APPLIED STATISTICAL METHODS

Mathematics Institute

Faculty of Mathematics and Computer Science

Wroclaw University – Fall 2015

LECTURE 4

TESTING HYPOTHESES - IDEAS and TESTS OF GOODNESS OF FIT

Wroclaw, 11 - 15 October, 2015

- **A Hypothesis** is a statement about a population. We have two hypotheses: H_o-null hypothesis, and H_a-alternative hypothesis. We evaluate null hypothesis in the context of the alternative.
- **Goal** is to decide if the null hypothesis is true or not.
- **Process: Step1.** State null and alternative hypotheses: H_o and H_a, respectively.
- **Step 2.** Compute a test statistic T (a function of data).
- **Step 3.** Make decision based on the value of the test statistic: either reject H_o if T falls into the *critical region* (CR) or do not reject H_o when T falls outside of the CR.

The main ideas of testing hypotheses theory

The ideas for the process of testing hypotheses are called *Neyman-Pearson framework* in honor of Jerzy Neyman and Egon Pearson who developed this area.

Critical region and Type I error. The CR is a subset of possible values of the test statistic. It is chosen so that

$$P(\underbrace{T \in CR \mid H_o \; true}_{\text{Type I error}}) = P(\text{ reject } H_o \text{ when(given) } H_o \text{ is true }) = small = \alpha.$$

- The probability of Type I error is called significance level of the test and it is denoted by α.
- It should be set by the researcher before any data is collected. It is a measure of error that the researcher can live with.

Power of a test and Type II error

Type II error is to not reject a false null hypothesis:

$$P(\underbrace{T \notin CR \mid H_o\ false}_{\text{Type II error}}) = P(\text{ do not reject } H_o \text{ when(given) } H_o \text{ is false }) = \beta.$$

Power of a test is the probability of rejecting a false null hypothesis:

$$\text{Power } = 1 - \beta = P(T \in CR | H_o\ false) = P(\text{reject } H_o \text{ when(given) } H_o \text{ is false })$$

$$= P(\text{a correct decision})$$

Testing considerations

- We want to maximize the power (minimize probability of Type II error) and minimize the level of significance (probability of Type I error).
- Usually, there is no way to do both, and in practice we tradeoff the probability of one error for the other.
- Most tests keep the level of significance α on a given level, and maximize power given that α.

EXAMPLE 1 TRADEOFF - Analogy to the US legal system

In the US legal system, an accused is presumed innocent until proven guilty beyond a reasonable doubt.

- Ho: A person is convicted (guilty verdict) vs.
- Ha: A person is set free (not guilty verdict)

How to make the decision?

- Type I error: Reject Ho—Ho true i.e. Verdict not guilty — a person is guilty (guilty person goes free);
- Type II error: Do not reject Ho — Ho false i.e. Verdict guilty — a person is not guilty (innocent person is convicted).

Two problems

- If we make it extremely difficult to convict criminals because we do not want to incarcerate any innocent people we would probably have a legal system in which no one gets convicted.

  That would mean a decision rule like this: Convict the accused only if his/her fault is proven beyond any shadow of a doubt (we are certain of their being guilty).

- On the other hand, if we make it very easy to convict, then we will have a legal system in which many innocent people end up behind bars.

  That would require a decision rule like this: Convict an accused if there is any evidence of their guilt.

Example, contd.

TRADEOFF: Legal system that does not require a guilty verdict to be *beyond a shadow of a doubt* (i.e., complete certainty) but *beyond a reasonable doubt.*

The decision rule takes the following into consideration:

- We set the probability of type I error (reject Ho when it is true, that is verdict not guilty when in fact the accused is guilty) on a small level (level of significance).
- Then, we make sure that the procedure makes the probability of Type II error (verdict guilty when in fact the accused is not guilty) as small as possible (convicted an innocent person).

The system minimizes the chances of convicting an innocent person.

EXAMPLE 2 TRADEOFF - Quality Control

A company purchases chips for its smart phones, in batches of 50,000. The company is willing to live with a few defects per 50,000 chips. How many defects?

- If the firm randomly samples 100 chips from each batch of 50,000 and rejects the entire shipment if there are ANY defects, it may end up rejecting too many shipments (error of rejection).
- If the firm is too liberal in what it accepts and assumes everything is *sampling error*, it is likely to make the error of acceptance.

TRADEOFF: This is why government and industry generally work with Type I error of .05 (level of significance $\alpha = 0.05$).

P-value of a test. Let the computed value of the test statistic be T^*. Then, the $p - value$ of this test is

$$\text{p-value } = P(T \text{ "at least as contradictory" to } H_o \text{ as } T^* | H_o \text{ true}).$$

A small p-value indicates that data is contradictory to H_o.

Idea:

- Often, small values of the test statistic support Ho (T^* small supports Ho true), and
- large values of the test stat support Ha (T^* large supports Ho false).
- Then, p-value is the probability of test stat being as small or smaller than observed if Ho is true.
- Small p-values support Ha.

Example: z-test (or t-test) about the mean of a population.

Construction of tests

Test on a given level of significance α is constructed so that

$$T \in CR \leftrightarrow \text{ p-value } < \alpha.$$

Thus, we reject H_o when $T \in CR$ or equivalently p-value $< \alpha$.

Testing if the model fits the data: Goodness-of-fit (GOF) tests

Goal: Decide if the model fits the data.

H_o : Data comes from a distribution with cdf F.

H_1 : Data does not come from a distribution with cdf F.

We will consider two tests based on the empirical cdf:

1. Kolmogorov-Smirnov GOF test (K-S), and
2. Anderson-Darling GOF test (A-D).

EDF statistic

Definition. Let $\hat{F}_n(x)$ be empirical cdf of a random sample $X_1, X_2, \ldots, X_n$. Let F be a cdf. Then, statistic $|\hat{F}_n(x) - F(x)|$ is called *EDF statistic*. It measures the distance between $\hat{F}_n(x)$ and $F(x)$.

Tests utilizing the EDF statistic:

1. Supremum statistic: $D = sup_x|\hat{F}_n(x) - F(x)|$ leads to the K-S test;
2. Quadratic statistic

$$Q = n \int_R |\hat{F}_n(x) - F(x)|^2 w(x) dF(x)$$

leads to the A-D test for the weight function $w(x) = \frac{1}{F(x)(1-F(x))}$.

Goodness of fit tests based on the EDF statistic

Goodness of fit tests based on the EDF statistic, the hypotheses:

H_o : Data comes from a distribution with cdf F.

H_1 : Data does not come from a distribution with cdf F.

Note: Distribution F has to be completely specified in the test setting above. That means that all the parameters have to be specified.

KS and A-D tests

K-S test: test statistic is

$$D = sup_x |\hat{F}_n(x) - F(x)|,$$

and the distribution of D under the null hypothesis is available in tables or in stat packages. Typically stat packages provide the p-value for this test.

A-D test. Test statistic is

$$A^2 = n \int_R |\hat{F}_n(x) - F(x)|^2 \frac{1}{F(x)(1 - F(x))} dF(x).$$

Distribution of the test statistic D under the null hypothesis is also specified in the tables, and stat packages typically provide the p-value for the A-D test.

Goodness of fit tests - testing for family of distributions

What if we want to test if the data came from a family of distributions?

That could mean checking if the data came from a normal distribution or from an exponential distribution, etc. In such cases the exact distribution of the model (F) is not fully specified. That is we only specify the type of the model distribution F (e.g. normal, exponential, etc.), but not the specific parameters.

We can still use GOF tests (K-S and A-D) but the distribution of the test statistic is adjusted for the estimated parameters. Note, that this is a different test than the one when the distribution of F is completely specified in the null hypothesis.

Testing for a family of distributions

H_o : Data comes from a normal distribution.

H_1 : Data does not come from any normal distribution.

We may also use GOF tests to check if two samples come from the same or different distributions. In those cases, the hypotheses would look like this:

H_o : DataA and DataB come from the same distribution.

H_1 : DataA and DataB do not come from the same distribution.

Note. In practice we use stat software to perform the tests. Also, there are many more GOF tests.