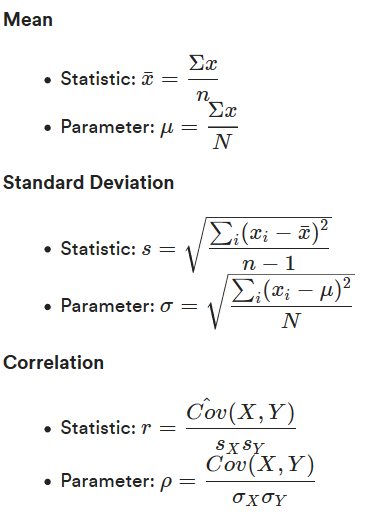
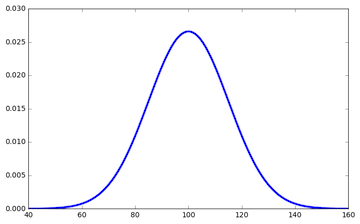
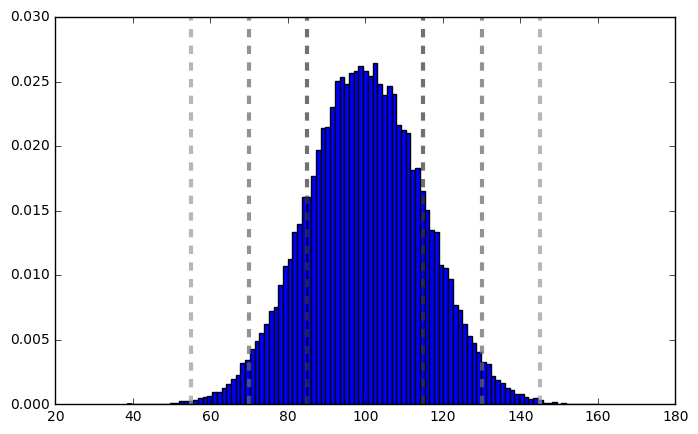
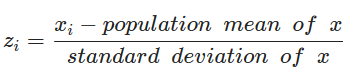
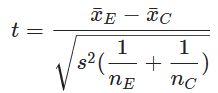
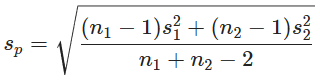
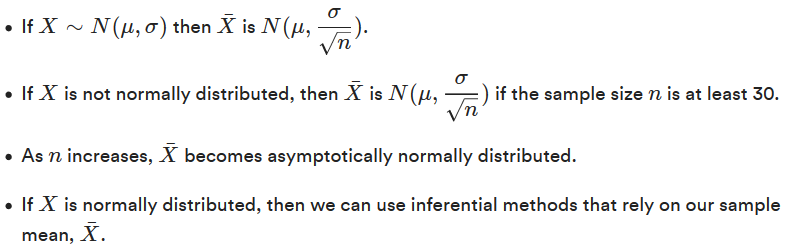
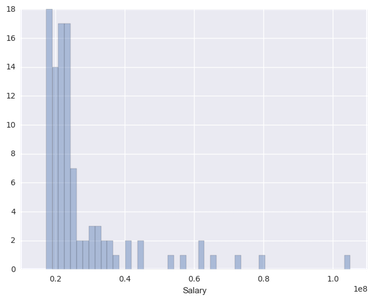
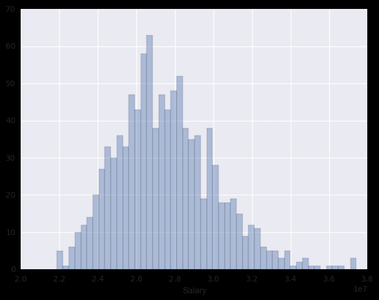
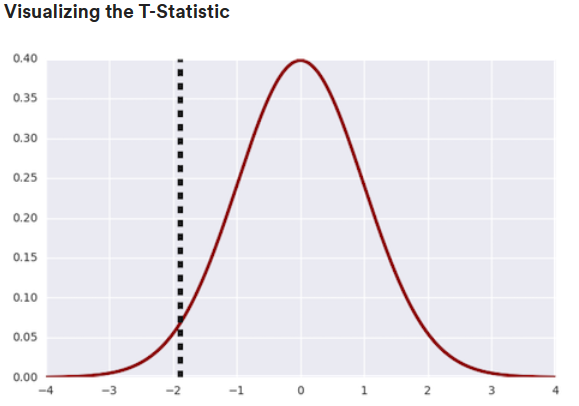
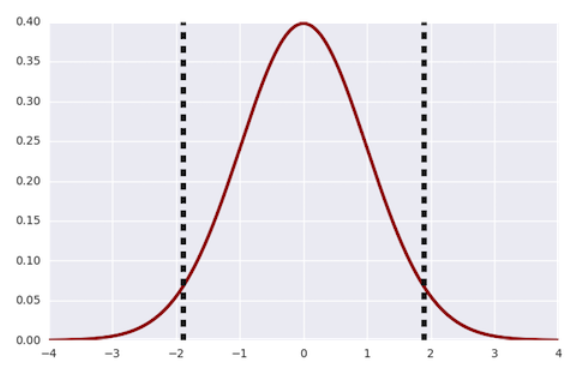
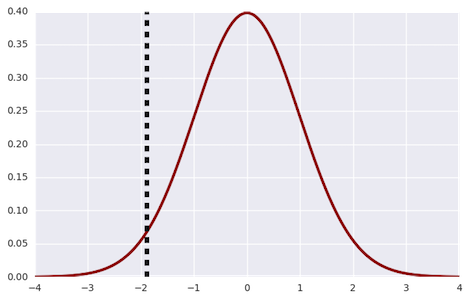
Unit 3-2 Conducting a T-Test and Calculating P Values

* Drug Efficacy Experiment
  + Imagine that we're involved with the following experiment:
  + We're interested in testing the efficacy of a new drug, which is supposed to lower blood pressure.
  + We randomly select 50 people to be in the placebo control condition and 50 people to receive the treatment.
  + We know our sample is selected from the broader, unknown population pool.
  + We're interested in the average difference in blood pressure levels between the treatment and control groups.
* Review Sample Statistics and Population Parameters
  + A sample is a small group that's representative of a larger population of interest.
  + We use sample statistics to estimate population parameters. Our goal is to calculate sample statistics and then rely on the properties of a random sample to make inferences we can apply to a larger population of interest.
  + If a sample is truly randomized, then its distribution should be representative of the population. Meanwhile, a statistic is a function of the data, while a parameter is a characteristic of the population.
  + Let's recall some example sample statistics and population parameters:
    - 
* The Normal Distribution
  + **Normal distribution** is the most commonly used distribution in statistics. Normality is an assumption that underlies many statistical tests and is a convenient model for the distribution of many variables. Two parameters are used for the normal distribution: the mean and the standard deviation.
  + Example: Intelligence Quotient
    - Intelligence quotient (IQ) follows a normal distribution by design. IQ is normally distributed with a mean of 100 and a standard deviation of 15. You might see this specified like:
      * IQ∼Normal(100,15)
      * Or
      * IQ∼N(100,15)
    - Here we've plotted a normal distribution of IQ with a mean of 100 and a standard deviation of 15:
      * import numpy as np
      * import scipy.stats as stats
      * import matplotlib.pyplot as plt
      * # Generate points on the x axis:
      * xpoints = np.linspace(40, 160, 500)
      * # Use stats.norm.pdf to get values on the probability density function for the normal distribution:
      * ypoints = stats.norm.pdf(xpoints, 100, 15)
      * # Initialize a matplotlib "figure:"
      * fig, ax = plt.subplots(figsize=(8,5))
      * # Plot the lines using Matplotlib's plot function:
      * ax.plot(xpoints, ypoints, linewidth=3)
      * plt.show()
      * 
* Standard Deviation
  + Remember that a standard deviation is used to quantify the amount of variation in or the dispersion of a set of values. Low standard deviations mean that the data tend to be clustered close to the mean, while high standard deviations indicate that the data points are spread in a wider distribution around the mean.
* 68-95-99.7 Rule
  + It’s useful to know how extreme a particular observation is. By extreme, we mean how far it is from the mean (expected value).
  + For example, an extreme stock price might mean a big shift in the market (and might indicate whether or not it's a good idea to buy or sell). An extreme drop in air pressure might indicate a weather event. Quantifying how extreme these values are may indicate whether or not we need to take action.
  + For a normal distribution:
    - 68 percent of observations from a population will fall within 1 standard deviation of the population mean.
    - 95 percent of observations from a population will fall within 2 standard deviations of the population mean.
    - 99.7 percent of observations from a population will fall within 3 standard deviations of the population mean.
    - Here is a visual representation of the 68-95-99.7 rule on the IQ distribution:
      * mu, sigma = 100, 15
      * xpoints=np.random.normal(mu, sigma, 50000)
      * avg=np.mean(xpoints)
      * std=np.std(xpoints)
      * print(avg,std)
      * std1 = avg + std
      * std1\_neg = avg - std
      * std2 = avg + 2\*std
      * std2\_neg = avg - 2\*std
      * std3 = avg + 3\*std
      * std3\_neg = avg - 3\*std
      * fig, ax = plt.subplots(figsize=(8,5))
      * ax.axvline(std1\_neg, ls='dashed', lw=3, color='#333333', alpha=0.7)
      * ax.axvline(std1, ls='dashed', lw=3, color='#333333', alpha=0.7)
      * ax.axvline(std2\_neg, ls='dashed', lw=3, color='#666666', alpha=0.7)
      * ax.axvline(std2, ls='dashed', lw=3, color='#666666', alpha=0.7)
      * ax.axvline(std3, ls='dashed', lw=3, color='#999999', alpha=0.7)
      * ax.axvline(std3\_neg, ls='dashed', lw=3, color='#999999', alpha=0.7)
      * ax.hist(xpoints,normed=True, bins=100)
      * 
* Z-Score
  + The 68-95-99.7 rule is helpful, but we can get even more specific about quantifying how extreme an observation is.
  + The z-score of an observation quantifies how many standard deviations the observation is away from the population mean.
  + 
  + Z is called the standard normal distribution because it has a mean of 0 and standard deviation of 1.
  + **We denote that a random variable X has a normal distribution as X∼N(μ,σ), where μ is the mean and σ is the standard deviation. The z-score distribution is Z∼N(0,1)**
  + The scipy.stats.zscore() function will convert a vector of values to its respective z-scores.
    - stats.zscore([1,2,3,4,5])
    - array([-1.41421356, -0.70710678, 0. , 0.70710678, 1.41421356])
* T-Score
  + For our drug efficacy experiment, we measured the blood pressure of the subjects in each group (drug treatment and placebo control).
  + We find that the 50 subjects in the **control group** have an mean systolic blood pressure of **121.38**.
  + The 50 subjects in the treatment group have an mean systolic blood pressure of **111.56**.
  + The difference between the means of the treatment and control groups is **-9.82** points. How confident are we that this measured difference is real? Was this measurement simply the result of chance? We can perform a **t-test** to help us answer these questions.
  + First, we'll calculate a t-statistic. The **t-statistic** is a measure of the degree to which our groups differ, standardized by the variance of our measurements.
  + We'll then calculate a **p value**. The p value is a metric indicating the probability that our measured difference was the result of random chance in the sampling of the subjects.
  + Let's set up Python NumPy arrays for the control and experimental observations:
    - control = np.array([166, 165, 120, 94, 104, 166, 98, 85, 97, 87, 114, 100, 152, 87, 152, 102, 82, 80, 84, 109, 98, 154, 135, 164, 137, 128, 122, 146, 86, 146, 85, 101, 109, 105, 163, 136, 142, 144, 140, 128, 126, 119, 121, 126, 169, 87, 97, 167, 89, 155])
    - experimental = np.array([ 83, 100, 123, 75, 130, 77, 78, 87, 116, 116, 141, 93, 107, 101, 142, 152, 130, 123, 122, 154, 119, 149, 106, 107, 108, 151, 97, 95, 104, 141, 80, 110, 136, 134, 142, 135, 111, 83, 86, 116, 86, 117, 87, 143, 104, 107, 86, 88, 124, 76])
  + Let’s look back at the null and alternative hypotheses for our experiment:
    - H0: The difference in systolic blood pressure between the experimental and control groups is 0.
    - H1: The difference in systolic blood pressure between the experimental and control groups is not 0.
  + Let's look at the mean systolic blood pressure of both groups and the difference between the two:
    - print(np.mean(control))
    - print(np.mean(experimental))
    - print(np.mean(experimental) - np.mean(control))
    - 121.38 111.56 -9.82
      * Our measured difference was -9.82. Is this extreme enough to reject our null hypothesis?
  + As frequentists, we want to know P(data∣meandifference). That is, what is the probability that we observed this data given a specified mean in blood difference?
  + We don’t know the true mean difference (because the point of the experiment is to evaluate this drug). So, we'll assume that the true mean difference is zero. We assume the null hypothesis, HO, to be true. P(data∣meandifference=0)
* T-Statistic
  + The **t-statistic** is a classic metric for quantifying the difference between groups. It's a standardized version of the difference between groups, where the standardization adjusts for the variance between measurements.
  + 
  + The numerator is the difference between the group means. E​ is the mean of the experimental group sample measurements, and C​ is the mean of the control group sample measurements. Remember that we assumed the mean difference is zero for our null hypothesis.
  + The denominator is the square root of the pooled sample variance, divided by the sample size. This is the **standard error of the mean**. *nE* and *nC*​ are the observations in each group. *s2* is our sample variance. In this version of the t-test, we're assuming equal variances in our experimental and control groups in the overall population. There is another way to calculate the t-test where equal variance is not assumed, but in our case this is a reasonable assumption to make.
  + The sample variance is calculated as
    - 
    - This combines the variance of the two groups' measurements into a single, pooled metric.
  + Let's set up the mean difference as a Python variable:
    - mean\_difference = np.mean(experimental) - np.mean(control)
    - print(mean\_difference)
    - -9.82
  + Next, we'll calculate the t-statistic using Python functions. First, create a function to calculate the pooled sample variance. Remember that this combines the variance of the two groups' measurements into a single, pooled metric:
    - def sample\_variance(sample1, sample2):
    - # Calculate the sample sizes for the two samples.
    - n1 = len(sample1)
    - n2 = len(sample2)
    - # Calculate the standard deviations for the two samples.
    - dev1 = np.sum([(x - np.mean(sample1))\*\*2 for x in sample1])
    - dev2 = np.sum([(x - np.mean(sample2))\*\*2 for x in sample2])
    - # Calculate the pooled sample variance.
    - s2 = float(dev1 + dev2) / (n1 + n2 - 2)
    - return s2
  + Now, set up a Python function to calculate the t-statistic. This function uses the sample\_variance() function to calculate the pooled sample variance. Then, it calculates the mean of the experimental group and the mean of the control group. It then calculates the denominator of the t-statistic, which standardizes the differences in the means by adjusting for the sample variances.
    - def t\_statistic(expr, ctrl):
    - # Use the sample\_variance function to calculate the pooled sample variance.
    - s2 = sample\_variance(expr, ctrl)
    - # Calculate the mean of the experimental group.
    - mean1 = np.mean(expr)
    - # Calculate the mean of the control group.
    - mean2 = np.mean(ctrl)
    - # The denominator standardizes by adjusting for variances.
    - std = np.sqrt(s2 \* (1./len(expr) + 1./len(ctrl)))
    - return float(mean1 - mean2) / std
    - t\_stat = t\_statistic(experimental, control)
    - print ('Manual t-statistic:', t\_stat)
    - Manual t-statistic: -1.89154629662
  + **scipy.stats can also perform this calculation.**
    - stats.ttest\_ind(experimental, control)
    - Ttest\_indResult(statistic=-1.8915462966190273, pvalue=0.061504240672530394)
* The Central Limit Theorem
  + **Normality** underlies many frequentist statistical inference techniques. Let's say we have a population (represented by the random variable X) and we take from it a sample of size n. We find the mean of that sample.
    - We can take another sample from that population of the same size, n, and take the mean of that sample. We can do this over and over until we've calculated the mean of every possible sample of size n.
    - If we plot every sample mean on a histogram, we get the sample distribution of .
    - **The sampling distribution of will be normally distributed, even if X is not.**
  + Here are some properties that arise from the central limit theorem:
    - 
  + Let's say we're interested in measuring the trees in a forest. The traditional measurement for this is the diameter at breast height (DBM), which is the diameter of the tree measured 4.5 feet above the ground. We can't measure every single tree in the forest, but let's say we measure many smaller samples and compare the means of each.
    - If we measured five trees per each of these repeated samples, we would find that the distribution of the samples means for all groups would be somewhat bell-shaped (normal). If we increase the size of each sample taken (to 20, and then to 200 trees sampled per group), we would notice that the spread of the distribution of the sample mean would decrease. That is, the distribution of the average tree DBM would become more and more normal.
      * The averages of samples have approximately normal distributions.
      * As the sample size increases, this distribution becomes narrower and more normal.
      * This works even if the distribution of tree DBMs in the forest is not normal to begin with.
  + Below, we've plotted the distribution of data on today's highest-paid athletes. As we can see, these are far from normally distributed. Despite this, the distribution of the sample means for these variables will be normally distributed (according to the central limit theorem).
    - 
    - We then take a random sample size of 30 from this salary list 1,000 times, calculate the means of the samples, and return the sample means as a new list. Here's the distribution of the sample means of size 30:
    - 
    - As we can see, this looks much more normally distributed. If we took more samples to increase the sample size, we'd see the spread of the sample mean distribution decrease.
* The Connection Between the T-Test/Z-Test and The Central Limit Theorem
  + From the central limit theorem, we know that (with asymptotically infinite samples) the distribution of sample means drawn from a population is normal. In the case of smaller sample sizes, we can adjust this to be the more conservative t-distribution. The t-distribution is like the normal distribution with fatter tails.
  + 
  + Here, we plotted the t-distribution for our drug efficacy experiment. It's centered on zero. This zero value corresponds to our null hypothesis (the difference between the means is zero). We plotted a vertical line with our measured t-statistic of -1.89.
* P Values
  + Now that we have a t-statistic, we can calculate the statistical significance of the test using a p value. Statistical significance indicates our degree of confidence to infer a parameter about the overall population from a statistic calculated using our sample.
  + **Beware**: P values are often communicated and perceived incorrectly. It's important to remember that the p value does NOT relate to the probability of the parameter. In frequentist statistics, we assume the parameter (such as the mean difference) is fixed. This probability is associated with the data.
  + The p value is the probability that, given the null hypothesis (H0) is true, we could have ended up with a statistic that was at least as extreme as the one measured from our random sample of data from the true population.
  + In our example, we've measured a difference in blood pressure of -9.82 between the experimental and control groups. We then calculated a t-statistic associated with this difference of -1.89. The p value is the probability that, given that there is a 0.0 true difference in blood pressure between experimental and control conditions, we were able to calculate a t-statistic at least as extreme as -1.89 from a random sample of data from the true population.
    - In other words, how easily could we have gotten the value -1.89 by chance? Does it indicate that there is indeed a difference between the experimental and control groups?
* One-Tailed Vs. Two-Tailed Test
  + Our null hypothesis states that there is no difference between groups. A t-statistic for no difference between groups would be zero.
  + Our alternative hypothesis is that the difference between groups is not zero. This could mean that the difference is greater than or less than zero — we didn’t specify which one. This is known as the **two-tailed t-test**.
  + 
  + Here, we test the t-distribution with our t-statistic of -1.89 as well as 1.89, because we are doing a two-tailed hypothesis test.
  + Our p value corresponds to the area under the curve of the distribution where the magnitude of the t-statistic is greater than or equal to the one we measured (shown with dotted lines). Our p value for the two-tailed test is 0.062.
  + If we'd indicated that we were only interested in whether or not the drug lowered blood pressure (and didn't raise it), we'd reframe our hypotheses.
    - H0: The difference in systolic blood pressure between the experimental and control groups is greater than or equal to 0.
    - H1: The difference in systolic blood pressure between the experimental and control groups is less than 0.
    - This would be a one-tailed t-test.
  + 
    - Here, we plot the t-distribution with our t-statistic of -1.89 for the one-tailed hypothesis test.
    - Our p value corresponds to the area under the curve of the distribution where the magnitude of the t-statistic is less than or equal to the one we measured (shown with dotted lines). Our p value for the one-tailed test is 0.031.
* Significance Thresholds
  + We say our p value has statistical significance when it's less than some predetermined significance threshold α. Typically, a level of 5 percent is used.
  + If our p value is less than our significance threshold, we conclude that our data were very unlikely to occur given the null hypothesis.
  + The significance level α is the probability of the study rejecting the null hypothesis, given it was true (type I error). We'll discuss type I errors in a later lesson.
  + In our drug efficacy experiment, our p value of 0.062 is not less than the α of .05, so we fail to reject our null hypothesis. That is, we can't mathematically conclude that the drug made a difference in systolic blood pressures.
* Limitations of the T-Test
  + The t-test can only be used when comparing two sample means. When you want to compare the means of more than two samples, an analysis of variance test (ANOVA) is preferred. Using a t-test to compare more than two means greatly increases the chance of errors.
  + The t-test also assumes that each observation is independent of the others. If you have paired values, you can use a paired t-test. The t-test also assumes that the dependent variable is continuous and that the dependent variable has a normal distribution with equal variances in each group.