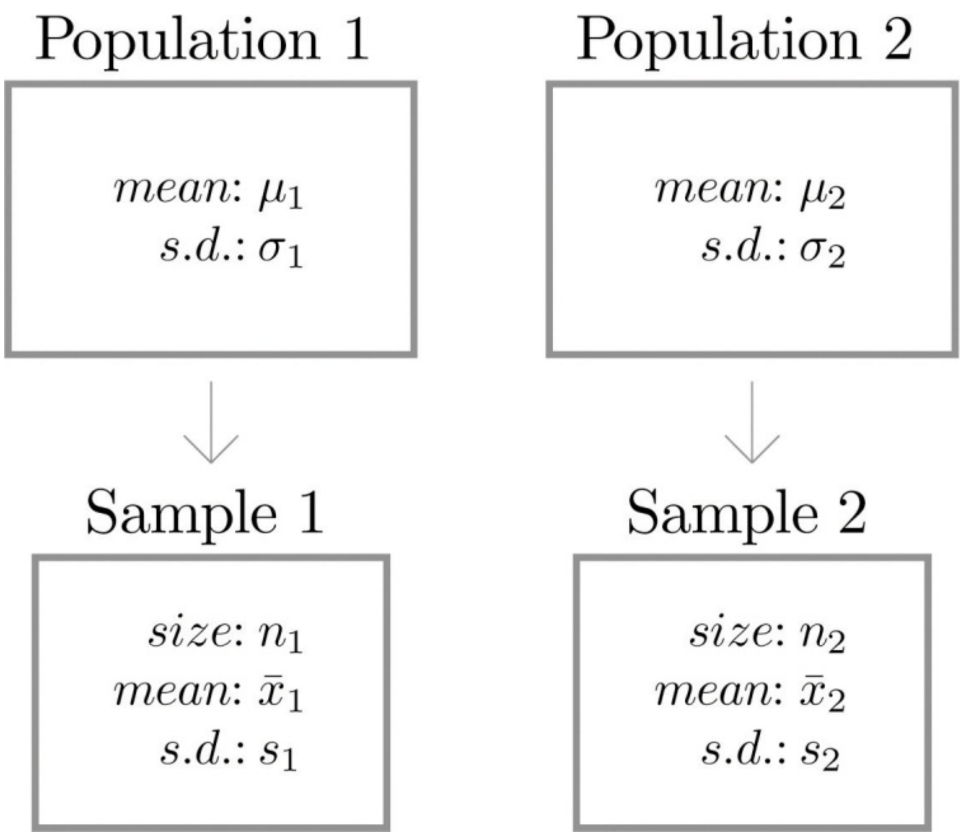
Two-Sample Problems

Previously, we made inference using a one sample, but imagine you want to compare two samples from two different populations.

Random samples from two populations

- Samples from two distinct populations are independent if each one is drawn without reference to the other, and has no connection with the other.



Confidence interval estimation

- Suppose we have a sample from both populations with both n>=30. Than sample mean $\bar{x}_1$ and sample mean $\bar{x}_2$ are both good estimate do the population means μ_1 and μ_2.
- So, we could subtract them to get a point estimate. Or we subtract both sample distributions for an interval estimation.
- If the samples are independent and n_1 and n_2 >= 30 then, can construct the difference of both sample distributions
- And estimate the interval accordingly: $(\bar{x}_1 - \bar{x}_2) \pm z_{\alpha/2}\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}$

Example

To compare customer satisfaction levels of two competing cable television companies, 174 customers of Company 1 and 355 customers of Company 2 were randomly selected and were asked to rate their cable companies on a five-point scale, with 1 being least satisfied and 5 most satisfied. The survey results are summarized in the following table:

Company 1	Company 2
$n_1=174$	$n_2=355$
$\bar{x}_1=3.51$	$\bar{x}_2=3.24$
$s_1=0.51$	$s_2=0.52$

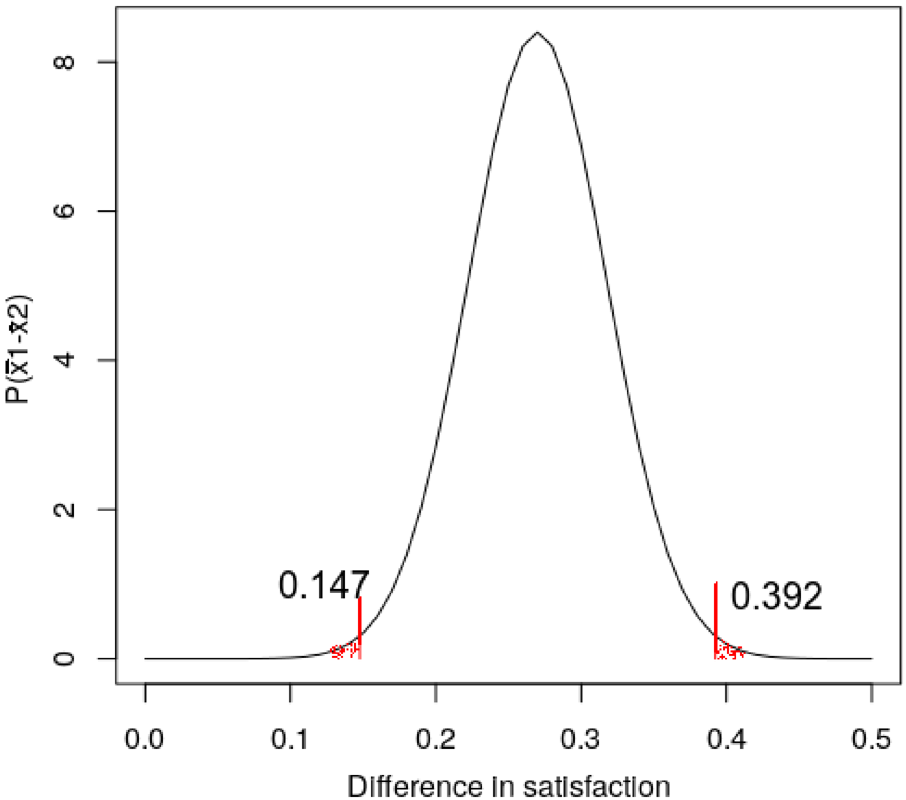
Construct a point estimate and a 99% confidence interval for $\mu_1 - \mu_2$, the difference in average satisfaction levels of customers of the two companies as measured on this five-point scale.

Example

The point estimate of $\mu_1 - \mu_2$ is: $\bar{x}_1 - \bar{x}_2 = 3.51 - 3.24 = 0.27$.

The average costumer satisfaction of company1 is 0.27 points higher.

We are 99% confident that the difference in the population means lies in the interval [0.15,0.29], in the sense that in repeated sampling 99% of all intervals constructed from the sample data in this manner will contain $\mu_1 - \mu_2$. We say that we are 99% confident that the average level of costumer satisfaction for company 1 is between 0.15 and 0.39 higher.



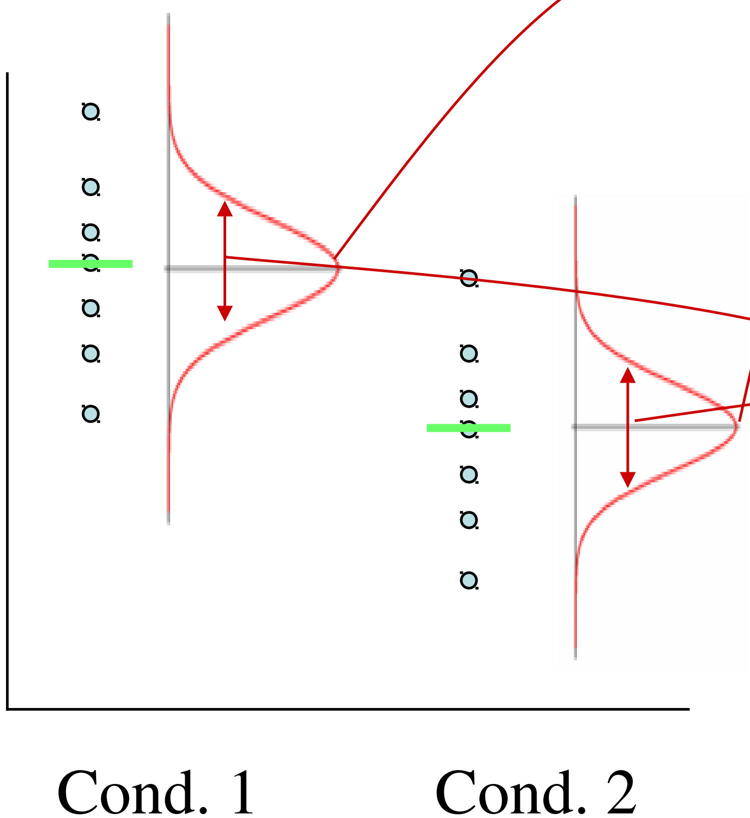
Hypothesis testing

- Hypotheses concerning the relative sizes of the means of two populations are tested using the same critical value and p-value procedures that were used in the case of a single population.
- Thus, the null hypothesis will always be written H_0:$\mu_1 - \mu_2 = D_0$, where D_0 is a number that is deduced from the statement of the situation.
- The alternative hypothesis can take one of the three forms, with the same terminology:

Form of H_a	Terminology
$H_a: \mu_1 - \mu_2 < D_0$	Left-tailed
$H_a: \mu_1 - \mu_2 > D_0$	Right-tailed
$H_a: \mu_1 - \mu_2 \neq D_0$	Two-tailed

Test statistic for large samples

- With both samples are larger than n>=30, we can use the normal test statistic:



$$Z = \frac{\bar{x}_1 - \bar{x}_2}{s_{\bar{x}_1 - \bar{x}_2}}$$

$$s_{\bar{x}_1 - \bar{x}_2} = \sqrt{\frac{s_1^{\,2}}{n_1} + \frac{s_2^{\,2}}{n_2}}$$

Example

- Continuing the initial example, does company1 has a higher mean satisfaction? Test at a 1% significance level.
- Because we assume that the customer mean satisfaction of company1 is higher:

  $H_0 = \mu_1 - \mu_2 = 0$

  $H_a = \mu_1 - \mu_2 > 0$

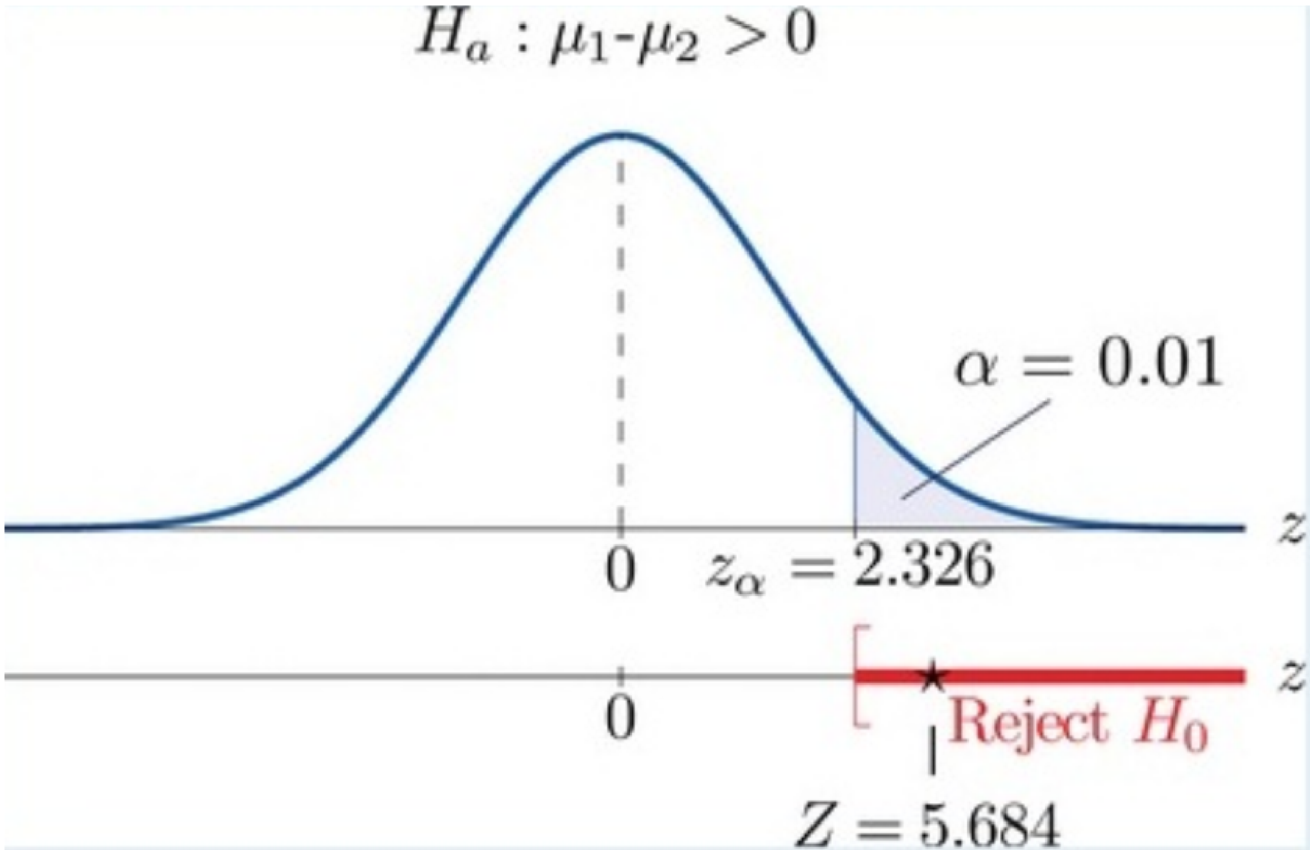
$$Z = \frac{(\bar{x}_1 - \bar{x}_2) - D_0}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}} = \frac{(3.51 - 3.24) - 0}{\sqrt{\frac{0.51^2}{174} + \frac{0.52^2}{355}}} = 5.684$$

- Estimate test statistic:
- Estimate critical region:

  since H_a assumes > its right tailed.

- Also because we want $\mu_1 - \mu_2$ to lie very far to the right away from 0.
- With α=0.01 which the rejection region is [2.32,∞)

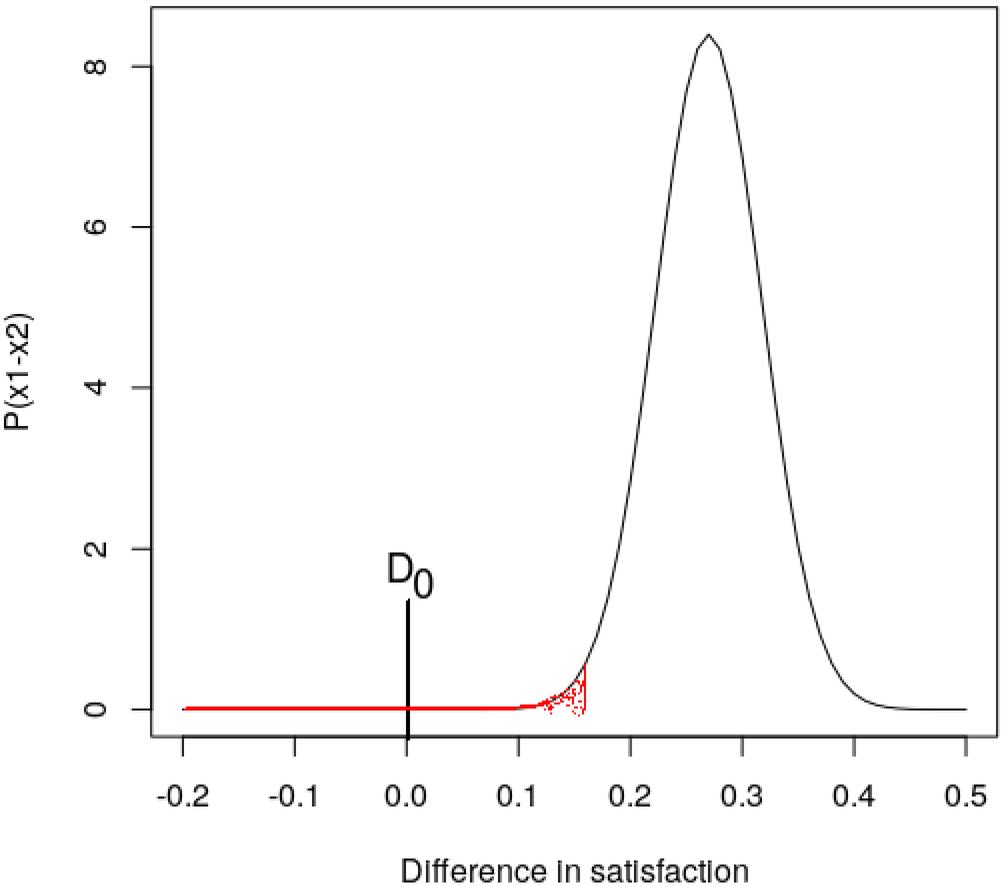
Example



- The decision is to reject H_0. The data provide sufficient evidence, at the 1% level of significance, to conclude that the mean customer satisfaction for Company 1 is higher than that for Company 2

Example

- What if we don't standardize ?
- We really want the difference to be bigger than 0 or D_0, if you think about it.



Small, independent samples

- When one of the sample sizes is smaller than 30, the Central limit theorem does not apply.
- If we assume that variance for population 1 is about the same as that of population 2, we can estimate the common variance by pooling information from samples from population 1 and population 2

  The test statistic is $t = \frac{\bar{X}_1 - \bar{X}_2}{s_p \cdot \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}$, where $s_p = \sqrt{\frac{(n_1 - 1)\, s_{X_1}^2 + (n_2 - 1)\, s_{X_2}^2}{n_1 + n_2 - 2}}.$

  and with df=n_1+n_2-2

- We we assume the variance of both populations to be different we can use a Welch-Test. The test statistic approximately follows a t distribution:

$$T = \frac{\bar{X} - \bar{Y} - \omega_0}{\sqrt{\frac{S_X^2}{n} + \frac{S_Y^2}{m}}} \approx t_\nu.$$

with df $= \frac{\left(\frac{s_x^2}{n} + \frac{s_y^2}{m}\right)^2}{\frac{1}{n-1}\left(\frac{s_x^2}{n}\right)^2 + \frac{1}{m-1}\left(\frac{s_y^2}{m}\right)^2}.$

Comparison of two paired samples population means

- If we have two samples from the some object under different conditions, we use a test for paired samples.
- Two sorts of gasoline are used for the same car. Instead of independent random samples, pairs are select.
- We aim to estimate if the fuel that was used in car1, has a higher fuel economy.

Make and Model	Car 1	Car 2
Buick LaCrosse	17.0	17.0
Dodge Viper	13.2	12.9
Honda CR-Z	35.3	35.4
Hummer H 3	13.6	13.2
Lexus RX	32.7	32.5
Mazda CX-9	18.4	18.1
Saab 9-3	22.5	22.5
Toyota Corolla	26.8	26.7
Volvo XC 90	15.1	15.0

Paired samples test statistic

- To estimate if one fuel is superior to the other, we estimate the difference between each sample pair. As indicated here:

Make and Model	Car 1	Car 2	Difference
Buick LaCrosse	17.0	17.0	0.0
Dodge Viper	13.2	12.9	0.3

- We than use the mean $\bar{d}$ and standard deviation s_d of the differences to estimate the test statistic. with df = n-1

$$T = \frac{\bar{d} - D_0}{s_d / \sqrt{n}}$$

- The difference in the samples can be considered as a random sample selected from a population with mean $\mu_d = \mu_1 - \mu_2$. This essentially transforms the paired two-sample problem into a one-sample problem.

Example

- Since the differences were computed in the order car1− car2, higher fuel economy with Type1 fuel corresponds to $\mu = \mu_1 - \mu_2 > 0$, with α=0.05

  H_0: $\mu_d = 0$

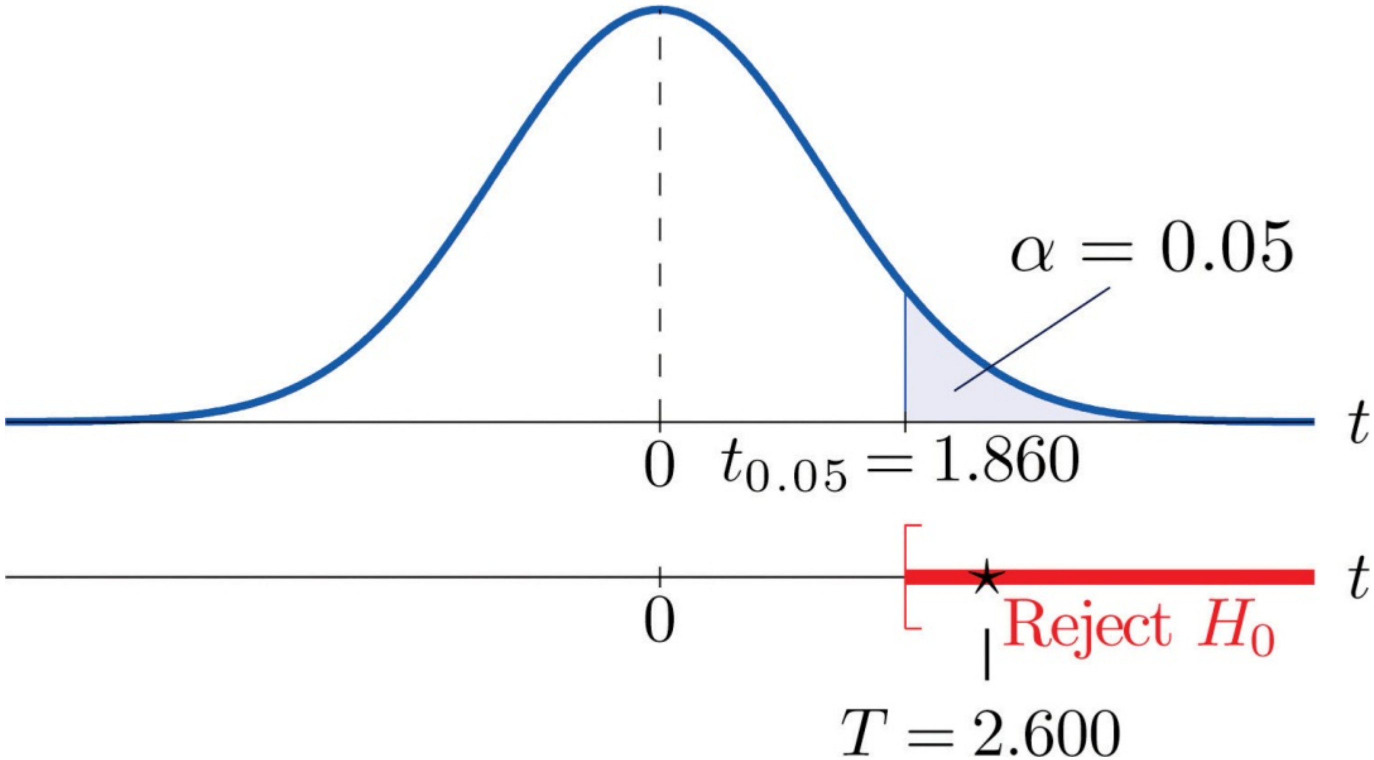
  H_a: $\mu_d > 0$

  $$T = \frac{\bar{d}}{sd/\sqrt{n}} = \frac{0.14}{0.16/3} = 2.6$$

- Rejection region with df=8 [1.86,∞)

Example

$$H_a : \mu_d > 0$$



The data provide sufficient evidence, at the 5% level of significance, to conclude that the mean fuel economy provided by Type 1 gasoline is greater than that for Type 2 gasoline.