Anna Janicka

Mathematical Statistics 2020/2021
Lecture 8

1. Hypothesis testing – introduction

During the previous lectures, we have learned how to provide assessments of the values of distribution parameters. Frequently, based on the outcome of these estimates, we may wish to verify some statements concerning the distribution. In order to be able to verify such types of claims with the use of statistical methods, we will first need to introduce some definitions. The scheme that is proposed here has been developed in the second half of the XXth century, as a "compromise" based on works of two "fractions": on the one hand, led by Ronald Fisher, and on the other – Jerzy Neyman and Egon Pearson. This procedure is referred to as "Null Hypothesis Significance Testing" (NHST).

1.1. **A Statistical Hypothesis.** First of all, we need a hypothesis to be tested. This is the name for a(ny) statement regarding the probability distribution governing the phenomenon of interest to the researcher (the random variable observed). We will want to conclude about the validity of this statement based on observed values of the random variable. Possible hypothesis include statements such as:

- $X_1, X_2, \ldots, X_n$ are a sample from an exponential distribution.
- $X_1, X_2, \ldots, X_n$ are a sample from a normal distribution (this part is assumed to be known) with parameters 5 and 1 (this part is uncertain).
- $\mathbb{E}X_i = 7$ (the expected value of the distribution is 7).
- $\mathrm{Var}X_i > 1$ (the variance of the distribution exceeds 1).
- $X_1, X_2, \ldots, X_n$ are independent.
- $\mathbb{E}X_i = \mathbb{E}Y_j$ ($X_1, X_2, \ldots, X_n$ and $Y_1, Y_2, \ldots, Y_m$ have the same expected value).

Hypotheses may concern the values of distribution parameters (**parametric hypotheses** – for example, that the value of parameter λ in a Poisson distribution is equal to 2) or other properties of the distribution (**non-parametric hypotheses** – for example, that political preferences and music preferences of individuals are independent). Hypotheses may also be classified depending on the distribution they specify – as simple (if a single distribution is specified, for example exponential with parameter 3) or composite (if more than one distribution is specified, for example exponential with parameter greater than 3).

Hypotheses may also be classified based on the role they play in the testing procedure, either as the null or as the alternative hypothesis. The **null hypothesis**, denoted H_0, is the "basic" hypothesis under consideration. It is deemed true until "proven otherwise" – in our case, until we find that empirical data are very hard to reconcile with the theory. In many cases, the main goal of the researcher is to disprove the null hypothesis (that a factor under consideration is not important, that a received/existing theory is true etc.). The hypothesis, in many cases, has a speculative character. The **alternative** hypothesis, usually denoted H_a or H_1, is the statement that we turn towards if the null hypothesis is rejected – a new theory, a different value of the parameter considered etc. In the testing procedure, we always specify both: the null and the alternative hypotheses. Testing combinations may include, for example:

- testing $H_0 : \lambda = 1$ against $H_1 : \lambda = 2$;
- testing $H_0 : \lambda = 1$ against $H_1 : \lambda \neq 1$;
- testing $H_0 : \lambda = 1$ against $H_1 : \lambda \geq 1$...

1.2. **A Statistical Test.** A statistical test is a procedure, which – for any possible set of observed values, i.e. for any sample of observations – leads to one of two possible decisions: reject the null hypothesis H_0 (in favor of the alternative H_1) or do not reject the null hypothesis (when there are no grounds to reject H_0). A formal specification of a test requires specification of a statistical model first, with a vector of observations $X = (X_1, X_2, \ldots, X_n)$ and a family of

distributions $X \sim P_\theta$, where $\theta \in \Theta$. The easiest way to specify null and alternative hypotheses is with the use of ranges for the parameter:

$$H_0 : \theta \in \Theta_0, \qquad H_1 : \theta \in \Theta_1,$$

where $\Theta_0 \subset \Theta$ and $\Theta_1 \subset \Theta$ are nonempty sets such that $\Theta_0 \cap \Theta_1 = \emptyset$. Please note that the null and alternative hypothesis must be mutually exclusive, but they needn't fill the whole range Θ.

Once the data and the hypotheses are specified, a **statistical test** of H_0 against H_1 is provided in one of the following ways:

- By specifying a statistic $\delta : X \to \{0, 1\}$ such that for any given observation, if the statistic takes on the value of 1, this is interpreted as a rejection of the null hypothesis in favor of the alternative (for this data), and if the value of the statistic is 0 – no grounds to reject the null;
- By specifying the **region of rejection** (also called the **critical region**) $C = \{x \in X : \delta(x) = 1\}$ – the set of observations for which we reject the null hypothesis;
- By specifying the **region of acceptance** $A = \{x \in X : \delta(x) = 0\}$ – the set of observations for which we do not have grounds to reject the null hypothesis.

The critical region of a test is frequently provided in the form

$$C = \{x \in X : T(x) > c\},$$

where T is a **test statistic** and c is a **critical value**. All specification forms (providing δ, providing C, providing T and c) are equivalent. The range of values of the test statistic T which lead to the rejection of the null hypothesis may also be referred to as the critical region.

Note that there are many ways of testing a null hypothesis H_0 against an alternative H_1 – just like there were many ways of estimating the unknown value of a distribution parameter. Some testing procedures are more intuitive than others, however, and some of the tests will have *better* properties than others. We will study the ways of designing tests and test properties later on. At this point we will concentrate on the *testing philosophy*.

Example:

(1) Let us assume that we want to verify whether a coin is symmetric. Let's say that we design the following experiment: we toss the coin $n = 400$ times and we record H, the number of heads obtained. In such a case, H has a binomial distribution with the number of trials equal to 400 and the probability of success in a single trial equal to p. p is an unknown parameter, the probability of obtaining heads on the coin. In this case, the statistical model may be specified as follows: $H \sim B(400, p)$, where $p \in (0, 1)$ is an unknown parameter.

Usually, when we say that we "want to verify whether a coin is symmetric", this means that we want to test the null hypothesis $H_0 : p = \frac{1}{2}$ against the alternative $H_1 : p \neq \frac{1}{2}$ (but we may also specify other alternative hypotheses, depending on our specific goals, fears etc.).

Please note that when thinking about the null hypothesis that $p = \frac{1}{2}$, we see that some experiment outcomes (such as $H = 200, H = 198, H = 205....$) are likely, while others ($H = 10, H = 300$) are not very likely. This may lead us to us the following testing procedure: we will reject the null hypothesis if the number of heads is "too far" from the value that we would expect if the coin was indeed symmetric (namely, $400 \cdot \frac{1}{2} = 200$). In this case, our test would adopt the following form: we would specify a test statistic $T(x) = |x - 200|$, and we would reject the null (in favor of the alternative) if $T(x) > c$ for some constant c. But how do we choose a specific value of c?

When designing tests, we must always take into account the fact that our data are random, which means that there is always a possibility of committing an error of some type. Two basic types of errors may be defined:

- when the null hypothesis is in fact true, but our testing procedure leads us to rejecting it. This is called a **type I error**. In our coin tossing example, this could happen in the following scenario: we have a symmetric coin, we toss it 400 times and we get 50 heads. This result is highly unlikely, so we treat it as an indication that the null is not true and we decide to reject it (committing an error, because the coin was in fact symmetric).
- when the null hypothesis is in fact false, but our testing procedure does not lead us to a rejecting decision. This is called the **type II error**. In our coin tossing example, this could happen in the following scenario: we have a biased coin, for which the probability of obtaining heads is $\frac{3}{4}$. We toss it 400 times and we get 205 heads (not likely, but possible). This result would be perfectly possible if the coin was symmetric, so we do not see anything suspicious and do not reject the null hypothesis that the coin is symmetric (although in reality it isn't).

Formally, if we have a test with critical region C, the probabilities of the errors of the I and II type can be calculated as follows:

- $P_\theta(C)$ for $\theta \in \Theta_0$ – the probability of committing an error of the first type;
- $P_\theta(A)$ for $\theta \in \Theta_1$ – the probability of committing an error of the second type.

Obviously, when designing a test we are interested in minimizing these two types of errors. However, there is a tradeoff between errors of the first and second type: it's impossible to minimize both simultaneously. This is because if we want to minimize the error of the Ist type, we need to avoid situations when we reject the null hypothesis when it is in fact true. In such cases we will try to treat even unlikely situations as not suspicious (not leading to a rejection of the null). But this also means that we expand the range of acceptable values in such a way that many more results which are typical to alternative hypotheses are treated as not suspicious – do not lead to rejection of the null. This means that the error of the second type becomes larger.

The probabilities of committing errors of the first and second type depend on the true value of the unknown parameter. If we want to control them, we will need to take into consideration the *worst case scenario*. In case of the error of the first type, we will define the **significance level** of the test; a test is said to have a significance level α, if for any $\theta \in \Theta_0$ we have $P_\theta(C) \leq \alpha$ (in other words, the probability of committing an error of the first type does not exceed α). On the other hand, we may also define the **power of the test** (for the alternative hypothesis) as the probability of not committing an error of the second type (assuming that the alternative is true), which amounts to $P_\theta(C)$ for $\theta \in \Theta_1$. In such a case, we may talk about the function of the power of a test $1-\beta$: $1-\beta(\theta) = P_\theta(C)$.

Usually, controlling the magnitude of the error of the first type (α) is more important to the researcher than error of the second type (β). Therefore, in the majority of cases, when looking for the best test, we will look for tests for a specific significance level α (depending on the conservatism of the researcher, this might be 0.01, 0.05 etc.) and the smallest error of the second type possible (the highest power of the test possible).

Going back to our coin tossing example:

(1) Let's say that we would like our test to have a significance level of $\alpha = 0.01$. This means that we want the probability of incorrectly rejecting the null hypothesis not to exceed 0.01. This means that if we want our test to have a critical region of the form

$$C = \{x : |x - 200| > c\},$$

(i.e. reject the null if the outcome is too large or too small, symmetrically) for this significance level, we need to find c such that

$$P_{\frac{1}{2}}(|H - 200| > c) = 0.01$$

From the de Moivre-Laplace theorem we have that $P_{\frac{1}{2}}(|H - 200| > c) \approx 2\Phi(-c/10)$; upon equating this value to 0.01, we see that we need $c \approx 25.8$ (that is, at least 26 – since our variable has discrete values).

This way, we construct a test with a significance level approximately equal to 0.01. In case of this test, we reject the null hypothesis if the number of tails is lower than 175 or higher than 225. The critical region of this test is $C = \{0, 1, \ldots, 174\} \cup \{226, 227, \ldots, 400\}$; it is two-sided, i.e. we treat too small and too large outcomes as equally suspicious.

This is nt the only way in which we can define a test statistic for the specified null hypothesis. We may want to adopt a different approach when we are not equally bothered by $p > \frac{1}{2}$ and $p < \frac{1}{2}$ (for example, we use the coin we are testing to decide who pays for a meal and we are not concerned with a bias in our favor, only concerned with the bias being against us).

(2) In such a case, we might want to test the null $p = \frac{1}{2}$ against the alternative $p > \frac{1}{2}$ (if we are concerned about p being larger than $\frac{1}{2}$). In such a setting, we might wish to use a test statistic $T(x) = x - 200$ and reject the null (in favor of the alternative that $p > \frac{1}{2}$) if the value is excessively large. Looking for the critical value for this one-sided test, for the $\alpha = 0.01$ significance level, we would have looked for c such that

$$P_{\frac{1}{2}}(H - 200 > c) = 0.01,$$

and would have adopted $c \approx 23.3$. Note that this critical value for the one-sided test is smaller than the critical value for the two-sided test for the same significance level. In case of the one-sided alternative, for a significance level $\alpha = 0.01$ we would reject the null if the number of heads amounted to 224 or more; the critical region in this case would be $C = \{224, 227, \ldots, 400\}$.

Please note that formally, it's not interdicted to use the $T(x) = |x - 200|$ test statistic for testing against the alternatives that $p > \frac{1}{2}$, $p < \frac{1}{2}$, $p = \frac{1}{3}$, or $p = \frac{3}{4}$; neither is using the statistic $T(x) = x - 200$ for testing against any of these possibilities. The significance levels of the tests remain the same in all the cases. But this does not mean that all of these tests are equally sensible. For example, if we tested the null that $p = \frac{1}{2}$ against the alternative that $p = \frac{1}{3}$ using the statistic $T(x) = x - 200$, we would have rejected the null if the number of heads was too large for $p = \frac{1}{2}$. We would have then adopted the alternative $p = \frac{1}{3}$ instead (even though this would not make any sense!). The significance level would remain the same, but the error of the second type would become large (or, equivalently, the power of the test would become small) if the alterative hypothesis were misspecified.

Let us now look at the calculations of the power of the test for the alternative hypothesis. When calculating the critical value, we only look at the null hypothesis: we did not take into account the exact form of the alternative hypothesis. On the other hand, when calculating the power of a test, we need to look at all possible values from the Θ_1 range. For the sake of illustration purposes, at this point we will just look at two specific values: $p_1 = \frac{3}{4}$ and $p_2 = 0.51$.

(1) When testing with a test with critical region $C = \{|X - 200| \geq 26\}$, the power for $p_1 = \frac{3}{4}$ is:

$$1 - \beta(\frac{3}{4}) = P_{\frac{3}{4}}(|X - 200| \geq 26) = P_{\frac{3}{4}}(X \leq 174 \cup 226 \leq X) = P_{\frac{3}{4}}(X \leq 174) + P_{\frac{3}{4}}(226 \leq X)$$

$$= P_{\frac{3}{4}}\left(\frac{X - 400 \cdot \frac{3}{4}}{\sqrt{\frac{3}{4} \cdot \frac{1}{4} \cdot 400}} \leq \frac{174 - 400 \cdot \frac{3}{4}}{\sqrt{\frac{3}{4} \cdot \frac{1}{4} \cdot 400}}\right) + P_{\frac{3}{4}}\left(\frac{X - 400 \cdot \frac{3}{4}}{\sqrt{\frac{3}{4} \cdot \frac{1}{4} \cdot 400}} \geq \frac{226 - 400 \cdot \frac{3}{4}}{\sqrt{\frac{3}{4} \cdot \frac{1}{4} \cdot 400}}\right) \approx$$

which, from the de Moivre-Laplace theorem, may be approximated as

$$\Phi\left(\frac{-126}{\sqrt{\frac{3}{4} \cdot \frac{1}{4} \cdot 400}}\right) + 1 - \Phi\left(\frac{-74}{\sqrt{\frac{3}{4} \cdot \frac{1}{4} \cdot 400}}\right) \approx 0 + 1 - 0 \approx 1,$$

which means that the probability of the error of the second type is for $p_1 = \frac{3}{4}$ almost equal to 0. This means that this test has an almost perfect ability of distinguishing a

value of $p = \frac{1}{2}$ from a value of $p = \frac{3}{4}$: if the true value of the parameter were $p = \frac{3}{4}$, the value of the test statistic would have almost always fallen into the critical region and we would have almost always rejected the null.

(2) If, on the other hand, we use the same test with a critical region $C = \{|X-200| \geq 26\}$, but we look at the power of this test for $p_2 = 0.51$, we have:

$$1-\beta(0.51) = P_{0.51}(|X-200| \geq 26) = P_{0.51}(X \leq 174 \cup 226 \leq X) = P_{0.51}(X \leq 174) + P_{0.51}(226 \leq X)$$

$$= P_{0.51}\left(\frac{X - 400 \cdot 0.51}{\sqrt{0.51 \cdot 0.49 \cdot 400}} \leq \frac{174 - 400 \cdot 0.51}{\sqrt{0.51 \cdot 0.49 \cdot 400}}\right) + P_{0.51}\left(\frac{X - 400 \cdot 0.51}{\sqrt{0.51 \cdot 0.49 \cdot 400}} \geq \frac{226 - 400 \cdot 0.51}{\sqrt{0.51 \cdot 0.49 \cdot 400}}\right) \approx$$

which, from the de Moivre-Laplace theorem, may be approximated as

$$\Phi\left(\frac{-30}{\sqrt{0.51 \cdot 0.49 \cdot 400}}\right) + 1 - \Phi\left(\frac{22}{\sqrt{0.51 \cdot 0.49 \cdot 400}}\right) \approx 0.001 + 1 - 0.986 \approx 0.015,$$

which means that the probability of the error of the second type is in this case almost equal to 1. This means that the test we are using has almost no ability of distinguishing a value of $p = \frac{1}{2}$ from a value of $p = 0.51$: if the true value of the parameter were $p = 0.51$, the value of the test statistic would have fallen into the critical region with probability similar to the significance level α and we would have almost never rejected the null.

The example above underlined an important property of hypothesis testing: while constructing a test for a given significance level, it needn't be necessary to take into account the specific form of the alternative hypothesis (or rather: we might come up with the same test for different alternative hypotheses). However, the constructed test will have different properties (in terms of the error of the second type and the power of the test) depending on the exact specification of the alternative hypothesis. Also, the fact that we reject the null hypothesis in such a procedure does not imply that the alternative hypothesis is true, nor even "sensible" – it is up to the researcher to make sure that the alternative makes sense in case that the null is rejected.

1.3. **The Testing Procedure.** To sum up, the testing procedure we have proposed includes the following steps:

(1) Define the statistical model
(2) Pose the hypotheses (H_0 and H_1)
(3) Choose significance level α
(4) Choose test statistic T (formula) and critical value c /define critical region C
(5) Decide based on whether the value of the test statistic for the observed data falls into the critical region

A slightly different approach to the testing procedure might also be applied. In this alternative scenario, we modify the last two steps of the procedure. In step (4), we do not specify the critical value c, but rather pose the following question: what would be the significance level of a test for which the critical value would be equal to the value of the test statistic that we see for the data? (in other words – what is the chance of obtaining a result as "extreme" – in terms of the testing scheme – as the one observed?) The answer to this question is denoted as the ***p*-value** of a result. We may then verify the null hypothesis (in step (5)) based on this p-value: if the p-value is smaller than the adopted significance level, we reject the null. If it is larger, we do not have grounds to do so.

Going back to our coin tossing example:

(1) Let us assume that we decided to verify the null hypothesis that $p = \frac{1}{2}$ against the alternative that $p \neq \frac{1}{2}$ with a test with critical region $C = \{|H - 200| > c\}$. Let us assume that we observe $H = 220$ heads in 400 trials. What should that tell us in terms of the null hypothesis? We have that $|220 - 200| = 20$; since we have that

$$P_{\frac{1}{2}}(|H - 200| > 20) \approx 0.05$$

(from the de Moivre-Laplace theorem), we have that the p-value of our observation (obtaining 220 heads) is approximately 0.05. This value is higher than the adopted significance level of 0.01 (it is less unlikely in view of the null hypothesis – less extreme), so we do not have grounds to reject the null hypothesis.

(2) Let us assume that we decided to verify the null hypothesis that $p = \frac{1}{2}$ against the alternative that $p \geq \frac{1}{2}$ with a test with critical region $C = \{H - 200 > c\}$. Let us assume that we observe $H = 220$ heads in 400 trials. What should that tell us in terms of the null hypothesis? We have that $220 - 200 = 20$; since we have that

$$P_{\frac{1}{2}}(H - 200 > 20) \approx 0.025$$

(from the de Moivre-Laplace theorem), we have that the p-value of our observation (obtaining 220 heads) is approximately 0.025. This value is higher than the adopted significance level of 0.01 (it is less unlikely in view of the null hypothesis – less extreme), so we do not have grounds to reject the null hypothesis. Please note that the p-value in the one-sided case is half of the p-value for the two-sided case for the same outcome of 220 heads. This is due to the fact that when considering the results "as extreme as the one obtained", in the one-sided case we only have results larger than 220 heads, while in the two-sided case we have both those that are larger than 220 and those that are smaller than 180.