5.2 General Framework, Z-Tests

Just like with confidence intervals, we start with the case where the test statistic has a $N(0,1)$ distribution, so we can better understand the ideas.

Let θ be a target parameter and let $\overline{\theta}$ be an unbiased estimator for θ ($E(\overline{\theta}) = \theta$), with standard error $\sigma_{\overline{\theta}}$, such that, under certain conditions, it is known that

$$Z \;=\; \frac{\overline{\theta} - \theta}{\sigma_{\overline{\theta}}} \;\left(=\; \frac{\overline{\theta} - E(\overline{\theta})}{\sigma(\overline{\theta})}\right) \tag{5.1}$$

has an approximately Standard Normal $N(0,1)$ distribution. We design a hypothesis testing procedure for θ the following way: for a given level of significance $\alpha \in (0,1)$, consider the hypotheses

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

with one of the alternatives

$$H_1: \quad \begin{cases} \theta < \theta_0 \\ \theta > \theta_0 \\ \theta \neq \theta_0. \end{cases} \tag{5.2}$$

Recall that we want to determine the rejection region RR such that

$$P(\text{type I error}) \;=\; P(\text{ reject } H_0 \mid H_0) \;=\; P(TS \in RR \mid H_0) \;=\; \alpha. \tag{5.3}$$

We will use the test statistic $TS = Z$ given by (5.1).
The **observed value of the test statistic** from the sample data is

$$TS_0 \;=\; TS(\theta = \theta_0). \tag{5.4}$$

In our case, this is

$$Z_0 \;=\; TS(\theta = \theta_0) \;=\; \frac{\overline{\theta} - \theta_0}{\sigma_{\overline{\theta}}}.$$

How to design the rejection region RR? Let us start with the left-tailed case. We need to determine the RR such that (5.3) holds. Intuitively, we reject H_0 if the observed value of the test statistic is *far* from the value specified in H_0, "far" in the sense of the alternative H_1, in this case *far to the left* of θ_0. So, we determine a rejection region of the form

$$RR \;=\; \{Z_0 \mid Z_0 \le k_1\} \;=\; (-\infty, k_1].$$

We have

$$\begin{aligned} \alpha &= P(Z_0 \in RR \mid H_0) \\ &= P(Z_0 \leq k_1 \mid \theta = \theta_0) \\ &= P(Z_0 \leq k_1 \mid Z_0 \in N(0,1)). \end{aligned}$$

Now, we know that if $Z_0 \in N(0,1)$, $P(Z_0 \leq z_\alpha) = \alpha$, where z_α is the quantile of order α for the $N(0,1)$ distribution. Thus, we choose $k_1 = z_\alpha$ and

$$RR_{\text{left}} = \{Z_0 \leq z_\alpha\}. \tag{5.5}$$

Similarly, for a right-tailed test, we want to find a rejection region of the form

$$RR = \{Z_0 \mid Z_0 \geq k_2\} = [k_2, \infty),$$

so that

$$\begin{aligned} \alpha &= P(Z_0 \in RR \mid H_0) \\ &= P(Z_0 \geq k_2 \mid \theta = \theta_0) \\ &= P(Z_0 \geq k_2 \mid Z_0 \in N(0,1)) \\ &= 1 - P(Z_0 < k_2 \mid Z_0 \in N(0,1)). \end{aligned}$$

Since $P(Z_0 < z_{1-\alpha}) = 1 - \alpha$, then $P(Z_0 \geq z_{1-\alpha}) = \alpha$ and so we choose $k_2 = z_{1-\alpha}$, the quantile of order $1 - \alpha$ for the $N(0,1)$ distribution and

$$RR_{\text{right}} = \{Z_0 \geq z_{1-\alpha}\}. \tag{5.6}$$

Finally, for a two-tailed test, we reject the null hypothesis if the observed value of the test statistic is far away from θ_0 *on either side*. That is, the rejection region should be of the form $RR = \{Z_0 \mid Z_0 \leq k_1 \text{ or } Z_0 \geq k_2\} = (-\infty, k_1] \cup [k_2, \infty)$. The rejection region should be chosen such that

$$P(Z_0 \leq k_1 \text{ or } Z_0 \geq k_2 \mid \theta = \theta_0) = \alpha,$$

or, equivalently,

$$P(k_1 < Z_0 < k_2 \mid Z_0 \in N(0,1)) = 1 - \alpha.$$

We encountered such problems before in the previous sections, when finding (two-sided) confidence intervals. As we did then, we will choose $k_1 = z_{\frac{\alpha}{2}}$ and $k_2 = z_{1-\frac{\alpha}{2}}$, so

$$RR_{\text{two}} = \{Z_0 \leq z_{\frac{\alpha}{2}} \text{ or } Z_0 \geq z_{1-\frac{\alpha}{2}}\}, \tag{5.7}$$

or, since the distribution of Z is symmetric and $z_{1-\frac{\alpha}{2}} > 0$,

$$RR_{\text{two}} = \{Z_0 \leq -z_{1-\frac{\alpha}{2}} \text{ or } Z_0 \geq z_{1-\frac{\alpha}{2}}\} = \{|Z_0| \geq z_{1-\frac{\alpha}{2}}\}.$$

To summarize, the rejection regions for the three alternatives (5.2) are given by

$$RR: \begin{cases} \{Z_0 \leq z_\alpha\} \\ \{Z_0 \geq z_{1-\alpha}\} \\ \{Z_0 \leq z_{\frac{\alpha}{2}} \text{ or } Z_0 \geq z_{1-\frac{\alpha}{2}}\} = \{|Z_0| \geq z_{1-\frac{\alpha}{2}}\}. \end{cases} \tag{5.8}$$

Remark 5.1.
1. Since a test statistic $Z \in N(0,1)$ was used, these are commonly known as ***Z*-tests**.
2. We will derive hypothesis tests for common parameters (mean, proportion, difference of means, ratio of variances, difference of proportions). The test statistics and their distributions will change, but the ideas and the principles will remain the same, as for the case we just described.
3. Notice from our derivation of the rejection region for a two-tailed test, that there is a strong relationship between confidence intervals and rejection regions: The values θ_0 of a target parameter θ in a $100(1-\alpha)\%$ CI $(\alpha \in (0,1))$, are precisely the values for which the test statistic falls *outside* the RR, and hence, for which the null hypothesis $\theta = \theta_0$ is *not* rejected at the significance level α. We say that the $100(1-\alpha)\%$ two-sided CI consists of all the *acceptable* values of the parameter, at the significance level α.
4. **Caution!** This is **not** saying that the rejection region is the complement of the confidence interval! The RR contains values for the *test statistic* TS, while the CI consists of values of the *parameter* θ.

Example 5.2. 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. 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?

Solution. The question is about the *average* number of sales per month, so the test is for the population mean μ. Recall that if either the original population is approximately Normally distributed or the sample size is large (over 30) and σ is known, then

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

Since the sample size $n = 36 > 30$ and we know $\sigma = 4$, we can use a Z-test. The manager's suspicion is that the average is *less* than 20, which is supposed to be, so the two relevant hypotheses in this case are

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

a left-tailed test.

A type I error would mean concluding that the average number of monthly sales is less than 20, when in fact, it is not; a type II error would be deciding that the average number of monthly sales is 20 (or higher), but it actually is not. We allow for the probability of a type I error (the significance level) to be $\alpha = 0.05$. The population standard deviation is known, $\sigma = 4$ and the sample mean is $\overline{X} = 19$.

The observed value of the test statistic is

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

The rejection region is, by (5.8),

$$RR = (-\infty, z_\alpha] = (-\infty, -1.645].$$

Since $Z_0 \notin RR$, we *do not reject* H_0. The evidence obtained from the data is not sufficient to reject it. In the absence of sufficient evidence, by default, we accept the null hypothesis. So, at the 5% significance level, the data *does not* confirm the manager's suspicion.

■

5.3 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 alternatives (5.2), 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

$$P = \begin{cases} P(TS \leq TS_0 \mid H_0) & = \quad F(TS_0) \\ P(TS \geq TS_0 \mid H_0) & = \quad 1 - F(TS_0) \\ 2 \cdot \min\{P(TS \leq TS_0 \mid H_0), P(TS \geq TS_0 \mid H_0)\} & = \quad 2 \cdot \min\{F(TS_0), 1 - F(TS_0)\}. \end{cases} \tag{5.9}$$

Then the decision will be

$$\begin{aligned} &\text{if } \ P \leq \alpha, \text{ reject } \ H_0, \\ &\text{if } \ P > \alpha, \text{ do not reject } \ H_0. \end{aligned} \tag{5.10}$$

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 5.3.

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 5.4. For the problem in Example 5.2, 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

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

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.

■

5.4 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

$$\begin{aligned} H_0: &\quad \mu = \mu_0, \text{ versus one of} \\ H_1: &\quad \begin{cases} \mu < \mu_0 \\ \mu > \mu_0 \\ \mu \neq \mu_0, \end{cases} \end{aligned} \tag{5.11}$$

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

<u>**Case σ known**</u> (**ztest**)

We use the test statistic

$$TS = Z = \frac{\overline{X} - \mu}{\frac{\sigma}{\sqrt{n}}} \in N(0,1), \tag{5.12}$$

with observed value

$$Z_0 = \frac{\overline{X} - \mu_0}{\frac{\sigma}{\sqrt{n}}}. \tag{5.13}$$

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

$$RR: \begin{cases} \{Z_0 \leq z_\alpha\} \\ \{Z_0 \geq z_{1-\alpha}\} \\ \{|Z_0| \geq z_{1-\frac{\alpha}{2}}\} \end{cases} \tag{5.14}$$

and the P-value will be computed as

$$P = \begin{cases} P(Z \leq Z_0 \mid H_0) & = \Phi(Z_0) \\ P(Z \geq Z_0 \mid H_0) & = 1 - \Phi(Z_0) \\ P(|Z| \geq |Z_0| \mid H_0) & = 2\left(1 - \Phi(|Z_0|)\right), \end{cases} \tag{5.15}$$

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.

<u>Case σ unknown</u> (**ttest**)

In this case, we use the test statistic

$$TS \;=\; T \;=\; \frac{\overline{X}-\mu}{\frac{s}{\sqrt{n}}} \;\in\; T(n-1), \tag{5.16}$$

with observed value

$$T_0 \;=\; \frac{\overline{X}-\mu_0}{\frac{s}{\sqrt{n}}}. \tag{5.17}$$

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

$$RR: \;\begin{cases} \{T_0 \le t_\alpha\} \\ \{T_0 \ge t_{1-\alpha}\} \\ \{|T_0| \ge t_{1-\frac{\alpha}{2}}\}, \end{cases} \tag{5.18}$$

and compute the P-value by

$$P = \;\begin{cases} P(T \le T_0 \mid H_0) &=& F(T_0) \\ P(T \ge T_0 \mid H_0) &=& 1-F(T_0) \\ P(|T| \ge |T_0| \mid H_0) &=& 2\left(1-F(|T_0|)\right), \end{cases} \tag{5.19}$$

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

$$Z \;=\; \frac{\overline{p}-p}{\sqrt{\frac{p(1-p)}{n}}} \tag{5.20}$$

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

$$H_1 : \begin{cases} p < p_0 \\ p > p_0 \\ p \neq p_0. \end{cases}, \tag{5.21}$$

we use the test statistic $TS = Z$ from (5.20). Then, as before, at the $\alpha \in (0, 1)$ significance level, the rejection region for each test will be given by

$$RR : \begin{cases} \{Z_0 \leq z_\alpha\} \\ \{Z_0 \geq z_{1-\alpha}\} \\ \{|Z_0| \geq z_{1-\frac{\alpha}{2}}\}, \end{cases} \tag{5.22}$$

and the P-value will be computed as

$$P = \begin{cases} P(Z \leq Z_0 \mid H_0) & = & \Phi(Z_0) \\ P(Z \geq Z_0 \mid H_0) & = & 1 - \Phi(Z_0) \\ P(|Z| \geq |Z_0| \mid H_0) & = & 2\left(1 - \Phi(|Z_0|)\right), \end{cases} \tag{5.23}$$

since $N(0, 1)$ is symmetric, where $\Phi(x)$ is Laplace's function, the cdf for the Standard Normal $N(0, 1)$ distribution.

Example 5.5. 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

$$P \;=\; P(Z \geq Z_0) \;=\; 1 - P(Z \leq 0.943) \;=\; 1 - \Phi(0.943) \;=\; 0.173.$$

Since

$$\alpha \;=\; 0.05 \;<\; 0.173 \;=\; P,$$

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.

■

5.5 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

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

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

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

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 5.6. *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 5.7. *If the samples are large enough ($n_1 + n_2 > 40$), then parts a), b) and c) of Proposition 5.6 still hold.*

Fisher (F) Distribution

In many cases, two population *variances* need to be compared. Such inference is used for the comparison of accuracy, stability, uncertainty, or risks arising in two populations.

Consider, for instance, two mutual fund investments that promise the same expected return. However, one of them recorded a 10% higher volatility over the last 15 days. Is this a significant evidence for a conservative investor to prefer the other mutual fund? *Volatility* is essentially the standard deviation of returns. This is a case where we should be able to compare variances (or standard deviations) of two populations.
Moreover, recall (from the construction of CI's for the difference of means) that we have several cases, depending on whether or not the population variances are known (assumed) to be equal or not. Rather than "assuming" equality of the population variances, we can now test that assertion, by comparing them based on data from samples.

Comparison of variances can be accomplished using the **Fisher-Snedecor (F) distribution**. This distribution was first considered in 1918 by a famous English statistician and biologist, Sir Ronald Fisher (1890-1962) and developed and formalized in 1934 by an American mathematician George Snedecor (1881-1974). A random variable X follows a *Fisher (F)* distribution with parameters $m, n \in \mathbb{N}$ (degrees of freedom), if its density function is

$$f(x) \;=\; \frac{1}{\beta(\frac{m}{2}, \frac{n}{2})} \left(\frac{m}{n}\right)^{\frac{m}{2}} x^{\frac{m}{2}-1} \left(1 + \frac{m}{n}x\right)^{-\frac{m+n}{2}}, \; x > 0,$$

where $\beta(a,b) = \int\limits_0^1 x^{a-1}(1-x)^{b-1}dx,\; a, b > 0$, is Euler's Beta function. Its density has a right-skewed shape (see Figure 1).

Since this is asymmetric, we no longer have the same relationship between its quantiles, that we have seen for the Normal or Student distributions (i.e., $q_\alpha = -q_{1-\alpha}$). However, there is an important property of the F distribution:

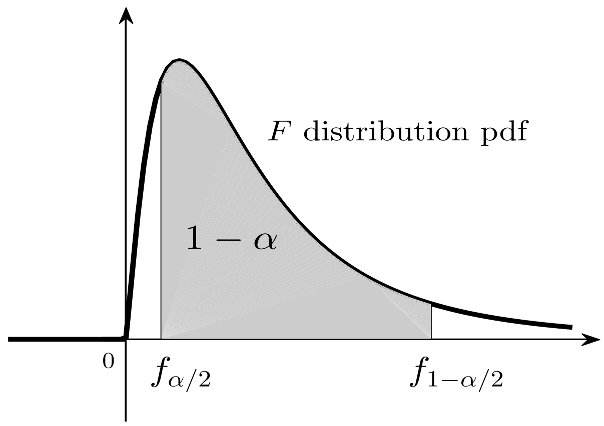

Fig. 1: Fisher (F) Distribution pdf and quantiles

Proposition 5.8. *If the variable X has a $F(m,n)$ distribution, then its reciprocal $\frac{1}{X}$ has a $F(n,m)$ distribution. As a consequence, the following relation holds for F-quantiles:*

$$f_{1-\alpha,m,n} = \frac{1}{f_{\alpha,n,m}},\ \forall\alpha\in(0,1), \tag{5.24}$$

where the quantile $f_{1-\alpha,m,n}$ refers to the $F(m,n)$ distribution and $f_{\alpha,n,m}$ is for the $F(n,m)$ distribution.

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

We test the hypotheses

$$\begin{array}{ll} H_0: & \mu_1-\mu_2=0, \\ H_1: & \begin{cases}\mu_1-\mu_2<0\\ \mu_1-\mu_2>0\\ \mu_1-\mu_2\neq 0,\end{cases}\end{array} \quad \text{equivalent to} \quad \begin{array}{ll} H_0: & \mu_1=\mu_2, \\ H_1: & \begin{cases}\mu_1<\mu_2\\ \mu_1>\mu_2\\ \mu_1\neq\mu_2,\end{cases}\end{array} \tag{5.25}$$

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

$$TS = Z = \frac{(\overline{X}_1-\overline{X}_2)-(\mu_1-\mu_2)}{\sqrt{\frac{\sigma_1^2}{n_1}+\frac{\sigma_2^2}{n_2}}} \in N(0,1), \tag{5.26}$$

with observed value

$$Z_0 \;=\; \frac{\overline{X}_1 - \overline{X}_2}{\sqrt{\dfrac{\sigma_1^2}{n_1} + \dfrac{\sigma_2^2}{n_2}}}. \tag{5.27}$$

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

$$RR: \begin{cases} \{Z_0 \le z_\alpha\} \\ \{Z_0 \ge z_{1-\alpha}\} \\ \{|Z_0| \ge z_{1-\frac{\alpha}{2}}\} \end{cases} \tag{5.28}$$

and the P-value will be computed as

$$P = \begin{cases} P(Z \le Z_0 \mid H_0) & = \;\Phi(Z_0) \\ P(Z \ge Z_0 \mid H_0) & = \;1 - \Phi(Z_0) \\ P(|Z| \ge |Z_0| \mid H_0) & = \;2\left(1 - \Phi(|Z_0|)\right), \end{cases} \tag{5.29}$$

since $N(0,1)$ is symmetric, where $\Phi(x)$ is Laplace's function, the cdf for the Standard Normal $N(0,1)$distribution.

Case $\sigma_1 = \sigma_2$ unknown (**ttest2**)
The test statistic is

$$TS \;=\; T \;=\; \frac{(\overline{X}_1 - \overline{X}_2) - (\mu_1 - \mu_2)}{s_p\sqrt{\dfrac{1}{n_1} + \dfrac{1}{n_2}}} \;\in\; T(n_1 + n_2 - 2), \tag{5.30}$$

with observed value

$$T_0 \;=\; \frac{\overline{X}_1 - \overline{X}_2}{s_p\sqrt{\dfrac{1}{n_1} + \dfrac{1}{n_2}}}. \tag{5.31}$$

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

$$RR: \begin{cases} \{T_0 \le t_\alpha\} \\ \{T_0 \ge t_{1-\alpha}\} \\ \{|T_0| \ge t_{1-\frac{\alpha}{2}}\}, \end{cases} \tag{5.32}$$

and compute the P-value by

$$P = \begin{cases} P(T \leq T_0 \mid H_0) & = \; F(T_0) \\ P(T \geq T_0 \mid H_0) & = \; 1 - F(T_0) \\ P(|T| \geq |T_0| \mid H_0) & = \; 2\,(1 - F(|T_0|))\,, \end{cases} \tag{5.33}$$

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

$$TS \;=\; T^* \;=\; \frac{(\overline{X}_1 - \overline{X}_2) - (\mu_1 - \mu_2)}{\sqrt{\dfrac{s_1^2}{n_1} + \dfrac{s_2^2}{n_2}}} \;\in\; T(n), \tag{5.34}$$

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

$$T_0^* \;=\; \frac{\overline{X}_1 - \overline{X}_2}{\sqrt{\dfrac{s_1^2}{n_1} + \dfrac{s_2^2}{n_2}}}. \tag{5.35}$$

The rejection regions and P-values for the three alternatives are again as in equations (5.32)-(5.33), with T_0 replaced by T_0^* from (5.35). The cdf F and the quantiles refer to the $T(n)$ distribution.

Remark 5.9. 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

$$\begin{array}{l} H_0: \ \dfrac{\sigma_1^2}{\sigma_2^2} = 1, \\ H_1: \ \begin{cases} \dfrac{\sigma_1^2}{\sigma_2^2} < 1 \\ \dfrac{\sigma_1^2}{\sigma_2^2} > 1 \\ \dfrac{\sigma_1^2}{\sigma_2^2} \neq 1, \end{cases} \end{array} \quad <=> \quad \begin{array}{l} H_0: \ \sigma_1^2 = \sigma_2^2, \\ H_1: \ \begin{cases} \sigma_1^2 < \sigma_2^2 \\ \sigma_1^2 > \sigma_2^2 \\ \sigma_1^2 \neq \sigma_2^2, \end{cases} \end{array} \quad <=> \quad \begin{array}{l} H_0: \ \sigma_1 = \sigma_2, \\ H_1: \ \begin{cases} \sigma_1 < \sigma_2 \\ \sigma_1 > \sigma_2 \\ \sigma_1 \neq \sigma_2. \end{cases} \end{array} \tag{5.36}$$

The test statistic used is

$$TS = F = \frac{s_1^2/\sigma_1^2}{s_2^2/\sigma_2^2} \in F(n_1 - 1, n_2 - 1), \tag{5.37}$$

with observed value

$$F_0 = \frac{s_1^2}{s_2^2}. \tag{5.38}$$

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

$$RR: \begin{cases} \{F_0 \le f_\alpha\} \\ \{F_0 \ge f_{1-\alpha}\} \\ \{F_0 \le f_{\frac{\alpha}{2}} \text{ or } F_0 \ge f_{1-\frac{\alpha}{2}}\}. \end{cases} \tag{5.39}$$

and the P-values given by

$$P = \begin{cases} P(F \le F_0 \mid H_0) & = F(F_0) \\ P(F \ge F_0 \mid H_0) & = 1 - F(F_0) \\ 2 \cdot \min\{P(F \le F_0 \mid H_0), P(F \ge F_0 \mid H_0)\} & = 2 \cdot \min\{F(F_0), 1 - F(F_0)\}, \end{cases} \tag{5.40}$$

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

Example 5.10. 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

$$H_0: \ \sigma_1^2 = \sigma_2^2$$
$$H_1: \ \sigma_1^2 \neq \sigma_2^2.$$

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

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

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 (5.35)

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

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.

■