Statistics

Week 4: Hypothesis Testing (Chapter 6)

ESD, SUTD

Term 5, 2017

SINGAPORE UNIVERSITY OF TECHNOLOGY AND DESIGN

Established in collaboration with MIT

Please complete the mid-term survey.

Outline

1. Hypothesis testing
   - p-value

Hypothesis

A **hypothesis** is a claim. In *hypothesis testing*, we attempt to answer the following:

Given some data from a sample, does it provide statistically significant evidence to prove (beyond reasonable doubt) a hypothesis about the population, or could it have arisen due to random chance?

As a generic example, a hypothesis could be that a particular treatment has a real effect (e. g. better than an existing treatment, or placebo, or doing nothing).

Null and alternative hypotheses

More specifically, using the sample data, we test the validity of a claim about the population, against a counter claim. We set up these two competing claims as follows:

- The **null** hypothesis, H_0, is the claim of no difference or no effect; usually, H_0 is the status quo.
- The **alternative** hypothesis, H_1, is the claim that there is a difference or effect (usually it is the claim you are interested to prove).

Rejecting the null hypothesis is a primary task in scientific research.

Exercise: write down H_0 and H_1 for the training technique example from last class.

'Proof' by contradiction

The standard approach is to first *assume H_0 is true*. Then, perform a calculation to determine whether the data *contradicts* this assumption beyond reasonable doubt.

- If Yes, then reject H_0. We may also accept H_1.
- If No, then do not reject H_0. We cannot rule out H_0 as an explanation for the data, but we have not proven it either. So we *do not accept either* hypothesis.

So if we fail to prove H_1, then it *may* be because H_0 is true, or it *may* be the case that H_1 is true, but there is *insufficient* information to rule out random chance as an alternative explanation for the data.

In this case, we take the conservative stance and 'do not reject' H_0 – the data, after all, may still be consistent with null hypothesis.

Analogies

Analogy 1: in most legal systems, a person is assumed innocent until proven guilty. The burden of proof is on the one who makes the (extraordinary) claim that the person is guilty.

H_0: innocent; H_1: guilty.

If there is not enough evidence to establish guilt, it does not prove that the person is innocent.

Analogy 2: in general, H_0 is usually a negative statement, such as 'telepathy does not exist', and it is very hard to prove negative statements. However, a person who makes the (extraordinary) claim that he is telepathic (H_1) needs to prove it.

'Extraordinary claims require extraordinary evidence.'

Example

Example: a sample of 50 tins of tomatoes are tested, to see if their average weight deviates from the acceptable value of $\mu_0 = 350\text{g}$. State the hypotheses.

Answer: $H_0 : \mu = \mu_0$; $H_1 : \mu \neq \mu_0$.

Suppose the weights satisfy $\sigma = 10$ and $\bar{x} = 355.2$. Take 'statistically significant' to mean 95% confidence.

Assuming that H_0 is true, we have

$$P\left(\mu_0 - 1.96\frac{\sigma}{\sqrt{n}} \leq \bar{X} \leq \mu_0 + 1.96\frac{\sigma}{\sqrt{n}}\right) = 0.95,$$

$$P\left(350 - 1.96\frac{10}{\sqrt{50}} \leq \bar{X} \leq 350 + 1.96\frac{10}{\sqrt{50}}\right) = 0.95.$$

Connection with CI

Assuming H_0, then with 95% probability, the sample mean lies between 347.2 and 352.8. As $\bar{x} = 355.2$, we reject H_0 (at the 5% significance level) and accept H_1.

Note that the inequalities on the last slide are *equivalent* to those involved in the confidence interval calculation for μ. This relationship also holds for one-sided tests and one-sided CIs.

Hypothesis test for μ

We reject H_0 at *significance level* α if and only if μ_0 falls outside the appropriate $(1-\alpha)$-level CI for μ.

Meaning of α: type I error

The significance level α is the (maximum) probability of accepting H_1 when H_0 is in fact true.

This type of error is known as a **type I error**, or a false positive.

Examples: (1) An innocent person is convicted to be guilty.

(2) A test shows a patient to have a rare disease when in fact she does not have it.

(3) A spam filter wrongly classifies a legitimate email as spam.

During an experimental set up, and before any hypothesis test is performed, we need to clearly specify H_0, H_1, as well as α.

Type II error and power

A **type II error** occurs when a test fails to reject H_0 when H_1 is actually true. It is also known as a false negative. Its probability is denoted by β.

Examples: (1) Baggage screening in airport security fails to pick up explosives.

(2) A person is guilty but the courtroom fails to identify it.

Exercises: (a) Is one type of error always more serious than the other?

(b) What does $(1-\beta)$ represent?

$(1-\beta)$ is called the *power* of a test. Usually a power of 80% is acceptable; 90% is desirable.

p-value

We have seen how to perform a hypothesis test using a CI.

Another approach to hypothesis testing is to ask the question: What is the probability of observing a sample statistic *at least as extreme* as the one observed, assuming H_0 is true?

Intuition for using 'at least as extreme': think of it as an area outside a confidence interval.

This probability is known as the p-value. If the p-value $\leq \alpha$, then reject H_0.

We have already computed a p-value back in Week 1.

Exercise: Compute the p-value for the tomatoes example.

p-value, properties

- The smaller the p-value, the more significant is the test result. Therefore, it is a good practice to quote the p-value after you perform a hypothesis test.
- The p-value is also the smallest α at which H_0 can be rejected.
- The p-value computation may be one- or two-sided, depending on the hypotheses.
- Sometimes the p-value is quoted as a number of standard deviations away from the mean in a normal distribution.

  For example, the 2012 discovery of the Higgs boson has a significance of 5 sigma (p-value $\approx 1/3.5$ million); $n \approx 300$ trillion proton-proton collisions were analyzed.