Special tests

t-test for the mean

Assume a sample $X_i \sim N(\mu, \sigma^2)$ $i = 1, 2, \ldots, n$ independent and σ^2 unknown

We have 3 different possible tests:

1) $H_0: \mu \leq \mu_0$ vs $H_1: \mu > \mu_0$

2) $H_0: \mu \geq \mu_0$ vs $H_1: \mu < \mu_0$

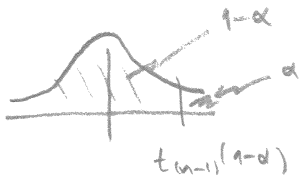
3) $H_0: \mu = \mu_0$ vs $H_1: \mu \neq \mu_0$

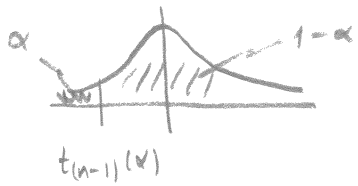
As we have seen in the examples, for cases 1 and 2 is enough to use $H_0: \mu = \mu_0$, that way we make possible to determine the null distribution

The test statistic is $T = \frac{\bar{X} - \mu_0}{}$ where $\bar{X} = \frac{1}{n} \sum_{i=1}^{n} X_i$ and

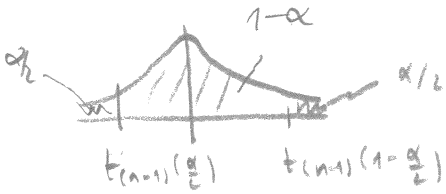
If we fix in advance a significance level α we have the following rules to reject $H_0: \mu = \mu_0$

H_1	Reject H_0 if
$\mu > \mu_0$	$T_{obs} > t_{(n-1)}(1-\alpha)$
$\mu < \mu_0$	$T_{obs} < t_{(n-1)}(\alpha)$
$\mu \neq \mu_0$	$\lvert T_{obs} \rvert > t_{(n-1)}(1-\frac{\alpha}{2})$





As in the previous example, if σ^2 is known, the rejection rule is the same but using the percentiles of $N(0,1)$



Example: Test $H_0: \mu = 15$ From a sample of size $n=30$ we get $\bar{X} = 14$

$H_1: \mu \neq 15$

and $S = 3$. Then $T = \frac{\bar{X} - \mu_0}{S/\sqrt{n}} = \frac{\bar{X} - 15}{S/\sqrt{n}} \sim t_{(30-1)} = t_{(29)}$

The observed test statistic is $T_{obs} = \frac{14 - 15}{3/\sqrt{30}} = -1.83 \Rightarrow |T| = 1.83$

If $\alpha = 0.05$ $t_{(29)}(1 - \frac{0.05}{2}) = t_{(29)}(0.975) = 2.045$ so we don't reject

$$p\text{-value} = 2\,P_r(T > |T_{obs}|) = 2\,P_r(T > 1.83) =$$

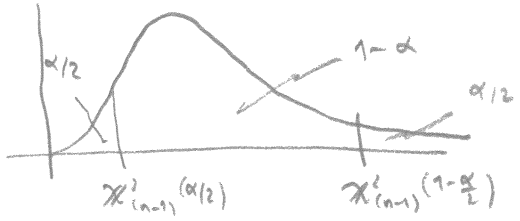
$$= 2[1 - P_r(T < 1.83)] = 2[1 - 0.9612] = 0.0775$$

Test for the variance

Assume $X_i \sim N(\mu, \sigma^2)$ $i = 1, 2, \ldots, n$ independent

we want to test $H_0: \sigma^2 = \sigma_0^2$ against three alternatives as before. The test statistic is $T = \frac{(n-1)S^2}{\sigma_0^2} \sim \chi^2_{(n-1)}$

H_1	Reject H_0 if
$\sigma^2 > \sigma_0^2$	$T_{obs} > \chi^2_{(n-1)}(1-\alpha)$
$\sigma^2 < \sigma_0^2$	$T_{obs} < \chi^2_{(n-1)}(\alpha)$
$\sigma^2 \neq \sigma_0^2$	$T_{obs} < \chi^2_{(n-1)}(\frac{\alpha}{2})$ or $T_{obs} > \chi^2_{(n-1)}(1-\frac{\alpha}{2})$ non-symmetry of χ^2 distribution



Example Test $H_0: \sigma^2 = 1$ from a sample of size $n=25$ that yield
$H_1: \sigma^2 \neq 1$

$s^2 = 1.408$ (assume the sample fulfills the requirements of the test)

Set $\alpha = 0.05$. $T_{obs} = \frac{(n-1)s^2}{\sigma_0^2} = \frac{(25-1)\cdot 1.408}{1} = 33.7920$

$\chi^2_{(24)}(0.05/2) = 12.40$, $\chi^2_{(24)}(1-0.05/2) = 39.36$

As $12.40 \leq T_{obs} \leq 39.36$ we don't reject H_0.

Test for two independent samples

In many experiments, two samples may be regarded as independent, such as two different signals, or medical studies where a sample has been treated with some drug and the other is a control. In such a case it's of interest to test if the drug is effective or not and one way of doing it is by comparing the mean values of both samples.

Assume $X_i \sim N(\mu_X, \sigma^2)$ $i=1,2,\dots,n$ and $Y_j \sim N(\mu_Y, \sigma^2)$ $j=1,2,\dots,m$ independent. The effect of the treatment can be measured by $\mu_X - \mu_Y$, and a natural estimate for it is $\bar{X} - \bar{Y} \sim N\left[\mu_X - \mu_Y, \sigma^2\left(\frac{1}{n} + \frac{1}{m}\right)\right]$

If σ^2 is known $Z = \frac{(\bar{X} - \bar{Y}) - (\mu_X - \mu_Y)}{\sigma\sqrt{\frac{1}{n} + \frac{1}{m}}} \sim N(0,1)$

However, it'll usually be unknown and would be estimated from the data as $s_p^2 = \frac{(n-1)s_X^2 + (m-1)s_Y^2}{m+n-2}$ where $s_X^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})^2$ and

$$s_Y^2 = \frac{1}{m-1}\sum_{j=1}^{m}(Y_j - \bar{Y})^2$$

Then $T = \frac{(\bar{X} - \bar{Y}) - (\mu_X - \mu_Y)}{s_p\sqrt{\frac{1}{n} + \frac{1}{m}}} \sim t_{(m+n-2)}$

Now we are ready to outline the details of a test. As we have already said, we want to test $H_0: \mu_X = \mu_Y$ which is equivalent to $H_0: \mu_X - \mu_Y = 0$

H_1	Reject H_0 if	
	σ^2 known	σ^2 unknown
$\mu_1 - \mu_2 > 0$	$z_{obs} > z(1-\alpha)$	$T_{obs} > t_{(m+n-2)}(1-\alpha)$
$\mu_1 - \mu_2 < 0$	$z_{obs} < z(\alpha)$	$T_{obs} < t_{(m+n-2)}(\alpha)$
$\mu_1 \neq \mu_2$	$\lvert z_{obs}\rvert > z(1-\frac{\alpha}{2})$	$\lvert T_{obs}\rvert > t_{(m+n-2)}(1-\frac{\alpha}{2})$

Example. Two different methods where used to determine latent heat fusion of ice.

Method A	Method B
$n = 13$	$m = 8$
$\bar{X}_A = 80.02$	$\bar{X}_B = 79.98$
$S_A = 0.024$	$S_B = 0.031$

We test $H_0: \mu_A = \mu_B$, $H_1: \mu_A \neq \mu_B$

$$S_p^2 = \frac{(n-1)S_A^2 + (m-1)S_B^2}{m+n-2} = \frac{12\cdot(0.024)^2 + 7(0.031)^2}{13+8-2}$$

$$= 0.0007178 \Rightarrow S_p = 0.027$$

$$T = \frac{(\bar{X}_A - \bar{X}_B) - (\mu_A - \mu_B)}{S_p\sqrt{\frac{1}{n}+\frac{1}{m}}} = \frac{80.02 - 79.98}{0.027\sqrt{\frac{1}{13}+\frac{1}{8}}} = 3.3333$$

. Set $\alpha = 0.05$, then

$t_{(m+n-2)}(1-\frac{\alpha}{2}) = t_{(19)}(0.975) = 2.093$ as $|T_{obs}| > 2.093$ we reject H_0

Confidence intervals and hypothesis tests

Let $X_1, X_2, \ldots, X_n$ be a random sample from $N(\mu, \sigma^2)$ where μ an unknow and σ^2 is known. consider the test $H_0: \mu = \mu_0$, $H_1: \mu \neq \mu_0$. According to our rule

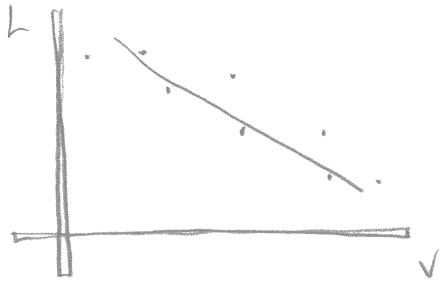
we accept H_0 (at significance level α) if $|z_{obs}| = \left|\frac{\bar{X} - \mu_0}{\sigma/\sqrt{n}}\right| < z(1-\frac{\alpha}{2}) \Leftrightarrow |\bar{X} - \mu_0| < \sigma/\sqrt{n}\, z(1-\frac{\alpha}{2})$

$$\Leftrightarrow -\frac{\sigma}{\sqrt{n}} z(1-\tfrac{\alpha}{2}) < \bar{X} - \mu_0 < \frac{\sigma}{\sqrt{n}} z(1-\tfrac{\alpha}{2}) \Leftrightarrow \bar{X} - \frac{\sigma}{\sqrt{n}} z(1-\tfrac{\alpha}{2}) < \mu_0 < \bar{X} + \frac{\sigma}{\sqrt{n}} z(1-\tfrac{\alpha}{2})$$

which is a $100(1-\alpha)\%$ confidence interval for μ_0, so a confidence interval consists precisely of the values μ_0 for which $H_0: \mu = \mu_0$ is accepted

Linear Regression

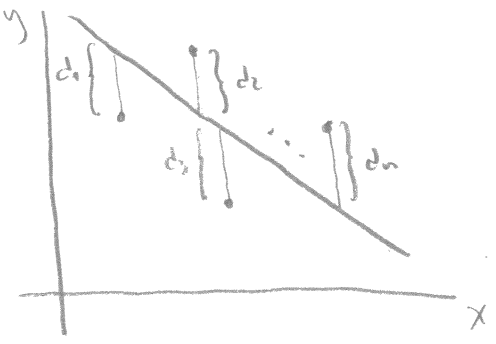
Frequently, data reveal relationship between two or more variables and it's desired to express this relationship in mathematical form. For example, we could collect data on the lifetime of bulbs as a function of the applied voltage. The data may look as in the figure



The purpose of this section is to show how to obtain the line shown in the plot which represents the best fit for the data

Consider two variables x and y, usually we consider x the independent or prediction variable and y the dependent or response variable. Consider a sample of points $(x_1, y_1), (x_2, y_2), \ldots (x_n, y_n)$. The general problem of finding a mathematical equation to represent the data is called curve fitting and the resulting curve is called regression curve. There are many types of curves that can fit the data (in fact n points can always be interpolated with a polynomial of degree $n-1$ or less, but this is not desired here since that method doesn't provide smoothing of the data). Here we will focus on linear functions, which is the reason why the technique is called linear regression. The linear regression equation is $y = \beta_0 + \beta_1 x$.

The problem is thus to find the linear form $y = \beta_0 + \beta_1 x$ that fits best our data and, of course, the problem is to define what we mean by best.



$$d_1 = y_1 - (\beta_0 + \beta_1 x_1)$$
$$d_2 = y_2 - (\beta_0 + \beta_1 x_2)$$
$$\vdots$$
$$d_n = y_2 - (\beta_0 + \beta_1 x_1)$$

One way of doing it is to find the line that minimizes the distances (difference) between the line and the data points. Since some of them are negative, we can square them to add them up and get an overall (square) difference. To get the best fit we just minimize this sum, that is,

$$d_1^2 + d_2^2 + \cdots + d_n^2 = \text{minimum}$$

This method is called least-squares and leads to a least-squares regression curve.