TO HYPOTESIS TESTING AND PARAMETRIC TESTS

Marianna Bolla, DSc

Institute of Mathematics, BME

2023. október 16.

1. Continuous distributions

1.1. One-sample, two-sided z-test

Problem: There are complaints that the bread (written 1 kg on it) weights less. We randomly select n breads, their weights are $X_1, \dots, X_n$, $n = 25$. The sample average is 0.98 kg. What to do? The 0.02 kg difference can be caused by randomness. Even if $X_1, \dots, X_n \sim \mathcal{N}(\mu_0, \sigma_0)$ with μ_0= 1 kg and $\sigma_0 = 0.05$, there are fluctuations.

Investigate the alternative:

$$H_0 : \mu = \mu_0 (= 1 \text{ kg}) \quad \text{versus} \quad H_1 : \mu \neq \mu_0.$$

By the assumption of innocence, the jury assumes H_0 and we must have enough evidence to prove the contrary: H_1.

If we construct, say, a 95% confidence interval for the population mean, the hypothetical μ_0=1 kg should be in it with high probability. If not, then

- either the complementary event happens, but this has small 5% probability;
- we rather suspect, that the population mean is not 1 kg, and reject H_0.

We can only rule the Type I error probability (now 0.05): we reject H_0 if it is true (the jury sentences an innocent). The Type II error probability is opposite: we accept H_0 if not true (the jury acquits someone who is guilty). Its probability increases if we decrease the Type I error probability, so selecting the Type I error probability (called significance, and denoted by α) raises ethical issues.

To simplify things, first we construct the test statistic:

$$Z = \frac{\bar{X} - \mu_0}{\sigma_0} \sqrt{n}$$

and select a significance (now $\alpha = 0.05$), then find the *critical value*

$$z_{\alpha/2} = \Phi^{-1}\left(1 - \frac{\alpha}{2}\right)$$

from the standard normal distribution table. When we constructed confidence interval last week, we saw that

$$\mathbb{P}\left(\mu_0 \in \left(\bar{X} - \frac{z_{\alpha/2}\sigma_0}{\sqrt{n}},\, \bar{X} + \frac{z_{\alpha/2}\sigma_0}{\sqrt{n}}\right)\right) = \mathbb{P}\left(|Z| < z_{\alpha/2}\right) = 1 - \alpha\,.$$

Therefore, our decision is: if $|z| < z_{\alpha/2}$, then we accept, and if $|z| \geq z_{\alpha/2}$, then we reject H_0 with significance α. The region

$$R = \{\mathbf{x} \,:\, |z(\mathbf{x})| \geq z_{\alpha/2}\}$$

is called *rejection or critical region.*

In our numerical example: $\bar{x} = 0.98$, $\mu_0 = 1$, $n = 25$ and let $\sigma_0 = 0.05$. So $z = -2$. With sign. $\alpha = 0.05$ $z_{\alpha/2} = 1.96$, therefore, with sign. 0.05 we reject H_0 and the decision of the lower court: the shop is guilty. They work with sign. 0.05, i.e., give 0.05 prob. to the event that an innocent is convicted (jury basically defends the innocents, and they must have enough evidence to convict someone).

Then the shop goes to the higher court. They work with sign. $\alpha = 0.01$, because they are more strict and better defend the innocents (give only 0.01 prob. to the event that an innocent is convicted). Now $z_{\alpha/2} = 2.58$, our $|z| = 2 < 2.58$, so the higher court cannot reject H_0 and acquits the shop.

From the table we can see that with $\alpha = 0.0456$, $z_{\alpha/2} = 2$, so 0.0456 is the smallest possible sign., at which we can reject H_0. Program packages output this sometimes called *P-value* and if it is small enough, we can reject H_0.

Investigate the Type I error prob. (α) and the Type II error prob. (β), which also depends on the true value of μ:

$$\alpha = \mathbb{P}_{\mu_0}\left(|Z| \geq z_{\alpha/2}\right)$$

per def, and if $\mu \neq \mu_0$:

$$\beta(\mu) = \mathbb{P}\left(|Z| < z_{\alpha/2} \,|\, \mu\right).$$

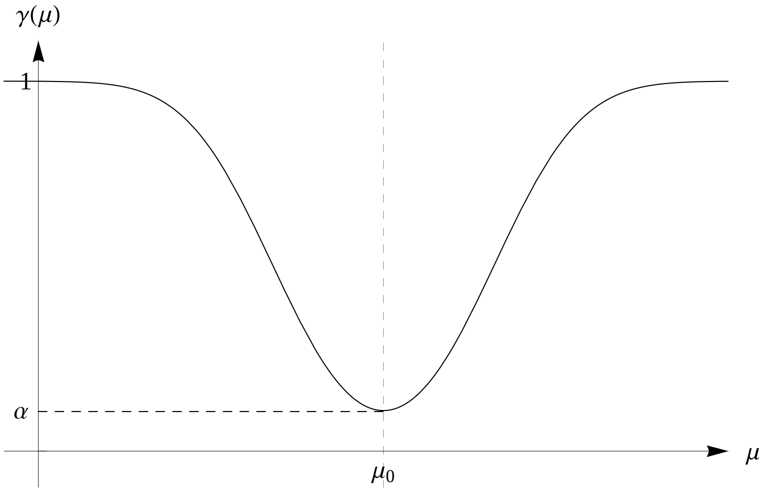
Sometimes the so-called *power function* γ is used, which is

$$\gamma(\mu) = 1 - \beta(\mu) = \mathbb{P}\left(|Z| \geq z_{\alpha/2} \,|\, \mu\right).$$

The theory guarantees the existence of a Uniformly Most Powerful (UMP) test in this situation (Neyman-Pearson): given α, the UMP test has the largest possible power (so the smallest possible Type II error) for any μ. This means, that if the jury convicts innocents with given prob, then they acquit any criminal with the smallest possible prob.

In our example,

$$\begin{aligned}
\gamma(\mu) &= 1 - \mathbb{P}\left(-z_{\alpha/2} < \frac{\bar{X} - \mu_0}{\sigma_0}\sqrt{n} < z_{\alpha/2} \,|\, \mu\right) = \\
&= 1 - \left(-z_{\alpha/2} - \Delta_n < \frac{\bar{X} - \mu}{\sigma_0}\sqrt{n} < z_{\alpha/2} - \Delta_n\right) = \\
&= 1 - \Phi(z_{\alpha/2} - \Delta_n) + \Phi(-z_{\alpha/2} - \Delta_n) = \\
&= 2 - \Phi(z_{\alpha/2} - \Delta_n) - \Phi(z_{\alpha/2} + \Delta_n),
\end{aligned}$$



1. ábra. Power function of the one-sample two-sided z-test

where

$$\Delta_n = \frac{\mu - \mu_0}{\sigma_0}\sqrt{n}$$

and $\frac{\bar{X}-\mu}{\sigma_0}\sqrt{n} \sim \mathcal{N}(0,1)$, if μ is the true population mean.

From this form, it is easy to see that if $n \to \infty$ (more and more witnesses) or μ gets farther and farther from μ_0 (more and more guilty), then $\gamma(\mu) \to 1$ and $\beta(\mu) \to 0$, see Fig. 1.

1.2. One-sample, one-sided z-test

The shop example is better formulated as

$$H_0 : \mu \geq \mu_0 \quad \text{versus} \quad H_1 : \mu < \mu_0$$

(the accuse is only about smaller breads, if larger, no problem). Then the rejection region is

$$R = \{\mathbf{x} : z(\mathbf{x}) \leq -z_\alpha\},$$

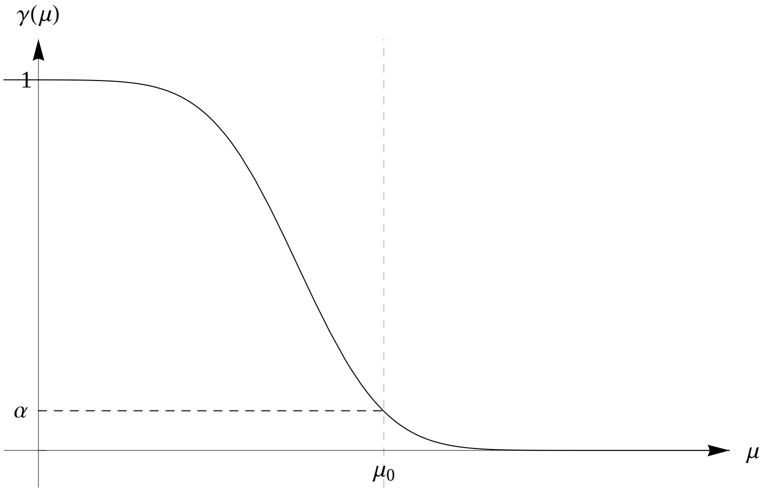
see the formulas to hypothesis testing.

Then with sign. 0.05 and 0.025 we reject H_0. From $-z_\alpha = -2$ we get that the smallest possible α (P-value) at which we can reject H_0 is $\alpha = 0.0228$. (In the two-sided situation the P-value was the double of this: 0.0456.) So, in the one-sided, modified accuse case, we can sooner convict the shop than in the two-sided situation.

The power function, fixing n and α is

$$\gamma(\mu) = \mathbb{P}(Z \leq -z_\alpha|\mu) = \mathbb{P}_\mu(\frac{\bar{X}-\mu}{\sigma_0}\sqrt{n} + \Delta_n \leq -z_\alpha) = \Phi(-z_\alpha - \Delta_n).$$

Here $\gamma(\mu)$ decreases in μ, see Fig. 2.



2. ábra. Power function of the one-sample one-sided z-test

2. Discrete distributions

In the reference book (Chapter 6), the Problem (cure-rate, p. 167) deals with the one-sided and Ex. 6.1 (cat-food) with the two sided problem (underlying Bernoulli distribution with binomial test statistic). These are small sample exercises. In the large sample case, we can use z-test for the population proportions, and get surprisingly other results.

2.1. One-sided alternative

Problem: Experience has shown that the cure rate for a given disease using standard medication is 60%. The cure rate of a new drug is anticipated to be better than the standard mediacation. Suppose that the new drug is to be tried on a sample of 20 patients and that the number cured X in the 20 is to be recorded. *Is there substancial evidence that the new drug has a higher cure rate than the standard medication?*

Consider the following rejection regions (strategies):

$$\mathrm{A}: \; R = \{X \geq 15\}, \quad \mathrm{B}: \; R = \{X \geq 18\}, \quad \mathrm{C}: \; R = \{X \geq 14\}.$$

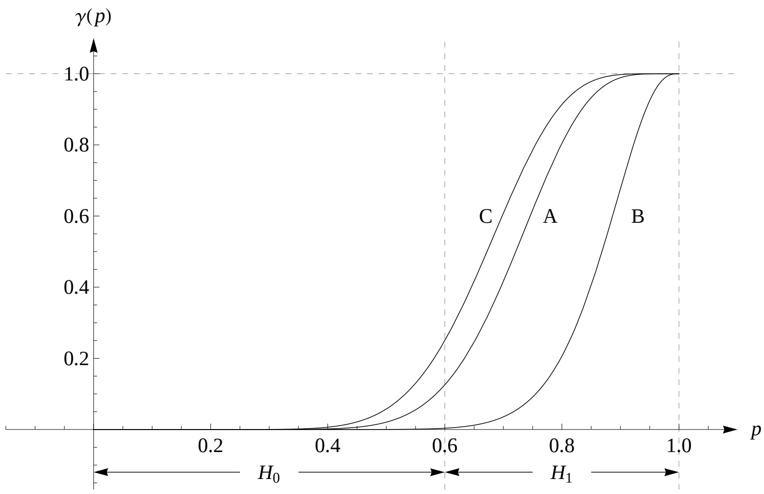
Solution. $X_1, \ldots, X_{20}$ are i.i.d. Bernoulli r.v.'s with parameter p $(0 < p < 1)$, where $X_i = 1$, if patient i is cured by the new drug and 0 if not. The test statistic is $X = \sum_{i=1}^{20} X_i \sim \mathcal{B}_{20}(p)$, based on which we decide about the alternative

$$H_0 : \; p \leq 0.6 \quad \text{versus} \quad H_1 : \; p > 0.6$$

and we are glad if can reject it with a relatively „small" significance.

Let us investigate the power function. With strategy (A) this is

$$\gamma(p) = \mathbb{P}(X \geq 15) = 1 - \sum_{k=0}^{14} \binom{20}{k} p^k (1-p)^{20-k} = 1 - F_p^+(14),$$



3. ábra. The $\gamma(p)$ function with the three strategies

where F_p^+ is the right continuous c.d.f. of the $\mathcal{B}_{20}(p)$ distribution (see Tables). The $\gamma(p)$ function is strictly monotone increasing in p, and its value at $p = 0.6$ is 0.126. Table 1 shows some values of the γ function with the three strategies.

p	0.3	0.4	0.5	0.6	0.7	0.8	0.9
$(A)\ \gamma(p) = P(X \geq 15)$	0.000	0.002	0.021	0.126	0.416	0.804	0.989
$(B)\ \gamma(p) = P(X \geq 18)$	0.000	0.000	0.000	0.004	0.035	0.206	0.677
$(C)\ \gamma(p) = P(X \geq 14)$	0.000	0.006	0.058	0.250	0.608	0.913	0.998

1. táblázat. Some $\gamma(p)$ values with the three strategies

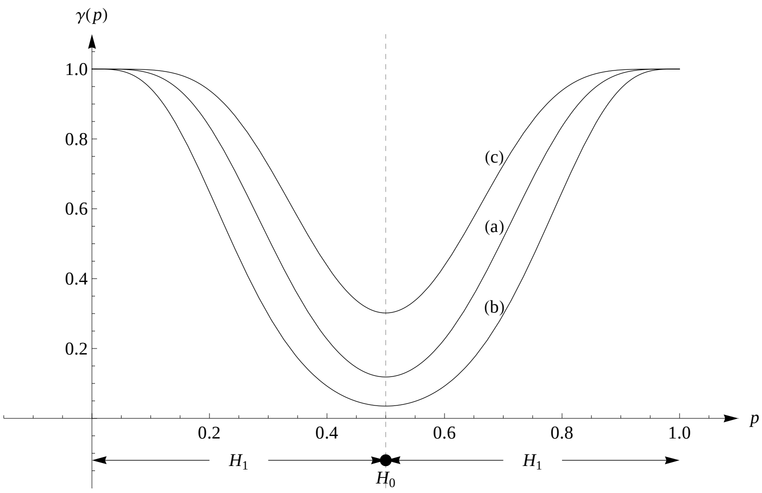
As H_0 is now composite, there are several Type I. errors, the maximum of which is $\max_{p \leq 0.6} \gamma(p) = \gamma(0.6) = 0.126$ in case (A). Therefore, 0.126 is the smallest significance (P-value) with which we can reject H_0 under strategy (A). For $p > 0.6$, the power function starts and it is strictly monotone increasing; it becomes 1 at $p = 1$.

The γ function shows similar behavior under the more strict strategy (B) and the more mild strategy (C) too (see Figure 3), and their significances are 0.004 and 0.250, respectively. So the most evidence is provided by strategy (B) for the rejection of H_0, but it requires too much: at least 18 must recover out of the 20 (it can also produce a high Type II. error). However, the significance 0.250 of strategy (C) is too large (it means that we state the effectiveness of the new drug without sufficient reason, the probability that we are wrong is 1/4. Finally, we can choose strategy (A) with an acceptable significance.

$\square$

2.2. Two-sided alternative

Problem: 15 cats are invited to taste A and B foods. Is there any evidence that te two flavors appeal differently to te cats?



4. ábra. The power function based on 15 cats under the three strategies, $\alpha = \gamma(0.5)$.

In case of 15 cats, Fig. 4 shows the power curves under strategies

$$\text{a}: R = \{X \le 4 \quad \text{or} \quad X \ge 11\}, \quad \text{b}: R = \{X \le 3 \quad \text{or} \quad X \ge 12\},$$

and the smallest rejection region, containing our evidence that 5 cats eat A:

$$\text{c}: R = \{X \le 5 \quad \text{or} \quad X \ge 10\}.$$

Now $X_1, \dots, X_{15}$ is i.i.d. Bernoulli sample with parameter p $(0 < p < 1)$, where $X_i = 1$ if cat i eats A, and 0, otherwise. The test statistic is $X = \sum_{i=1}^{15} X_i \sim \mathcal{B}_{15}(p)$, and the underlying alternative:

$$H_0 : p = 0.5 \quad \text{versus} \quad H_1 : p \neq 0.5.$$

Here, strategy (b) has the smallest significance (Type I error): $\alpha = \gamma(0.5) = 0.036$ and strategy (c) has the largest: 0.302. However, for $p \neq 0.5$, the Type II error $\beta(p) = 1 - \gamma(p)$ is the smallest in strategy (c).

So the divison 5:10 of the cats is not enough evidence to reject H_0, but in case of 150 cats, this is a strong evidence as follows.

2.3. Testing the population proportion for large samples, two-sided alternative

Now $n = 150 \ge 30$, and so, by the special case of the CLT (Moivre–Laplace theorem), $X = \sum_{i=1}^{n} X_i \sim \mathcal{B}_n(p)$ is approximately $\mathcal{N}(np, \sqrt{np(1-p)})$. Therefore, the population proportion $\bar{X}$ is approximately $\mathcal{N}(p, \sqrt{\frac{p(1-p)}{n}})$, where for p, $r = \bar{X}$, and for the standard deviation, $\sqrt{\frac{r(1-r)}{n}}$ are efficient estimators. Therefore, for the alternative

$$H_0 : p = 0.5 \quad \text{versus} \quad H_1 : p \neq 0.5$$

the test statistic

$$Z = \frac{r - 0.5}{\sqrt{r(1-r)}}\sqrt{n}$$

is approximately standard normal under H_0. The rejection region is the same as that of the two-sided z-test:

$$R = \{|z| \geq z_{\alpha/2}\}.$$

If 50 out of the 150 cats eat A, then $r = \frac{1}{3}$ and $z = -4.33$. This is on the boundary of R such that $z_{\alpha/2} = |-4.33|$. From the standard normal table, the corresponding α is near 0 (with many 0 decimals). So the P-value is practically 0, we can reject H_0 wit a very small significance (Type I error, that we state the difference of the foods without any reason, is very small, indeed). Also, we need not worry about the possibly larges Type II error, as it is always 'small' if n is 'large'. So 50 out of 150 is a strong evidence that the foods appeal differently to cats. Of course, 40:110 or 30:120 are much stronger.

Note that accepting H_0 is equivalent that the hypothetical $p_0 = 0.5$ is within the confidence interval of level $1 - \alpha$ constructed for p:

$$r \pm z_{\alpha/2}\sqrt{\frac{r(1-r)}{n}}. \tag{1}$$

2.4. Testing the population proportion for large samples, one-sided alternative

We revisit the patient recovery exercise with $n = 200$ patients trying the new pill.

Now $X_1, \dots, X_n$ is i.i.d. Bernoulli sample with parameter p ($0 < p < 1$); $X_i = 1$ if patient i recovers from the pill, and 0 if not. Since n is large, by the CLT (Moivre–Laplace theorem), $X = \sum_{i=1}^n X_i \sim \mathcal{B}_n(p)$ and so, $\bar{X}$ is approximately normal, as before. Here, for the alternative

$$H_0 : p \leq 0.6 \quad \text{versus} \quad H_1 : p > 0.6$$

the test statistic is

$$Z = \frac{r - 0.6}{\sqrt{r(1-r)}}\sqrt{n},$$

that is approximately standard normal if $p = 0.6$ (boundary of H_0). The rejection region is:

$$R = \{z \geq z_\alpha\}.$$

Because of the monotonic nature of the power function, this is good for composite H_0 too.

If 140 out of 200 patients recover, then $r = \frac{140}{200}$ and $z = 5.09$. This is on the boundary of R with $z_\alpha = 5.09$; so, α is again 0, practically. Therefore, 140 recover out of 200 is a strong evidence to prove that the new pill is more efficient than the old one. Note that 14 out of 20 was not strong enough, but this is the law of large numbers.

2.5. Comparing two population proportions for large samples

Now, we have two independent Bernoulli samples, with sizes $n_1 \geq 30$, $n_2 \geq 30$, and population proportions r_1, r_2. For example, we want to compare the recovery rate in two patient groups. Then

$$Z = \frac{r_1 - r_2}{\sqrt{\frac{r_1(1-r_1)}{n_1} + \frac{r_2(1-r_2)}{n_2}}}$$

under H_0 (that $p_1 = p_2$) is approximately $\mathcal{N}(0, 1)$.

Note that in this two-sample, two-sided case, the acceptance of H_0 is equivalent to the fact that $p_1 - p_2$ is within the confidence interval

$$r_1 - r_2 \pm z_{\alpha/2} \sqrt{\frac{r_1(1-r_1)}{n_1} + \frac{r_2(1-r_2)}{n_2}}.$$

Remark: instead of the standard deviation, $\sqrt{\frac{\hat{r}(1-\hat{r})}{n}}$ can as well be used, where $\hat{r} = \frac{n_1 r_1 + n_2 r_2}{n}$ and $n = n_1 + n_2$. With this pooled s.d.,

$$Z' = \frac{r_1 - r_2}{\sqrt{\hat{r}(1-\hat{r})}\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}},$$

resembling the formula of the independent sample t-test in the next section.

3. Two-sample t-test

Let $X_1, \ldots, X_{n_1} \sim \mathcal{N}(\mu_1, \sigma)$ be i.i.d. sample, and independently of it, let $Y_1, \ldots, Y_{n_2} \sim \mathcal{N}(\mu_2, \sigma)$ be another i.i.d. sample. Here $n_1, n_2 < 30$ and the unknown s.d. σ is assumed to be the same in the two samples. For this, first we perform an F-test, see the next section.

First test the following two-sided alternative:

$$H_0 : \mu_1 = \mu_2 \quad \text{vers.} \quad H_1 : \mu_1 \neq \mu_2 \tag{2}$$

We construct a statistic the distribution of which under H_0 is Student t. Indeed, under H_0

$$\bar{X} - \bar{Y} \sim \mathcal{N}\left(0, \sqrt{\frac{n_1 + n_2}{n_1 n_2}}\sigma\right),$$

standardize it, and put it into the numerator of the t-statistic to be constructed.

By the Lukács's theorem,

$$(n_1 - 1){S_X^*}^2/\sigma^2 \sim \chi^2(n_1 - 1)$$

and independently of this,

$$(n_2 - 1){S_Y^*}^2/\sigma^2 \sim \chi^2(n_2 - 1),$$

therefore,

$$\frac{(n_1 - 1){S_X^*}^2 + (n_2 - 1){S_Y^*}^2}{\sigma^2} \sim \chi^2(n_1 + n_2 - 2).$$

The squareroot of this divided by the d.f. n_1+n_2-2 is put into the denominator. The numerator and denominator are independent r.v.'s by the Lukács's theorem. Therefore, the test statistic is

$$\frac{\frac{\bar{X}-\bar{Y}-0}{\sqrt{\frac{n_1+n_2}{n_1 n_2}}\sigma}}{\sqrt{\frac{(n_1-1){S_X^*}^2+(n_2-1){S_Y^*}^2/\sigma^2}{n_1+n_2-2}}} \sim t(n_1+n_2-2),$$

where the unknown (but same) σ cancels, and we get that

$$t = \frac{\bar{X}-\bar{Y}}{s_{pooled}}\sqrt{\frac{n_1 n_2}{n_1+n_2}} = \frac{\bar{X}-\bar{Y}}{s_{pooled}\sqrt{\frac{1}{n_1}+\frac{1}{n_2}}},$$

where

$$s_{pooled} = \sqrt{\frac{(n_1-1){S_X^*}^2+(n_2-1){S_Y^*}^2/\sigma^2}{n_1+n_2-2}}$$

is the *pooled s.d.*, and the pooled variance s^2_{pooled} is unbiased estimator of σ^2.

For the two-sided alternative (4) the rejection region defining the α-sign. test is:

$$R = \{|t| \geq t_{\alpha/2}(n_1+n_2-2)\}.$$

. Likewise, for the one-sided alternative

$$H_0 : \mu_1 \leq \mu_2 \quad \text{vers.} \quad H_1 : \mu_1 > \mu_2, \tag{3}$$

the test statistic is the same, but the rejection region defining the α-sign. test is:

$$R = \{t \geq t_\alpha(n_1+n_2-2)\}.$$

If in (3) we want to test the opposite direction, we simply interchange the role-cast of the $X-Y$ samples.

Note that the acceptance of H_o in (4) is equivalent that 0 is within the confidence interval of level $1-\alpha$:

$$\bar{x}-\bar{y} \pm t_{\alpha/2}(n_1+n_2-2)s_{pooled}\sqrt{\frac{1}{n_1}+\frac{1}{n_2}}.$$

This was the independent sample t-test. IMPORTANT: in case of PARED (MATCHED) SAMPLES, one-sample test should be used for the differences $X_i - Y_i$, $i = 1, \dots, n$.

Note tat when the equality of variances is rejected (see the upcoming F-test), then a modified t-test, e.g., the Welch-test, should be used. In case of the paired sample case, it is not needed as we have only one sample $D_i = X_i - Y_i$, $i = 1, \dots, n$.

4. F-test

Let $X_1, \dots, X_{n_1} \sim \mathcal{N}(\mu_1, \sigma_1)$ be i.i.d. sample, and independently of it, let $Y_1, \dots, Y_{n_2} \sim \mathcal{N}(\mu_2, \sigma_2)$ be another i.i.d. sample.

We test the following two-sided alternative:

$$H_0 : \sigma_1 = \sigma_2 \quad \text{vers.} \quad H_1 : \sigma_1 \neq \sigma_2 \tag{4}$$

We construct a statistic the distribution of which under H_0 is Fischer F (by definition, this is the distribution of the ratio of two independent χ^2-distributed r.v.'s, each divided with its own d.f.).

Because of the above considerations,

$$F = \frac{{S_X^*}^2}{{S_Y^*}^2}$$

follows $\mathcal{F}(n_1 - 1, n_2 - 1)$ distribution under H_0. However in the F-table only values at least 1 can be seen. So our test statistic is actually

$$F^* = \max\{\frac{{S_X^*}^2}{{S_Y^*}^2}, \frac{{S_Y^*}^2}{{S_X^*}^2}\} \geq 1$$

and the rejection region is

$$R = \{F^* \geq F_{\alpha/2}(f_1 - 1, f_2 - 1)\},$$

where f_1 is the sample size of the sample having the largest, while f_2 is the size of the sample having the smallest empirical variance.

5. Sample size, confidence interval for the population standard deviation, and testing the Pearson's correlation coefficient

Problem: Determine the sample size so that to estimate the populaton mean μ with precision ε and confidence $1 - \alpha$. (Assume that the population standard deviation σ_0 is given.)

Solution. For „large" n, the radius of the confidence interval around $\bar{X}$ satisfies

$$\frac{z_{\alpha/2}\sigma_0}{\sqrt{n}} \leq \varepsilon,$$

from where $n \geq (\frac{z_{\alpha/2}\sigma_0}{\varepsilon})^2 = \frac{z_{\alpha/2}^2\sigma_0^2}{\varepsilon^2}$.

Note that the Chebyshev inequality gives a much larger bound for n:

$$\mathbb{P}(|\bar{X} - \mu| \leq \varepsilon) \geq 1 - \frac{\mathrm{Var}(\bar{X})}{\varepsilon^2} = 1 - \frac{\sigma_0^2}{n\varepsilon^2} \geq 1 - \alpha,$$

from where $n \geq \frac{\sigma_0^2}{\alpha\varepsilon^2}$.

But $z_{\alpha/2}^2 < \frac{1}{\alpha}$ for „small" α's. Note that if the precision ε is decreased or the confidence $1 - \alpha$ is increased (α is decreased), then the sample size n is increased.

□

Problem: Construct confidence interval of level $1-\alpha$ for the population standard deviation.

Solution. By Lukács' theorem, $\frac{(n-1)S_n^{*2}}{\sigma^2} \sim \chi^2(n-1)$, so

$$\mathbb{P}\left(\chi^2_{1-\alpha/2}(n-1) < \frac{(n-1)S_n^{*2}}{\sigma^2} < \chi^2_{\alpha/2}(n-1)\right) = 1-\alpha,$$

where $\chi^2_\alpha(n)$ is the $(1-\alpha)$ quantile value of the $\chi^2(n)$ distribution. Therefore,

$$\mathbb{P}\left(\frac{(n-1)S_n^{*2}}{\chi^2_{\alpha/2}} < \sigma^2 < \frac{(n-1)S_n^{*2}}{\chi^2_{1-\alpha/2}}\right) = 1-\alpha$$

and

$$\mathbb{P}\left(\sqrt{\frac{(n-1)S_n^{*2}}{\chi^2_{\alpha/2}}} < \sigma < \sqrt{\frac{(n-1)S_n^{*2}}{\chi^2_{1-\alpha/2}}}\right) = 1-\alpha,$$

from where the bounds are obvious. □

For testing independence of two Gaussian rv's, we test their Pearson's correlation coefficient, based on the sample correlation coefficient R_n.

$$H_0 \,:\, r = 0 \quad \text{versus} \quad H_1 \,:\, r \neq 0.$$

The test statistic

$$t = \sqrt{n-2}\,\frac{R_n}{1-R_n^2}$$

follows Student t-distribution with $df = n-2$ under H_0. Therefore a test, similar to the classical t-test can be performed.