5 Intoduction to hypothesis testing

A statistical hypothesis test is a method of making statistical decisions using experimental data. These decisions are almost always made using null-hypothesis tests; that is, ones that answer the question: "Assuming that the null hypothesis is true, what is the probability of observing a value for the test statistic that is at least as extreme as the value that was actually observed?" The critical region of a hypothesis test is the set of all outcomes which, if they occur, cause the null hypothesis to be rejected and the alternative hypothesis accepted. One use of hypothesis testing is deciding whether experimental results contain enough information to cast doubt on conventional wisdom. This means that we often introduce the null hypothesis as that proposition which we actually wish to disprove. For instance using experimental data the researcher wants to examine "new wisdom" that passive smoking is detrimental to health. So the null hypothesis is that the passive smoking is not detrimental to health and the alternative hypothesis is stated as smoking is detrimental to health.

Hypothesis testing is aimed to find conclusion of "reject" or "do not reject" the null hypothesis.

Definition 5.1
Let $X_1, \ldots, X_n$ be a random sample from distribution $L(\vartheta)$, where parameter $\vartheta \in \Theta$ is unknown, let $h(\vartheta)$ be a parametric function and $c \in \mathbf{R}$ is a constant.

(i.) $H_0 : h(\vartheta) = c \quad$ versus $\quad H_1 : h(\vartheta) \neq c$

The assertion $H_0 : h(\vartheta) = c \quad$ is called simple null hypothesis
the assertion $H_1 : h(\vartheta) \neq c \quad$ is called composite two-sided alternative hypothesis.
We speak about *two-tailed test*.

(ii.) $H_0 : h(\vartheta) \geq c \quad$ versus $\quad H_1 : h(\vartheta) < c$

The assertion $H_0 : h(\vartheta) \geq c \quad$ is called composite right-sided null hypothesis,
the assertion $H_1 : h(\vartheta) < c \quad$ is called composite left-sided alternative hypothesis.
We speak about *left-tailed test*.

(iii.) $H_0 : h(\vartheta) \leq c \quad$ versus $\quad H_1 : h(\vartheta) > c$

The assertion $H_0 : h(\vartheta) \leq c \quad$ is called composite left-sided null hypothesis,
the assertion $H_1 : h(\vartheta) > c \quad$ composite right-sided alternative hypothesis.
We speak about *right-tailed test*.

The Hypothesis testing H_0 versus H_1 is a decision procedure based on random sample $X_1, \ldots, X_n$, which leads to the result: to reject the null hypothesis in favor of alternative one or not to reject (which is not the same as accept).
The criterion for rejecting or not rejecting the null hypothesis on the basis of sample evidence is not a guarantee of arriving at a correct conclusion. Let us now consider in detail the kinds of error that could be made.

Definition 5.2
Testing H_0 versus H_1 we may commit one error of two types:
Error Type I is made when H_0 is incorrectly rejected, though H_0 is in fact true. The α stands for the probability of making the Error Type I.
Error Type II is made by failing to reject H_0, though H_0 is in fact false. The β stands for the probability of making the Error Type II.

	H_0 is not rejected	H_0 is rejected
H_0 is true	correct decision $P(H_0 \text{ is not rejected}\|H_0 \text{ is true}) = 1 - \alpha$	Type I error $P(H_0 \text{ is rejected}\|H_0 \text{ is true}) = \alpha$
H_0 is false	Type II error $P(H_0 \text{ is not rejected}\|H_0 \text{ is false}) = \beta$	correct decision $P(H_0 \text{ is rejected}\|H_0 \text{ is false}) = 1 - \beta$

The probability of making the Error Type I α is called *significance level of a test* .
The value $1 - \beta$ stands for the probability that the false null hypothesis is correctly rejected. This value is called *power of a test* .
[The statisticians wish to have a power of a test $1 - \beta$ as great as possible and simultaneously to have a level of significance α as small as possible. However by decreasing the probability α it increases the probability β and the power of the test goes down. The standard practice is to fix first the low level of significance α and then to use a test statistic (if there are more test statistics) that would make the probability β as small as possible. For a given significance level, the test with the greatest power $1 - \beta$ is called most powerful test. In statistical testing, the only way in which we could reduce the probabilities of both kinds of error at the same time is by increasing the sample size - assuming that the test statistic used is the best that can be devised.]

Remark 5.3
There are three methods of testing hypothesis H_0 versus H_1 at a level of significance α. a) classical method using critical region, b) confidence interval method, c) p-value method. It is essential to understand the first one, the others may be easily derived.

Definition 5.4
Let $X_1, \ldots, X_n$ be a random sample. Let the rejection or non-rejection of H_0 is decided by numerical realization of the statistic $T_0 = T_0(X_1, \ldots, X_n)$. Then T_0 is called *test statistic* (It is a random variable whose value varies from sample to sample). Consider a set of all realization of the test statistic T_0. This can be partioned into two subsets. The *critical region*, sometimes called the *region of rejection*, is a subset such that if the value of the test statistics falls in it, the null hypothesis is rejected. It is denoted as W.
Similarly the *region of nonrejection* is a subset such that if the value of the test statistics falls in it, the null hypothesis is not rejected. It is denoted as V.
The boundary between the rejection and nonrejection regions is called the *critical value* which is determined by prior information concerning the distribution of the test statistic and by the specification of the alternative hypothesis; it is tabulated.
[The test statistic serves us as an indicator of discrepancy between tested null hypothesis and the observed data. The realizations of the test statistic in the critical region means that this discrepancy is not more acceptable.]

If the numerical realization t_0 of the test statistic T_0 falls within critical region W, then H_0 is rejected at the level of significance α in favor of alternativ hypothesis H_1, which is accepted. This is the real negation of H_0.
If the numerical realization t_0 of the test statistic T_0 falls within nonrejection region V, then H_0 is not rejected at the level of significance α. But it does not mean, that H_0 is true; we do not have arguments for rejection only. The probabilities of Type Error I and Type Error II can be expressed as follows:
$P(T_0 \in W | H_0 \text{ platí}) = \alpha$
$P(T_0 \in V | H_1 \text{ platí}) = \beta$

Remark 5.5
The values of the test statistic, which give evidence in favor of alternative hypothesis, are concentrated in critical region. According to the form of the alternative hypothesis the list of corresponding critical regions (for common test statistics) follows :

$$W_1 = (t_{min}, \ K_{\alpha/2}(T)\rangle \ \cup \ \langle K_{1-\alpha/2}(T), t_{max}) \quad \text{for two-tailed test, where } H_1 : \ h(\vartheta) \neq c$$
$$W_2 = \phantom{(t_{min}, \ K_{\alpha/2}(T)\rangle \ \cup} \langle K_{1-\alpha}(T), t_{max}) \quad \text{for right-tailed test, where } H_1 : \ h(\vartheta) > c$$
$$W_3 = (t_{min}, \ K_{\alpha}(T)\rangle \phantom{\cup \ \langle K_{1-\alpha/2}(T)} \quad \text{for left-tailed test, where } H_1 : \ h(\vartheta) < c \quad ,$$

and where t_{min} stands for the minimal value which the test statistic T_0 may take,
t_{max} stands for the maximal value which the test statistic T_0 may take and finally
$K_{\alpha}(T)$ stands for α-quantile of the test statistic T_0.

The procedure for testing hypothesis can be succinctly described in the following steps:

- State the null hypothesis and the alternative hypothesis. (It is common practice to choose as the null hypothesis that claim which more nearly represents the status quo and as the alternative hypothesis that claim which represents "new wisdom".)

- Choose the level of significance α. The common levels of significance are $\alpha = 0.05$ or 0.01 or 0.1, whereas the most common level is $\alpha = 0,05$

- Select the test statistic and determine the distribution of the test statistic under the null hypothesis.

- Determine the critical region.

- Draw a sample and evaluate the test statistic.

- Make a decision and reach a conclusion.

Now let us move on to confidence interval method of hypothesis testing. The selection of appropriate pivotal statistic (doing confidence intervals) and the selection of appropriate test statistic (doing testing) are inter-related. This relation will be demonstrated in seminars.

Theorem 5.6
Let (d, h) be an empirical confidence interval for parametric function $h(\vartheta)$. If this interval does contain the constant c, then H_0 is not rejected at the significance level α. If it does not contain the constant c, then H_0 is rejected at the level α.

1. In case of two-tailed test, where $H_1 : \ h(\vartheta) \neq c$, we form two-sided confidence interval (d, h).

2. In case of left-tailed test, where $H_1 : \ h(\vartheta) < c$ we form right-sided confidence interval $(-\infty, h)$.

3. In case of right-tailed test, where $H_1 : \ h(\vartheta) > c$ we form left-sided confidence interval (d, ∞).

Doing testing using statistical software it is common that input does not require level of significance α and the output produces so-called p-value instead. The principle of p-value is analogous to significance level α, but generally it gives more information about the test result.

Definition 5.7
The smallest level of significance, at which the given sample observations and related value of the test statistic lead to rejection of the null hypothesis, is called the *p-value*.

Theorem 5.8

Consider the hypothesis testing at significance level α. The decision about the null hypothesis follows:

If $p - \text{value} \leq \alpha$, then we reject H_0.

[Thus the realization of the test statistic, for which the $p-$value was calculated, falls within the critical region corresponding to significance level α.]

If $p - \text{value} > \alpha$, then we do not reject H_0.

[Thus the realization of the test statistic, for which the $p-$value was calculated, does not fall within the critical region corresponding to the significance level α.]

According to the form of the alternative hypothesis we select one of following ways of calculation the $p-$value:

$p = 2\min\{P(T_0 \leq t_0), P(T_0 \geq t_0)\}$ for two-tailed test, where $H_1: \ h(\vartheta) \neq c$

$p = P(T_0 \geq t_0)$ for right-tailed test, where $H_1: \ h(\vartheta) > c$

$p = P(T_0 \leq t_0)$ for left-tailed test, where $H_1: \ h(\vartheta) < c$

The same decision is reached regardless of which of the three methods is used.