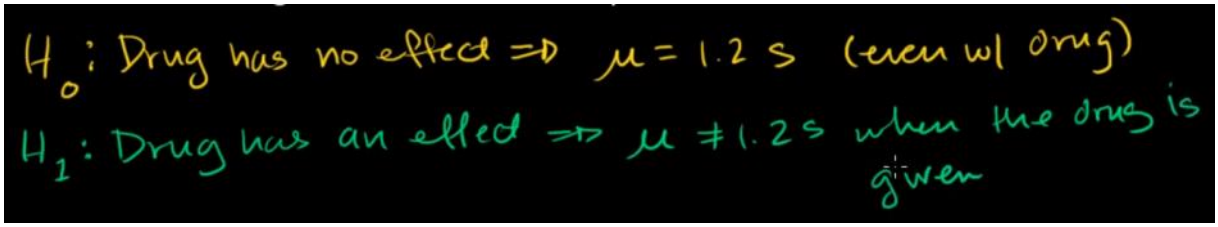
Tuesday, May 14, 2019 7:38 PM

Hypothesis testing:

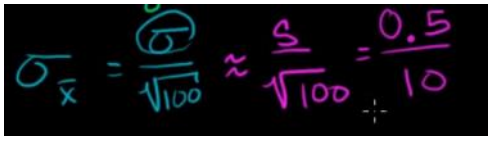
A neurologist is testing the effect of a drug on response time by injecting 100 rats with a unit dose of the drug, subjecting each to neurological stimulus, and recording its response time. The neurologist knows that the mean response time for rats not injected with the drug is 1.2 seconds. The mean of the 100 injected rats' response times is 1.05 seconds with a sample standard deviation of 0.5 seconds. Do you think that the drug has an effect on response time?

H0: drug has no effect

H1: drug has an effect



Let's assume the H0(null hypothesis) is true:

- The mean of our sample distribution is the same as the mean of the population distribution(1.2 s)
- The standard deviation of our sampling distribution should be equal to 1/(square root of population distribution)
- We don't know the standard deviation of the population, so we estimate it as the standard deviation of the sample, which is 0.5s (notice that the sd of the sample is not the sd of the sample distribution)
- 

So if the sample(drug has no effect) is come from the population, we need to consider what is the probability for the sample to get a mean of 1.05, or what is the standard deviation for the population

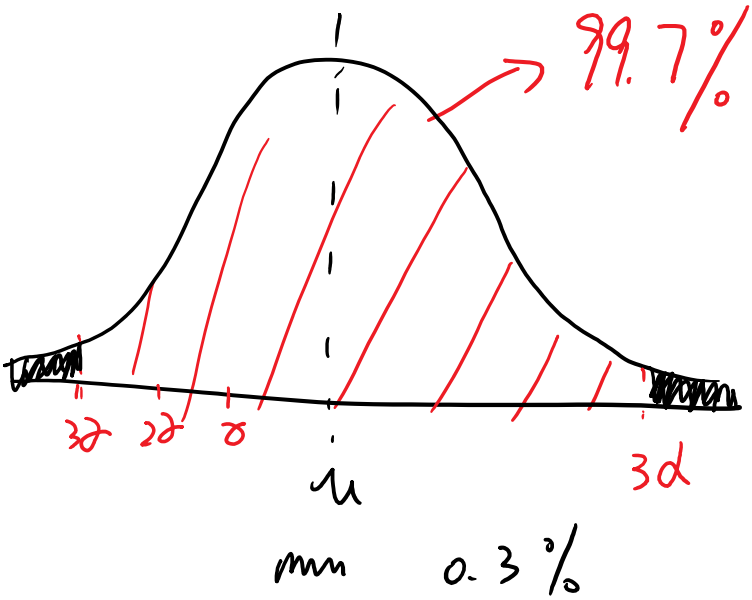
with a mean of 1.2 s to get a sample mean of 1.05s

Z score:

Z = (1.2-1.05)/0.05=3

So the sample mean is 3 sd away from the population mean

The probability of within 3 sd is 99.7% (two-tail)

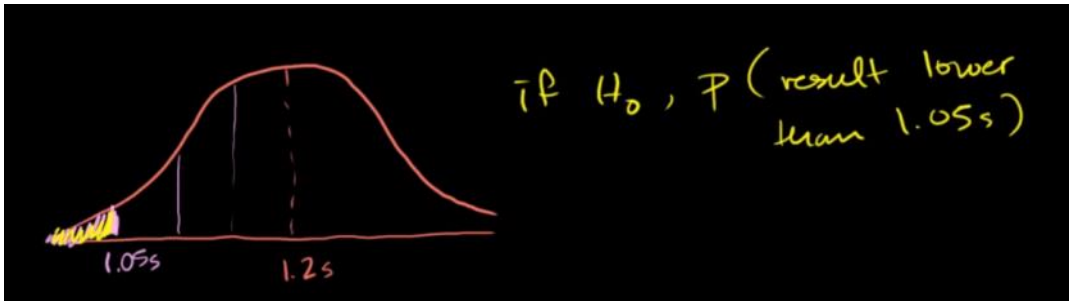


We are going to reject the null hypothesis (two tail test)

One tail test:

H0: drug has no effect

H1: durg lowers response time



Z-statistics & t-statistics:

$$Z = \frac{\bar{x} - \mu_{\bar{x}}}{\sigma_{\bar{x}}}$$

$\bar{x}$: sample mean; $\mu_{\bar{x}}$: population mean

$\sigma_{\bar{x}}$: sample distribution SD

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

n → sample size

σ: SD of population

↓

SS

S

SD of sample

Here is the problem:
We can only assume the sd of the population is approximately the sd of the sample, when the sample size is relatively big.

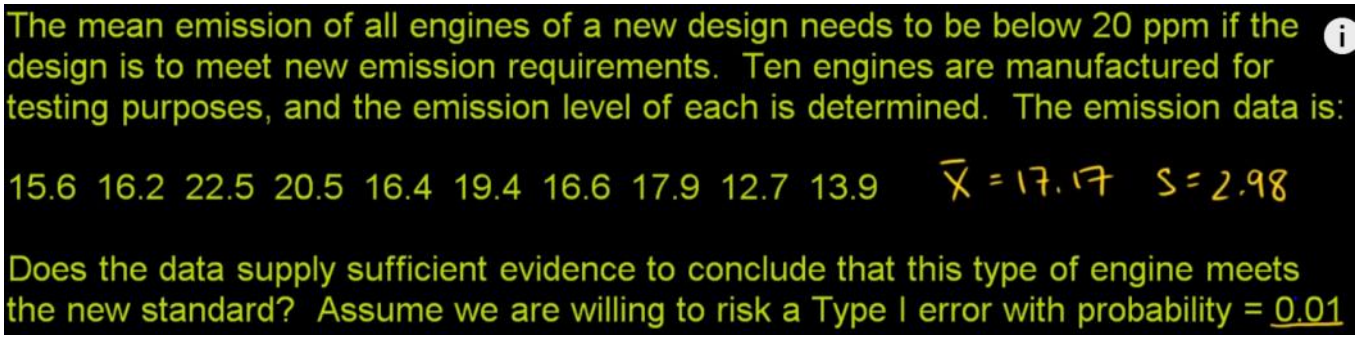
And only in this situation, can we apply z score test

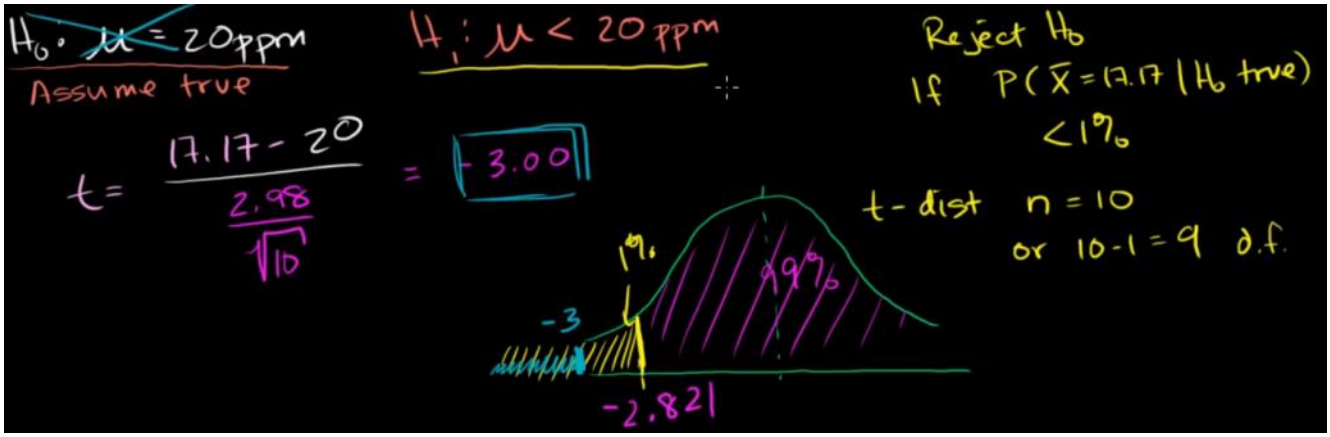
In most of the case, we need our sample size to be greater than 30.

If the sample size is small (<30), we use **t test**, look up the t table.

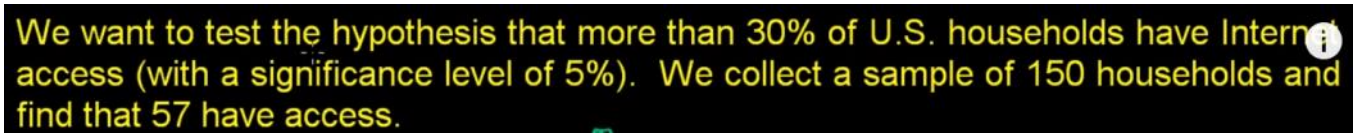
Type 1 error:

Rejecting H0 even though it is true





And vice visa, you can define the confidence interval, and calculate the new requirement of engines

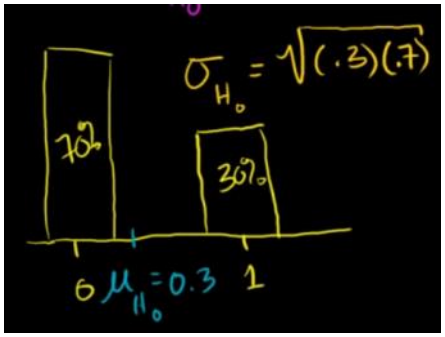


H0: p <= 30%

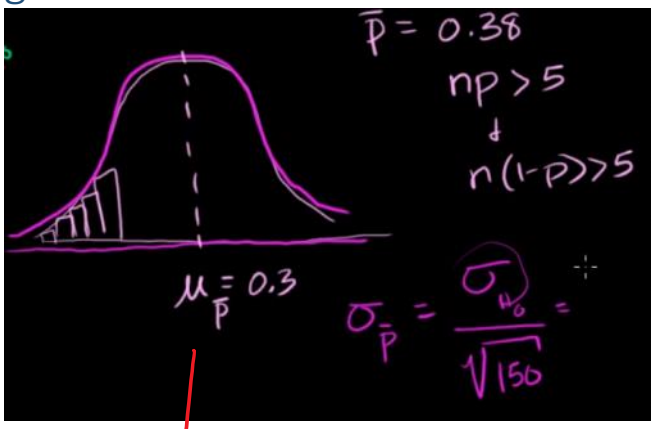
H1: p>30%

Assuming H0 true

This is a Bernoulli distribution, the SD = square root of (p.(1-p))



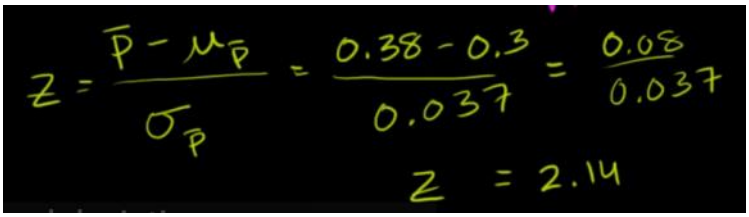
In Bernoulli distribution, you can assume the sample distribution is a normal distribution when np>5 and n(1-p) are greater than 5

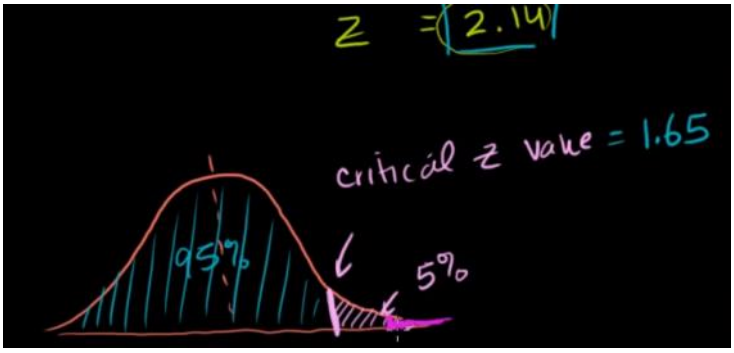


Sample distribution graph

population mean

Sample distribution SD





Variance of differences of random variables:

$$Var(X) = E\left((X-\mu_X)^2\right) = \sigma_X^2$$

$$Var(Y) = E\left((Y-\mu_X)^2\right) = \sigma_Y^2$$

$$Z = X + Y$$

$$A = X - Y$$

$$Var(Z) = Var(X) + Var(Y)$$

$$Var(A) = Var(X) + Var(Y)$$

Proving:
https://en.wikipedia.org/wiki/Variance

$$\operatorname{Var}(X) = \operatorname{E}\left[(X - \mu)^2\right].$$

This definition encompasses random variables that are generated by processes that are discrete, continuous, neither, or mixed. The variance can also be thought of as the covariance of a random variable with itself:

$$\operatorname{Var}(X) = \operatorname{Cov}(X, X).$$

The variance is also equivalent to the second cumulant of a probability distribution that generates X. The variance is typically designated as $\operatorname{Var}(X)$, σ_X^2, or simply σ^2 (pronounced "sigma squared"). The expression for the variance can be expanded:

$$\begin{aligned}\operatorname{Var}(X) &= \operatorname{E}\left[(X - \operatorname{E}[X])^2\right] \\ &= \operatorname{E}\left[X^2 - 2X\operatorname{E}[X] + \operatorname{E}[X]^2\right] \\ &= \operatorname{E}\left[X^2\right] - 2\operatorname{E}[X]\operatorname{E}[X] + \operatorname{E}[X]^2 \\ &= \operatorname{E}\left[X^2\right] - \operatorname{E}[X]^2\end{aligned}$$

https://stats.stackexchange.com/questions/31177/does-the-variance-of-a-sum-equal-the-sum-of-the-variances

To see this let $X_1, \ldots, X_n$ be random variables (with finite variances). Then,

$$\operatorname{var}\left(\sum_{i=1}^{n} X_i\right) = E\left(\left[\sum_{i=1}^{n} X_i\right]^2\right) - \left[E\left(\sum_{i=1}^{n} X_i\right)\right]^2$$

Now note that $\left(\sum_{i=1}^{n} a_i\right)^2 = \sum_{i=1}^{n}\sum_{j=1}^{n} a_i a_j$, which is clear if you think about what you're doing when you calculate $(a_1 + \ldots + a_n) \cdot (a_1 + \ldots + a_n)$ by hand. Therefore,

$$E\left(\left[\sum_{i=1}^{n} X_i\right]^2\right) = E\left(\sum_{i=1}^{n}\sum_{j=1}^{n} X_i X_j\right) = \sum_{i=1}^{n}\sum_{j=1}^{n} E(X_i X_j)$$

similarly,

$$\left[E\left(\sum_{i=1}^{n} X_i\right)\right]^2 = \left[\sum_{i=1}^{n} E(X_i)\right]^2 = \sum_{i=1}^{n}\sum_{j=1}^{n} E(X_i)E(X_j)$$

so

$$\operatorname{var}\left(\sum_{i=1}^{n} X_i\right) = \sum_{i=1}^{n}\sum_{j=1}^{n}\left(E(X_i X_j) - E(X_i)E(X_j)\right) = \sum_{i=1}^{n} \sum_{j=1}^{n} \operatorname{cov}(X_i, X_j)$$

To prove the initial statement, it suffices to show that

$$\operatorname{Var}(X+Y) = \operatorname{Var}(X) + \operatorname{Var}(Y).$$

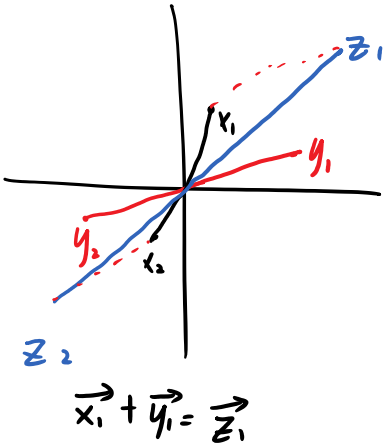
The general result then follows by induction. Starting with the definition,

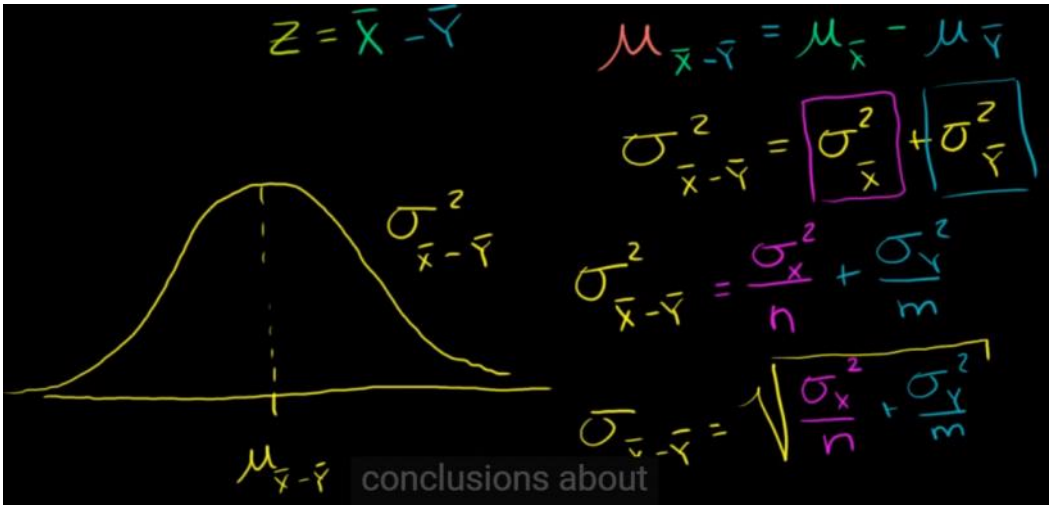
$$\begin{aligned}\operatorname{Var}(X+Y) &= \mathrm{E}[(X+Y)^2] - (\mathrm{E}[X+Y])^2 \\ &= \mathrm{E}[X^2 + 2XY + Y^2] - (\mathrm{E}[X] + \mathrm{E}[Y])^2.\end{aligned}$$

Using the linearity of the expectation operator and the assumption of independence (or uncorrelatedness) of *X* and *Y*, this further simplifies as follows:

$$\begin{aligned}\operatorname{Var}(X+Y) &= \mathrm{E}[X^2] + 2\,\mathrm{E}[XY] + \mathrm{E}[Y^2] - (\mathrm{E}[X]^2 + 2\,\mathrm{E}[X]\,\mathrm{E}[Y] + \mathrm{E}[Y]^2) \\ &= \mathrm{E}[X^2] + \mathrm{E}[Y^2] - \mathrm{E}[X]^2 - \mathrm{E}[Y]^2 \\ &= \operatorname{Var}(X) + \operatorname{Var}(Y).\end{aligned}$$

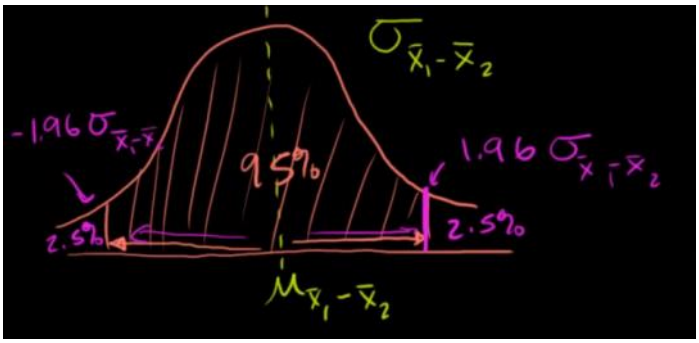
The square root of variance is the standard deviation

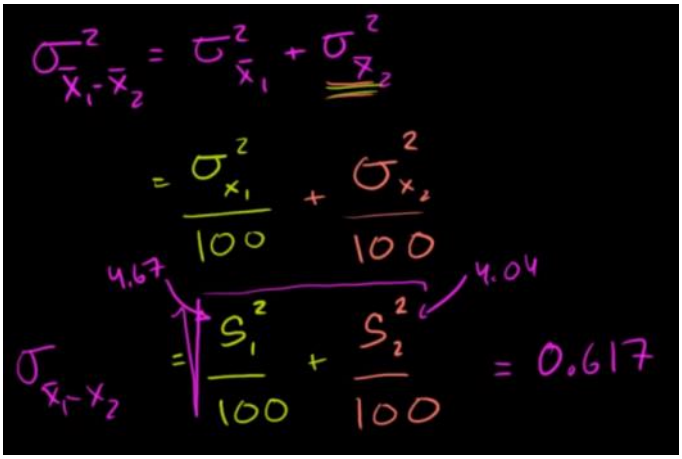




The standard deviation looks like a distance formula...

We're trying to test whether a new, low-fat diet actually helps obese people lose weight. 100 randomly assigned obese people are assigned to group 1 and put on the low fat diet. Another 100 randomly assigned obese people are assigned to group 2 and put on a diet of approximately the same amount of food, but not as low in fat. After 4 months, the mean weight loss was 9.31 lbs. for group 1 (s=4.67) and 7.40 lbs. (s=4.04) for group 2.





1.91+-1.21

0.7-3.12

We are 95% confident the differences between the two groups(normal diet and low fat diet) is between 0.7 and 3.21

Which indicate that the low fat diet does do something on weight diet, because the lower limit is also positive

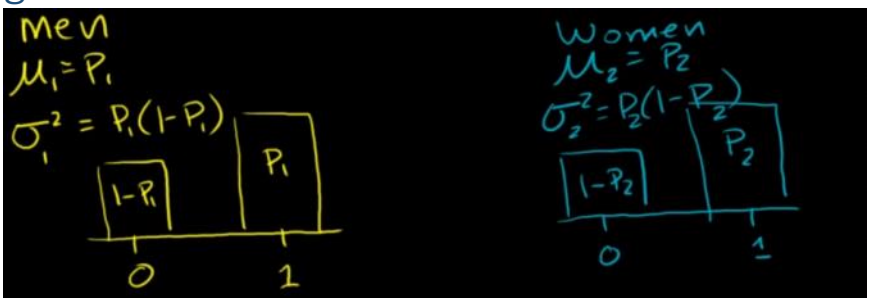
We can also do this using hypothesis test

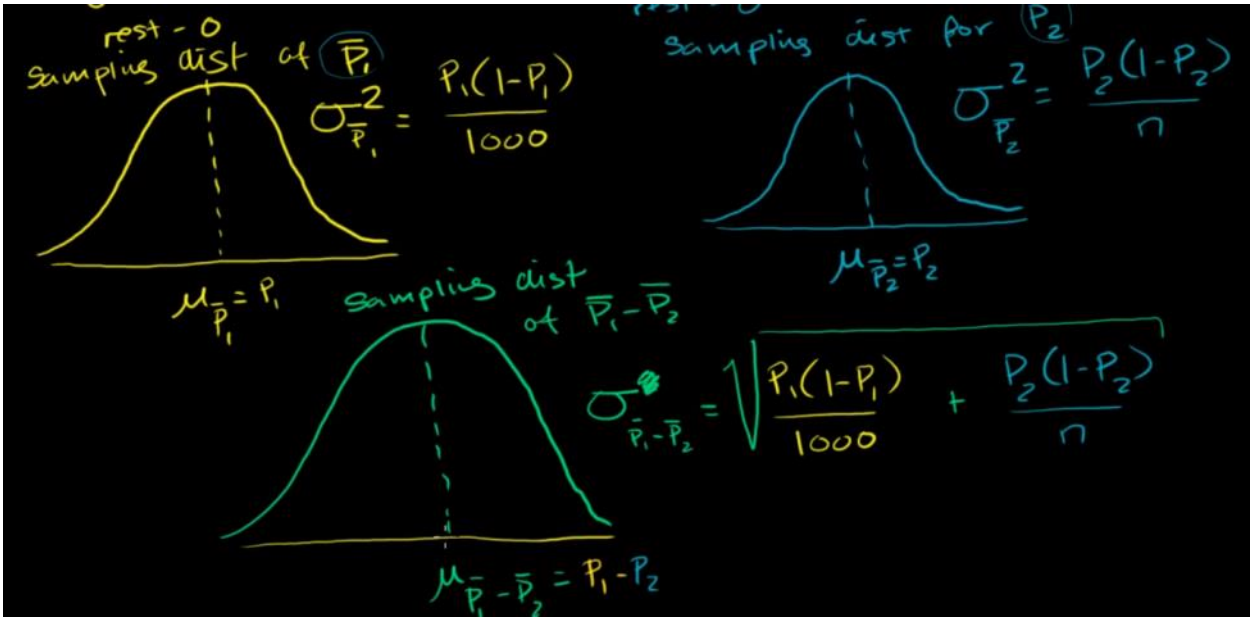
H0: u1-u2 = 0

H1: u1-u2!= 0

Example :

Look for whether there exists difference on the voting rates of different gender





The 95% confidence interval for p1-p2 is 0.008 - 0.094

This problem can also be solved by hypothesis test