CmpE 343 Lecture Notes
10: Hypothesis Testing

Ethem Alpaydın

December 31, 2014

1 Introduction

HYPOTHESIS testing is different from estimation in that we are not really interested in estimating the value of an unknown population parameter but rather, we are given a claim, conjecture, assertion, or hypothesis about the value of the population parameter, and we are asked whether the sample is consistent with it or not.

What we would like to test is written as a *null hypothesis* and we calculate the probability of a statistic assuming that the null hypothesis holds. If this probability—called the *p value*—is very small, we *reject* the null hypothesis in favor of the alternative hypothesis; otherwise we accept the null hypothesis. The fact that we reject a hypothesis does not imply that it is wrong, it just shows that the sample does not favor it, and we may be unlucky and could have drawn a very rare sample. Or, the claim may be wrong but we may accept it. Both of these are unwanted and we want to make our accept/reject decisions such that those two types of errors do not happen frequently.

The four possibilities are:

	Truth	
Decision	H_0 true	H_0 wrong
Accept	Correct	Type II error (β)
Reject	Type I error (α)	Power ($1-\beta$)

Type I error is the probability of rejecting a true hypothesis and type II error is the probability of accepting a wrong hypothesis. We want both to be as small as possible. Power is the probability of rejecting a wrong hypothesis and we want it to be large.

Let us see an example.

2 Testing the Mean of a Single Population

Let us say somebody makes the claim that the mean of a population is 5. The way we proceed is to calculate the point estimator to the mean, namely the sample average, and reject the claim if it is far away from 5. We know that $\overline{X}$ will never be exactly 5 but somewhere close. How close can $\overline{X}$ be to μ is given by the sampling distribution of $\overline{X}$. Actually the $(1-\alpha)100\%$ confidence interval tells us where $\overline{X}$ lies with respect to μ in $(1-\alpha)100\%$ of the time. So we accept the claim that the mean is 5 if 5 is in the $(1-\alpha)100\%$ confidence interval, and we reject if it lies outside. If we decide this way, we know that the probability of wrongly rejecting, that is, the type I error, is α.

Now let us say actually μ is 7, but $\overline{X}$ can still be close to 5. In such a case, the probability that 5 falls in the confidence interval (even though the sample is drawn from a population with $\mu = 7$) is the probability of Type II error; note that to be able to calculate the type II probability we need to know what the real value is. We see that the type II error decreases if real μ goes further and further away from 5. The probability that 5 lies outside of this interval is the probability of correctly rejecting a wrong hypothesis, namely, the power of the test.

To summarize, for

$$H_0 : \mu = \mu_0 \text{ vs. } H_1 : \mu \neq \mu_0$$

for some particular value of μ_0, we use the fact that (under the null hypothesis that $\mu = \mu_0$)

$$P\left(-z_{\alpha/2} < \frac{\overline{X} - \mu_0}{\sigma/\sqrt{n}} < z_{\alpha/2}\right) = 1 - \alpha \tag{1}$$

and we accept H_0 if $\mu_0 \in \left(\overline{X} - z_{\alpha/2}\frac{\sigma}{\sqrt{n}}, \overline{X} + z_{\alpha/2}\frac{\sigma}{\sqrt{n}}\right)$, and reject otherwise.

Because the alternative hypothesis is $H_1 : \mu \neq \mu_0$, this is called a *two-sided* test; we reject $H_0 : \mu = \mu_0$ if $\overline{X}$ is much smaller than μ_0 or much larger. The decision rule above implies that we reject if $|(\overline{X} - \mu_0)/(\sigma/\sqrt{n})| > z_{\alpha/2}$, or equivalently, if $2P(Z > (\overline{X} - \mu_0)/(\sigma/\sqrt{n})) < \alpha$. Actually, this last probability, $2P(Z > (\overline{X} - \mu_0)/(\sigma/\sqrt{n}))$, is called the *p value* and is the probability of seeing a sample whose sample average is $\overline{X}$ or larger when the population mean is μ_0. *p value*

In hypothesis testing, either someone specifies α (as an upper bound for type I error) and we make a accept/reject decision depending on respectively whether the p value is larger/smaller than α, or no α is specified and we just report the p value. This latter case is more informative because the p value is an indicator of how much the sample supports the claim—the smaller the p value is, the more proof we have that the null hypothesis is wrong.

Sometimes, the test is *one-sided*. If the claim is something of the sort "better than, faster than, superior to," and so on, and in such a case, we reject only for differences in one direction, or the reject region is only one tail of the distribution (still with an upper bound of α).

For example for $H_0 : \mu = \mu_0$ vs. $H_1 : \mu > \mu_0$, the p value is $P(Z > (\overline{X} - \mu_0)/(\sigma/\sqrt{n}))$, and we reject if this probability is smaller than α. For $H_0 : \mu = \mu_0$ vs. $H_1 : \mu < \mu_0$, the p value is $P(Z < (\overline{X} - \mu_0)/(\sigma/\sqrt{n}))$, and we reject if this probability is smaller than α. We need to be careful which of the two one-sided tests we use: Whatever is claimed and hence will be interesting or different or surprising should be in H_1 and H_0 corresponds to the status quo—it is the rejection that is informative.

3 Generalization to Other Tests

The framework we discussed above can be generalized and made applicable to different scenario, which we list below. The development of tests for these are straightforward given the content of Lecture 9 on confidence intervals, and omitted to avoid repetition of very similar material:

- For the case above, if σ is unknown, we use s instead and the t distribution.
- The same approach can be used to devise tests for two populations where a two-sided test compares the equality of the two means and one-sided tests compares the two means. Similarly a paired comparison test can also be devised.
- One can test for the proportion of a single population, or compare the proportions of two populations by using the central limit theorem and testing means.
- One can decide on the sample size n given bounds for α and β for a given real value for the parameter. In Lecture 9, we discussed how to calculate n for a given α and error; here β for a given real parameter value replaces the error.
- One can devise tests for variance of a single population or compare the variances of two populations.

4 The Goodness-of-Fit Test

In the tests we discussed above, we tested for the value of a population parameter; those are called *parametric* in the sense that they make an assumption about the distribution (generally normal) and test for its parameters only. Now we discuss a *nonparametric* test where we make no assumptions about the distribution. This makes it applicable in a wider domain but keep in mind that a nonparametric test is not as good as a parametric test; that is, if a parametric test is available for a certain task, use it rather than the corresponding nonparametric test because the parametric test will almost always have smaller type I and II errors and higher power. But the catch is that there are cases where no known distribution is appropriate and hence no parametric test can be devised. Below we discuss the goodness-of-fit test which is a widely-used nonparametric test.

This test uses cells where for each cell, there is a condition that should be satisfied and contains a subset of the observations from the sample that satisfy its condition. These conditions are mutually exclusive and exhaustive. According to the null hypothesis, the cells are expected to contain a certain percentage of the distribution and if we multiply these percentages by the sample size, for each cell, we can calculate an expected count for each cell. The goodness-of-fit test uses the sum of (normalized) differences between these observed and expected counts.

For k cells where O_i and E_i are the observed and expected counts in cell i, the statistic

$$\sum_{i=1}^{k} \frac{(O_i - E_i)^2}{E_i} \sim \chi^2_{k-1} \tag{2}$$

and as usual, either we are given an α and we reject if this sum is greater than $\chi^2_{k-1,\alpha}$, or we calculate and report the p value.

- The goodness-of-fit test tests the whole distribution or some property of the whole distribution, so in that sense, it can be used for a variety of aims, e.g., test for uniformity, homogeneity, independence, and so on.

- The test uses discrete cells and so if the population is a continuous distribution, it should be discretized. Note that there is no requirement that the cells contain equal (or roughly equal) number of observations—there is a normalizer in the denominator. But cells should not contain zero or very few (less than five) observations; if this is the case, one can merge cells. Too much merging may smooth too much and lose information though. So one should be careful in choosing k and defining the cell conditions.

- Although the observed values are always integers (because they care counts), the expected values can be real-valued.

- If the cells are organized in a two dimensional grid (e.g., as in a contingency table) rather than in one dimension, the degrees of freedom of the distribution should be adjusted to reflect the structure. For r rows and c columns, the degrees of freedom is $(r-1) \cdot (c-1)$.