in a row. According to a study by the Harvard School of Public Health, 44% of college students engage in binge drinking, 37% drink moderately, and 19% abstain entirely. Another study, published in the American Journal of Health Behavior, finds that among binge drinkers aged 21-34, 17% have been involved in an alcohol-related automobile accident, while among non-binge drinkers of the same age, only 9% have been involved in alcohol-related accidents.

Create a contingency table to organize the information provided in the paragraph. Note that some labels are provided in the table, but you will need to choose the other labels before filling in the table. Once the table is complete, find the following probabilities.

	binge	non-binge	total
total			

Solution:

	binge	non-binge	total
accident	$0.17 \times 0.44 = 0.0748$	$0.09 \times 0.56 = 0.0504$	0.1252
no accident	$0.83 \times 0.44 = 0.3652$	$0.91 \times 0.56 = 0.5096$	0.8748
total	0.44	0.56	

1. Probability that a randomly selected student is a binge drinker
2. Probability that a randomly selected binge drinker has been in an accident
3. Probability that a randomly selected student is not a binge drinker (either drinks moderately or abstains)
4. Probability that a student who drinks moderately or abstains has been in an accident
5. Probability that someone who has been in an accident is a binge drinker
6. Probability that someone who has been in an accident is not a binge drinker

▲

2.5 Lab 3 - Conditional probability

Reading Assignment 2.27. Read Lab 3 before class ▲

Preview Activity 2.28. Download Lab 3 and complete the first question before class. Submit your answers on Moodle.

Activity 2.29. Lab 3 focused on conditional probability using the scenarios of

- The relationship between a child's choice to attend college and their parents' educational status
  Data set available in Lab3DataSet
- The relationship between seat belts and rates of serious injuries
- Surveys with sensitive questions: A survey is described in which some participants are asked if they use marijuana, while other participants are asked a non-threatening question about their phone number. Because the participant is assured that the researcher will not know which question they are answering, the participants may be more comfortable answering truthfully. In this task, students work backwards from the results to determine the rates of marijuana use within the sample.
- The role of conditional probability in simulations:
  This task relies on ideas from Lab 2. The task is included here to keep those simulation ideas fresh in mind and to connect sampling distributions to probability.

▲

2.6 Hypothesis testing and sampling distributions

Reading Assignment 2.30. In the OpenIntro stats book, read Sect 2.3, pages 68 to 76 about the language of hypothesis testing. Then answer the preview questions about null and alternative hypotheses in context problems. ▲

Preview Activity 2.31. Hypothesis testing

1. A major wind storm sweeps through the Helena area and damages many houses. A news reporter claims that 40% of the residential houses on the west side of Helena are going to need new shingles after the storm. An insurance broker hears this report and begins to panic! Then he realized that he has statistics on his side! His conjecture is that fewer than 40% of the houses on the west side will need new shingles.

   (a) What is the null hypothesis?

   (b) What is the alternative hypothesis?

   (c) Let's assume that the insurance broker uses a proper sampling technique. He finds a p-value of 0.073 for the hypothesis test. What is the conclusion?

2. The Yummy-Math-O Cereal Company claims that the weight of the average box of Stat-Loops Kids Cereal is 18.5 ounces. An independent quality control group gathers a sample of Stat-Loops boxes and finds an average of 17.1 ounces. They would like to do a hypothesis test to determine if the Yummy-Math-O Cereal Company is falsely advertising their average weight.

   (a) What is the null hypothesis?

   (b) What is the alternative hypothesis?

   (c) The quality control group finds a p-value of 0.012. What is the conclusion?

Activity 2.32. In Statsville county, the population eligible for jury selection is 75% white and 25% black. In a particular trial there 11 white jurors and 1 black juror. This is only 8.3%, as opposed to the expected 25%. Is this statistical evidence of a bias in jury selection?

1. Make a prediction.

2. State the null and alternative hypotheses in clear language.

3. Model the situation in TinkerPlots to test your prediction. Use 100 samples of size 12. (You may want to save this TinkerPlots file to modify in the next activity.)

4. Create a graph of the results of your simulation. Clearly indicate the cutoff for 1 or fewer black jurors.

5. What portion of randomly selected juries have 1 or fewer black jurors?

6. What is your statistical conclusion?

▲

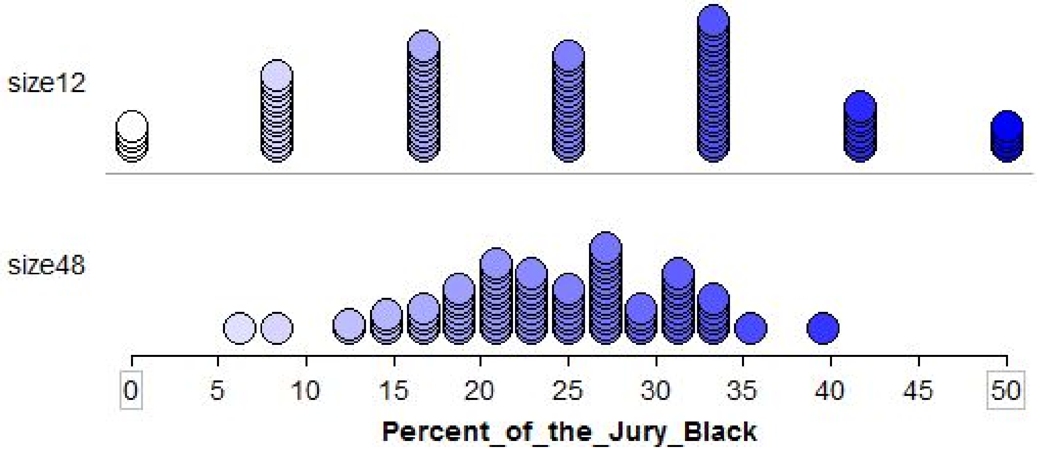
Activity 2.33. Juries are generally selected from juror panels. For example, 48 randomly selected adults from Statsville county receive a letter informing them that they have been selected for jury duty on a particular day. The 48 adults arrive at the courthouse and a subset become jurors. On a particular day, only 4 black citizens were summoned to the court house (this is 8.3% of the jury panel, just like the prior activity.) What is the likelihood that 4 or fewer members of the juror panel will be black (assuming that the county's adult population is 25% black)?

1. Make a prediction.
2. State the null and alternative hypotheses in clear language.
3. Tinker with your model from the prior activity. What do you adjust from the prior problem? Save this TinkerPlots file as a new simulation so we can compare to the previous one.
4. Create a graph of the results of your simulation. Clearly indicate the cutoff for 4 or fewer black jurors.
5. What portion of randomly selected juries have 4 or fewer black jurors?
6. What is your statistical conclusion?

▲

Activity 2.34. Put the plots from your simulations for the 12 person juries and the 48 person jury panels next to each other.

1. What is the center of each distribution?



2. Compare the spreads of the distributions. What do you notice?
3. How does sample size affect the shape of the sampling distribution?

▲

Activity 2.35. School Racial Bias

A school district is being sued for racial bias in their suspension practices. An advocacy group says that students of color make up 20% of the district but 30% of all suspensions. They base this claim on the fact that out of 60 suspensions last year, 18 of them were students of color.

1. Part 1: Create a simulation to determine whether there is evidence of racial bias in the suspension practices.
   (a) State proper hypotheses
   (b) Create a simulation for this scenario with samples of size 60 and the proportion that matches the null hypothesis.
   (c) State your conclusions clearly, including a p-value.
2. Part 2: A lawyer decides to dig a bit deeper and randomly selects 240 suspensions over the past dozen years and still finds that roughly 30% of all suspensions are students of color, although students of color have consistently made up only 20% of the district.
   (a) What do you expect will happen to the p-value when we shift from the smaller data set of 60 to investigate the lawyer's larger data set?
   (b) When you plot all of your new simulated percents, what do you expect the center of the distribution to be?
   (c) When you plot all of your simulated percents, what can you say (qualitatively) about the spread of the distribution for the larger sample size?
   (d) Verify your ideas with a simulation.

▲

A wrap-up: Suppose we have a really large bin of marbles where 70% of the marbles are red and 30% are yellow. In scenario A, samples of size $N = 20$ are drawn from the bin. In scenario B, samples of size $N = 200$ are drawn from the bin. Then even without running a simulation or calculating anything with formulas, we can make the following claims:

- In order to compare the two sampling distributions, we should use proportions rather than counts. Why?
- Both sampling distributions will have the same center, which will be $p = 0.7$, because this is the population center.
- The larger samples will tend to have proportions more similar to the population proportion of 0.7, while the smaller samples will vary more.
- The sampling distribution for scenario A will be wider (larger standard deviation) while the sampling distribution for scenario B will be narrower (smaller standard deviation.)

2.7 Hypothesis testing and errors

Reading Assignment 2.36. Read pages 151 - 154 in the Cartoon Guide to Statistics about the errors that can occur during a hypothesis test. ▲

Preview Activity 2.37. Online

1. We can make errors with hypothesis tests. These errors are beyond our control (sometimes) but we need to know that they are there and we need to know the terminology associated with these errors (hence the reading). Drag the correct words into the spots below. Choose the correct phrase to finish each sentence below.

   (a) If we fail to reject the null hypothesis in a hypothesis test when, in fact, the null hypothesis is actually false, then this is _________

   (b) If we fail to reject the null hypothesis in a hypothesis test when, in fact, the null hypothesis is actually true, then this is _________

   (c) If we reject the null hypothesis in a hypothesis test when, in face, the null hypothesis is actually true, then this is _________

not an error	Type I error	Type II error

2. In a jury trial we have the baseline assumption (the null hypothesis) that the defendant is innocent.

   (a) If we convict an innocent person, then this is _________

   (b) If we convict a guilty person, then this is _________

   (c) If we fail to convict a guilty person, then this is _________

not an error	Type I error	Type II error

Activity 2.38. Multiple choice clicker questions about hypothesis testing

1. Put the following steps in order:

   (a) Write the alternative hypothesis

   (b) Create the sampling distribution (using simulation for now)

   (c) Gather data

   (d) Either reject or fail to reject the null hypothesis

   (e) Write the null hypothesis

   (f) Decide if your data is extreme

2. The average age of first marriage in the U.S. is 28. Does completing college affect this average age? You gather data from 100 randomly selected college graduates who are or have been married and you find the average age of first marriage to be 27.6 years. Which of these is a correct way to state the null hypothesis? (select all that are correct)

(a) $H_O = 28$

(b) $H_O : \mu = 27.6$

(c) $H_O : \bar{x} = 27.6$

(d) $H_O : \mu = 28$

(e) $H_O : \mu < 28$

(f) H_O : College graduates will have the same average first marriage age as non-college graduates.

3. The average age of first marriage in the U.S. is 28. Does completing college affect this average age? You gather data from 100 randomly selected college graduates who are or have been married and you find the average age of first marriage to be 27.6 years. Which of these is a correct way to state the null hypothesis? (select all that are correct)

(a) $H_A > 28$

(b) H_A : College graduates have a higher average first marriage age than non-graduates.

(c) H_A : College graduates have a different average first marriage age than non-graduates.

(d) $H_A : \bar{x} \neq 27.6$

(e) $H_A : \mu \neq 28$

(f) $H_A : \mu < 28$

4. According to the National Center for Education Statistics, about 80% of students in the US change their major at least once. Do freshmen entering college with a health science major have a lower chance of changing their major? A random sample of 200 college students who had started as a health science major found that 70% changed their major at least once. Which of these is a correct way to state the null hypothesis? (select all that are correct)

(a) $H_O = 80\%$

(b) $H_O : \mu = 160$

(c) $H_O : p = 0.8$

(d) $H_O : p = 0.7$

(e) $H_O : p = 140$

Which of these is a correct way to state the alternate hypothesis? (select all that are correct)

(a) $H_A : p \neq 140$

(b) $H_A : p \neq 0.70$

(c) $H_A : p < 0.80$

(d) $H_A : p \neq 0.80$

(e) H_A : Health science majors have a lower chance of changing their majors than students with other majors.

▲

What if we are wrong?
There are two different ways to be wrong when we conclude a hypothesis test.

- First, we could reject the null hypothesis when the null hypothesis is actually true. This is called a Type I error, and is sometimes denoted α (alpha).

  When setting up a hypothesis test, the researcher generally chooses the α level when they decide how strong the evidence needs to be. This α is also called the significance level. While $\alpha = 0.05$ is common, α can be set at 0.01 or 0.10 or whatever value is acceptable in the specific field in which you are investigating.

- Second, we could fail to reject the null hypothesis, when the null hypothesis is actually not true. This is called a Type II error, and is sometimes denoted β (beta).

	Test conclusion	
	You have not rejected H_0	You have rejected H_0
Actually H_0 is true	good	Type I Error
Actually H_A is true	Type II Error	good

A US court considers two possible claims about a defendant: she is either innocent or guilty. If we set these claims up in a hypothesis framework, which would be the null hypothesis and which the alternative?

The null hypothesis is that the person is innocent. Only if we have sufficient evidence should she be declared guilty. If we have some evidence, but not enough (beyond a shadow of doubt), we might not think she's really innocent, but we do not declare her guilty. If the evidence isn't strong enough against her, she is considered "not guilty". Statisticians describe this situation with a double negative. We say "Don't reject the null hypothesis" (not guilty), but we do not say "Accept the null hypothesis" (innocent).

Activity 2.39. Questions about Errors

1. In the criminal justice system, what is a Type 1 error and what is a Type 2 error?

2. True or False: It is easier to reject the null hypothesis if the researcher uses a smaller alpha level. Defend your answer.

3. True or False: You are more likely to make a Type I error when using a small sample than when using a large sample. Defend your answer.

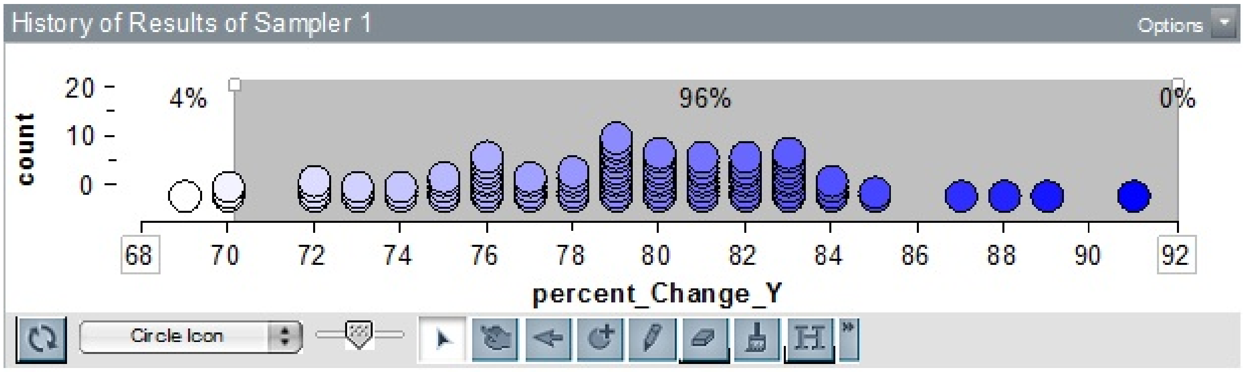
4. A gambler is trying to determine if a coin is weighted, but unknown to him the coin is actually fair. He flips the coin 100 times, find the percent heads, and determines that the probability of finding that many heads on a fair coin is 0.04. Using an alpha level of 0.05 does the gambler make an error? If so, what type?

▲

If we reject the null hypothesis, what do we do next?

Recall the earlier example: According to the National Center for Education Statistics, about 80% of students in the US change their major at least once. Do freshmen entering college with a health science major have a lower chance of changing their major? A random sample of 100 college students who had started as a health science major found that 70% changed their major at least once.

A sampling distribution for this scenario is provided below, based on a simulation from TinkerPlots. With only 4% of the simulated samples at 70% or lower, this is strong enough evidence to reject the null hypothesis using a significance level of 0.05. In other words, if the proportion of health science students changing majors really is 80%, then the chance of getting a random sample as low as ours is rare (only 4%).



What should we do next? Well, if we reject someone else's claim, we should follow-up with a better claim of our own. This claim is stated as a confidence interval that indicates where we think the population parameter is. In this case, we would make a claim about the proportion of freshmen health science majors who will change their majors.

Our best guess is 70%, because that is what our sample data showed us. But it would be better to provide a range of values.

- Point estimate: 70% of health science majors will change majors.
- Interval estimate: Between ____ % and ____ % of health science majors will change majors.

  In order to fill in the blanks for the interval, we can return to our simulator and reset the spinner to 70%. After collecting many samples, we can find the interval that contains 95% of our data. This will approximate our 95% confidence interval for the true proportion of freshmen health science majors who will change their majors. This concept is the focus of the next section.

2.8 Introducing confidence intervals with simulations

Reading Assignment 2.40. Read pages 111 - 116 of the Cartoon Guide to Statistics. Preview quiz on idea of expected intervals ▲

Preview Activity 2.41. In the reading you no doubt ran into the archery example. These questions are follow-ups to that.

1. If the archer hits the 10cm bull's eye 95% of the time, and if he is very unlikely to shoot a wild shot, would the circle encompassing 99% of his shots be wider or narrower than 10cm?
2. The "brave detective" is sitting behind the bulls eye. Why? In other words, what analogy is the author trying to make about statistics here?
3. If the "brave detective" draws a circle of 10cm around each arrow hole (from behind the target), what percent of the circles would likely contain the actual bull's eye?
4. If we gather a sample of N=1000 voters and learn that 55% of them will vote a certain way, what other piece of statistical information would we need to build a confidence interval (also known as a "range estimate")?

Activity 2.42. What happens when we reject H_0?

Create a simulation to answer this question: A new cardiac drug is given to patients after open heart surgery. The pharmaceutical company claims that 20% of people that take this drug suffer from headaches. A doctor has been using this drug on many patients and notices that more than 20% of his patients have complained of headaches. He has the hospital's statistician gather a random sample of 50 patients that have used this drug and finds that 40% complain of headaches while on the drug. Does the doctor have evidence that the 20% advertised side effect rate is incorrect?

- Null Hypothesis: $H_0 : p = 0.2$
  (where p is the proportion of patients that have headaches)
- Alternative Hypothesis: $H_A : p > 0.2$
- P-Value =
- Decision =

Since we are rejecting H_0, we know that the proportion of patients with headaches is not 0.2. If the proportion isn't 0.2, then what is it? When we reject the null hypothesis, it will be our responsibility to report a better estimate for the population parameter.

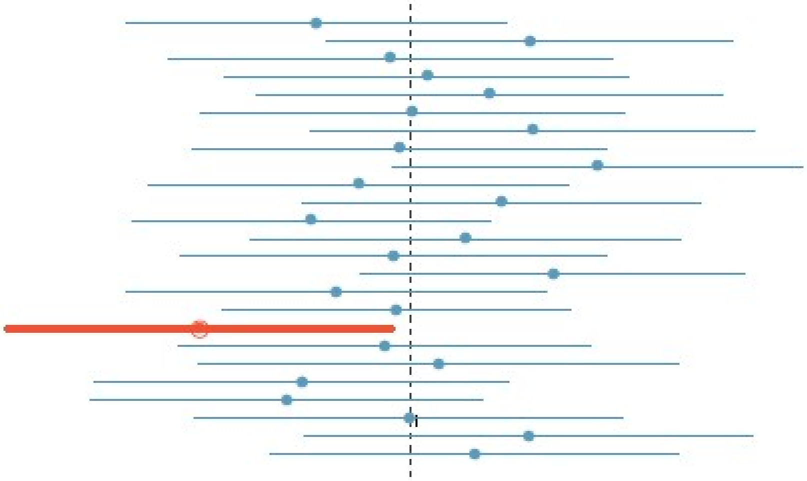
The best estimate we have is our data point ($\hat{p} = 0.4$), so will use this statistic along with a simulator to get a sense for the level of variance in the scenario. This will allow us to create a confidence interval for the population parameter of the proportion of patients with headaches.

Teacher Note: Students tend to struggle with this switch from using p to using $\hat{p}$. It was correct to use the hypothesized proportion of 20% when testing our initial null hypothesis. The spinner in TinkerPlots was set at 20% / 80%. But now that we have rejected the null hypothesis, it is no longer appropriate to use a spinner set at the rejected 20%. Instead we use our sample proportion of $\hat{p} = 0.4$.

- Build a simulation centered around your data point $\hat{p} = 0.4$.
- Use the divider to gather 95% of the simulated results
- Make a statement about where you think the true population proportion could be. We are 95% confident that the true proportion of people that get headaches on this drug is between _____ and _____.
- Notice that your interval does not contain the pharmaceutical company's claimed 20%. Does that seem appropriate?

▲

Confidence intervals usually capture the population parameter you are estimating (e.g. p or μ). A 95% confidence interval has a 95% chance of capturing the population parameter. As the image shows, that means about 1 in 20 confidence intervals does not capture the population parameter.



Confidence Intervals, image from OpenIntro Statistics by Diez et al.

Activity 2.43. What if we simply had no prior knowledge? Or no prior claims to evaluate?

While hiking the Pacific Crest Trail (from Mexico to Canada along the Sierra Nevada and Cascade mountains) Jerry notices that many fellow hikers have adopted a trail name. Many do this to maintain anonymity on the trail and some do it just to embrace the full experience of the trail. Jerry polls 40 randomly selected PCT hikers and finds that 65% of them have adopted trail names (26 out of 40). He knows that it is unlikely that exactly 65% of all Pacific Crest hikers have adopted trail names, but he has no other information.

1. Build a simulation to estimate a 95% confidence interval for the proportion of PCT hikers that have adopted a trail name.

2. Using the same simulated data, create a 90% confidence interval by capturing the middle 90% of data points.

3. Using the same simulated data, create a 80% confidence interval by capturing the middle 80% of data points.

4. As the confidence level goes lower, what happens to the width of the interval? Will this always happen when you reduce your confidence level?

▲

Activity 2.44. A physical therapy researcher does a baseline study to understand if a new test for an ACL tear is working. She gathers a random sample of 100 patients that she knows have an ACL tear (usually indicated by an MRI scan). Of those 100, the new test gives a positive result on 82 patients. Estimate a 95% confidence interval for the likelihood that this test will give a true positive result for an ACL tear. ▲

Activity 2.45. Open the *videoGamesXBoxPS3* data set from Moodle. Recall that this data set is a randomly selected subset of video games, taken from *kaggle.com/datasets*. This data set is limited to games on the XBox and PS3 platforms.

1. Find a 95% confidence interval for the proportion of games rated E for everyone. State your conclusions in a complete sentence related to the context of video games.

2. Find a 95% confidence interval for the proportion of games rated M for mature. State your conclusions in a complete sentence related to the context of video games.

▲

2.9 Lab 4

Reading Assignment 2.46. Find Lab 4 on Moodle and read it before coming to class. ▲

Preview Activity 2.47. Preparing for Lab 4

1. Read the first page of Lab 4, paying close attention to the worked example (in red). Then answer the 4 parts of problem 1 based on finding areas in the graph. Provide all of your answers as decimals, accurate to 4 decimal places.

   (a) What is the probability that you will have to wait between 1 and 5 minutes?

   (b) What is the probability that you will have to wait longer than 5 minutes?

   (c) What is the probability that the bus will arrive in the first 30 seconds?

   (d) What is the probability that you will have to wait longer than 9 minutes?

2. Follow-up question: The total area under the curve for the uniform distribution on page 1 is [_____]

Activity 2.48. Lab 4: P-values from Distributions

1. Students learn how to find probabilities using the concept of area under the curve for a distribution. This is done graphically by counting boxes or using $area = length \times width$. This is a non-calculus approach, to provide some intuition for working with the normal distribution going forward.

2. Lab includes a uniform distribution function (bus wait time), a histogram (ages of students seeking an advisor), an exponential distribution (call wait time), and a normal distribution (weights of dogs).

3. Lab includes a brief overview and practice with the 68%, 95%, 99.7% rule.

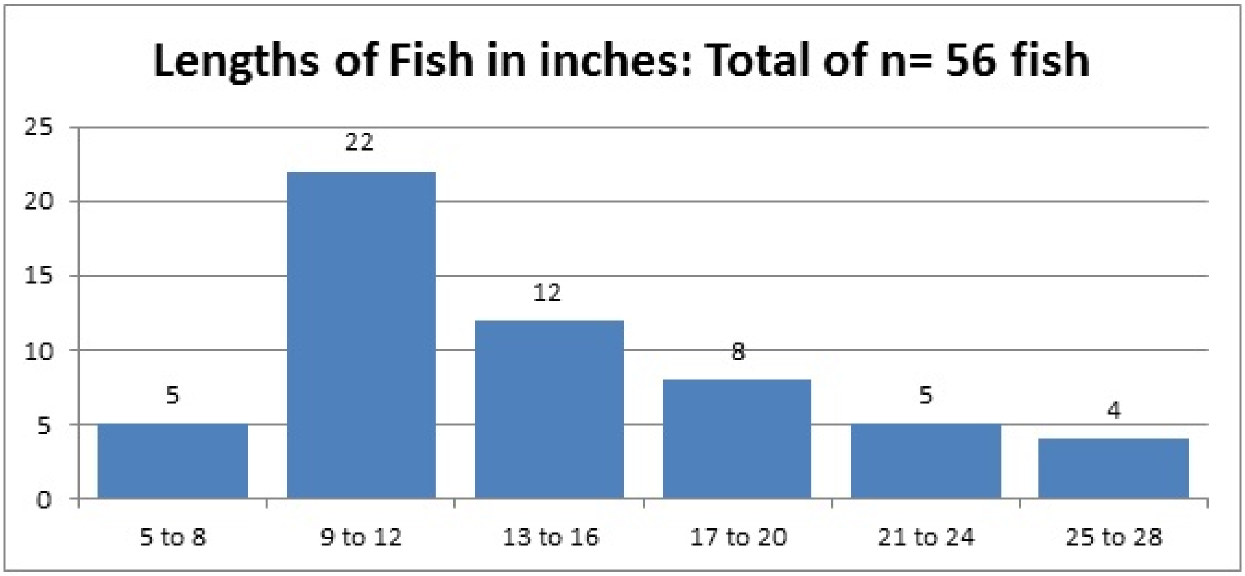
▲

2.10 Review day

Reading Assignment 2.49. ▲

Preview Activity 2.50.

Activity 2.51. Review: Reading a histogram



1. Some students notice that the tallest bar is 22 units high and then say that the longest fish must be in that bar. Why does that not make sense? Where are the longest and shortest fish located in the histogram?

2. Which interval / section of the histogram contains the minimum? The lower quartile? The median? The upper quartile? The maximum?

3. Why can't we find the mean and standard deviation for the above histogram?

▲

Activity 2.52. Each baby that is born has a 50% chance of being female and 50% chance of being male. A small rural hospital has noticed a jump in baby girls. Last year, 60% of the babies born were girls (27 out of 45)

1. State the null hypothesis

   State the alternative hypothesis

2. Create a simulator to determine whether the hospital has evidence that something strange is going on or whether this is a typical amount of variation.

3. Find the p-value. The p-value is the probability of *something.* Explain what that *something* is in this context.

4. What are your conclusions?

5. How extreme would the gender proportion have to be before you would claim that there is evidence of something unusual at the rural hospital?

6. That same year, a larger city hospital found that 60% of their 400 newborns were female. How much would you expect the p-value to change in this situation? Would you expect the conclusions to be different at the city hospital? [Test out your theory with a TinkerPlots simulation]

7. Why does the size of the sample matter in a hypothesis test?

▲

Activity 2.53. Finding summary data in an Excel spreadsheet.

- Download the Counties data (data set from OpenIntro Stats)
- Create a pivot table and put states in the rows.
- Find the mean, standard deviation, and count for the population of Montana counties in 2010.
- Find a state with a larger mean for county populations.
- Find a state with a larger standard deviation for county populations.

▲

Activity 2.54. Have you been doing your reading assignments this term?

In order to get an accurate sense of what portion of stats students are doing their reading, we use the following method to protect students from self-incrimination. 180 stats students from this year were asked the following pair of questions

1. Flip a coin. If it's heads, answer "no." If it's tails, answer the following question instead.

2. If you flipped your coin and a got tails, answer "yes" if you honestly read 90% of the readings and answer "no" if you did not.

Under this method, if a student says "no", they aren't incriminating themselves since we don't know which question they answered. For this experiment, 40% of people said yes and 60% said no.

Does this mean that 60% of students in this survey did not do the reading? If so, back this up with a contingency table or tree diagram. If not, first estimate whether you think the percent is higher or lower than 60% and why. Then use a contingency table or tree to find an accurate rate of reading. ▲

Activity 2.55. Create your own false-positive style conditional probability problem. Show how to solve your problem using a contingency table or tree. ▲

2.11 Exam 1

2.12 Introduction to normal distributions

Reading Assignment 2.56. Read the first part of Section 2.6 in the OpenIntro online textbook. Read sections 2.6.1 to 2.6.5 (Pages 85 to 94) ▲

Preview Activity 2.57. The reading introduced several notions related to the normal distribution. In particular, you have likely already noticed that the simulations that we've done in class all seem to follow the same "bell shaped" curve. This is no coincidence!! We have been examining the normal distribution all semester.

1. The textbook authors give some shorthand notation for the normal distribution. What does **N(0,1)** mean?

2. The textbook authors next give the notion of the **z-score**. This is really just a measure of how far a particular value is from the mean. You can think of it as "how many standard deviations is x away from the mean?" The z-score allows us to use the standard deviation as a measuring stick for all normal distribution. If someone's ACT score was 27 and we know that the mean ACT mean is 21 with a standard deviation of 5, then what is the person's z-score?

3. The third part of the reading deals with finding probabilities from the normal distribution. There is a handy rule of thumb that works well for estimating probabilities on the normal curve:

   (a) If you go out 1 standard deviation from the mean you will capture about what percent of the data?

   (b) If you go out 2 standard deviations from the mean you will capture about what percent of the data?

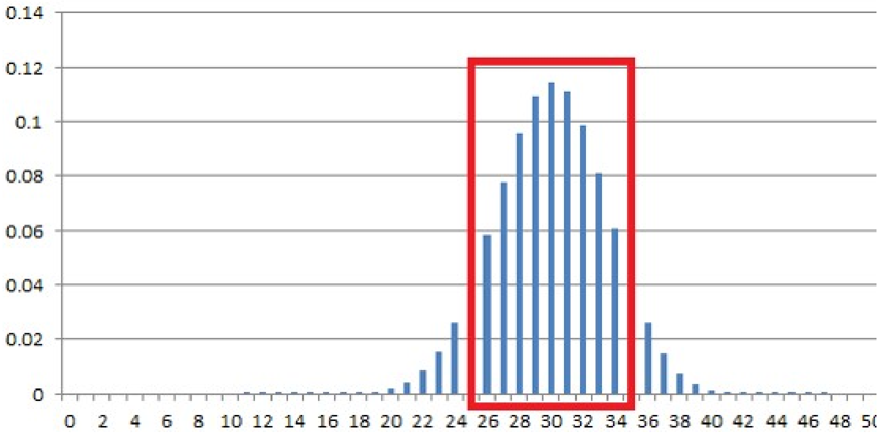
   (c) And if you go out 3 standard deviations form the mean you will capture about what percent of the data?

4. Given the rule of thumb, since SAT score closely follow the normal model N(1500,300), about what percent of test takers score between 1200 and 1800?

Big ideas about normal distributions

- In a probability distribution, the total area represented in the graph is always 1 (or 100%) and the probability of being in a given section of the graph is the same as the area of that section of the graph.

- This area represents 88.87% of the data, so if a data point is picked at random, there is an 88.87% chance of a data point ending up in this area.



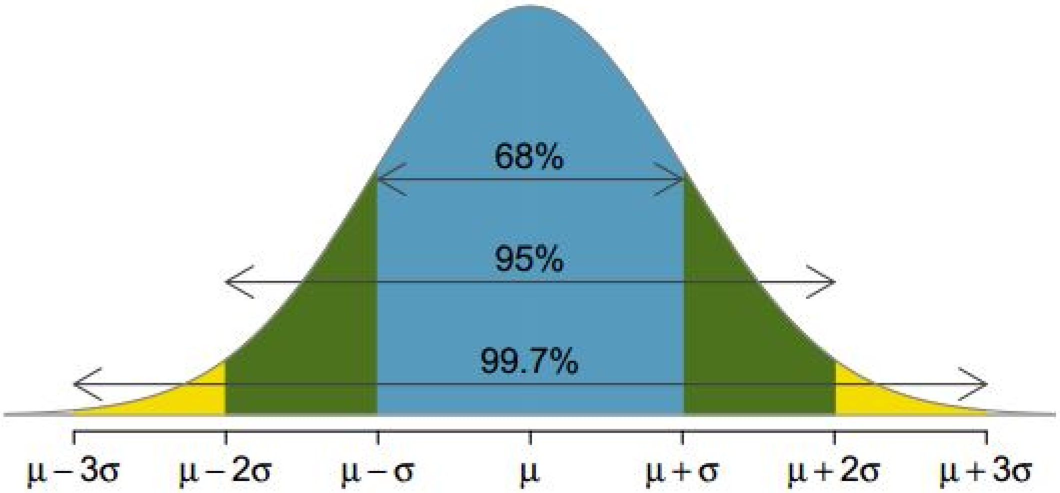
- Many real world situations result in probability distributions in the shape of a bell curve / mound shaped distribution. Examples:
  - Heights of 5 year olds
  - Weights of parakeets
  - Scores on standarized tests
  - Sampling distributions

  In each case, there are many minor independent causes for change underlying the variable, which causes most data to clump toward the center.

- Curves of this type are so common in statistics, that this curve is called the *normal distribution.*
  - The normal distribution can be built with any mean and standard deviation you like.
  - The mean is denoted μ (mu) and the standard deviation is σ (sigma). The normal distribution can be noted in shorthand as $N(\mu, \sigma)$. e.g. If we say a distribution is $N(400, 100)$, then it is a normal distribution with mean 400 and standard deviation 100.
  - If you pick the mean of $\mu = 0$ and the standard deviation of $\sigma = 1$, then this is even more special and it is called the *standard normal distribution,* i.e. $N(0, 1)$.

- All normal distributions are the same shape, except for scale.

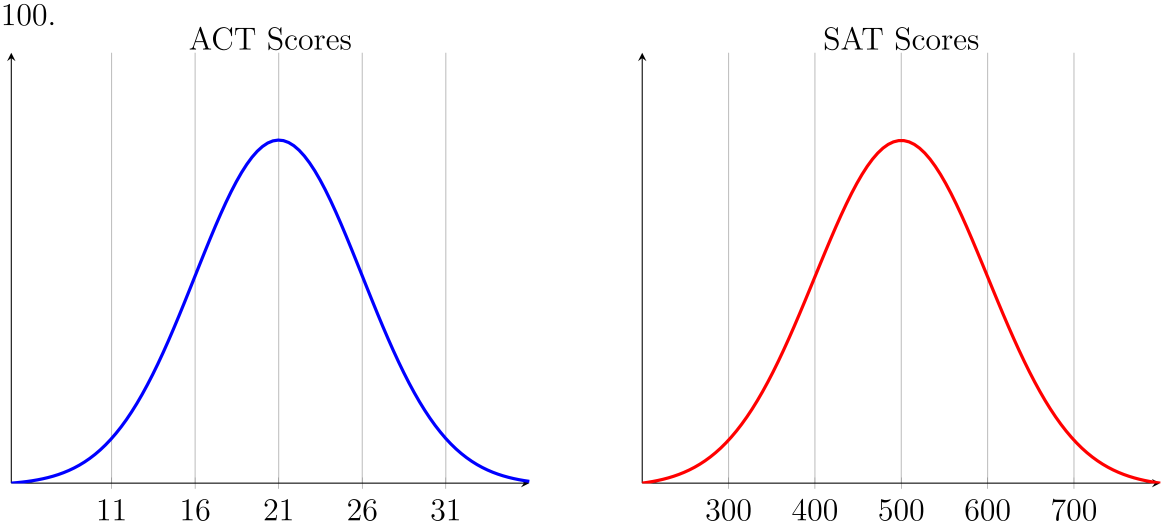
- Whatever the mean and standard deviation, for a normal distribution, we always have
  - about 68% of data within 1 standard deviation σ of the mean μ
  - about 95% of data within 2 standard deviations σ of the mean μ
  - about 99.7% of data within 3 standard deviation σ of the mean μ
  - If your data is more than 3 standard deviations from the mean, it is really rare / weird.



68%, 95%, 99.7% Rule, image from OpenIntro Statistics by Diez et al.

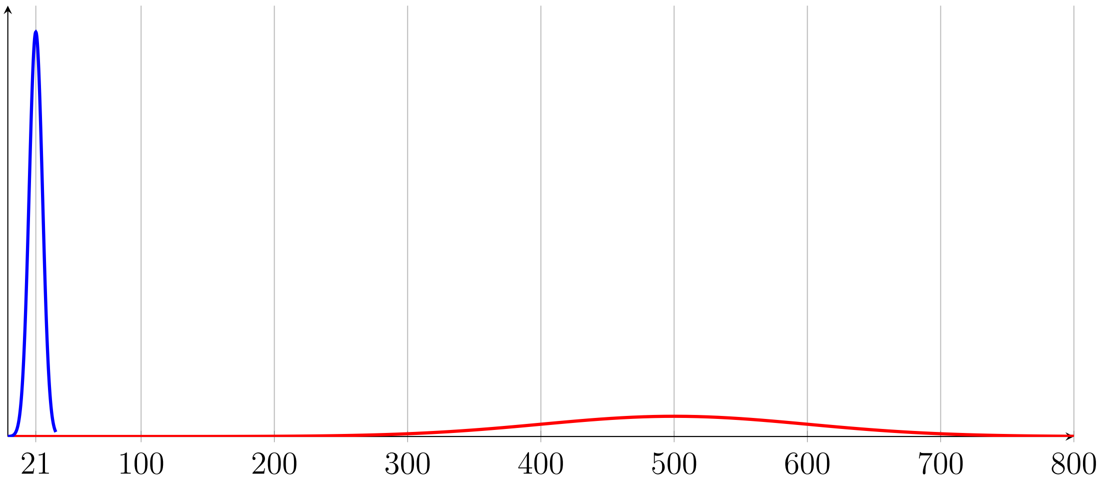
Activity 2.58. Bell curves in ACTs, SATs, and IQ scores

Consider the scores on two different standardized tests: the ACTs and SATs. If we were to plot everyone who took the ACT in a given year we would get a bell curve with a mean of 21 points and a standard deviation of about 5 points. If we were to do the same with the SAT math scores we would get a bell curve with a mean of 500 and a standard deviation of about 100.



The graphs above are both normal distributions. They have the same shape but they do differ in one important way. Take a closer look at the axes. Notice that one test maxes out at 36 points while the other maxes out at 800 points. If we were to place both exams on the same axis (below), the curves look very different. In order to compare scores on two different standardized tests, we can describe how many standardized deviations each score is from the mean. For example, it is more useful to state that a person scored 1.4 standard deviations below average than to say they scored 7 points below average.

SAT Scores



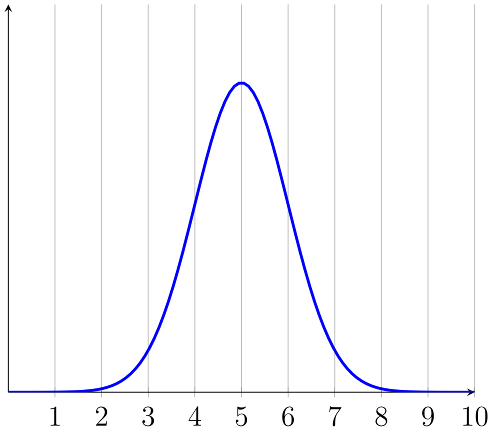
Nearpod response questions: One image, 3 open response, 10 multiple choice, and 5 multiple choice/draw

1. An IQ test is normally distributed with mean 100 and standard deviation 15, $N(100, 15)$. Sketch a normal distribution curve for the IQ test and label the x axis with appropriate values.
2. Earning a 21 on the ACT is similar to earning a 500 on the SAT Math. What SAT Math score is similar to a 16 on the ACT?
3. What ACT score is similar to a 700 on the SAT?
4. How rare is it to earn a 31 or higher on the ACT?

▲

Activity 2.59. A set of multiple choice questions about normal distributions.

1. Approximate the standard deviation on the blue normal distribution.



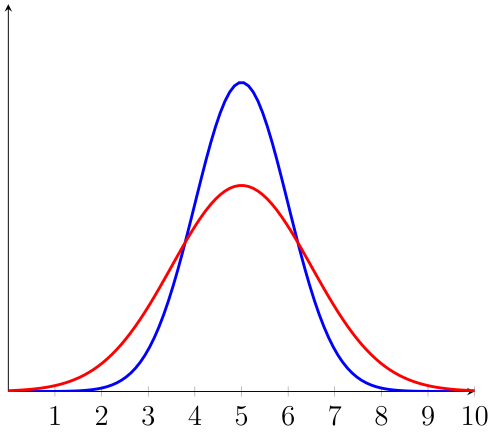
(A) $\mu = 5$ and $\sigma = 1$

(B) $\mu = 5$ and $\sigma = 2$

(C) $\mu = 5$ and $\sigma = 3$

(D) $\mu = 5$ and $\sigma = 4$

2. Pick the correct comparison



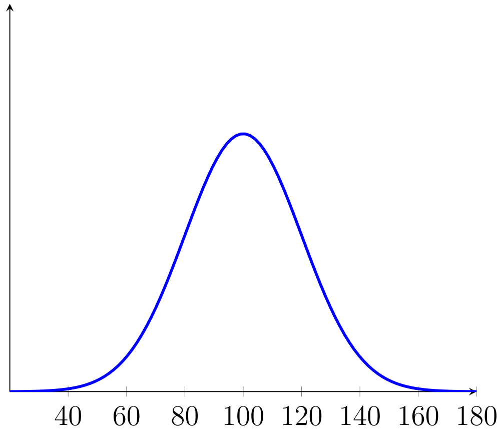
(A) $\sigma_{red} = \sigma_{blue}$

(B) $\sigma_{red} > \sigma_{blue}$

(C) $\sigma_{red} < \sigma_{blue}$

(D) Not enough information

3. In the following normal distribution we have $\mu = 100$ and $\sigma = 20$.
Question: 68% of the data described by this distribution is between which two values?



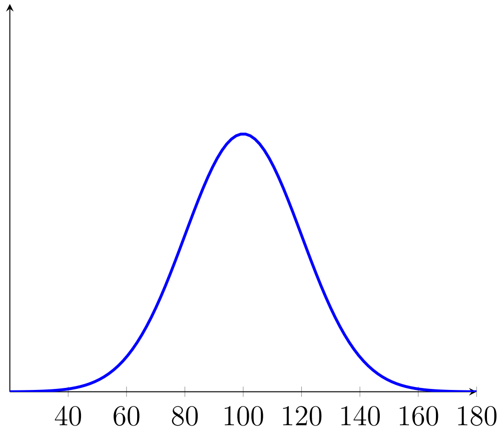
(A) 99 and 101

(B) -80 and 120

(C) 80 and 120

(D) 60 and 140

4. In the following normal distribution we have $\mu = 100$ and $\sigma = 20$.
Question: 95% of the data described by this distribution is between which two values?



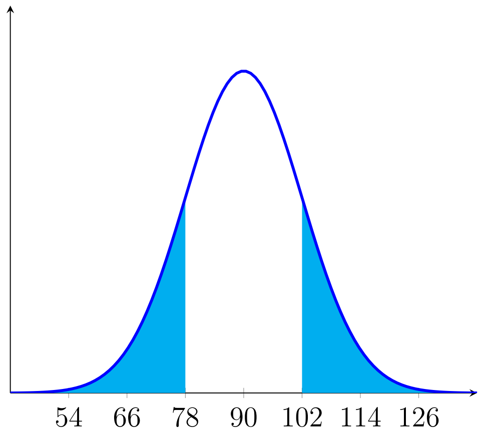
(A) 99 and 101

(B) -80 and 120

(C) 80 and 120

(D) 60 and 140

5. In the following normal distribution we have $\mu = 90$ and $\sigma = 12$.
What is the approximate percent of the data that is either greater than 102 or less than 78?



(A) 68%

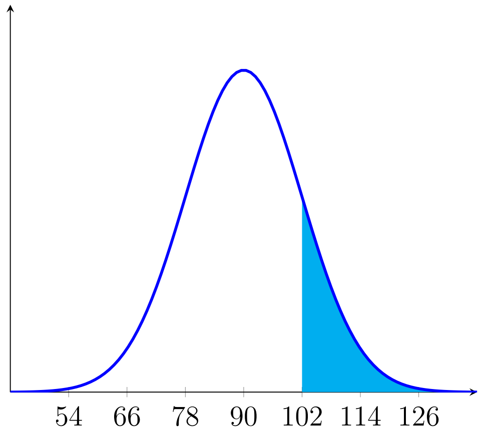
(B) 95%

(C) 32%

(D) 5%

(E) Not enough information

6. In the following normal distribution we have $\mu = 90$ and $\sigma = 12$.
   What is the approximate percent of the data that is greater than 102?



(A) 32%

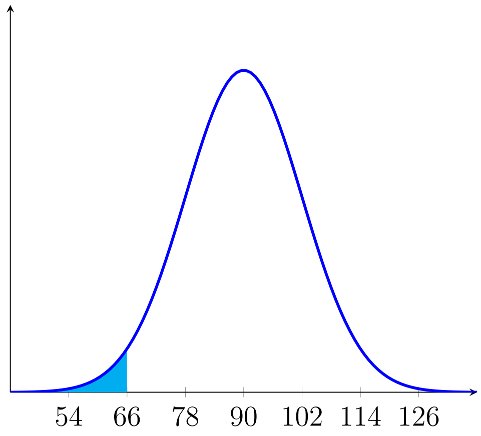
(B) 68%

(C) 16%

(D) 5%

(E) Not enough information

7. In the following normal distribution we have $\mu = 90$ and $\sigma = 12$.
   What is the approximate percent of the data that is less than 66?



(A) 95%

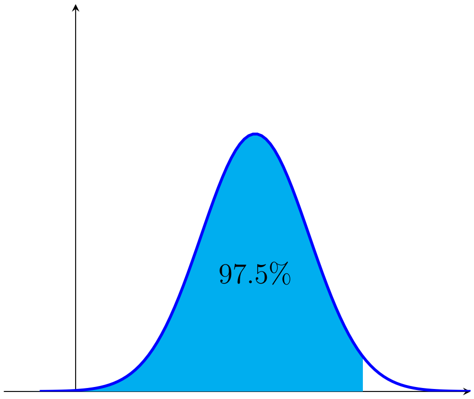
(B) 5%

(C) 2.5%

(D) 10%

(E) Not enough information

8. In the following normal distribution we have $\mu = 5$ and $\sigma =??$.
   If 97.5% of the data falls below 8 then what is the approximate standard deviation for the distribution?



(A) $\sigma \approx 1$

(B) $\sigma \approx 1.5$

(C) $\sigma \approx 2$

(D) $\sigma \approx 2.5$

(E) $\sigma \approx 3$

Definition 2.60. The *z-score* for a value from a normal distribution is the number of standard deviations the value is away from the mean.

9. In a normal distribution with mean $\mu = 6$ and standard deviation $\sigma = 2$, what is the z score for $x = 5.5$?

   (A) $z = -0.5$
   (B) $z = -0.25$
   (C) $z = 0.5$
   (D) $z = 0.25$

   Note: The distance from the mean $= x - \mu$, while the number of standard deviations from the mean is $z = \frac{x-\mu}{\sigma}$

10. A number 1.5 standard deviations below the mean has a z score of

    (A) 1.5
    (B) -1.5
    (C) 3
    (D) -3
    (E) not enough information

    Draw the associated normal distribution plot.

▲

For the prior problems, we leveraged the 68%, 95%, 99.7% rule. This allowed us to shift from z-scores to probabilities and vice versa. This rule works well when the value of interest is a whole number of standard deviations from the center. However, this is often not the case. What proportion of the data is 1.5 σ above μ or 2.34 σ below μ? Fortunately, we can use Excel to find these exact proportions. The next set of tasks provide the Excel commands for translating between data values and probabilities and between z-scores and probabilities.

Activity 2.61. Using Excel to find probabilities for the normal distribution

1. Excel With Normal Distributions: Given Score Find Area

Mathematical Question	Excel Command
Prob. getting less than x	`=NORM.DIST(`x`,mean,stdev,1)`
Prob. getting greater than x	`=1-NORM.DIST(`x`,mean,stdev,1)`
Prob. getting exactly x	not possible

The distribution of heights of American women aged 18 to 24 is approximately normally distributed with a mean of 65.5 inches and a standard deviation of 2.5 inches. What percent of these women is less than 5'8" (68 inches)?

(A) $P(x < 68) \approx 0.841$

(B) $P(x < 68) \approx 0.159$

(C) $P(x < 68) \approx 0.097$

(D) $P(x < 68) \approx 0.903$

Draw the associated normal distribution plot.

2. The distribution of heights of American women aged 18 to 24 is approximately normally distributed with a mean of 65.5 inches and a standard deviation of 2.5 inches. What percent of these women is greater than 5'8" (68 inches)?

(A) $P(x > 68) \approx 0.841$

(B) $P(x > 68) \approx 0.159$

(C) $P(x > 68) \approx 0.097$

(D) $P(x > 68) \approx 0.903$

Draw the associated normal distribution plot.

3. Excel With Normal Distributions: Given Area Find Score

Mathematical Question	Excel Command
Prob. getting less than x is P. Find x	`=NORM.INV(P,mean,stdev)`
Prob. getting greater than x is P. Find x	`=NORM.INV(1-P,mean,stdev)`

A group of students at Carroll takes a statistics quiz. The distribution is normal with a mean of 25 and a standard deviation of 4. Everyone who scores in the top 30% of the distribution gets a certificate. What is the lowest score someone can get and still earn a certificate?

(A) 29

(B) 25

(C) 27

(D) 23

Draw the associated normal distribution plot.

4. A group of students at Carroll takes a statistics quiz. The distribution is normal with a mean of 25 and a standard deviation of 4. The top 5% of the scores get to compete in a statewide statistics contest. What is the lowest score someone can get and still go on to compete with the rest of the state?

   (A) 31

   (B) 32

   (C) 18

   (D) 19

   Draw the associated normal distribution plot.

5. Assume a normal distribution with a mean of 70 and a standard deviation of 12. What limits would include the middle 65% of the cases?

   (A) Bottom Score = `norm.inv(58,70,12)`,
   Top Score = `norm.inv(82,70,12)`

   (B) Bottom Score = `norm.inv(5,70,12)`,
   Top Score = `norm.inv(135,70,12)`

   (C) Bottom Score = `norm.inv(.175,70,12)`,
   Top Score = `norm.inv(.825,70,12)`

   (D) Bottom Score = `norm.inv(.475,70,12)`,
   Top Score = `norm.inv(.975,70,12)`

   (E) not enough information

   Draw the associated normal distribution plot.

▲

2.13 Lab 5 - Normal Distributions

Reading Assignment 2.62. Watch the Excel related videos linked on Moodle.

1. Using Excel for the Normal Distribution (Part 1)

2. Using Excel for the Normal Distribution (Part 2)

▲

Preview Activity 2.63. Practicing with the Normal Distribution

The weights of newborn babies in the US are normally distributed with mean 7.57 pounds and standard deviation 1.06 pounds.

1. What is the probability of randomly selecting a baby from the US and finding the weight to be less than 5 pounds?

2. What is the probability of randomly selecting a baby from the US and finding the weight to be above 7.5 pounds?

3. What is the probability of randomly selecting a baby from the US and finding the weight to be between 7 and 10 pounds?

4. What is the cutoff weight for the smallest 25% of US babies?

5. What is the cutoff weight for the largest 10% of US babies?

6. What are the cutoff weights for the middle 90% of US babies?

Teacher Note: It may be helpful to bring printouts of the blank normal distribution curves. This lab requires students including images. Last year students tried to do this in a separate file or exclude altogether. However, the directions are clear that images are required to be copied into an appropriate location in each file.

Lab focuses on normal distributions and finding area under the curve for specific intervals of a distribution. The lab introduces the Excel commands Norm.Dist, Norm.S.Dist, Norm.Inv, and Norm.S.Inv. There are videos to accompany this. The lab requires that students include shaded drawings for the normal distribution. This can be one within paint or other software, or done by hand and scanned and clipped into the correct places in the lab.

2.14 Summary

Summary 2.64. Student learning outcomes from Chapter 2

1. Students understand how to create a one-proportion sampling distribution using simulation software, such as TinkerPlots. Students are able to use the sampling distribution to describe "typical" and "weird" outcomes.

2. Students have a conceptual understanding of p-values and confidence intervals and are able to compute them for one-proportion scenarios using simulations.

3. Students use the language of hypothesis testing appropriately, including stating null and alternative hypotheses and summarizing whether a hypothesis should be rejected or not. Students know the difference between Type I and Type II errors.

4. Students have a conceptual understanding of normal distributions and the skills to use Excel to compute probabilities and intervals associated with normal distributions.

5. Students are able to use pivot tables to summarize qualitative data in Excel.

6. Students are able to compute basic and conditional probabilities. In particular, students can make sense of scenarios with false positives and false negatives. They can also compute probabilities from sampling distributions created by simulations.