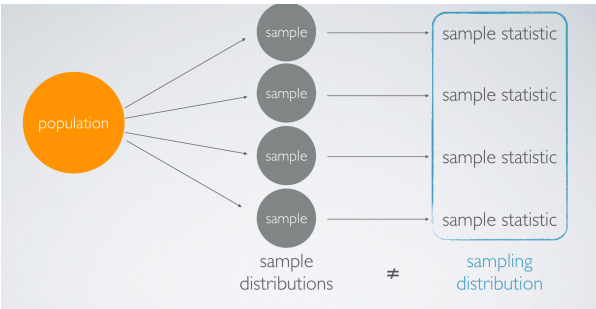
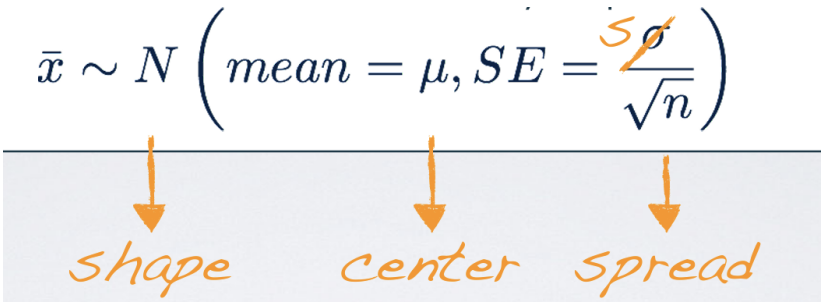
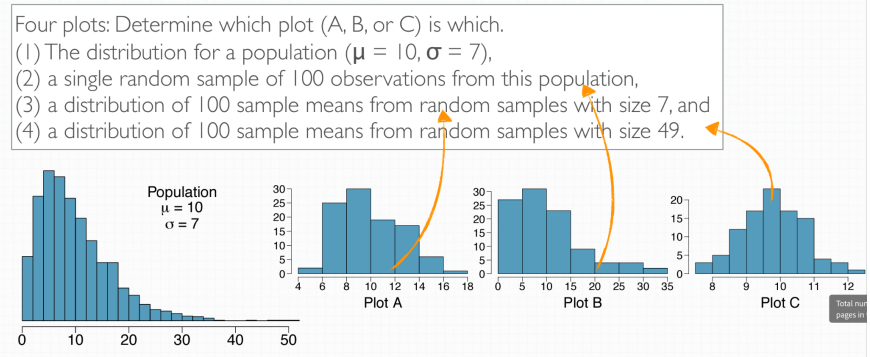
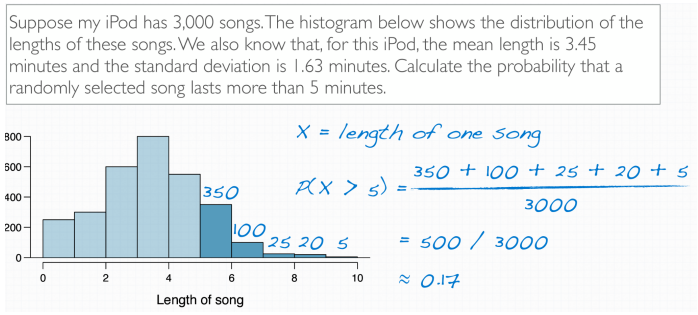
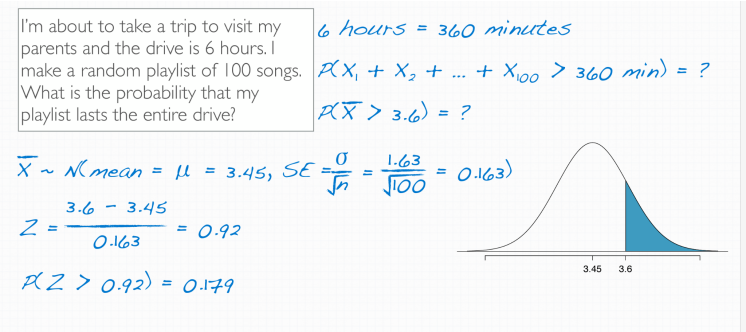
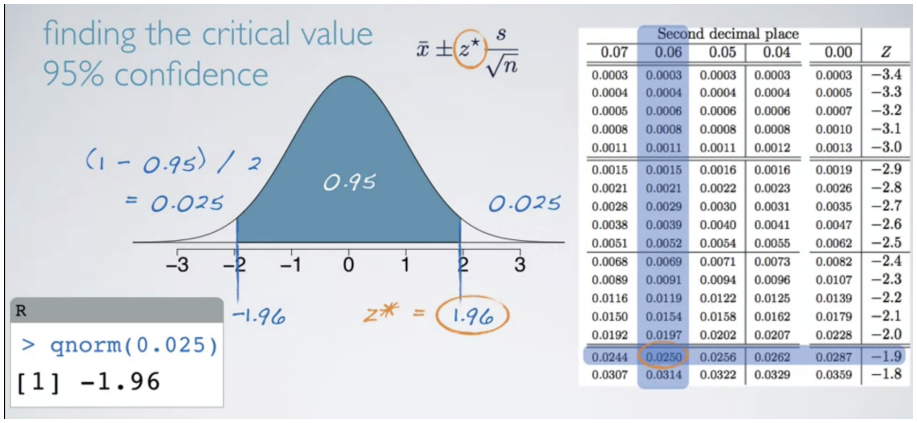
CLT and Sampling
https://gallery.shinyapps.io/CLT_mean/

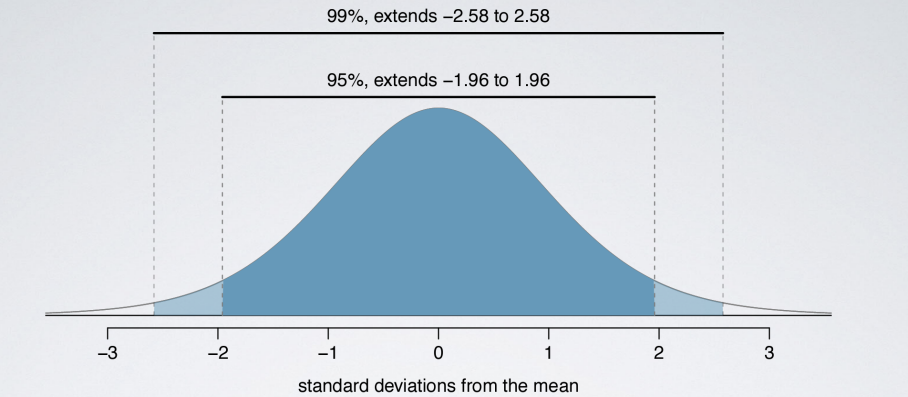
Sample vs Sampling Distribution	Sample: - Distribution of data from the samples Sampling - Distribution of a statistic from several samples (ex. Average of all possible samples) 
Central Limit Theorem; Because this theorem is “central” to statistic logics	CLT: - Distribution of sample statistics is nearly normal, centered at the population mean, with standard error equal to population standard deviation divided by square root of sample size - N means the shape of the distribution - Mean = center - SE = spread of the graph  - Note: if we don’t have the population standard deviation (sigma), then we use the sample standard deviation (s) - Note: we are re-calibrating a population distribution by taking a sample and calibrating around the sample statistic (mean).
Central Limit Theorem Conditions:	1. Independent: a. Random sample/assignment b. If sampling without replacement, n<10% of population* 2. Sample size/skew:

	a. Either the population distribution is normal, or if the population distribution is skewed, the sample size if large n>30 *More on sampling without replacement: - Sampling without replacement - don't take more than 10% of the population to avoid dependence (ex. Population has you and your extended family, if you take 60% of the population, chances are you and your extended family are in there and then the sample would not be independent)
Central Limit In Practice	1. Distribution of a population 2. Distribution of a sample from the population (distribution most resembles 1) 3. Distribution of sample mean from size of 7 of sample (more normal than 2 given CLT, but more skewed than 4) 4. Distribution of sample mean from size of 100 (distribution is most normal given CLT since size if bigger) *sample statistics is more normally distributed than the sample itself 
Central Limit Probability Example:	

	 **Step 1:** Population distribution is shown above as a bar chart **Step 2:** To find the probability of a sample of 100 songs with total playtime of more than 6 hours, we create a distribution of sample statistics of the mean using CLT then find the z-score to calculate the probability that the average is greater than 3.6min Given: - n=3000 - mean=3.45 - sd=1.63 Find: - P(x-bar > 3.6) - Z-score = (observed value - mean)/standard error - Why standard error rather than deviation? Because we are calculating for the sample statistic (mean) and the standard deviation of 1.63 applies to the length of songs.

Confidence Intervals (for a mean)	
Confidence Interval	Plausible range of values for the population parameter Remember: x-bar is sample mean u is population mean
Confidence Interval and Central Limit Theorem:	The standard error lets us know how close the sample statistic's distribution is to the population mean. "The mean of the sample will be within x-standard errors of the population mean" Approximate confidence interval of 64% = mean +- SE Approximate confidence interval of 95% = mean +- 2SE

	Approximate confidence interval of 99.7% = mean +- 3SE **Confidence interval for a population mean:** Computed as the sample mean plus/minus a margin of error (critical value corresponding to the middle XX% of the normal distribution times the standard error of the sampling distribution). $\bar{x} \pm z^\star \frac{s}{\sqrt{n}}$
Confidence Interval Conditions: (sample as CLT)	**Conditions for this confidence interval:** 1. ***Independence:*** Sampled observations must be independent. ▸ random sample/assignment ▸ if sampling without replacement, n < 10% of population 2. ***Sample size/skew:*** n ≥ 30, larger if the population distribution is very skewed.
Finding the Confidence with Critical Values	Critical Value = the specific standard deviation amount (z*) 1. Find the confidence of 95% (margin of error 5%) (Margin of Error)/2 (1-0.95)/2 2. Find corresponding critical value for z a. Using z-table b. Using R (qnorm(0.025) 3. Resulting z is the specific standard deviation of 1.96 
Confidence Level Accuracy vs Precision	The higher the confidence level, the wider the spread to “capture more range of data” so we can have more confidence that we captured the right number.

	 However, this means decreased precision, aka the usefulness of the data. For example, a larger weather range may be more accurate, but the extremes in the range means we don't know what temperature to dress for. CL ↑ width ↑ accuracy ↑ precision ↓ For higher precision and higher accuracy, increase sample size
Calculate the required sample size for target margin of error	$ME = z^\star \frac{s}{\sqrt{n}} \rightarrow n = \left(\frac{z^\star s}{ME}\right)^2$
Calculating sample size for target margin of error Example:	A group of researchers want to test the possible effect of an epilepsy medication taken by pregnant mothers on the cognitive development of their children. As evidence, they want to estimate the IQ scores of three-year-old children born to mothers who were on this medication during pregnancy. Previous studies suggest that the SD of IQ scores of three-year-old children is 18 points. How many such children should the researchers sample in order to obtain a 90% confidence interval with a margin of error less than or equal to 4 points? ME ≤ 4 pts CL = 90% z* = 1.65 σ = 18 $4 = 1.65 \frac{18}{\sqrt{n}} \rightarrow n = \left(\frac{1.65 \times 18}{4}\right)^2 = 55.13$ We need at least 56 such children in the sample to obtain a maximum margin of error of 4 points.

	z* is calculated using the chart or by r where qnorm(0.05)
Interpreting Confidence Interval Examples:	The General Social Survey asks: "For how many days during the past 30 days was your mental health, which includes stress, depression, and problems with emotions, not good?" Based on responses from 1,151 US residents, the survey reported a 95% confidence interval of 3.40 to 4.24 days in 2010. Interpret this interval in context of the data. We are 95% confident that Americans on average have 3.40 to 4.24 bad mental health days per m The General Social Survey asks: "For how many days during the past 30 days was your mental health, which includes stress, depression, and problems with emotions, not good?" Based on responses from 1,151 US residents, the survey reported a 95% confidence interval of 3.40 to 4.24 days in 2010. In this context, what does a 95% confidence level mean? 95% of random samples of 1,151 Americans will yield CIs that capture the true population mean of number of bad mental health days per month.
Confidence Interval Calculation	A sample of 50 college students were asked how many exclusive relationships they've been in so far. The students in the sample had an average of 3.2 exclusive relationships, with a standard deviation of 1.74. In addition, the sample distribution was only slightly skewed to the right. Estimate the true average number of exclusive relationships based on this sample using a 95% confidence interval. Step 1: Does this meet the Conditions of Confidence Intervals? 1. random sample & 50 < 10% of all college students We can assume that the number of exclusive relationships one student in the sample has been in is independent of another. 2. n > 30 & not so skewed sample We can assume that the sampling distribution of average number of exclusive relationships from samples of size 50 will be nearly normal. $n = 50$ $\bar{x} = 3.2$ $s = 1.74$ Step 2: Find CI CI = mean +- margin of error Step 3: Find margin of error then put it together to find CI $SE = \frac{s}{\sqrt{n}} = \frac{1.74}{\sqrt{50}} \approx 0.246$ $\bar{x} \pm z^* SE = 3.2 \pm 1.96(0.246)$ $= 3.2 \pm 0.48$ $= (2.72, 3.68)$ Step 5: Interpret the Answer

	We are 95% confident that college students on average have been in 2.72 to 3.68 exclusive relationships.