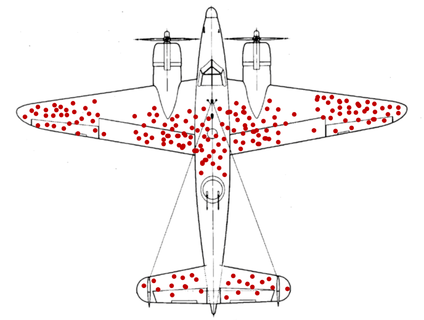
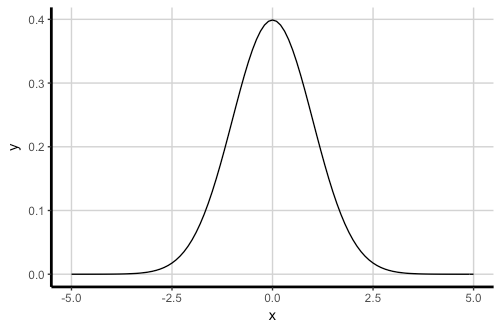
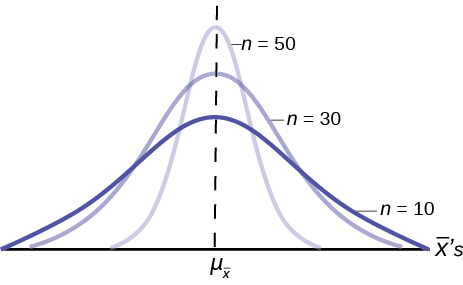
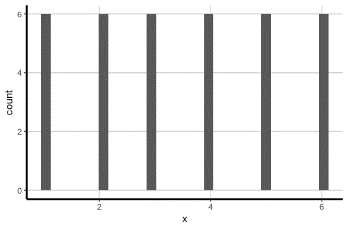
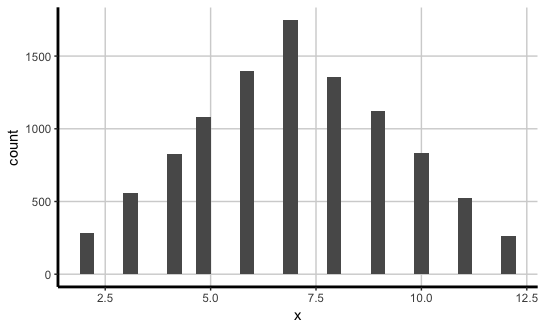
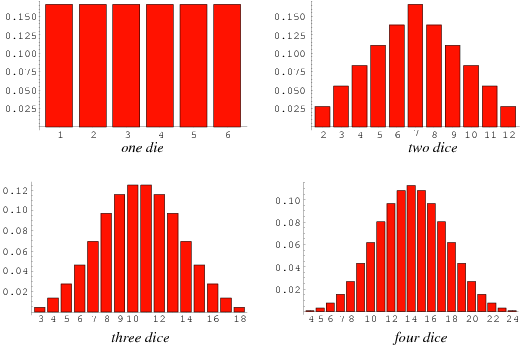
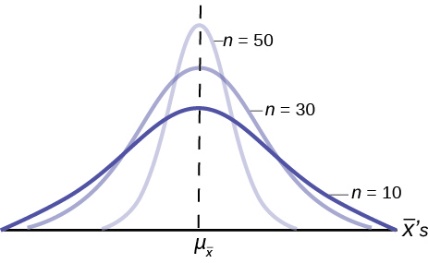
Unit 1-3 Central Limit Theorem

* Inferential Statistics
  + Let's say your client, a food maven and entrepreneur, has hired you to determine if a product under consideration for mass market —anchovy-stuffed brussel sprout bites — will be a hit. If launched, the product would be sold in supermarkets across the United States. Clearly you can't survey the tastes of the entire population of the USA, so you go with a carefully selected sample group of 100 people. Is your data complete? Negative. But you being the savvy data scientist that you are, make an inference based on the data you do have. Marvel at our newest trick: inferential statistics.
  + In most real-world cases you won't be working with a complete dataset (think election polling), so inferential statistics is a pretty handy technique.
* Descriptive vs. Inferential Statistics
  + Before we dive in to inferential statistics, let's review the difference between descriptive and inferential statistics:
    - **Descriptive statistics** focus on summarizing, describing, and understanding data we observe.
    - **Inferential statistics** focus on generalizing results from a sample to a larger population.
* Samples vs. Populations
  + A population is a group of interest. Two examples for you: (1) all individuals who will vote in an upcoming election, (2) all individuals with cancer.
  + A sample is any subset of a population. Two examples for you here as well: (1) a group of individuals responding to questions about their voting preferences, (2) a group of patients selected to try a new cancer treatment.
* Selection Bias
  + When making decisions about data sampling, it's important to take into account selection bias.
  + A classic example of selection bias can be seen in the image below, which represents an analysis of where British fighter jets that returned to base were targeted by German bombers during World War II.
  + When the British mapped out where planes were getting hit, this image began to appear. Naturally, the first instinct was to better protect those parts of the plane (covered by red dots).
  + 
  + However, when this same image was analyzed by statistician Abraham Wald, his suggestion was to better protect the areas not covered by red dots. Why do you think that is?
  + Statistician Abraham Wald suggested that the areas not covered by red dots represented missing data from downed planes. In his analysis, areas not covered by red dots were places where the planes may have been weakest. Fire affecting those points meant planes got shot down and never returned to safety!
* Sample Mean
  + Oftentimes, we'll want to use the mean of a sample to learn about the mean of a population. For example, say you'd like to measure the miles per gallon for all cars on the road to learn about which types of cars are more or less efficient than the average car.
  + You may not have data for every single type of car made in the world, but calculating the mean mpg of a subset of all types of cars will help you make your determinations.
* Normal Distribution
  + Lots of real-world phenomena follow a normal distribution. However, there's a theorem called the **central limit theorem** that, in many cases, ensures this assumption of normality is satisfied rather than just believing it to be true.
  + 
* Sampling Distribution
  + Recall that a **distribution** is *the set of all values a variable can take on and how frequently each value occurs.*
  + Suppose we take a sample of *n* observations from a population and record the mean of this sample. We then replace all observations, pull another sample of *n* observations from the same population, and record the mean of this sample.
  + Suppose we do this over and over again until we've recorded the mean of every possible sample of size *n* from the population. This is known as the sampling distribution of .
  + While it would be impractical to construct this distribution for moderately large populations, there is a benefit to understanding how the variable behaves.
  + When we calculate a sample mean, , we treat it as one observation from the distribution of all possible sample means .
  + If we know that follows a normal distribution, we can then rely on this when conducting inference on a sample mean.
  + When Does a Sample Have a Normal Distribution?
    - The sampling distribution of has a normal distribution in two cases:
      * If the original variable X follows a normal distribution, then follows a normal distribution exactly.
      * If the original variable X does not follow a normal distribution or is unknown, then is approximately normal for sample size *n*>30.
    - This second bullet point relies on the **central limit theorem**.
* Central Limit Theorem
  + The central limit theorem states that, regardless of the distribution of the original data X, the sampling distribution of will approach a normal distribution as the sample size *n* gets larger and larger.
  + For the sake of practicality, we'll say that if n is larger than 30, then is sufficiently close to normal. In this case, n>30 is a rule of thumb and may not always be the minimum sample size needed in all cases. However, the larger the value of n, the better. Generally, if n<30 our data won't be statistically significant enough to make decisions off of.
  + When dealing with larger datasets, typically *using a sample of 10% of the population should be sufficient*. **As the number of samples increases, the margin of error will decrease**.
  + Below, we can see that, as the sample size gets larger, the graph approaches a more normal distribution:
    - 
* Visualizing the Central Limit Theorem
  + Suppose we roll a standard six-sided die exactly once and then calculate the average of the roll. (It seems silly to calculate the average because we're only rolling once, but this idea will be important later.)
    - We roll a 5. Obviously, our average here is 5.
      * We will denote this by 1​, which indicates the average of our first sample.
      * The mean of our sample is 1=5.
    - Let's say we do this again with a brand new sample and roll a 3. Then, the mean of this sample is 2=3. (We write 2​ because this is our second sample.)
    - We want to do this for every possible sample of size 1. In this case, that means rolling a 1, 2, 3, 4, 5, and 6 exactly once each.
    - We use the random variable (upper-case!) to represent the sampling distribution (all possible values of and how frequently do we observe these values of of the sample mean.
    - If we visualize this with a histogram, we get a flat histogram, where the values 1 through 6 are all equally likely.
    - 
  + Let's do this again but instead with sample size n=2. In this case, we will roll the die twice and find the average.
    - *Sample* 1: {1,1} => 1= 1
    - *Sample* 2: {1,2} => 2= 1.5
    - *Sample* 3: {1,3} => 3= 2
    - *…*
    - *Sample* 36: {6,6} => 36= 6
    - If we visualize the distribution of the sample mean for all possible samples of size 2 for this experiment, what do we expect this distribution to look like?
    - Here is the distribution of the sample mean for all possible samples of size 2 for this experiment. Does this distribution reflect what you initially thought?
    - 
  + Here we have four graphs. The first is the distribution of the sample mean for all possible samples of size 1, the second is the distribution for samples of size 2, the third is the distribution for samples of size 3, and the fourth is the distribution for samples of size 4.
    - 
    - As n increases, what do we notice?
* Applying the Central Limit Theorem
  + Suppose X is a random variable reflecting the systolic blood pressure of everyone in a country. You take a sample of 50 individuals, measure their blood pressure, and then calculate the mean of this sample, of
  + If we know that systolic blood pressure (X) follows a normal distribution, then we know that the sampling distribution of follows a normal distribution as well. (We don't use the central limit theorem here!)
  + If we don't know that systolic blood pressure (X) follows a normal distribution, we can assume that the sampling distribution of is approximately normally distributed because our sample size is greater than 30. (We do rely on the central limit theorem here.)
* Revisiting the Central Limit Theorem
  + Let's go back to the original statement of the central limit theorem.
  + The central limit theorem states that, regardless of the distribution of the original data X, the sampling distribution of will approach a normal distribution as the sample size n increases.
  + See how it applies to our example?
* Simulating the Central Limit Theorem, Continued
  + Now that we've visualized the central limit theorem, let's talk about how you can simulate it.
    - Step 1: Obtain a large number of quantitative observations, NNN.
  + For example: N=1000. You can gather this data in any fashion. Try pulling these from a data set or by generating random numbers.
    - Hint: you can do this in Python!
    - Step 2: Pick some fixed sample size, n.
    - Step 3: Make a list of every possible combination of n observations from the original population of N. For each combination, record the average.
    - Step 4: Plot these averages on a histogram. This histogram is the sampling distribution of the sample mean for sample size n.
    - Note: If you create different histograms for different values of n, you'll note that the distribution looks more like a normal distribution as n increases.
    - This is the central limit theorem at play! As the sample size increases, the behavior of approaches a normal distribution.
    - 
* Real-World Recommendation
  + As you can imagine, it doesn't make sense for us to actually construct the sampling distribution of . It requires access to the entire population, at which point statistical inference becomes irrelevant because we could just directly calculate a parameter (with no uncertainty).
  + **However, knowing how of behaves allows us to understand under what conditions we can rely on the normal distribution. In our case, if our sample size is at least 30, then we can use the normal distribution when conducting inference on population means or proportions — even if our original data N is not normally distributed**!
* Experimental Design
  + When running an experiment, it's important to ensure that the survey or means of collecting data is done in an appropriate manner. This is called **experimental design**, the process of which ensures that:
    - The question that the experiment is trying to answer is clearly defined.
    - Sources of variability in the experiment are identified.
    - Descriptions exist of how participants are allocated — i.e., if they randomly chosen or other methods were used.
  + In data science, **sample protocol** is an important part of dealing with survey data. Statistical significance should always be used when making decisions based on the data, but it's also important to consider the types of questions being answered and who is answering them.
  + For example, customers with bad experiences may be more likely to write reviews on a website than those with good experiences. Therefore, this does not necessarily mean that users are having more negative experiences than positive ones! Instead, it suggests that the results are biased toward angry customers.
  + If you know your dataset contains biased answers, it's up to you to decide how to deal with the data so that you can still get meaningful results.
* Practice 1
  + We want to estimate the average income in our country. We sample 20 individuals and ask them about their annual income. We do not know whether or not the population of income would follow a normal distribution. Now determine: A) if we can rely on a normal distribution for inference and B) if the central limit theorem applies.
  + Hold off on advancing to the next slide until after you've made your determinations.
  + Practice 1 Answer
    - Remember, we're trying to decide: A) if we can rely on a normal distribution for inference and B) if the central limit theorem applies.
      * We want to estimate the average income in our country. We sample 20 individuals and ask them about their annual income. We do not know whether or not the population of income would follow a normal distribution.
    - A) We cannot rely on a normal distribution; our sample size is fewer than 30 so we don't know whether our original data follows a normal distribution.
    - B) We do not have a large enough sample size to apply the central limit theorem in this situation.
* Practice 2
  + Let's say we frequently make use of the ride-sharing service, Uber, and we want to learn about our average Uber ride time in minutes.
  + We begin by looking at a set of 50 previous Uber trips on our phone, including specific ride times for each. We do not know whether or not our set of Uber ride times follow a normal distribution.
  + Now determine:
    - A) if we can rely on a normal distribution for inference, and
    - B) if the central limit theorem applies.
  + Hold off on advancing to the next slide until after you've made your determinations.
  + Practice 2 Answer
    - A) Yes, we can rely on a normal distribution because our sample size is greater than 30. We don't need to know the distribution of the original data because n>30.
    - B) Yes, the central limit theorem applies in this case because we have a large enough sample size!
* Practice 3
  + We seek to test whether or not 66 inches is a reasonable guess for average height at our company. We sample 15 people. We believe that heights do follow a normal distribution.
  + Now determine:
    - A) if we can rely on a normal distribution for inference and
    - B) if the central limit theorem applies.
  + Hold off on advancing to the next slide until after you've made your determinations.
  + Practice 3 Answer
    - Remember, we're trying to decide: A) if we can rely on a normal distribution for inference and B) if the central limit theorem applies.
      * We seek to test whether or not 66 inches is a reasonable guess for average height at our company. We sample 15 people. We believe that heights do follow a normal distribution.
    - A) We can rely on a normal distribution as our original data follows a normal distribution.
    - B) We do not need to apply the central limit theorem. Because our original data are normally distributed, the sampling distribution will be exactly normally distributed, regardless of sample size!
* Recap
  + Here's what we covered in this lesson:
    - The central limit theorem states that, regardless of the distribution of the original data X, the sampling distribution of will approach a normal distribution as the sample size n increases.
    - If our original data follow a normal distribution, then we can use the normal distribution for inferential purposes (i.e., confidence intervals, hypothesis tests, etc.).
    - If the distribution of the original data is not normal or is unknown, then we can still use the normal distribution for inferential purposes, as long as our sample size is at least 30.