Lecture 11

4.3 Significance Testing, *P*-Values

Recall that we test the *null* hypothesis, always simple, i.e.

$$H_0 : \theta = \theta_0,$$

against one of the *alternative* hypotheses

$$\begin{aligned} H_1 : &\quad \theta < \theta_0 \ \text{(left-tailed test)}, \\ H_1 : &\quad \theta > \theta_0 \ \text{(right-tailed test)}, \\ H_1 : &\quad \theta \neq \theta_0 \ \text{(two-tailed test)}. \end{aligned} \tag{4.1}$$

There is a problem that might occur in hypothesis testing: We preset α, the probability of a type I error, and henceforth determine a rejection region. We get a value of the test statistic that *does not belong* to it, so we cannot reject the null hypothesis H_0, i.e. we accept it as being true.
However, when we compute the probability of getting that value of the test statistic under the assumption that H_0 is true, we find it is very small, comparable with our preset α. So, we accept H_0, yet considering it to be true, we find that it is very unlikely (very improbable) that the test statistic takes the observed value we found for it.

That makes us wonder if we set our RR right and if we didn't "accept" H_0 too easily, by hastily dismissing values of the test statistic that did not fall into our RR. So we should take a look at how "far-fetched" does the value of the test statistic seem, under the assumption that H_0 is true. If it seems really implausible to occur by chance, i.e. if its probability is *small*, then maybe we should reject the null hypothesis H_0.

To avoid this situation, we perform what is called a **significance test**: for a given random sample (i.e. sample variables $X_1, \ldots, X_n$), we still set up H_0 and H_1 as before and we choose an appropriate test statistic.

Then, we compute the probability of observing a value at least as extreme (in the sense of the test conducted) of the test statistic TS as the value observed from the sample, TS_0, under the assumption that H_0 is true.

This probability is called the critical value, the descriptive significance level, the probability of the test, or, simply the ***P*-value** of the test. If it is small, we reject H_0, otherwise we do not reject it.

The P-value is a numerical value assigned to the test, it depends only on the sample data and its distribution, but *not* on α.

In general, for the three alternatives (4.1), if TS_0 is the value of the test statistic TS under the assumption that H_0 is true and F is the cdf of TS, the P-value is computed by

$$P = \begin{cases} P(TS < TS_0|H_0) & = F(TS_0) \\ P(TS > TS_0|H_0) & = 1 - F(TS_0) \\ 2\min\{P(TS < TS_0|H_0), P(TS > TS_0|H_0)\} & = 2\min\{F(TS_0), 1 - F(TS_0)\} \end{cases} \tag{4.2}$$

Then the decision will be

$$\begin{array}{l} \text{if } \ P \leq \alpha, \ \text{reject } \ H_0, \\ \text{if } \ P > \alpha, \ \text{do not reject } \ H_0. \end{array} \tag{4.3}$$

So, more precisely, the P-value of a test is the smallest level at which we could have preset α and still have been able to reject H_0, or the lowest significance level that *forces* rejection of H_0, i.e. the minimum rejection level.

Remark 4.1.

1. Thus, we can avoid the costly computation of the rejection region (costly because of the quantiles) and compute the P-value instead. Then, we simply compare it to the significance level α. If α is above the P-value, we reject H_0, but if it is below that minimum rejection level, we can no longer reject the null hypothesis.

2. Hypothesis testing (determining the rejection region) and significance testing (computing the P-value) are two methods for testing the same thing (the same two hypotheses), so, of course, the outcome (the decision of rejecting or not H_0) will be the same, for the same data.

Example 4.2.

Recall the problem in Example 4.4 (Lecture 10): *The number of monthly sales at a firm is known to have a mean of* 20 *and a standard deviation of* 4 *and all salary, tax and bonus figures are based on these values. However, in times of economical recession, a sales manager fears that his employees do not average* 20 *sales per month, but less, which could seriously hurt the company. For a number of* 36 *randomly selected salespeople, it was found that in one month they averaged* 19 *sales. At the* 5% *significance level, does the data confirm or contradict the manager's suspicion?*

Now, let us perform a significance test.

Solution.

We tested a left-tailed alternative for the mean

$$\begin{aligned} H_0: &\quad \mu = 20 \\ H_1: &\quad \mu < 20. \end{aligned}$$

The population standard deviation was given, $\sigma = 4$, and for a sample of size $n = 36$, the sample mean was $\overline{X} = 19$.

For the test statistic

$$Z = \frac{\overline{X} - \mu}{\sigma/\sqrt{n}} \in N(0,1),$$

the observed value was

$$Z_0 = \frac{\overline{X} - \mu_0}{\sigma/\sqrt{n}} = \frac{19-20}{4/6} = -1.5.$$

Now, we compute the P-value

$$P = P(Z < Z_0) = P(Z < -1.5) = 0.0668.$$

Since

$$\alpha = 0.05 < 0.0668 = P,$$

(is below the minimum rejection level), we do not reject H_0, so, at the 5% significance level, we conclude that the data contradicts the manager's suspicion.
But, for example, at the 7% significance level, we would have rejected it.

4.4 Tests for the Parameters of One Population

Let X be a population characteristic, with pdf $f(x;\theta)$, mean $E(X) = \mu$ and variance $V(X) = \sigma^2$. Let $X_1, X_2, \ldots, X_n$ be sample variables.

Tests for the mean of a population, $\theta = \mu$

We test the hypotheses

$$H_0: \quad \mu = \mu_0, \text{ versus one of the alternatives}$$

$$H_1: \begin{cases} \mu < \mu_0 \\ \mu > \mu_0 \\ \mu \neq \mu_0, \end{cases} \tag{4.4}$$

under the assumption that either X is approximately Normally $N(\mu, \sigma)$ distributed or that the sample is large ($n > 30$).

Case σ known (ztest)

We use the test statistic

$$TS = Z = \frac{\overline{X} - \mu}{\frac{\sigma}{\sqrt{n}}} \in N(0,1), \tag{4.5}$$

with observed value

$$Z_0 = \frac{\overline{X} - \mu_0}{\frac{\sigma}{\sqrt{n}}}. \tag{4.6}$$

Then, as before, at the $\alpha \in (0,1)$ significance level, the rejection region for each test will be given by

$$RR: \begin{cases} \{Z_0 \leq z_\alpha\} \\ \{Z_0 \geq z_{1-\alpha}\} \\ \{|Z_0| \geq z_{1-\frac{\alpha}{2}}\} \end{cases} \tag{4.7}$$

and the P-value will be computed as

$$P = \begin{cases} P(Z \leq Z_0 \mid H_0) & = & \Phi(Z_0) \\ P(Z \geq Z_0 \mid H_0) & = & 1 - \Phi(Z_0) \\ P(|Z| \geq |Z_0| \mid H_0) & = & 2\left(1 - \Phi(|Z_0|)\right), \end{cases} \tag{4.8}$$

since $N(0,1)$ is symmetric, where

$$\Phi(x) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{x} e^{-\frac{t^2}{2}} dt$$

is Laplace's function, the cdf for the Standard Normal $N(0,1)$ distribution.

Case σ unknown (**ttest**)

In this case, we use the test statistic

$$TS = T = \frac{\overline{X} - \mu}{\frac{s}{\sqrt{n}}} \in T(n-1), \tag{4.9}$$

with observed value

$$T_0 = \frac{\overline{X} - \mu_0}{\frac{s}{\sqrt{n}}}. \tag{4.10}$$

Similarly to the previous case, we find the rejection region for the three alternatives as

$$RR: \begin{cases} \{T_0 \leq t_\alpha\} \\ \{T_0 \geq t_{1-\alpha}\} \\ \{|T_0| \geq t_{1-\frac{\alpha}{2}}\}, \end{cases} \tag{4.11}$$

and compute the P-value by

$$P = \begin{cases} P(T \leq T_0 \mid H_0) &= F(T_0) \\ P(T \geq T_0 \mid H_0) &= 1 - F(T_0) \\ P(|T| \geq |T_0| \mid H_0) &= 2\left(1 - F(|T_0|)\right), \end{cases} \tag{4.12}$$

where the cdf F and the quantiles refer to the $T(n-1)$ distribution.

Tests for the variance of a population, $\theta = \sigma^2$ (vartest)

Assuming that X has a Normal $N(\mu, \sigma)$ distribution, we test the hypotheses

$$H_0: \quad \sigma^2 = \sigma_0^2, \qquad H_1: \begin{cases} \sigma^2 < \sigma_0^2 \\ \sigma^2 > \sigma_0^2 \\ \sigma^2 \neq \sigma_0^2, \end{cases} \quad \text{equivalent to} \quad H_0: \quad \sigma = \sigma_0, \qquad H_1: \begin{cases} \sigma < \sigma_0 \\ \sigma > \sigma_0 \\ \sigma \neq \sigma_0. \end{cases} \tag{4.13}$$

The test statistic will be

$$TS \quad = \quad V \quad = \quad \frac{(n-1)s^2}{\sigma^2} \quad \in \quad \chi^2(n-1), \tag{4.14}$$

with observed value

$$V_0 \quad = \quad \frac{(n-1)s^2}{\sigma_0^2}. \tag{4.15}$$

Even though the $\chi^2(n-1)$ distribution is not symmetric, we use the same line of reasoning and computations to find the rejection region for the three alternatives:

$$RR: \quad \begin{cases} \{V_0 \leq \chi^2_{\alpha}\} \\ \{V_0 \geq \chi^2_{1-\alpha}\} \\ \{V_0 \leq \chi^2_{\frac{\alpha}{2}} \text{ or } V_0 \geq \chi^2_{1-\frac{\alpha}{2}}\}. \end{cases} \tag{4.16}$$

Same goes for the computation of the P-values:

$$P = \begin{cases} P(V \leq V_0 \mid H_0) & = F(V_0) \\ P(V \geq V_0 \mid H_0) & = 1 - F(V_0) \\ 2 \cdot \min\{P(V \leq V_0 \mid H_0), P(V \geq V_0 \mid H_0)\} & = 2 \cdot \min\{F(V_0), 1 - F(V_0)\}, \end{cases} \tag{4.17}$$

where the cdf F and the quantiles refer to the $\chi^2(n-1)$ distribution.

Example 4.3.

Let us consider again the problem in Example 4.2: The number of monthly sales at a firm is known to have a mean of 20 and a standard deviation of 4 and all salary, tax and bonus figures are based on these values. Suppose that for the sample considered (of 36 randomly selected salespeople), the standard deviation is found to be $s = 4.5$. Assuming that the number of monthly sales at that firm is Normally distributed, at the 5% significance level, does the assumption on σ seem to be correct?

Solution.

We are now testing the variance. We want to know if the value $\sigma = 4$ is correct or not, so, this will be a two-tailed test.

$$\begin{aligned} H_0 : &\quad \sigma = 4 \\ H_1 : &\quad \sigma \neq 4, \end{aligned}$$

i.e.,

$$\begin{aligned} H_0 : &\quad \sigma^2 = 16 = \sigma_0^2 \\ H_1 : &\quad \sigma^2 \neq 16 = \sigma_0^2. \end{aligned}$$

We have $n = 36$ and $s^2 = (4.5)^2 = 20.25$. The observed value of the test statistic is

$$V_0 = \frac{(n-1)s^2}{\sigma_0^2} = \frac{35 \cdot 20.25}{16} = 44.2969.$$

The significance level is $\alpha = 0.05$ and the two quantiles for the $\chi^2(35)$ distribution are

$$\chi^2_{0.025} = 20.5694, \quad \chi^2_{0.975} = 53.2033.$$

Then the rejection region is

$$RR = (-\infty, 20.5694] \cup [53.2033, \infty),$$

which does not include the value V_0. Therefore, the decision is to not reject the null hypothesis, i.e. to conclude that the assumption $\sigma = 4$ is correct.

On the other hand, the P-value is

$$P \;=\; 2\cdot\min\{P(V \le V_0), P(V \ge V_0)\} \;=\; 2\cdot\min\{0.8652, 0.1348\} \;=\; 0.2697.$$

Since

$$\alpha \;=\; 0.05 \;<\; 0.2697 \;=\; P,$$

the decision is to not reject the null hypothesis.

Notice that the significance test tells us more! Since the P-value is so large (remember, it is comparable to a probability of an *error*, so a *small* quantity), not only at the 5% significance level we decide to accept H_0, but at any reasonable significance level, the decision would be the same. That means that the data strongly suggests that H_0 is true and should not be rejected. Even though the *sample* standard deviation *is not* equal to 4, still, statistically, the data strongly suggests that the *population* standard deviation *is* 4.
We should be careful not to extrapolate the property of one sample to the entire population (data from a sample may be misleading, if it is not used properly ...)

4.5 Tests for Comparing the Parameters of Two Populations

Assume we have two population characteristics $X_{(1)}$ and $X_{(2)}$, with means and variances $E(X_{(1)}) = \mu_1, V(X_{(1)}) = \sigma_1^2$ and $E(X_{(2)}) = \mu_2, V(X_{(2)}) = \sigma_2^2$, respectively. We draw two independent random samples

$$X_{11}, \ldots, X_{1n_1} \quad \text{and} \quad X_{21}, \ldots, X_{2n_2},$$

with sample means

$$\overline{X}_1 = \frac{1}{n_1} \sum_{i=1}^{n_1} X_{1i}, \quad \overline{X}_2 = \frac{1}{n_2} \sum_{j=1}^{n_2} X_{2j}$$

and sample variances

$$s_1^2 = \frac{1}{n_1 - 1} \sum_{i=1}^{n_1} \left(X_{1i} - \overline{X}_1\right)^2, \quad s_2^2 = \frac{1}{n_2 - 1} \sum_{j=1}^{n_2} \left(X_{2j} - \overline{X}_2\right)^2,$$

respectively.

In addition, we have

$$s_p^2 = \frac{\sum_{i=1}^{n_1} \left(X_{1i} - \overline{X}_1\right)^2 + \sum_{j=1}^{n_2} \left(X_{2j} - \overline{X}_2\right)^2}{n_1 + n_2 - 2} = \frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{n_1 + n_2 - 2}$$

the pooled variance of the two samples, a variance that considers the sample data from both samples.

Recall that when comparing the means of two populations, we estimate their difference and when comparing the variances, we estimate their ratio.

The formulas for testing hypotheses for the difference of means $\mu_1 - \mu_2$ and for the ratio of variances $\frac{\sigma_1^2}{\sigma_2^2}$ are based on the same results (which follow either from properties of random variables, or are the consequence of some CLT) that were used for finding confidence intervals (see Propositions 3.1 and 3.2 in Lecture 10).

Tests for the difference of means, $\theta = \mu_1 - \mu_2$

We test the hypotheses

$$H_0: \quad \mu_1 - \mu_2 = 0, \qquad H_1: \quad \begin{cases} \mu_1 - \mu_2 < 0 \\ \mu_1 - \mu_2 > 0 \\ \mu_1 - \mu_2 \neq 0, \end{cases} \qquad \text{equivalent to} \qquad \begin{aligned} H_0: &\quad \mu_1 = \mu_2, \\ H_1: &\quad \begin{cases} \mu_1 < \mu_2 \\ \mu_1 > \mu_2 \\ \mu_1 \neq \mu_2, \end{cases} \end{aligned} \tag{4.18}$$

under the assumption that either $X_{(1)}$ and $X_{(2)}$ have approximately Normal distributions or that the samples are large enough ($n_1 + n_2 > 40$).

Case σ_1, σ_2 known

We use the test statistic

$$TS = Z = \frac{(\overline{X}_1 - \overline{X}_2) - (\mu_1 - \mu_2)}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}} \in N(0,1), \tag{4.19}$$

with observed value

$$Z_0 = \frac{\overline{X}_1 - \overline{X}_2}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}}. \tag{4.20}$$

The rejection regions and P-values for the three alternatives have the same form as the ones for the mean (case σ known), with Z_0 from (4.20).

Case $\sigma_1 = \sigma_2$ unknown (**ttest2**)

The test statistic is

$$TS = T = \frac{(\overline{X}_1 - \overline{X}_2) - (\mu_1 - \mu_2)}{s_p \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} \in T(n_1 + n_2 - 2), \quad (4.21)$$

with observed value

$$T_0 = \frac{\overline{X}_1 - \overline{X}_2}{s_p \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}. \quad (4.22)$$

The rejection regions and P-values for the three alternatives have the same form as the ones for the mean (case σ unknown), where T_0 is given in (4.22) and the cdf F and the quantiles refer to the $T(n_1 + n_2 - 2)$ distribution.

Case σ_1, σ_2 unknown (ttest2)

We now use the test statistic

$$TS = T^* = \frac{(\overline{X}_1 - \overline{X}_2) - (\mu_1 - \mu_2)}{\sqrt{s_1^2/n_1 + s_2^2/n_2}} \in T(n), \tag{4.23}$$

where $\frac{1}{n} = \frac{c^2}{n_1 - 1} + \frac{(1-c)^2}{n_2 - 1}$ and $c = \frac{\frac{s_1^2}{n_1}}{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}$.

The observed value of the test statistic is

$$T_0^* = \frac{\overline{X}_1 - \overline{X}_2}{\sqrt{s_1^2/n_1 + s_2^2/n_2}}. \tag{4.24}$$

The rejection regions and P-values for the three alternatives are the same as in the previous case, with T_0 replaced by T_0^* from (4.24). The cdf F and the quantiles refer to the $T(n)$ distribution.

Remark 4.4.

The same Matlab command **ttest2** performs a T-test for the difference of two population means, when the variances are not assumed equal, with the option *vartype* set on "unequal" (the default being "equal", when it can be omitted).

Tests for the ratio of variances, $\theta = \dfrac{\sigma_1^2}{\sigma_2^2}$ (**vartest2**)

Assuming that both $X_{(1)}$ and $X_{(1)}$ have Normal distributions, we test the hypotheses

$$H_0: \ \sigma_1^2/\sigma_2^2 = 1, \qquad H_0: \ \sigma_1^2 = \sigma_2^2, \qquad H_0: \ \sigma_1 = \sigma_2,$$
$$H_1: \begin{cases} \sigma_1^2/\sigma_2^2 < 1 \\ \sigma_1^2/\sigma_2^2 > 1 \\ \sigma_1^2/\sigma_2^2 \neq 1, \end{cases} <=> H_1: \begin{cases} \sigma_1^2 < \sigma_2^2 \\ \sigma_1^2 > \sigma_2^2 \\ \sigma_1^2 \neq \sigma_2^2, \end{cases} <=> H_1: \begin{cases} \sigma_1 < \sigma_2 \\ \sigma_1 > \sigma_2 \\ \sigma_1 \neq \sigma_2. \end{cases} \tag{4.25}$$

The test statistic used is

$$TS \quad = \quad F \quad = \quad \frac{s_1^2/\sigma_1^2}{s_2^2/\sigma_2^2} \quad \in \quad F(n_1-1, n_2-1), \tag{4.26}$$

with observed value

$$F_0 \quad = \quad s_1^2/s_2^2. \tag{4.27}$$

The $F(n_1-1, n_2-1)$ distribution is not symmetric, but as before, we find

$$RR: \quad \begin{cases} \{F_0 \leq f_\alpha\} \\ \{F_0 \geq f_{1-\alpha}\} \\ \{F_0 \leq f_{\frac{\alpha}{2}} \text{ or } F_0 \geq f_{1-\frac{\alpha}{2}}\} \end{cases}, \tag{4.28}$$

$$P = \begin{cases} P(F \leq F_0 \mid H_0) & = F(F_0) \\ P(F \geq F_0 \mid H_0) & = 1 - F(F_0) \\ 2 \cdot \min\{P(F \leq F_0 \mid H_0), P(F \geq F_0 \mid H_0)\} & = 2 \cdot \min\{F(F_0), 1 - F(F_0)\}, \end{cases} \tag{4.29}$$

where the cdf F and the quantiles refer to the $F(n_1-1, n_2-1)$ distribution.

Example 4.5.

Suppose the strengths to a certain load of two types of material, $M1$ and $M2$, are studied, knowing that they are approximately Normally distributed. The more weight they can resist to, the stronger they are. Two independent random samples are drawn and they yield the following data.

$M1$			$M2$		
n_1	$=$	25	n_2	$=$	16
$\overline{X}_1$	$=$	380	$\overline{X}_2$	$=$	370
s_1^2	$=$	537	s_2^2	$=$	196

a) At the 5% significance level, do the variances of the two populations seem to be equal or not?
b) At the same significance level, does the data suggest that on average, $M1$ is stronger than $M2$?
(In both parts, perform both hypothesis and significance testing).

Solution.

a) First, we compare the variances of the two populations, so we know which way to proceed for comparing the means. We want to know if they are equal or not, so it is a two-tailed test. Hence, our hypotheses are

$$\begin{aligned} H_0 &: \quad \sigma_1^2 = \sigma_2^2 \\ H_1 &: \quad \sigma_1^2 \neq \sigma_2^2. \end{aligned}$$

The observed value of the test statistic is

$$F_0 = \frac{s_1^2}{s_2^2} = \frac{537}{196} = 2.7398.$$

For $\alpha = 0.05$, $n_1 = 25$ and $n_2 = 16$, the quantiles for the $F(24, 15)$ distribution are

$$f_{\frac{\alpha}{2}} = f_{0.025} = 0.4103, \quad f_{1-\frac{\alpha}{2}} = f_{0.975} = 2.7006.$$

Thus, the rejection region for our test is

$$RR = (-\infty, 0.4103] \cup [2.7006, \infty)$$

and clearly, $F_0 \in RR$. Thus we reject H_0 in favor of H_1, i.e. we conclude that the data suggests that the population variances are *different*.

Let us also perform a significance test.

The P-value of this (two-tailed) test is

$$\begin{aligned} P &= 2 \cdot \min\{P(F \leq F_0), P(F \geq F_0)\} \\ &= 2 \cdot \min\{0.9765, 0.0235\} = 0.0469. \end{aligned}$$

Since our $\alpha > P$, the "minimum rejection significance level", we reject H_0.

Note. We now know that for instance, at 1% significance level (or any level less than 4.69%), we would have *not* rejected the null hypothesis.
This goes to show that the data can be "misleading". Simply comparing the values of the sample functions does not necessarily mean that the same thing will be true for the corresponding population parameters. Here, s_1^2 is much larger than s_2^2, yet at 1% significance level, we would have concluded that the population variances seem to be equal.

b) Next we want to compare the population means. If $M1$ is to be *stronger* than $M2$ on average, than we must perform a right-tailed test:

$$\begin{aligned} H_0 : &\quad \mu_1 = \mu_2 \\ H_1 : &\quad \mu_1 > \mu_2 \end{aligned}$$

Which one of the tests for the difference of means should we use? The answer is in part a). At this significance level, the variances are unknown and different.

Then the value of the test statistic is, by (4.24)

$$T_0^* = \frac{\overline{X}_1 - \overline{X}_2}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}} = \frac{380 - 370}{\sqrt{\frac{537}{25} + \frac{196}{16}}} = 1.7218.$$

To find the rejection region, we compute

$$c = 0.6368, \quad n = 38.9244 \approx 39$$

and the quantile for the $T(39)$ distribution

$$t_{1-\alpha} = t_{0.95} = 1.6849.$$

Then the rejection region of the test is

$$RR = [1.6849, \infty),$$

which includes the value T_0^*, so we reject H_0 in favor of H_1. Hence, we conclude that yes, the data suggests that material $M1$ is, on average, stronger than material $M2$.

On the other hand, the P-value of this test is

$$P = P(T^* \geq T_0^*) = 1 - F(T_0^*) = 1 - F(1.7218) = 0.0465,$$

where F is the cdf of the $T(39)$ distribution. Again, the P-value is lower than $\alpha = 0.05$, which forces the rejection of H_0.

■

Remark 4.6.

1. As mentioned before, both hypothesis and significance testing lead to the same conclusion. From the implementation point of view, significance testing is more efficient, since it avoids the inversion of a cdf (i.e. computation of quantiles), which is often a complicated improper integral. This is the reason why, although the main tests are implemented in Matlab, the rejection regions are not computed.

2. Many tests (and formulas for CI's) work under the assumption of Normality of the population from which the sample was drawn. In practice, when there are outliers in the data, that is rarely the case. How important is this assumption of Normality and how affected are the results of these tests by small departures from model assumptions?
Z-tests and T-tests work well even when the underlying population is not quite Normally distributed. From this point of view, they are called **robust** tests. χ^2-tests and F-tests, however, are not robust, they perform very poorly when the assumption of Normality is breached. In modern Statistics there is an ongoing search for finding robust methods of estimation for variances.

4.6 Summary of hypothesis and significance testing

We can use data to verify statements and test hypotheses. Essentially, we measure the evidence provided by the data against the null hypothesis H_0. Then we decide whether it is sufficient for rejecting it or not. Given a significance level $\alpha \in (0, 1)$, we can construct acceptance and rejection regions, compute a suitable test statistic, and make a decision depending on which region it belongs to.

Alternatively, we may compute a P-value of the test. It shows how *significant* the evidence against H_0 is. Low P-values suggest rejection of the null hypothesis. The P-value of a test is the boundary between levels α-to-reject and α-to-accept. It also represents the probability of observing the same or more extreme sample than the one that was actually observed.

We already mentioned that in practice, significance testing is preferred, i.e., computing the P-value and comparing it to the significance level α (and that is how hypothesis testing is implemented in any software). That is much more efficient from the computational perspective, as computation of the quantiles can be rather expensive.

In fact, in practice, a significance level α is hardly ever specified. Instead, just the P-value is computed.

Since the null hypothesis is *always* in the form of an equality

$$H_0 : \theta = \theta_0,$$

whichever alternative we are testing (left-, right-, or two-tailed), to reject H_0 (when P is "small") means that the data shows that there are significant differences (statistically speaking) from what it states. How "significant"? That depends on how small the P-value is. The following levels are customary for how "significant" the differences are:

$$\begin{aligned} P &> 0.05 &&\Rightarrow && \textbf{not significant}, \\ 0.01 < P &\leq 0.05 &&\Rightarrow && \textbf{(moderately) significant}, \\ 0.001 < P &\leq 0.01 &&\Rightarrow && \textbf{distinctly significant}, \\ P &\leq 0.001 &&\Rightarrow && \textbf{very significant}. \end{aligned}$$