MUNI
ECON

Lecture 5

One-Sample Inference for Normal Data

Pavel Morcinek
474416@mail.muni.cz

Department of Mathematics and Statistics, Masaryk University

March 18, 2024

Many of the random variables which can be met in the practise follow a normal distribution. Moreover, using the *central limit theorem*, we can approximate other distributions by a normal distribution under fairly general conditions. Therefore, it is very important to pay attention to the samples from the normal distribution.

The normal distribution is characterized by two parameters, the **mean** μ and the **variance** σ^2. We will therefore solve problems involving these parameters. These tasks consist of constructing estimates and testing hypotheses.

The following statement gives the distribution of commonly used test criteria for tests for one random sample from a normal distribution.

Pivot and Test Statistics

Let $X_1, \ldots, X_n$ be a random sample from normal distribution $N(\mu, \sigma^2)$. Then the following applies:

1. Sample mean $M = \frac{1}{n} \sum_{i=1}^{n} X_i$ and the sample variance $S^2 = \sum_{i=1}^{n} (X_i - M)^2$ are independent.

2. $Z = \frac{M-\mu}{\frac{\sigma}{\sqrt{n}}} \sim N(0,1), \quad$ thus $\quad M \sim N(\mu, \frac{\sigma^2}{n})$
   (The pivot statistic Z is used to solve problems **about** μ when we **know** σ^2.)

3. $K = \frac{(n-1)S^2}{\sigma^2} \sim \chi^2(n-1)$
   (The pivot statistic K is used to solve problems **about** σ^2 when μ is **unknown**.)

4. $T = \frac{M-\mu}{\frac{S}{\sqrt{n}}} \sim t(n-1)$
   (The pivot statistic T is used to solve problems **about** μ when σ^2 is **unknown**.)

5. $\frac{\sum_{i=1}^{n}(X_i-\mu)^2}{\sigma^2} \sim \chi^2(n)$
   (This pivot statistic is used to solve problems **about** σ^2 when we **know** μ.)

Distribution parameter properties:

$$z_\alpha = -z_{1-\alpha} \qquad t_\alpha(\nu) = t_{1-\alpha}(\nu) \qquad F_\alpha(\nu_1, \nu_2) = \frac{1}{F_{1-\alpha}(\nu_1, \nu_2)}$$

Confidence Intervals

A common task of a statistician is to derive confidence intervals for unknown parameters. In the case of a normal distribution, these are the parameters μ and σ^2. Thus, four situations can occur. We look for a confidence interval:

1. for μ when we know σ^2;
2. for σ^2 when μ is unknown;
3. for μ when σ^2 is unknown;
4. for σ^2 when we know μ.

m and *s* denote the realizations of the sample mean *M* and sample standart deviation *S*.

When constructing confidence intervals, we need to know which pivot statistic is appropriate for which of these four options. It is then easy to derive the lower and upper estimates respectively. This derivation for the first option – for μ when σ^2 is known is part of the previous lecture. How the estimates for the remaining options would turn out using the same procedure is given in the following statement.

Let $X_1, \ldots, X_n$ be a random sample from a normal distribution $N(\mu, \sigma^2)$. Assume a $100(1-\alpha)\%$ confidence interval.

Remind that M and S denote random variables (theoretical models), while m and s denote their concrete realizations (computed from the realization $x_1, \ldots, x_n$ of our random samples $X_1, \ldots, X_n$).

Confidence Intervals

1. We derive a confidence interval for μ when we know σ^2 from the pivot statistic:

$$Z = \frac{M - \mu}{\frac{\sigma}{\sqrt{n}}} \sim N(0, 1).$$

Then the boundaries of CI are:

two-sided CI: $(l, u) = \left(m - \frac{\sigma}{\sqrt{n}} \cdot z_{1-\alpha/2},\ m + \frac{\sigma}{\sqrt{n}} \cdot z_{1-\alpha/2}\right)$

left-sided CI: $(l, \infty) = \left(m - \frac{\sigma}{\sqrt{n}} \cdot z_{1-\alpha},\ \infty\right)$

right-sided CI: $(-\infty, u) = \left(-\infty,\ m + \frac{\sigma}{\sqrt{n}} \cdot z_{1-\alpha}\right)$

2. We derive a confidence interval for σ^2 when μ is unknown from the pivot statistic:

$$K = \frac{(n-1)S^2}{\sigma^2} \sim \chi^2(n-1).$$

Then the boundaries of CI are:

two-sided CI: $$(l, u) = \left(\frac{(n-1)s^2}{\chi^2_{1-\alpha/2}(n-1)}, \frac{(n-1)s^2}{\chi^2_{\alpha/2}(n-1)} \right)$$

left-sided CI: $$(l, \infty) = \left(\frac{(n-1)s^2}{\chi^2_{1-\alpha}(n-1)}, \infty \right)$$

right-sided CI: $$(-\infty, u) = \left(-\infty, \frac{(n-1)s^2}{\chi^2_{\alpha}(n-1)} \right)$$

3. We derive a confidence interval for μ when σ^2 is unknown from the pivot statistic:

$$T = \frac{M - \mu}{\frac{S}{\sqrt{n}}} \sim t(n-1).$$

Then the boundaries of CI are:

two-sided CI: $$(l, u) = \left(m - \tfrac{s}{\sqrt{n}} \cdot t_{1-\alpha/2}(n-1),\ m + \tfrac{s}{\sqrt{n}} \cdot t_{1-\alpha/2}(n-1)\right)$$

left-sided CI: $$(l, \infty) = \left(m - \tfrac{s}{\sqrt{n}} \cdot t_{1-\alpha}(n-1),\ \infty\right)$$

right-sided CI: $$(-\infty, u) = \left(-\infty,\ m + \tfrac{s}{\sqrt{n}} \cdot t_{1-\alpha}(n-1)\right)$$

4. We derive a confidence interval for σ^2 when we know μ from the pivot statistic:

$$\frac{\sum_{i=1}^{n}(X_i-\mu)^2}{\sigma^2} \sim \chi^2(n).$$

Then the boundaries of CI are:

$$\textbf{two-sided CI:} \quad (l,u) = \left(\frac{\sum_{i=1}^{n}(x_i-\mu)^2}{\chi^2_{1-\alpha/2}(n)} , \frac{\sum_{i=1}^{n}(x_i-\mu)^2}{\chi^2_{\alpha/2}(n)} \right)$$

$$\textbf{left-sided CI:} \quad (l,\infty) = \left(\frac{\sum_{i=1}^{n}(x_i-\mu)^2}{\chi^2_{1-\alpha}(n)} , \infty \right)$$

$$\textbf{right-sided CI:} \quad (-\infty,u) = \left(-\infty , \frac{\sum_{i=1}^{n}(x_i-\mu)^2}{\chi^2_{\alpha}(n)} \right)$$

Reminder

For computations of quantiles of different distributions we use the built-in functions in **R** starting with `q`:

for **normal** quantiles:	`qnorm(`*prob*, *mean*, *sd*`)`
for **t** quantiles:	`qt(`*prob*, *df*`)`
for **chi-squared** quantiles:	`qchisq(`*prob*, *df*`)`
for **F** quantiles:	`qf(`*prob*, *df1*, `df2)`

For other properties (like density computation, cumulative distribution function or random number generating) we use functions starting with `d` (density), `p` (distribution), `r` (random).

Example 1

A certain constant μ was measured 10 times independently. The measurement results were:

$$2 \quad 1.8 \quad 2.1 \quad 2.4 \quad 1.9 \quad 2.1 \quad 2 \quad 1.8 \quad 2.3 \quad 2.2$$

These results are considered as realizations of a random sample $X_1, \ldots, X_n$ from the distribution $N(\mu, \sigma^2)$ where μ, σ^2 are unknown. Find 95% confidence interval for a parameter μ:

1. two-sided
2. left-sided
3. right-sided

Example 1

Solution

This is the confidence interval for μ when σ^2 is unknown. We use the statistic $T = \frac{M-\mu}{\frac{S}{\sqrt{n}}} \sim t(n-1)$ to derive the boundaries. Quantiles can be found using function `qt()`.

$$n = 10, \quad \alpha = 0.05, \quad m = 2.06, \quad s^2 = 0.0404$$

1. $l = m - \frac{s}{\sqrt{n}} \cdot t_{1-\alpha/2}(n-1) = 2.06 - \frac{0.2011}{\sqrt{10}} \cdot 2.2622 = 1.92$
   $u = m + \frac{s}{\sqrt{n}} \cdot t_{1-\alpha/2}(n-1) = 2.06 - \frac{0.2011}{\sqrt{10}} \cdot 2.2622 = 2.20$
2. $l = m - \frac{s}{\sqrt{n}} \cdot t_{1-\alpha}(n-1) = 2.06 - \frac{0.2011}{\sqrt{10}} \cdot 1.9331 = 1.94$
3. $u = m + \frac{s}{\sqrt{n}} \cdot t_{1-\alpha}(n-1) = 2.06 + \frac{0.2011}{\sqrt{10}} \cdot 1.8331 = 2.18$

So far we have looked at confidence intervals for the parameters of the normal distribution, now we will look at hypothesis testing on the parameters μ and σ^2. We will look at testing using the critical region, other ways of testing can be easily derived.

Hypothesis Testing

The procedure is the same as in the previous weeks: we will use the (most common) critical-value approach. Firstly, we will find the right test statistic (which consists of appropriate known and unknown terms), compute its quantiles (to quantify how probable our realization of the test statistic is) and compare them with our realization of the test statistic.

The choice of the appropriate test statistic for the chosen test is similar to the choice of the appropriate pivot random variable, i.e. for the z-test we choose the test criterion T_0 derived from the Z statistic, for the t-test from the T statistic and for the test of variance from the K statistic.

Let $X_1, \ldots, X_n \sim N(\mu, \sigma^2)$ be our random sample, $c \in \mathbb{R}$, $n \geq 2$. There are 3 types of tests according to the parameters we know or we want to test.

Hypothesis Testing

1. Let $X_1, \ldots, X_n$ be a random sample from $N(\mu, \sigma^2)$ where σ^2 is known. The hypothesis are:

   $H_0 : \mu = c$
   $H_1 : \mu \neq c$ (or one-sided versions)

   We reject H_0 in favour of the alternative H_1 when the test criterion: $T_0 = \frac{M-c}{\frac{\sigma}{\sqrt{n}}}$ is realized in the domain W, where:

   for **two-tailed**: $W = (-\infty, -z_{1-\alpha/2}\rangle \cup \langle z_{1-\alpha/2}, \infty)$

   for **left-tailed**: $W = (-\infty, -z_{1-\alpha}\rangle$

   for **right-tailed**: $W = \langle z_{1-\alpha}, \infty)$

2. Let $X_1, \ldots, X_n$ be a random sample from $N(\mu, \sigma^2)$ where σ^2 is unknown. The hypothesis are:

   $H_0 : \mu = c$
   $H_1 : \mu \neq c$ (or one-sided versions)

   We reject H_0 in favour of the alternative H_1 when the test criterion: $T_0 = \frac{M-c}{\frac{S}{\sqrt{n}}}$ is realized in the domain W, where:

   for **two-tailed**: $W = (-\infty, -t_{1-\alpha/2}(n-1)\rangle \cup \langle t_{1-\alpha/2}(n-1), \infty)$

   for **left-tailed**: $W = (-\infty, -t_{1-\alpha}(n-1)\rangle$

   for **right-tailed**: $W = \langle t_{1-\alpha}(n-1), \infty)$

3. Let $X_1, \ldots, X_n$ be a random sample from $N(\mu, \sigma^2)$ where μ is unknown. The hypothesis are:

$H_0 : \sigma^2 = c$
$H_1 : \sigma^2 \neq c$ (or one-sided versions)

We reject H_0 in favour of the alternative H_1 when the test criterion: $T_0 = \frac{(n-1)S^2}{c}$ is realized in the domain W, where:

for **two-tailed**: $W = (0, \chi^2_{\alpha/2}(n-1)\rangle \cup \langle \chi^2_{1-\alpha/2}(n-1), \infty)$

for **left-tailed**: $W = (0, \chi^2_{\alpha}(n-1)\rangle$

for **right-tailed**: $W = \langle \chi^2_{1-\alpha}(n-1), \infty)$

Therminology

The test of $H_0 : \mu = c$ against $H_1 : \mu \neq c$ (or one-sided alternatives) with known σ is called *z*-test.

The test of $H_0 : \mu = c$ against $H_1 : \mu \neq c$ (or one-sided alternatives) with unknown σ is called one-sample *t*-test.

The test of $H_0 : \sigma^2 = c$ against $H_1 : \sigma^2 \neq c$ (or one-sided alternatives) with unknown μ is called test of variance.

Example 2

According to the information on the chocolate packaging, the net weight of the chocolate should be 125 g. The producer has received several complaints from buyers claiming that the weight of the chocolates is lesser than the declared 125 g.

For this reason, the control department has randomly selected 50 chocolates and found that their average weight was 122 (sample mean realization) g and the standard deviation was 8.6 g (sample sd realization).

Assuming that the weight of the chocolates follows a normal distribution, can we consider the buyers complaints justified at the $\alpha = 0.01$ level of significance?

Example 2

Solution

$X_1, \ldots, X_{50} \sim N(\mu, \sigma^2)$. The parameter σ^2 is unknown, thus we will use one-sample t-test.

$H_0 : \mu \geq 125$, $H_1 : \mu < 125$ (left-tailed alternative, "producer is lying")

Test statistic	Its realization	Critical region
$T_0 = \frac{M-c}{\frac{S}{\sqrt{n}}}$	$t_0 = \frac{122-125}{\frac{8.6}{\sqrt{50}}} = -2.4667$	$W = (-\infty, -t_{1-\alpha}(n-1)\rangle$ $= (-\infty, -t_{0.99}(49)\rangle$ $= (-\infty, -2.4049\rangle$

Because $t_0 \in W$, we **reject** H_0 at significance level 0.01. The **producer is cheating** with a probability of mistake in our conclusion no more than 1%.

Paired Tests

Let's assume we are observing two independant variables (the first might represent men, the second women measurements etc.) – so we are observing two random samples $(X_1, \ldots, X_n)$ and $(Y_1, \ldots, Y_n)$, which are assumed to be **normally distributed**:

$$X_i \quad \sim \quad N(\mu_1, \sigma_1^2), \qquad Y_i \quad \sim \quad N(\mu_2, \sigma_2^2).$$

The commom situation we meet in the practise is to compare these two random samples (for example average salaries of men and women), so to compare μ_1 and μ_2.

In the case of the same sample sizes n $(= n_1 = n_2)$ we can use the knowledge from the previous course: the **sum** (and the **difference** as well) of two normally distributed random variables is again a **normally distributed** random variable.

So the following holds:

$$X_i - Y_i \quad \sim \quad N(\mu_1 - \mu_2, \sigma_1^2 + \sigma_2^2)$$

Thus, we can test the hypothesis and construct CI's for $\mu_1 - \mu_2$ using this new random sample: $Z_i = X_i - Y_i$. For example, the boundaries of $100(1-\alpha)\%$ CI are:

$$(l, u) = (m_1 - m_2) - \frac{\sqrt{\sigma_1^2 + \sigma_2^2}}{\sqrt{n}} \cdot z_{1-\alpha/2}$$

$$(m_1 - m_2) + \frac{\sqrt{\sigma_1^2 + \sigma_2^2}}{\sqrt{n}} \cdot z_{1-\alpha/2}$$

for **known variances**. If we **do not know** the variances σ_1^2 and σ_2^2, we would replace them by the sample realizations s_1^2 and s_2^2 and the normal quantiles $z_{1-\alpha/2}$ by the t-quantiles $t_{1-\alpha/2}(n-1)$.

Example 3

We are observing the concentration of the specific chemical substance in the blood of the patients. We have measured 5 men and 5 women.

patient	1	2	3	4	5
men	2.3	1.9	2.1	2.4	2.6
women	2.4	2.0	2.0	2.3	2.5

Assuming that the data are from a two-dimensional normal distribution, construct a 90% empirical confidence interval for the difference in expected values of the results for men and women.

Example 3

Solution

We will work with a new random sample Z_i of difference $X_i - Y_i$: $Z = (-0.1, -0.1, 0.1, 0.1, 0.1)$. We wil construct demanded CI for this normally distributed random sample $Z_i \sim N(\mu_1 - \mu_2, \sigma_1^2 + \sigma_2^2)$. We do not know variances σ_1^2, σ_2^2, so we have to use t-quantiles.

$$m = m_1 - m_2 = 0.02, \quad s^2 = s_1^2 + s_2^2 = 0.012, \quad n = 5$$

$$l = m - \frac{s}{\sqrt{n}} \cdot t_{1-\alpha/2}(n-1) = 0.02 - \frac{0.109545}{\sqrt{5}} \cdot 2.1318 = -0.0844$$
$$u = m + \frac{s}{\sqrt{n}} \cdot t_{1-\alpha/2}(n-1) = 0.02 + \frac{0.109545}{\sqrt{5}} \cdot 2.1318 = 0.1244$$

With the probability at least 0.90 the following applies:

$$-0.0844 < \mu_1 - \mu_2 < 0.1244.$$

Example 4

The following table shows the return achieved by 12 randomly selected companies when investing in international business (X) and national business (Y):

company	1	2	3	4	5	6	7	8	9	10	11	12
X	10	12	14	12	12	17	9	15	9	11	7	15
Y	11	14	15	11	13	16	10	13	11	17	9	19

The return is expressed as a percentage and represents the proportion of profit on the investment made in a year. Assuming that the data come from a two-dimensional normal distribution at the 0.1 significance level, test the hypothesis that there is no difference between investing in national and international business. Conduct the test using (a) a confidence interval, (b) a critical region.

Example 4

Solution

First we move on to the difference random sample $Z \sim N(\mu, \sigma^2)$. The realizations of its characteristics are then $m = m_1 - m_2 = -1.33$, $s^2 = s_1^2 + s_2^2 = 4.78$.

We test the hypothesis $H_0 : \mu = 0$ against $H_1 : \mu \neq 0$. We will need the quantile $t_{0.95}(11) = 1.7959$

1. $l = m - \frac{s}{\sqrt{n}} \cdot t_{1-\alpha/2}(n-1) = -1.33 - \frac{\sqrt{4.78}}{\sqrt{12}} \cdot 1.7959 = -2.4677$
   $u = m + \frac{s}{\sqrt{n}} \cdot t_{1-\alpha/2}(n-1) = -1.33 + \frac{\sqrt{4.78}}{\sqrt{12}} \cdot 1.7959 = -0.1987$

   Because $0 \notin (-2.4677, -0.1989)$, we **reject** H_0 at the significance level 0.1 (there is a difference in the investments).

Example 4

2. We will conclude the same using the critical region approach.

Test criterion: $T_0 = \frac{M-c}{\frac{S}{\sqrt{n}}}$

Its realization: $t_0 = \frac{-1.33-0}{\frac{\sqrt{4.78}}{\sqrt{12}}} = -2.11085$

The critical region: $W = (-\infty,\ -t_{1-\alpha/2}(n-1)\rangle \quad \cup$

$$\cup \quad \langle t_{1-\alpha/2}(n-1),\ \infty) =$$

$$= (-\infty,\ -1.7959\rangle \cup \langle 1.7959,\ \infty)$$

Because $t_0 \in W$, we **reject** H_0 at the significance level 0.1.

Example 4

3. Use the built-in **R** function `t.test()`:

```
> t.test(x, y, mu = 0, paired = TRUE)
```

```
        Paired t-test

data:  x and y
t = -2.1109, df = 11, p-value = 0.05849
alternative hypothesis: true mean difference is not equal to 0
95 percent confidence interval:
 -2.72359986  0.05693319
sample estimates:
mean difference
      -1.333333
```