TESTING STATISTICAL HYPOTHESES

- hypoteheses about the parameters of a population distribution
- non-parametric hypotheses - goodness of fit test

Example 3

We choose ten people from a population and for each individual measure his or her height. We obtain the following results in cm:

178, 180, 158, 166, 190, 180, 177, 178, 182, 160

It is known that the the random variable describing the height of an individual from that particular population has a normal distribution $N(\mu, 100)$

Someone claims that the population average height amounts to 176. How could we use the above population sample top either reject or prove such a hypothesis?

Using the point estimator $\overline{X}$ we see that an estimate of the sample arithmetic mean is 174.9 cm, while according to the hypothesis, the population expectancy is supposed to be 176 cm. We can try to reconcile this difference by asking with what probability a sample arithmetic mean can be 174.9 cm provided that the population expectancy is 176 cm. We can agree beforehand that if this probability is sufficiently high, say greater than an $\alpha > 0$, we will not reject the hypothesis.

One way of solving this problem is using the confidence interval from Example 2:

$$\left[\overline{X} - \frac{\sigma}{\sqrt{n}} u_{1-\alpha/2}, \overline{X} + \frac{\sigma}{\sqrt{n}} u_{1-\alpha/2}\right]$$

which gives us an interval estimate

$$168.7 \le \mu \le 181.1$$

saying that, given that the sample arithmetic mean is 174.9, the population expectancy lies within the above interval with a probability of 1 – 0.05 which is exactly the information we need not to reject the hypothesis that

$$\mu = 176\,\text{cm}$$

Generalizing, we could say that, denoting μ_0 the hypothesis claimed, which we call the null hypothesis $H_{0,}$ we do not reject the hypothesis if

$$\bar{x} - \frac{\sigma}{\sqrt{n}} u_{1-\alpha/2} < \mu_0 < \bar{x} + \frac{\sigma}{\sqrt{n}} u_{1-\alpha/2}$$

which is equivalent to

$$\frac{\sqrt{n}\left(\bar{x} - \mu_0\right)}{\sigma} \in \left[-u_{1-\alpha/2}, u_{1-\alpha/2}\right]$$

To test a null hypothesis H_0,

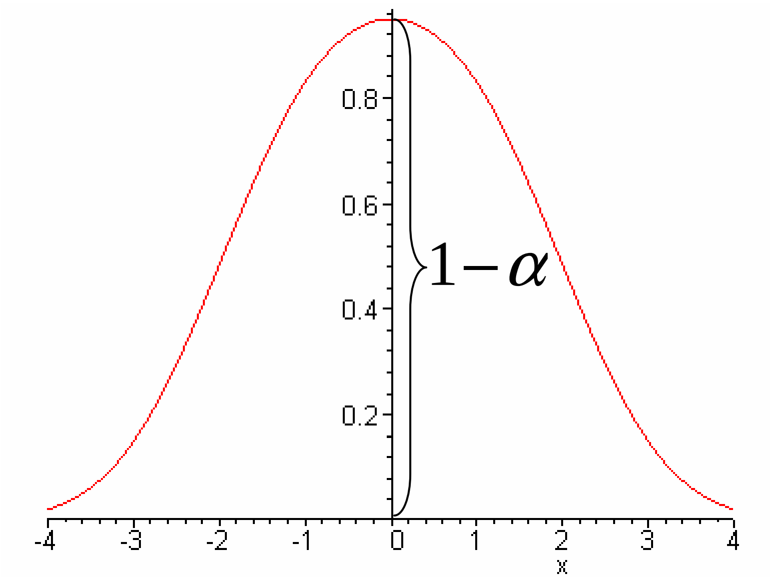
- calculate a testing statistic t
- establish a testing interval I_α
- if $t \notin I_\alpha$ reject the null hypothesis H_0 at the significance level α
- if $t \in I_\alpha$ do not reject H_0 at the significance level α

We call α the **significance level of the test**. It is the probability that we reject the null hypothesis if it is correct.

In this event we say that we have made a first type error.

The probability of the first type error α is given beforehand and is usually set at 0.05 or 0.01.

Not rejecting a hypothesis when it proves to be erroneous is an error of the second type. The probability of such an error generally is a function err(x) of the difference between the value of the null hypothesis and the real one. For instance in Example 3, $\mathrm{err}(x) = \Phi(x + u_{0.975}) - \Phi(x - u_{0.975})$



With a given α, the more "spiky" err(x) is the better the test.

Testing the null hypothesis $\mu = \mu_0$ for population distribution

$X \sim N(\mu, \sigma^2)$ where σ^2 is <u>known</u>:

$$t = \frac{\sqrt{n}(X - \mu_0)}{\sigma}$$

$$I_\alpha = \lfloor -u_{1-\alpha/2}, u_{1-\alpha/2} \rfloor$$

$u_{1-\alpha/2}$ is a quantile of the standardized normal distribution

Testing the null hypothesis $\mu = \mu_0$ for population distribution

$X \sim N\left(\mu, \sigma^2\right)$ where σ^2 is <u>unknown</u>:

$$t = \frac{\sqrt{n}\left(X - \mu_0\right)}{S}$$

$$I_\alpha = \left\lfloor -t_{1-\alpha/2}, t_{1-\alpha/2} \right\rfloor$$

$t_{1-\alpha/2}$ is a quantile of Student's t-distribution with $k = n - 1$ degrees of freedom

Testing the null hypothesis $\sigma^2 = \sigma_0^2$ for population distribution

$X \sim N(\mu, \sigma^2)$:

$$t = \frac{nS^2}{\sigma_0^2}$$

$$I_\alpha = \left[\chi_{\alpha/2}, \chi_{1-\alpha/2}\right]$$

$\chi_{\alpha/2}, \chi_{1-\alpha/2}$ are quantiles of chi squared distribution with $k = n - 1$ degrees of freedom

Testing the null hypothesis $\mu_1 - \mu_2 = \mu_0$ for population distributions

$$X_1 \sim N\left(\mu_1, \sigma^2\right) \quad X_1 \sim N\left(\mu_1, \sigma^2\right):$$

The variances as such are not known but are known to be identical.

$$t = \frac{\overline{X}_1 - \overline{X}_2 - \mu_0}{\sqrt{n_1 S_1^2 + n_2 S_2^2}} \sqrt{\frac{n_1 n_2 (n_1 + n_2 - 2)}{n_1 + n_2}}$$

$$I_\alpha = \left\lfloor -t_{1-\alpha/2}, t_{1-\alpha/2} \right\rfloor$$

$t_{1-\alpha/2}$ is a quantile of Student's t-distribution with $k = n_1 + n_2 - 2$ degrees of freedom

Testing the null hypothesis $\mu_1 - \mu_2 = \mu_0$ for population distributions

$$X_1 \sim N\left(\mu_1, \sigma_1^2\right) \qquad X_2 \sim N\left(\mu_2, \sigma_2^2\right)$$

The variances as such are not known but are known to be different.

$$t = \frac{\overline{X}_1 - \overline{X}_2 - \mu_0}{\sqrt{\dfrac{S_1^2}{n_1} + \dfrac{S_2^2}{n_2}}} \qquad \tau_{1-\alpha/2} = \frac{c_1 t^1_{1-\alpha/2} + c_2 t^2_{1-\alpha/2}}{c_1 + c_2}$$

$$I_\alpha = \left\lfloor -\tau_{1-\alpha/2}, \tau_{1-\alpha/2} \right\rfloor \qquad c_1 = \frac{S_1^2}{n_1 - 1} \quad c_2 = \frac{S_2^2}{n_2 - 1}$$

$t^1_{1-\alpha/2}$ is a quantile of Student's t-distribution with $k = n_1 - 1$ degrees of freedom

$t^2_{1-\alpha/2}$ is a quantile of Student's t-distribution with $k = n_2 - 1$ degrees of freedom

Testing the null hypothesis $\sigma_1^2 = \sigma_2^2$ for population distributions

$$X_1 \sim N\left(\mu_1, \sigma_1^2\right) \qquad X_2 \sim N\left(\mu_2, \sigma_2^2\right)$$

$$t = \frac{S_1^2 n_1 (n_2 - 1)}{S_2^2 n_2 (n_1 - 1)}$$

if $t < 0$, we must swap X_1 and X_2

$$I_\alpha = \left[0, F_{1-\alpha}\right]$$

$F_{1-\alpha}$ is a quantile of F-distribution with $k_1 = n_1 - 1$ and $k_2 = n_2 - 1$ degrees of freedom

Goodness-of-fit test (chi-squared test)

We have a frequency distribution of a random variable X over a population and want to test the hypothesis that X has a distribution given by a probability distribution $F(x)$, probability function $p(x)$ or density $f(x)$.

x	r_1	r_2	...	r_k
f	f_1	f_2	...	f_k

We can use the following test, which is called a goodness-of-fit test or chi-squared test.

Using the distribution, probability function, or density, we calculate ideal frequencies $\overline{f}_1, \overline{f}_2, \dots, \overline{f}_k$

With these frequencies we calculate a testing statistic

$$t = \sum_{i=1}^{k} \frac{\left(f_i - \overline{f}_i\right)^2}{\overline{f}_i} = \sum_{i=1}^{k} \frac{f_i^{\,2}}{\overline{f}_i} - n$$

Then we establish the testing interval $I_\alpha = \left[0, \chi^2_{1-\alpha}\right]$

$\chi^2_{1-\alpha}$ is a quantile of chi-squared distribution where the number k of degrees of freedom is calculated as follows

$$k = k - 1 - q$$

where q is the number of distribution parameters we had to replace by their estimates when calculating the ideal frequencies

Use the below sample to test at a 0.05 significance level whether the population from which the random sample was taken has a normal distribution $N(\mu, \sigma^2)$

5,854	6,598	5,791	6,207	8,054	5,081	6,195	7,234	7,545	7,425
7,357	6,703	6,475	4,852	6,319	8,081	6,662	5,301	4,717	5,990
7,650	7,856	4,083	6,507	7,423	5,564	6,262	8,685	5,862	6,352
6,954	5,865	6,747	7,287	7,772	5,669	6,240	6,512	6,420	7,105
4,898	4,105	9,074	6,692	8,191	8,080	7,766	7,819	6,071	5,856

Maximum data item in sample: 9.074

Minimum data item in sample: 4.083

Range: 4 - 10

Number of class intervals: 6

Frequency table

FROM	TO	REP	FREQ
4	5	4.5	5
5	6	5.5	10
6	7	6.5	17
7	8	7.5	12
8	9	8.5	5
9	10	9.5	1

Sample arithmetic mean: 6.6

Sample standard deviation - best point estimator: 1.199

Since the last two frequencies in the table are not greater than five, we have to amalgamate the last two table entries.

Modified frequency table

FROM	TO	REP	FREQ
4	5	4.5	5
5	6	5.5	10
6	7	6.5	17
7	8	7.5	12
8	10	9	6

Calculating expected frequencies $\overline{f}$

x	$\frac{x-\overline{x}}{s'}$	$\Phi\left(\frac{x-\overline{x}}{s'}\right)$	Δ	$\Delta \cdot n = \overline{f}$
$-\infty$	$-\infty$	0	0.091	5.487
5	– 1.334	0.091	0.217	9.202
6	– 0.500	0.308	0.322	17.942
7	0.333	0.631	0.248	11.622
8	1.167	0.878	0.122	5.922
∞	∞	1.000		

Testing statistic

$$t = \sum_{i=1}^{k} \frac{f_i^2}{\overline{f}_i} - n = 0.175$$

Number of degrees of freedom

$$k = \#\text{class intervals} - 1 - \#\text{parameters estimated} = 5 - 1 - 2 = 2$$

Testing interval

$$I_{0.05} = \left[0, \chi^2_{0.95}\right] = 5.991$$

Conclusion: Since $t \in I_\alpha$ we do not reject the hypothesis that the population has a normal distribution.