

## 7 The statistical inferences based on two independent samples from the normal distribution.

In this chapter we will be concerned with two independent samples, where the first sample follows a distribution  $N(\mu_1, \sigma_1^2)$  and the second one follows a distribution  $N(\mu_2, \sigma_2^2)$ . We will do inferences (interval estimation and hypothesis testing) about parametric a function  $\mu_1 - \mu_2$  or  $\frac{\sigma_1^2}{\sigma_2^2}$ .

Statistics derived from sample means and sample variances of mentioned two samples and assertions about their distributions are stated in following theorem.

### Theorem 7.1

Consider two independent samples. Let  $X_{11}, \dots, X_{1n_1}$  be a random sample from normal distribution  $N(\mu_1, \sigma_1^2)$  and  $X_{21}, \dots, X_{2n_2}$  be a random sample from normal distribution  $N(\mu_2, \sigma_2^2)$ , whereas  $n_1 \geq 2$ ,  $n_2 \geq 2$ . Let us denote  $M_1, M_2$  as sample means,  $S_1^2, S_2^2$  as sample variances and  $S_*^2 = \frac{(n_1-1)S_1^2 + (n_2-1)S_2^2}{n_1+n_2-2}$  as weighted mean of sample variances. Then :

1. Statistics  $(M_1 - M_2)$  and  $S_*^2$  are independent.
2.  $U = \frac{(M_1 - M_2) - (\mu_1 - \mu_2)}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}} \sim N(0, 1)$ , thus  $M_1 - M_2 \sim N\left(\mu_1 - \mu_2, \frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}\right)$   
[Pivotal statistic  $U$  is instrumental towards inferences about  $\mu_1 - \mu_2$ , when  $\sigma_1^2, \sigma_2^2$  is known.]
3. If  $\sigma_1^2 = \sigma_2^2 =: \sigma^2$ , then  
 $K = \frac{(n_1+n_2-2)S_*^2}{\sigma^2} \sim \chi^2(n_1 + n_2 - 2)$   
[Pivotal statistic  $K$  is instrumental towards inferences about common  $\sigma^2$ .]
4. If  $\sigma_1^2 = \sigma_2^2 =: \sigma^2$ , then  
 $T = \frac{(M_1 - M_2) - (\mu_1 - \mu_2)}{S_* \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} \sim t(n_1 + n_2 - 2)$   
[Pivotal statistic  $T$  is instrumental towards inferences about  $\mu_1 - \mu_2$ , when  $\sigma_1^2, \sigma_2^2$  are unknown, but equal.]
5.  $F = \frac{S_1^2/S_2^2}{\sigma_1^2/\sigma_2^2} \sim F(n_1 - 1, n_2 - 1)$   
[Pivotal statistic  $F$  is instrumental towards inferences about  $\sigma_1^2/\sigma_2^2$ .]

Using the above mentioned pivotal statistics we can construct confidence intervals for parametric functions  $\mu_1 - \mu_2$  and  $\sigma_1^2/\sigma_2^2$ . Estimating  $\mu_1 - \mu_2$  we have to differentiate whether variances are known or unknown. If they are unknown, then we have to find out whether they are equal or not. The equivalence of variances may be tested by means of F-test, which will be showed later.

The following theorem states the confidence limits for mentioned parametric functions.

### Theorem 7.2

Let us consider two independent samples. Let  $X_{11}, \dots, X_{1n_1}$  be a random sample from normal distribution  $N(\mu_1, \sigma_1^2)$  and  $X_{21}, \dots, X_{2n_2}$  be a random sample from normal distribution  $N(\mu_2, \sigma_2^2)$ , whereas  $n_1 \geq 2$ ,  $n_2 \geq 2$ . Further let us consider the confidence level  $100(1 - \alpha)\%$ . Then

1. The confidence interval for  $\mu_1 - \mu_2$ , when  $\sigma_1^2, \sigma_2^2$  are known, is derived from pivotal statistic  $U = \frac{(M_1 - M_2) - (\mu_1 - \mu_2)}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}} \sim N(0, 1)$ . Thus the limits are for:

$$\begin{aligned}
\text{two-sided c.i. } (d, h) &= \left( m_1 - m_2 - \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}} \cdot u_{1-\alpha/2} \quad , \quad m_1 - m_2 + \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}} \cdot u_{1-\alpha/2} \right) \\
\text{left-sided c.i. } (d, \infty) &= \left( m_1 - m_2 - \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}} \cdot u_{1-\alpha} \quad , \quad \infty \right) \\
\text{right-sided c.i. } (-\infty, h) &= \left( -\infty \quad , \quad m_1 - m_2 + \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}} \cdot u_{1-\alpha} \right)
\end{aligned}$$

2. The confidence interval for equal unknown variance  $\sigma^2$  is derived from pivotal statistic

$K = \frac{(n_1+n_2-2)S_*^2}{\sigma^2} \sim \chi^2(n_1+n_2-2)$ . Thus the limits are for:

$$\begin{aligned}
\text{two-sided c.i. } (d, h) &= \left( \frac{(n_1+n_2-2)s_*^2}{\chi_{1-\alpha/2}^2(n_1+n_2-2)} \quad , \quad \frac{(n_1+n_2-2)s_*^2}{\chi_{\alpha/2}^2(n_1+n_2-2)} \right) \\
\text{left-sided c.i. } (d, \infty) &= \left( \frac{(n_1+n_2-2)s_*^2}{\chi_{1-\alpha}^2(n_1+n_2-2)} \quad , \quad \infty \right) \\
\text{right-sided c.i. } (-\infty, h) &= \left( -\infty \quad , \quad \frac{(n_1+n_2-2)s_*^2}{\chi_{\alpha}^2(n_1+n_2-2)} \right)
\end{aligned}$$

3. The confidence interval for  $\mu_1 - \mu_2$ , when  $\sigma_1^2, \sigma_2^2$  are unknown but equal, is derived from pivotal statistic

$T = \frac{(M_1-M_2)-(\mu_1-\mu_2)}{S_*\sqrt{\frac{1}{n_1}+\frac{1}{n_2}}} \sim t(n_1+n_2-2)$ . Thus the limits are for:

$$\begin{aligned}
\text{two-sided c.i. } (d, h) &= \left( m_1 - m_2 - s_*\sqrt{\frac{1}{n_1} + \frac{1}{n_2}} \cdot t_{1-\alpha/2}(n_1+n_2-2) \quad , \right. \\
&\quad \left. m_1 - m_2 + s_*\sqrt{\frac{1}{n_1} + \frac{1}{n_2}} \cdot t_{1-\alpha/2}(n_1+n_2-2) \right) \\
\text{left-sided c.i. } (d, \infty) &= \left( m_1 - m_2 - s_*\sqrt{\frac{1}{n_1} + \frac{1}{n_2}} \cdot t_{1-\alpha}(n_1+n_2-2) \quad , \quad \infty \right) \\
\text{right-sided c.i. } (-\infty, h) &= \left( -\infty \quad , \quad m_1 - m_2 + s_*\sqrt{\frac{1}{n_1} + \frac{1}{n_2}} \cdot t_{1-\alpha}(n_1+n_2-2) \right)
\end{aligned}$$

4. The confidence interval for the ratio of variances  $\sigma_1^2/\sigma_2^2$  is derived from pivotal statistic

$F = \frac{S_1^2/S_2^2}{\sigma_1^2/\sigma_2^2} \sim F(n_1-1, n_2-1)$ . Thus the limits are for:

$$\begin{aligned}
\text{two-sided c.i. } (d, h) &= \left( \frac{s_1^2/s_2^2}{F_{1-\alpha/2}(n_1-1, n_2-1)} \quad , \quad \frac{s_1^2/s_2^2}{F_{\alpha/2}(n_1-1, n_2-1)} \right) \\
\text{left-sided c.i. } (d, \infty) &= \left( \frac{s_1^2/s_2^2}{F_{1-\alpha}(n_1-1, n_2-1)} \quad , \quad \infty \right) \\
\text{right-sided c.i. } (-\infty, h) &= \left( -\infty \quad , \quad \frac{s_1^2/s_2^2}{F_{\alpha}(n_1-1, n_2-1)} \right)
\end{aligned}$$

### Remark 7.3

If the assumption of equal variances does not hold in the point 3 of previous theorem, the approximative  $100(1-\alpha)\%$  confidence interval for  $\mu_1 - \mu_2$  may be constructed. In this case the test statistic  $T$  has an approximative student's distribution  $t(\nu)$ , where the degrees of freedom are calculated as follows:

$$\nu = \frac{(s_1^2/n_1 + s_2^2/n_2)^2}{\frac{(s_1^2/n_1)^2}{n_1-1} + \frac{(s_2^2/n_2)^2}{n_2-1}} \quad \text{so called Welch's approximation}$$

If  $\nu$  is not an integer number, then use linear interpolation.

### Example 7.4

In two tanks the chlorine content was tested (g/l). 25 specimens was drawn from the first tank, 10 specimen was drawn from the second one. The realization of sample statistics follow:  $m_1 = 34,48$ ,  $m_2 = 35,59$ ,  $s_1^2 = 1,7482$ ,  $s_2^2 = 1,7121$ . The values of specimens are assumed to be realization of two independent samples drawn from normal distributions  $N(\mu_1, \sigma^2)$  and  $N(\mu_2, \sigma^2)$ . Determine the 95% empirical confidence interval for the difference between expected values  $\mu_1 - \mu_2$ .

### Solution

We have to construct the confidence interval for  $\mu_1 - \mu_2$ , when the variances  $\sigma_1^2$ ,  $\sigma_2^2$  are unknown but equal. It can be derived from pivotal statistic

$$T = \frac{(M_1 - M_2) - (\mu_1 - \mu_2)}{S_* \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} \sim t(n_1 + n_2 - 2).$$

We will need the quantile  $t_{1-\alpha/2}(n_1 + n_2 - 2) = t_{0,975}(33) = 2,035$

and the realization of the weighted mean of sample variances  $s_*^2 = \frac{(n_1-1)s_1^2 + (n_2-1)s_2^2}{n_1+n_2-2} = \frac{24.1,7482+9.1,7121}{33} = 1,7384$ .

---

Thus the confidence limits are :

$$\begin{aligned} d &= m_1 - m_2 - s_* \sqrt{\frac{1}{n_1} + \frac{1}{n_2}} \cdot t_{1-\alpha/2}(n_1 + n_2 - 2) = \\ &= 34,48 - 35,59 - \sqrt{1,7384} \cdot \sqrt{1/25 + 1/10} \cdot 2,035 = -2,114 \end{aligned}$$

$$\begin{aligned} h &= m_1 - m_2 + s_* \sqrt{\frac{1}{n_1} + \frac{1}{n_2}} \cdot t_{1-\alpha/2}(n_1 + n_2 - 2) = \\ &= 34,48 - 35,59 + \sqrt{1,7384} \cdot \sqrt{1/25 + 1/10} \cdot 2,035 = -0,106 \end{aligned}$$

Hence  $\mu_1 - \mu_2 \in (-2,114 \text{ g/l}, -0,106 \text{ g/l})$  with the probability at least 0,95.

### Example 7.5

Consider previous example assuming that given independent samples are from distributions  $N(\mu_1, \sigma_1^2)$  and  $N(\mu_2, \sigma_2^2)$ . Determine the 95% empirical confidence interval for the ratio of variances.

### Solution

We have to construct the confidence interval for the ratio of variances  $\sigma_1^2/\sigma_2^2$ . It can be derived from pivotal statistic

$$F = \frac{S_1^2/S_2^2}{\sigma_1^2/\sigma_2^2} \sim F(n_1 - 1, n_2 - 1).$$

We will need the quantile  $F_{1-\alpha/2}(n_1 - 1, n_2 - 1) = F_{0,972}(24, 9) = 3,6142$  a

$$F_{\alpha/2}(n_1 - 1, n_2 - 1) = F_{0,025}(24, 9) = \frac{1}{F_{0,975}(9,24)} = \frac{1}{2,7027}$$

---

Thus the confidence limits are:

$$d = \frac{s_1^2/s_2^2}{F_{1-\alpha/2}(n_1-1, n_2-1)} = \dots = 0,28$$

$$h = \frac{s_1^2/s_2^2}{F_{\alpha/2}(n_1-1, n_2-1)} = \dots = 2,76$$

Hence  $P(\sigma_1^2/\sigma_2^2 \in (0,28 ; 2,76)) \geq 0,95$

### Definition 7.6

Let us consider two independent samples. Let  $X_{11}, \dots, X_{1n_1}$  be a random sample from normal distribution  $N(\mu_1, \sigma_1^2)$  and  $X_{21}, \dots, X_{2n_2}$  be a random sample from normal distribution  $N(\mu_2, \sigma_2^2)$ , whereas  $n_1 \geq 2$ ,  $n_2 \geq 2$ . Let  $c$  be a constant.

- (i.) Assume that  $\sigma_1^2, \sigma_2^2$  is known.  
Then the test  $H_0 : \mu_1 - \mu_2 = c$  versus  $H_1 : \mu_1 - \mu_2 \neq c$  (eventually  $H_1 : \mu_1 - \mu_2 < c$  eventually  $H_1 : \mu_1 - \mu_2 > c$ ) is called the *two-sample z-test*.
- (ii.) Assume that  $\sigma_1^2, \sigma_2^2$  is unknown, but the equation  $\sigma_1^2 = \sigma_2^2$  holds.  
Then the test  $H_0 : \mu_1 - \mu_2 = c$  versus  $H_1 : \mu_1 - \mu_2 \neq c$  (eventually  $H_1 : \mu_1 - \mu_2 < c$  eventually  $H_1 : \mu_1 - \mu_2 > c$ ) is called the *two-sample t-test*.
- (iii.) The test  $H_0 : \frac{\sigma_1^2}{\sigma_2^2} = 1$  versus  $H_1 : \frac{\sigma_1^2}{\sigma_2^2} \neq 1$  (eventually  $H_1 : \frac{\sigma_1^2}{\sigma_2^2} < 1$  eventually  $H_1 : \frac{\sigma_1^2}{\sigma_2^2} > 1$ ) is called the *F-test*.

### Remark 7.7

The selection of an appropriate test statistic corresponding to particular test is analogous to the selection of an appropriate pivotal statistic in 7.2, thus for two-sample z-test the test statistic  $T_0$  is derived from statistic  $U$ , for two-sample t-test it is derived from statistic  $T$  and for F-test it is derived from statistic  $F$ .

### Theorem 7.8

Let us consider two independent samples. Let  $X_{11}, \dots, X_{1n_1}$  be a random sample from normal distribution  $N(\mu_1, \sigma_1^2)$  and  $X_{21}, \dots, X_{2n_2}$  be a random sample from normal distribution  $N(\mu_2, \sigma_2^2)$ , whereas  $n_1 \geq 2, n_2 \geq 2$ . Let  $c$  be a constant.

1. Considering two-sample z-test at the significance level  $\alpha$  the null hypothesis  $H_0$  is rejected in favor of alternative hypothesis  $H_1$ , if the realization of the test statistic

$$T_0 = \frac{M_1 - M_2 - c}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}}$$

falls within the critical region  $W$ . According to the form of the alternative hypothesis the list of corresponding critical regions follows :

two-tailed test  $H_1 : \mu_1 - \mu_2 \neq c \quad W = (-\infty, -u_{1-\alpha/2}) \cup \langle u_{1-\alpha/2}, \infty \rangle$

left-tailed test  $H_1 : \mu_1 - \mu_2 < c \quad W = (-\infty, -u_{1-\alpha})$

right-tailed test  $H_1 : \mu_1 - \mu_2 > c \quad W = \langle u_{1-\alpha}, \infty \rangle$

2. Considering two-sample t-test at the significance level  $\alpha$  the null hypothesis  $H_0$  is rejected in favor of alternative hypothesis  $H_1$ , if the realization of the test statistic

$$T_0 = \frac{M_1 - M_2 - c}{S_* \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}$$

two-tailed  $H_1 : \mu_1 - \mu_2 \neq c \quad W = (-\infty, -t_{1-\alpha/2}(n_1 + n_2 - 2)) \cup \langle t_{1-\alpha/2}(n_1 + n_2 - 2), \infty \rangle$

left-tailed  $H_1 : \mu_1 - \mu_2 < c \quad W = (-\infty, -t_{1-\alpha}(n_1 + n_2 - 2))$

right-tailed  $H_1 : \mu_1 - \mu_2 > c \quad W = \langle t_{1-\alpha}(n_1 + n_2 - 2), \infty \rangle$

3. Considering F-test at the significance level  $\alpha$  the null hypothesis  $H_0$  is rejected in favor of alternative hypothesis  $H_1$ , if the realization of the test statistic

$$T_0 = \frac{S_1^2/S_2^2}{1}$$

two-tailed  $H_1 : \sigma_1^2/\sigma_2^2 \neq 1 \quad W = (0, F_{\alpha/2}(n_1 - 1, n_2 - 1)) \cup \langle F_{1-\alpha/2}(n_1 - 1, n_2 - 1), \infty \rangle$

left-tailed  $H_1 : \sigma_1^2/\sigma_2^2 < 1 \quad W = (0, F_{\alpha}(n_1 - 1, n_2 - 1))$

right-tailed  $H_1 : \sigma_1^2/\sigma_2^2 > 1 \quad W = \langle F_{1-\alpha}(n_1 - 1, n_2 - 1), \infty \rangle$

### Example 7.9

In a restaurant "White Pony" the servicing time was measured 20 times. The results follow: 6, 8, 11, 4, 7, 6, 10, 6, 9, 8, 5, 12, 13, 10, 9, 8, 7, 11, 10, 5. In a restaurant "Golden Lion" the same measurement was kept 15 times and the results come next: 9, 11, 10, 7, 6, 4, 8, 13, 5, 15, 8, 5, 6, 8, 7. Assuming that both samples are independent and normally distributed use 0.05 significance level and test equality of the mean values of the servicing time in both restaurants.

### Solution

At the significance level 0.05 we are testing  $H_0 : \mu_1 - \mu_2 = 0$  versus  $H_1 : \mu_1 - \mu_2 \neq 0$ , which is a two-sample z-test. This test can be used only if the assumption of equal variances holds. This equality has to be tested first. The F-test is a good instrumental towards it.

$$m_1 = 8, 25; \quad m_2 = 8, 13; \quad s_1^2 = 6, 307; \quad s_2^2 = 9, 41;$$

$$s_* = \frac{(n_1-1)S_1^2 + (n_2-1)S_2^2}{n_1+n_2-2} = \frac{19 \cdot 6,307 + 14 \cdot 9,41}{19+14} = 7,623$$

Thus, using significance level 0.05 we test hypothesis

•  $H_0 : \frac{\sigma_1^2}{\sigma_2^2} = 1$  proti  $H_1 : \frac{\sigma_1^2}{\sigma_2^2} \neq 1$ .

The test statistic:

$T_0 = \frac{S_1^2/S_2^2}{1}$  , the numerical realization  $t_0 = \frac{6,307}{9,41} = 0,6702$ .

The critical region:

$W = \langle 0, F_{\alpha/2}(n_1 - 1, n_2 - 1) \rangle \cup \langle F_{1-\alpha/2}(n_1 - 1, n_2 - 1), \infty \rangle =$   
 $\langle 0, F_{0,025}(19, 14) \rangle \cup \langle F_{0,975}(19, 14), \infty \rangle =$   
 $\langle 0; \frac{1}{F_{0,975}(14,19)} \rangle \cup \langle 2, 8607; \infty \rangle = \langle 0; 0,3778 \rangle \cup \langle 2, 8607; \infty \rangle$

$t_0 \notin W$ , thus  $H_0$  assuming equality of variances is not rejected at the significance level 0.05. The two-sample t-test may follow.

•  $H_0 : \mu_1 - \mu_2 = 0$  versus  $H_1 : \mu_1 - \mu_2 \neq 0$

Test statistic:

$T_0 = \frac{M_1 - M_2 - c}{S_* \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}$  , the numerical realization:  $t_0 = \frac{8,25 - 8,13}{\sqrt{7,623} \sqrt{\frac{1}{20} + \frac{1}{15}}} = 0,124$

The critical region:

$W = (-\infty, -t_{1-\alpha/2}(n_1 + n_2 - 2)) \cup \langle t_{1-\alpha/2}(n_1 + n_2 - 2), \infty \rangle =$   
 $= (-\infty, -t_{0,975}(33)) \cup \langle t_{0,975}(33), \infty \rangle = (-\infty, -2,035) \cup \langle 2,035, \infty \rangle$

Since  $t_0 \notin W$ ,  $H_0$  is not rejected at the significance level 0.05. [Thus the data does not give evidence against the equality of servicing time.]