Lecture 8
Testing Concepts.

1 Hypotheses

Hypotheses are some statements about population distribution, which are either true or untrue for the given population.

Example For example, let $X_1, ..., X_n$ be a random sample from distribution $N(\mu, \sigma^2)$ with σ^2 known and $\mu \in \mathcal{M}$. Suppose our hypothesis is that $\mu \in \mathcal{M}_1$ for some $\mathcal{M}_1 \subset \mathcal{M}$, i.e. $\mathcal{M}_1$ is some subset of $\mathcal{M}$. It is called the null hypothesis. It is denoted as $H_0 : \mu \in \mathcal{M}_1$. Then the alternative hypothesis is that $\mu \notin \mathcal{M}_1$, i.e. $\mu \in \mathcal{M} \backslash \mathcal{M}_1$. It is denoted as $H_a : \mu \notin \mathcal{M} \backslash \mathcal{M}_1$. For example, if $\mathcal{M}_1 = \{\mu : \mu \leq \mu_0\}$ and $\mathcal{M} = \mathbb{R}$, then $H_0 : \mu \leq \mu_0$ and $H_a : \mu > \mu_0$. Or, as another example, if $\mathcal{M}_1 = \mu_0$, then $H_0 : \mu = \mu_0$ and $H_a : \mu \neq \mu_0$.

If a hypothesis uniquely identifies the distribution of the data, it is called simple. Otherwise, the hypothesis is called composite. In the the second example above, the null hypothesis is simple, while the alternative is composite. It is customary to mention both the null and the alternative hypotheses since the full parameter space $\mathcal{M}$ is often unspecified.

2 Testing

We observe a sample from a population and, based on this sample, create a test. Our test is intended to decide whether we accept the null hypothesis or reject it in favor of the alternative. Some people argue that instead of word “accept” it is more appropriate to say “do not reject”. We are not going to emphasize this difference here.

2.1 Critical region

Let X denote our data. Then any test consists of the critical region C, which is a function of our null and alternative hypotheses, such that we accept the null hypothesis if $X \in C$ and reject it if $X \notin C$. For example, if our data is $X = (X_1, ..., X_n)$, then the critical region might be $C = \{\sum_{i=1}^n X_i < \delta\}$ for some $\delta \in \mathbb{R}$. The value δ in this example might depend both on the null and on the alternative.

In testing, four situations are possible. If H_0 is true and we accept it, then it is a correct decision. If H_0 is true but we reject it, then it is a type 1 error. If H_0 is false but we accept it, then it is a type 2 error. If

H_0 is false and we reject it, then it is a correct decision again. So, in addition to correct decisions, there are errors of two types.

2.2 Size and power trade-off

The probability of a type-1 error is called the *size* of the test.

Example (cont.) In the example above, suppose our null hypothesis is $H_0 : \mu = \mu_0$ and our alternative is $H_a : \mu > \mu_0$. Then the natural test is to accept the null hypothesis if the data belongs to the critical region $C = \{\sum_{i=1}^n X_i < \delta\}$. Then

$$P_{\mu_0}\{\sum_{i=1}^n X_i \geq \delta\} = P_{\mu_0}\{\sqrt{n}(\overline{X}_n - \mu_0)/\sigma \geq \sqrt{n}(\delta/n - \mu_0)/\sigma\} = 1 - \Phi(\sqrt{n}(\delta/n - \mu_0)/\sigma),$$

which is a decreasing function of δ. If δ is large, then size of the test is small, which is good. Please note that the size is calculated at the null value (often called "under the null").

What is the probability of a type-2 error? If true parameter value $\mu > \mu_0$, then

$$P_\mu(\sum_{i=1}^n X_i < \delta) = \Phi(\sqrt{n}(\delta/n - \mu)/\sigma).$$

First, notice that it is a function of true μ. Second, if δ is large, then the probability of a type-2 error is large as well, which is bad.

Thus, there is a trade-off between the probability of a type-1 error and the probability of a type-2 error. This trade-off exists in most practically relevant situations. Before we consider how one should choose the test in light of this trade-off, the introduction of some additional concepts is necessary.

The *Power* of the test is defined as the probability of correctly rejecting the null hypothesis. Thus, the power of the test is defined as 1 minus the probability of a type-2 error. Apparently, the power of the test depends on the true parameter value. So, power is usually considered to be a function of the true parameter value on the set of alternatives.

The size of the test also depends on the true parameter value when the null hypothesis is composite. But, instead of considering the size of the test as a function of the true parameter value, the concept of the level of the test is used. We say that the test has *level* α if for any true parameter value in the null hypothesis, the size is not greater than α. The level of the test is defined as the maximum of the size over all possible true parameter values in the null hypothesis. In the example above, the level of the test is $\sup_{\mu \in \mathcal{M}_1} \text{size}(\mu)$.

Once we have some notion of the power of the test and its level, let us consider how to choose the test. Common practice is to fix the level of the test (usually, it is 1, 5, or 10%) and then to choose a test with as much power as possible among all tests of a given level. In this sense the null and the alternative are not treated equally.

Example (cont.) Let us return to our example with gaussian sample with known variance, where $H_0 : \mu = \mu_0$ and $H_a : \mu > \mu_0$. Suppose we want a test with level 5%. We wish to reject the null when $\sum_i X_i$

is large (that is, the critical region is of the form $C = \{\sum_{i=1}^{n} X_i < \delta\}$). We equivalently may construct statistics as follows:

$$Z = Z(X, \mu_0) = \frac{\sqrt{n}}{\sigma}\left(\bar{X} - \mu_0\right) \sim N(0,1) \text{ under the null.}$$

This is often called a $Z-$statistic. Let $Z_{0.95}$ denote the 95%-quantile of a standard normal distribution. Then $\sqrt{n}(\delta/n - \mu_0)/\sigma = Z_{0.95}$ or, our test will accept the null if $Z < Z_{0.95}$. This is the test with exact (finite-sample) size 5%.

Since the power of the test depends on the true parameter value, it is possible that one test has maximal power among all tests with a given level at one parameter value, while another test has maximal power at some other parameter value. So it is possible that there is no *uniformly most powerful test*. In this situation the researcher should consider some additional criteria to choose a test. This observation explains a wide variety of tests suggested in the statistics and econometrics literature. However, we should note that there is an important class of problems where uniformly most powerful tests do exist. We will discuss it next time.

2.3 P-value

The result of any test is either acceptance or rejection of the null hypothesis. At the same time, it would be interesting to know to what extent we are sure about the result of the test. The concept of the p-value gives us such a measure. The *p-value* is the probability (calculated under the null) of obtaining a sample at least as adverse to the null hypothesis as given. Notice that the p-value is a random variable.

Example (cont.) Let $z = Z(x, \mu_0)$ be the value of the Z-statistic, as defined above, that we see in our data set

$$p - value = P\{N(0,1) > z\} = 1 - \Phi(z).$$

Note again, that it is a function of z, and thus is a random variable. By construction, our test rejects the null if the p-value is smaller than 0.05.

If the p-value is much smaller than 0.05, then we can be quite sure that the null hypothesis does not hold. If the p-value is close to 0.05, then we are not so sure. Moreover, reporting the p-value has the advantage that, once the p-value is reported, any researcher can decide for himself whether he or she accepts or rejects the null hypothesis depending on his/her own favorite level of the test.

Let us now emphasize some frequent misunderstandings about the concept of the p-value. First, a p-value is not the probability that the null is true. There is no such probability at all since parameters are not random according to the frequentist (classical) approach. Second, the p-value is not the probability of falsely rejecting the null. This probability is measured by the size of the test. Third, one minus p-value is not the probability of the alternative being true. Again, there is no such probability since parameters are not random. Finally, the level of the test is not determined by a p-value. Instead, once we know the p-value of the test, the level of the test determines whether we accept or reject the null hypothesis.

Example. Let $X_1, ...X_n$ be a random sample from an $N(\mu, \sigma^2)$ distribution. The null hypothesis, H_0, is that $\sigma^2 = \sigma_0^2$. The alternative hypothesis, H_a, is that $\sigma^2 < \sigma_0^2$. Note that both hypotheses are composite since they contain all possible values of μ. Let us construct a test based on sample variance s^2. We know that $(n-1)s^2/\sigma^2 \sim \chi^2(n-1)$. Since small values of $(n-1)s^2/\sigma_0^2$ are a sign in favor of the alternative, our critical region should take the form $C = \{(n-1)s^2/\sigma_0^2 > k\}$. Under H_0, $(n-1)s^2/\sigma_0^2 \sim \chi^2(n-1)$. Then a test with level, say, 5%, accepts the null hypothesis if $(n-1)s^2/\sigma_0^2 > \chi^2_{0.05}(n-1)$, where $\chi^2_{0.05}(n-1)$ denotes the 5%-quantile of $\chi^2(n-1)$. What is the power of this test? Let $\sigma^2 < \sigma_0^2$. Then

$$P_{\sigma^2}\{(n-1)s^2/\sigma_0^2 \leq \chi^2_{0.05}\} = P_{\sigma^2}\{(n-1)s^2/\sigma^2 \leq (\sigma_0^2/\sigma^2)\chi^2_{0.05}\} = F_{\chi^2(n-1)}((\sigma_0^2/\sigma^2)\chi^2_{0.05}),$$

where $F_{\chi^2(n-1)}$ denotes the cdf of $\chi^2(n-1)$. So the power of the test increases as σ^2 decreases. Suppose $n = 101$, $\sigma_0^2 = 1$, and we observe $s^2 = 0.9$. What is the p-value of our test? Let $A \sim \chi^2(n-1)$. Then the p-value equals

$$P\{A \leq (n-1)s^2/\sigma_0^2\} = F_{\chi^2(n-1)}((n-1)s^2/\sigma_0^2) = F_{\chi^2(100)}(100 \cdot 0.9/1) = F_{\chi^2(100)}(90) \approx 0.25.$$

Thus, the test with level 5% does not reject the null hypothesis.

3 Pivotal Statistics

By definition, a statistic is called *pivotal* if its distribution is independent of unknown parameters. Pivotal statistics are useful in testing because one can calculate quantiles of their distributions and, thus, critical values for tests based on these statistics. For example, $(n-1)s^2/\sigma_0^2$ from the example above is pivotal under the null since its distribution does not depend on μ.

3.1 Asymptotic tests

Example. Test of the mean. Let $X = (X_1, ..., X_n)$ be a random sample from some unknown distribution with finite second moment. The null hypothesis is that $H_0 : EX_i = \mu_0$. The alternative is that $H_a : EX_i \neq \mu_0$. Both hypotheses are composite. Let us construct a test based on $|\overline{X}_n - \mu_0|$. Large values of $|\overline{X}_n - \mu_0|$ are a sign in favor of the alternative. Thus, our critical region should take the form $C = \{|\overline{X}_n - \mu_0 \leq \delta\}$ for some $\delta > 0$. It is impossible to find the exact distribution of the test statistic in this case but possible to find an *asymptotically pivotal* test statistic, that is a test statistic with an asymptotic distribution that does not depend on any unknown parameter.

Under the null, $\sqrt{n}(\overline{X}_n - \mu_0) \Rightarrow N(0, Var(X_i))$. We also have a consistent estimate for $Var(X_i)$ which is s^2. Let us define

$$t = (\overline{X}_n - \mu_0)/\sqrt{s^2/n} \Rightarrow N(0,1) \text{ if the null is true.}$$

This is a well-known t-statistics. We reject the null at 5% level if $|t| > 1.96$ and $p - value = 2\Phi(-|t|)$. This test has *asymptotic size* 5%. That is,

$$P_{H_0}\{reject\} \rightarrow 0.05 \text{ as } n \rightarrow \infty.$$

Example As another example, let $X_1, ..., X_m$ and $Y_1, ..., Y_n$ be independent random samples from two different distributions. We want to test null hypothesis, H_0, that $EX_i = EY_i$. against the alternative, H_a, that $EX_i > EY_i$. A natural place to start is to note that if the null hypothesis is true, then $\overline{X}_m$ should be close to $\overline{Y}_n$ with high probability. But $\overline{X}_m - \overline{Y}_n \sim N(0, \sigma_x^2/m + \sigma_y^2/n)$ with σ_x^2 and σ_y^2 unknown variances of the two distributions. So consider

$$t = \frac{\overline{X}_m - \overline{Y}_n}{\sqrt{s_x^2/m + s_y^2/n}},$$

where s_x^2 and s_y^2 are sample variances. The exact distribution of the t-statistics here is not pleasant. Instead, let us use asymptotic theory. As an exercise, prove that if both n and m increase to infinity we have $t \Rightarrow N(0,1)$. So we can use the quantiles of a standard normal distribution to form a test with size approximately equal to that of the required level of the test. This gives us a test of "asymptotically the correct size".

3.2 Bootstrap

Example. Test of the variances. Let $X = (X_1, ..., X_n)$ be a random sample from some unknown distribution with finite fourth moment. The null hypothesis is that $H_0 : Var(X_i) = \sigma_0^2$. The alternative is that $H_a : Var(X_i) \neq \sigma_0^2$. Both hypotheses are composite. We obviously, want to use statistic based on the sample variance s^2 but without normality assumption the exact distribution of s^2 is unavailable. We know, that s^2 is consistent and is asymptotically gaussian (think why), but we may be lazy to figure out what is the asymptotic variance of it. Imagine we want to use statistic $z = \sqrt{n}(s^2 - \sigma_0^2)$ and reject when it is too large or too small. We may do the following:

- For $b = 1, ..., B$ repeat:
  - Draw i.i.d. sample $X_b^* = (X_{1b}^*, ..., X_{nb}^*)$ from a set of initial observations $\{X_1, ..., X_n\}$ with replacement;
  - Calculate s_b^2 to be a sample variance of X_b^*;
  - Calculate $z_b = \sqrt{n}(s_b^2 - s^2)$;
- Order z_b in ascending order: $z_{(1)} \leq ... \leq z_{(B)}$;
- For test of size α, if $z_{([\frac{\alpha}{2}B])} < z < z_{([(1-\frac{\alpha}{2})B])}$ accept the null, otherwise reject.

Think why this procedure would give asymptotically level-α test. We know that under the null $z = \sqrt{n}(s^2 - \sigma_0^2) \Rightarrow N(0, V)$, where the asymptotic variance V can be calculated using CLT and delta-method. It depends on the first three moments of random variable X_i (which are unknown).

For the bootstrapped sample the variance equals to the sample variance and four first moments are consistent estimators of the population first four moments. We have $z^* = \sqrt{n}((s^*)^2 - s^2) \Rightarrow N(0, V^*)$ and $V^* \to^p V^2$. So the bootstrapped distribution of z-statistic will asymptotically be the same as the true asymptotic distribution of z-statistic.