Lecture 9
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 $\mathcal{M}_1 = \mu_0$ and $\mathcal{M} = \{\mu : \mu \geq \mu_0\}$, then $H_0 : \mu = \mu_0$ and $H_a : \mu > \mu_0$. Or, as another example, if $\mathcal{M}_1 = \mu_0$ and $\mathcal{M} = \{\mu : \mu \in \mathbb{R}\}$, then $H_0 : \mu = \mu_0$ and $H_a : \mu \neq \mu_0$.

If a hypothesis includes only one parameter value, it is called simple. Otherwise, the hypothesis is called complex. In the examples above, the null hypothesis was simple while the alternative was complex. In principle, we can also allow for complex null hypotheses. For example, if $\mathcal{M}_1 = \mu_1 \cup \mu_2$ and $\mathcal{M} = \{\mu : \mu \in \mathbb{R}\}$, then $H_0 : \mu = \mu_1$ or $\mu = \mu_2$ and $H_a : \mu \neq \mu_1$ and $\mu \neq \mu_2$. 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 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 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, some additional concepts are 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 as 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, level of the test is $\sup_{\mu \in \mathcal{M}_1} \text{size}(\mu)$.

Once we have the notion of 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 with a given level. In this sense the null and the alternative are not treated equally.

Example (cont.) Let us return to our example where $H_0 : \mu = \mu_0$ and $H_a : \mu > \mu_0$. Suppose we want a test with level 5%. All tests based on the critical region $C = \{\sum_{i=1}^n X_i < \delta\}$ with k such that $1 - \Phi(\sqrt{n}(\delta/n - \mu_0)/\sigma) < 0.05$ has level 5%. Since the power is a decreasing function of δ in this example, one should choose δ such that $1 - \Phi(\sqrt{n}(\delta/n - \mu_0)/\sigma) = 0.05$. Let $Z_{0.95}$ denote 95%-quantile of standard normal distribution. Then $\sqrt{n}(\delta/n - \mu_0)/\sigma = Z_{0.95}$ or, equivalently, $\delta = n(\sigma Z_{0,95}/\sqrt{n} + \mu_0)$. So, our test is to accept the null if $\sum_{i=1}^n X_i < n(\sigma Z_{0,95}/\sqrt{n} + \mu_0)$.

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 use some additional criteria to choose a test. This observation explains a wide variety of tests suggested in the statistical and econometric literature. However, we should note that there is an important class of problems where uniformly most powerful tests 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.) We observe data $X = (X_1, ...X_n)$ from $N(\mu, \sigma^2)$ with known σ^2. $H_0 : \mu = \mu_0$ and $H_a : \mu > \mu_0$. Assume you have a realized sample $(x_1, ..., x_n)$, denote realized value of $\sum_{i=1}^n x_i$ by, say, A. Since large values of A is a sign in favor of the alternative, we should reject the null if A is large. In the previous section, we showed that the test of level 5% rejects the null if $A \geq n(\sigma Z_{0,95}/\sqrt{n} + \mu_0)$. The quantity on the right hand side of this inequality satisfies $P_{\mu_0}\{\sum_{i=1}^n X_i \geq n(\sigma Z_{0,95}/\sqrt{n} + \mu_0)\} = 0.05$. The samples $(X_1, ..., X_n)$ which are more adverse to the null than our sample are those for which $\sum_{i=1}^n X_i > A$. So, the

$$p - value = P_{\mu_0}\{\sum_{i=1}^n X_i > A\} = 1 - \Phi\left(\sqrt{n}\frac{A/n - \mu_0}{\sigma}\right)$$

Note again, that it is a function of A, and thus is a random variable. By construction, the p-value is smaller than 0.05 if and only if $A \geq n(\sigma Z_{0,95}/\sqrt{n} + \mu_0)$. So, if we have a test at the level 0.05, then 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 are 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 an 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 of 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 $N(\mu, \sigma^2)$ 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 complex since both of them 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.

Example As another example, let $X_1, ..., X_n$ be a random sample from distribution $N(\mu, \sigma^2)$ with unknown σ^2. The null hypothesis is that $H_0 : \mu = \mu_0$. The alternative is that $H_a : \mu \neq \mu_0$. Again, both hypotheses are complex since both of them contain all possible values of σ^2. 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 $\delta > 0$. Under the null, $|\overline{X}_n - \mu_0| \sim N(0, \sigma^2/n)$. Since we do not know σ^2 we cannot choose δ so that $P_{\mu_0}\{|\overline{X}_n - \mu_0| > \delta\} = 0.05$. It is because statistic $|\overline{X}_n - \mu_0|$ is not pivotal in this case. One way to proceed is to estimate σ^2 by, say s^2, and then use a pivotal statistic. We know that, under the null, $(\overline{X}_n - \mu_0)/\sqrt{s^2/n} \sim t(n-1)$. Again, a reasonable test should be based on the critical region $C = \{|\overline{X}_n - \mu_0|/\sqrt{s^2/n} \leq \delta\}$. Since the pdf of t-distribution is symmetric around zero, in particular $t_{0,975}(n-1) = -t_{0.025}(n-1)$, a test with level, say, 5% accepts the null hypothesis if

$|\overline{X}_n - \mu_0|/\sqrt{s^2/n} \leq t_{0.975}(n-1)$. Indeed,

$$\begin{aligned} P_{\mu_0}\left\{\frac{|\overline{X}_n - \mu_0|}{\sqrt{s^2/n}} \leq t_{0.975}(n-1)\right\} &= P_{\mu_0}\left\{\frac{\overline{X}_n - \mu_0}{\sqrt{s^2/n}} \leq t_{0.975}(n-1)\right\} - P_{\mu_0}\left\{\frac{\overline{X}_n - \mu_0}{\sqrt{s^2/n}} \leq t_{0.025}(n-1)\right\} \\ &= 0.975 - 0.025 \\ &= 0.95 \end{aligned}$$

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

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

where s_x^2 and s_y^2 are sample variances. Exact distribution of $s_x^2/m + s_y^2/n$ is not pleasant. Instead, let us use asymptotic theory. By the Law of large numbers

$$\begin{aligned} \frac{s_x^2/m + s_y^2/n}{\sigma_x^2/m + \sigma_y^2/n} &= \frac{(\sigma_x^2/m)(s_x^2/\sigma_x^2) + (\sigma_y^2/n)(s_y^2/\sigma_y^2)}{\sigma_x^2/m + \sigma_y^2/n} \\ &= \frac{(\sigma_x^2/m)\chi_{m-1}^2/(m-1) + (\sigma_y^2/n)\chi_{n-1}^2/(n-1)}{\sigma_x^2/m + \sigma_y^2/n} \\ &\to_p \frac{\sigma_x^2/m + \sigma_y^2/n}{\sigma_x^2/m + \sigma_y^2/n} \\ &= 1 \end{aligned}$$

Thus, by the Slutsky theorem, $t \Rightarrow N(0,1)$. So we can use quantiles of standard normal distribution to form a test with size approximately equal to the required level of the test. This gives us a test with "asymptotically correct size".

14.381 Statistical Method in Economics
Fall 2013