4.3 Significance Testing, P-Values

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 after all.

To avoid this situation, we perform what is called a **significance test**: for a given random sample $(X_1, \ldots, X_n)$, 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 (left-, right- and two-tailed) alternatives, 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

$$P = \begin{cases} P(TS \leq TS_0 \mid H_0) & = & F(TS_0) \\ P(TS \geq TS_0 \mid H_0) & = & 1 - F(TS_0) \\ 2 \cdot \min\{P(TS \leq TS_0 \mid H_0), P(TS \geq TS_0 \mid H_0)\} & = & 2 \cdot \min\{F(TS_0), 1 - F(TS_0)\}. \end{cases} \tag{4.1}$$

Then the decision will be

$$\begin{aligned} &\text{if } \ P \leq \alpha, \text{ reject } \ H_0, \\ &\text{if } \ P > \alpha, \text{ do not reject } \ H_0. \end{aligned} \tag{4.2}$$

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. Significance testing is preferable to hypothesis testing, especially from the computer implementation point of view, since it avoids the inversion of a cdf, which is, oftenly, a complicated improper integral.

Example 4.2. Recall the problem in Example 4.4 from last time: *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

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

the observed value was

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

Now, we compute the P-value

$$P \;=\; P(Z \leq Z_0) \;=\; P(Z \leq -1.5) \;=\; 0.0668.$$

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 (the *same* conclusion as last time).

■

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

$$\begin{aligned} H_0 :&\quad \mu = \mu_0, \text{ versus one of} \\ H_1 :&\quad \begin{cases} \mu < \mu_0 \\ \mu > \mu_0 \\ \mu \neq \mu_0, \end{cases} \end{aligned} \tag{4.3}$$

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

$$TS \;=\; Z \;=\; \frac{\overline{X} - \mu}{\frac{\sigma}{\sqrt{n}}} \;\in\; N(0,1), \tag{4.4}$$

with observed value

$$Z_0 \;=\; \frac{\overline{X} - \mu_0}{\frac{\sigma}{\sqrt{n}}}. \tag{4.5}$$

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

by

$$RR: \begin{cases} \{Z_0 \leq z_\alpha\} \\ \{Z_0 \geq z_{1-\alpha}\} \\ \{|Z_0| \geq z_{1-\frac{\alpha}{2}}\} \end{cases} \tag{4.6}$$

and the P-value will be computed as

$$P = \begin{cases} P(Z \leq Z_0 \mid H_0) & = & \Phi(Z_0) \\ P(Z \geq Z_0 \mid H_0) & = & 1 - \Phi(Z_0) \\ P(|Z| \geq |Z_0| \mid H_0) & = & 2\left(1 - \Phi(|Z_0|)\right), \end{cases} \tag{4.7}$$

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

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

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

$$TS \;\; = \;\; T \;\; = \;\; \frac{\overline{X} - \mu}{\frac{s}{\sqrt{n}}} \;\; \in \;\; T(n-1), \tag{4.8}$$

with observed value

$$T_0 \;\; = \;\; \frac{\overline{X} - \mu_0}{\frac{s}{\sqrt{n}}}. \tag{4.9}$$

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

$$RR: \begin{cases} \{T_0 \leq t_\alpha\} \\ \{T_0 \geq t_{1-\alpha}\} \\ \{|T_0| \geq t_{1-\frac{\alpha}{2}}\}, \end{cases} \tag{4.10}$$

and compute the P-value by

$$P = \begin{cases} P(T \leq T_0 \mid H_0) & = & F(T_0) \\ P(T \geq T_0 \mid H_0) & = & 1 - F(T_0) \\ P(|T| \geq |T_0| \mid H_0) & = & 2\left(1 - F(|T_0|)\right), \end{cases} \tag{4.11}$$

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

Tests for a population proportion, $\theta = p$

Let us recall that, when estimating a population proportion p, if the sample size is large enough ($n > 30$), then the variable

$$Z \;=\; \frac{\overline{p} - p}{\sqrt{\dfrac{p(1-p)}{n}}} \tag{4.12}$$

has an approximately $N(0,1)$ distribution, where $\overline{p}$ is the sample proportion. So this case fits the general Z-test framework.
To test

$$H_0: \;\; p = p_0,$$

with one of the alternatives

$$H_1: \begin{cases} p < p_0 \\ p > p_0 \\ p \neq p_0. \end{cases}, \tag{4.13}$$

we use the test statistic $TS = Z$ from (4.12). Then, as before, at the $\alpha \in (0,1)$ significance level, the rejection region for each test will be given by

$$RR: \begin{cases} \{Z_0 \leq z_\alpha\} \\ \{Z_0 \geq z_{1-\alpha}\} \\ \{|Z_0| \geq z_{1-\frac{\alpha}{2}}\}, \end{cases} \tag{4.14}$$

and the P-value will be computed as

$$P = \begin{cases} P(Z \leq Z_0 \mid H_0) & = & \Phi(Z_0) \\ P(Z \geq Z_0 \mid H_0) & = & 1 - \Phi(Z_0) \\ P(|Z| \geq |Z_0| \mid H_0) & = & 2\left(1 - \Phi(|Z_0|)\right) \end{cases}. \tag{4.15}$$

Example 4.3. A company is receiving a large shipment of items. For quality control purposes, they collect a sample of 200 items and find 24 defective ones in it.
a) The manufacturer claims that at most 1 in 10 items in the shipment is defective. At the 5% significance level, does the data confirm or contradict his claim?
b) Find the P-value of the test in part a).

Solution.
We have a sample of size $n = 200$ for which the sample proportion is

$$\overline{p} \;=\; \frac{24}{200} \;=\; \frac{3}{25} \;=\; 0.12.$$

a) The manufacturer claims that *at most* 1 in 10 items is defective, i.e. that $p \leq 0.1$. So, we are testing a *right*- tailed alternative

$$\begin{aligned} H_0 \;&:\; p = 0.1 \\ H_1 \;&:\; p > 0.1. \end{aligned}$$

If we decide to reject H_0, that means the data *contradicts* the manufacturer's claim, whereas if we do not reject it, it means the data is insufficient to contradict his claim, so we consider it to be true. We have a significance level $\alpha = 0.05$, so for the rejection region we need the quantile

$$z_{1-\alpha} \;=\; z_{0.95} \;=\; 1.645$$

and the rejection region is

$$RR \;=\; [1.645, \infty).$$

The test statistic is

$$Z \;=\; \frac{\overline{p} - p}{\sqrt{\dfrac{p(1-p)}{n}}}$$

and its observed value is

$$Z_0 \;=\; \frac{0.12 - 0.1}{\sqrt{\dfrac{0.1 \cdot 0.9}{200}}} \;=\; 0.943.$$

Since $Z_0 \notin RR$, we *do not* reject H_0 at this significance level, i.e. conclude that the data seems to confirm the manufacturer's claim that at most 10% of items are defective. Notice that even though the sample proportion was 0.12, *bigger* than 0.1, the inference on the *entire* population proportion

is that it *does not exceed* 0.1 (data from a sample may be misleading, if it is not used properly ...)

b) The P-value is

$$P \;=\; P(Z \geq Z_0) \;=\; 1 - P(Z \leq 0.943) \;=\; 1 - \Phi(0.943) \;=\; 0.174.$$

Since

$$\alpha \;=\; 0.05 \;<\; 0.174 \;=\; P,$$

the decision is to *not reject* the null hypothesis. i.e. accept the manufacturer's claim.
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. So, even more we see that we should be careful not to extrapolate the property of one sample to the entire population.

■

4.5 Tests for Comparing the Parameters of Two Populations

Assume we have two characteristics $X_{(1)}$ and $X_{(2)}$, relative to two populations, with means $\mu_1 = E(X_{(1)}), \mu_2 = E(X_{(2)})$ and variances $\sigma_1^2 = V(X_{(1)}), \sigma_2^2 = V(X_{(2)})$, respectively.
Recall that we draw from both populations random samples of sizes n_1 and n_2, respectively, that are **independent**. Denote the two sets of random variables by

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

Then we have two sample means and two sample variances, given by

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

and

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

respectively. In addition, denote by

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

the *pooled variance* of the two samples, i.e. a variance that considers ("pools") the sample data from both samples.

When comparing the means or proportions of two populations, we estimate their *difference*, whereas for comparing their variances, the *ratio* of the variances will be estimated.

We will use the following theoretical results.

Proposition 4.4. *Assume $X_{(1)} \in N(\mu_1, \sigma_1)$ and $X_{(2)} \in N(\mu_2, \sigma_2)$. Then*

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

b) $T = \dfrac{(\overline{X}_1 - \overline{X}_2) - (\mu_1 - \mu_2)}{s_p\sqrt{\dfrac{1}{n_1} + \dfrac{1}{n_2}}} \in T(n_1 + n_2 - 2)$;

c) $T^* = \dfrac{(\overline{X}_1 - \overline{X}_2) - (\mu_1 - \mu_2)}{\sqrt{\dfrac{s_1^2}{n_1} + \dfrac{s_2^2}{n_2}}} \in T(n)$, *where* $\dfrac{1}{n} = \dfrac{c^2}{n_1 - 1} + \dfrac{(1-c)^2}{n_2 - 1}$ *and* $c = \dfrac{\dfrac{s_1^2}{n_1}}{\dfrac{s_1^2}{n_1} + \dfrac{s_2^2}{n_2}}$;

d) $F = \dfrac{s_1^2/\sigma_1^2}{s_2^2/\sigma_2^2} \in F(n_1 - 1, n_2 - 1)$.

Proposition 4.5. *If the samples are large enough ($n_1 + n_2 > 40$), then parts a), b) and c) of Proposition 4.4 still hold.*

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

We test the hypotheses

$$H_0: \ \mu_1 - \mu_2 = 0, \qquad H_1: \begin{cases} \mu_1 - \mu_2 < 0 \\ \mu_1 - \mu_2 > 0 \\ \mu_1 - \mu_2 \neq 0, \end{cases} \quad \text{equivalent to} \quad \begin{array}{l} H_0: \ \mu_1 = \mu_2, \\ H_1: \begin{cases} \mu_1 < \mu_2 \\ \mu_1 > \mu_2 \\ \mu_1 \neq \mu_2, \end{cases} \end{array} \tag{4.16}$$

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

$$TS \;=\; Z \;=\; \frac{(\overline{X}_1 - \overline{X}_2) - (\mu_1 - \mu_2)}{\sqrt{\dfrac{\sigma_1^2}{n_1} + \dfrac{\sigma_2^2}{n_2}}} \;\in\; N(0,1), \tag{4.17}$$

with observed value

$$Z_0 \;=\; \frac{\overline{X}_1 - \overline{X}_2}{\sqrt{\dfrac{\sigma_1^2}{n_1} + \dfrac{\sigma_2^2}{n_2}}}. \tag{4.18}$$

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

$$RR: \;\begin{cases} \{Z_0 \le z_\alpha\} \\ \{Z_0 \ge z_{1-\alpha}\} \\ \{|Z_0| \ge z_{1-\frac{\alpha}{2}}\} \end{cases} \tag{4.19}$$

and the P-value will be computed as

$$P = \;\begin{cases} P(Z \le Z_0 \mid H_0) &=& \Phi(Z_0) \\ P(Z \ge Z_0 \mid H_0) &=& 1 - \Phi(Z_0) \\ P(|Z| \ge |Z_0| \mid H_0) &=& 2\,(1 - \Phi(|Z_0|)) \end{cases} . \tag{4.20}$$

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

The test statistic is

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

with observed value

$$T_0 \;=\; \frac{\overline{X}_1 - \overline{X}_2}{s_p\sqrt{\dfrac{1}{n_1} + \dfrac{1}{n_2}}}. \tag{4.22}$$

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

$$RR: \begin{cases} \{T_0 \le t_\alpha\} \\ \{T_0 \ge t_{1-\alpha}\} \\ \{|T_0| \ge t_{1-\frac{\alpha}{2}}\}, \end{cases} \tag{4.23}$$

and compute the P-value by

$$P = \begin{cases} P(T \le T_0 \mid H_0) &= F(T_0) \\ P(T \ge T_0 \mid H_0) &= 1 - F(T_0) \\ P(|T| \ge |T_0| \mid H_0) &= 2\left(1 - F(|T_0|)\right), \end{cases} \tag{4.24}$$

where 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

$$TS \;=\; T^* \;=\; \frac{(\overline{X}_1 - \overline{X}_2) - (\mu_1 - \mu_2)}{\sqrt{\dfrac{s_1^2}{n_1} + \dfrac{s_2^2}{n_2}}} \;\in\; T(n), \tag{4.25}$$

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

The observed value of the test statistic is

$$T_0^* \;=\; \frac{\overline{X}_1 - \overline{X}_2}{\sqrt{\dfrac{s_1^2}{n_1} + \dfrac{s_2^2}{n_2}}}. \tag{4.26}$$

The rejection regions and P-values for the three alternatives are again as in equations (4.23)-(4.24), with T_0 replaced by T_0^* from (4.26). The cdf F and the quantiles refer to the $T(n)$ distribution.

Remark 4.6. 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

$$\begin{array}{lll} H_0: \ \dfrac{\sigma_1^2}{\sigma_2^2} = 1, & & \\ H_1: \ \begin{cases} \dfrac{\sigma_1^2}{\sigma_2^2} < 1 \\ \dfrac{\sigma_1^2}{\sigma_2^2} > 1 \\ \dfrac{\sigma_1^2}{\sigma_2^2} \neq 1, \end{cases} & <=> & \begin{array}{ll} H_0: & \sigma_1^2 = \sigma_2^2, \\ H_1: & \begin{cases} \sigma_1^2 < \sigma_2^2 \\ \sigma_1^2 > \sigma_2^2 \\ \sigma_1^2 \neq \sigma_2^2, \end{cases} \end{array} \quad <=> \quad \begin{array}{ll} H_0: & \sigma_1 = \sigma_2, \\ H_1: & \begin{cases} \sigma_1 < \sigma_2 \\ \sigma_1 > \sigma_2 \\ \sigma_1 \neq \sigma_2. \end{cases} \end{array} \end{array} \tag{4.27}$$

The test statistic used is

$$TS \ = \ F \ = \ \frac{s_1^2/\sigma_1^2}{s_2^2/\sigma_2^2} \ \in \ F(n_1 - 1, n_2 - 1), \tag{4.28}$$

with observed value

$$F_0 \ = \ \frac{s_1^2}{s_2^2}. \tag{4.29}$$

The $F(n_1 - 1, n_2 - 1)$ distribution is not symmetric, but proceeding as before, we find the rejection region for the three alternatives as

$$RR: \ \begin{cases} \{F_0 \leq f_\alpha\} \\ \{F_0 \geq f_{1-\alpha}\} \\ \{F_0 \leq f_{\frac{\alpha}{2}} \text{ or } \ F_0 \geq f_{1-\frac{\alpha}{2}}\}. \end{cases} \tag{4.30}$$

and the P-values given by

$$P = \ \begin{cases} P(F \leq F_0 \mid H_0) & = \ F(F_0) \\ P(F \geq F_0 \mid H_0) & = \ 1 - F(F_0) \\ 2 \cdot \min\{P(F \leq F_0 \mid H_0), P(F \geq F_0 \mid H_0)\} & = \ 2 \cdot \min\{F(F_0), 1 - F(F_0)\}, \end{cases} \tag{4.31}$$

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

Example 4.7. 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

$$\begin{aligned} f_{\frac{\alpha}{2}} \quad &= \quad f_{0.025} \quad = \quad 0.4103 \\ f_{1-\frac{\alpha}{2}} \quad &= \quad f_{0.975} \quad = \quad 2.7006. \end{aligned}$$

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

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

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 (4.26)

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

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. Thus, 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.

■

Tests for the difference of means, paired data, $\theta = \mu_1 - \mu_2$ (ttest)

Recall that in many applications, we want to compare the means of two populations, when two random samples (one from each population) are available, which *are not* independent, where each observation in one sample is naturally or by design *paired* with an observation in the other sample.

In such cases, both samples have the same length, n:

$$X_{11}, \ldots, X_{1n} \text{ and } X_{21}, \ldots, X_{2n}$$

and we consider the sample of their differences,

$$D_1, \ldots, D_n,$$

where

$$D_i = X_{1i} - X_{2i},\ i = \overline{1, n}.$$

For this sample, we have

$$\begin{aligned} \overline{X}_d &= \frac{1}{n}\sum_{i=1}^{n} D_i, \text{ the sample mean and} \\ s_d^2 &= \frac{1}{n-1}\sum_{i=1}^{n}\left(D_i - \overline{X}_d\right)^2, \text{ the sample variance.} \end{aligned}$$

Then, it is known that when n is large enough ($n > 30$) or the two populations that the samples are drawn from have approximately Normal distributions $N(\mu_1, \sigma_1), N(\mu_2, \sigma_2)$, the statistic

$$T = \frac{\overline{X}_d - (\mu_1 - \mu_2)}{\frac{s_d}{\sqrt{n}}} \tag{4.32}$$

has a Student $T(n-1)$ distribution, so we can use it as a test statistic for testing the hypotheses

(4.16). Its observed value is

$$T_0 \;=\; \frac{\overline{X}_d}{\frac{s_d}{\sqrt{n}}}. \tag{4.33}$$

Then, as before, we determine the rejection region corresponding to the three alternatives to be

$$RR: \;\begin{cases} \{T_0 \le t_\alpha\} \\ \{T_0 \ge t_{1-\alpha}\} \\ \{|T_0| \ge |t_{1-\frac{\alpha}{2}}|\} \end{cases} \tag{4.34}$$

and compute the P-value by

$$P = \;\begin{cases} P(T \le T_0 \mid H_0) &= F(T_0) \\ P(T \ge T_0 \mid H_0) &= 1 - F(T_0) \\ P(|T| \ge |T_0| \mid H_0) &= 2\left(1 - F(|T_0|)\right), \end{cases} \tag{4.35}$$

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

Example 4.8. Information about ocean weather can be extracted from radar returns with the aid of special algorithms. A study is conducted to estimate the difference in wind speed as measured on the ground, at 12 specified times, using two methods simultaneously. These data result:

Times	1	2	3	4	5	6	7	8	9	10	11	12
Method I	4.46	3.99	3.73	3.29	4.82	6.71	4.61	3.87	3.17	4.42	3.76	3.3
Method II	4.08	3.94	5.00	5.2	3.92	6.21	5.95	3.07	4.76	3.25	4.89	4.8

Assuming the measurements taken by the two methods are approximately Normally distributed, at the 1% significance level, does the data suggest that, on average, the two sets of measurements differ?

Solution. By looking at the data, we see that at some times the measurement taken by the first method is higher, at others, the one given by the second. So we cannot say if, on average, these differences will cancel each other, to yield about the same mean value.

So, we want to test

$$\begin{aligned} H_0: &\;\; \mu_1 = \mu_2 \\ H_1: &\;\; \mu_1 \neq \mu_2, \end{aligned}$$

a two-tailed alternative. The samples yield the following data: sample size $n = 12$, sample mean

$\overline{X}_d = -0.4117$ and sample variance $s_d^2 = 1.2973$, so $s_d = 1.139$.
The observed value of the test statistic from (4.33) is

$$T_0 \;=\; \frac{\overline{X}_d}{\frac{s_d}{\sqrt{n}}} \;=\; -1.2521.$$

For $\alpha = 0.01$, the quantiles for the $T(11)$ distribution are

$$\begin{aligned} t_{\alpha/2} \;&=\; t_{0.005} \;=\; -3.1058, \\ t_{1-\alpha/2} \;&=\; -t_{\alpha/2} \;=\; 3.1058, \end{aligned}$$

so the rejection region is

$$RR \;=\; (-\infty, -3.1058] \,\cup\, [3.1058, \infty).$$

Since $T_0 \notin RR$, we cannot reject the null hypothesis, which means we decide that the two population means are approximately equal.
On the other hand, the P-value of this test is

$$P \;=\; 2\,(1 - F(|T_0|)) \;=\; 0.2365,$$

We have

$$\alpha \;=\; 0.01 \;<\; 0.2365 \;=\; P,$$

the minimum rejection level, so the decision is to *not reject* the null hypothesis. Notice that, again, the P-value is much larger than any conceivable significance level α, so that means that the data strongly suggests that H_0 should not be rejected, i.e., that the two population means *do not* differ.

■

Remark 4.9. The Matlab command **ttest** that performs a T-test for one population mean (in the general case, when σ is not known), can also be used for a paired T-test.

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 H_0 states. How "significant"? That depends on how small the P-value is. The following levels are customary for how "significant" the differences are:

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