14.30 Introduction to Statistical Methods in Economics
Spring 2009

14.30 Introduction to Statistical Methods in Economics
Lecture Notes 21

Konrad Menzel

May 5, 2009

Constructing Hypothesis Tests

If X_i has support S_X, then the sample $X = (X_1, \dots, X_n)$ has support S_X^n. The *critical region* of a test is a region $C_X \subset S_X^n$ of the support of the sample for which we reject the null.

The following example illustrates most of the important issues in a standard setting, so you should look at this carefully and make sure that you know how to apply the same steps to similar problems.

Example 1 *Suppose* $X_1, \dots, X_n$ *are i.i.d. with* $X_i \sim N(\mu, 4)$*, and we are interested in testing* $H_0 : \mu = 0$ *against* $H_A : \mu = 1$*. Let's first look at the case* $n = 2$*:*

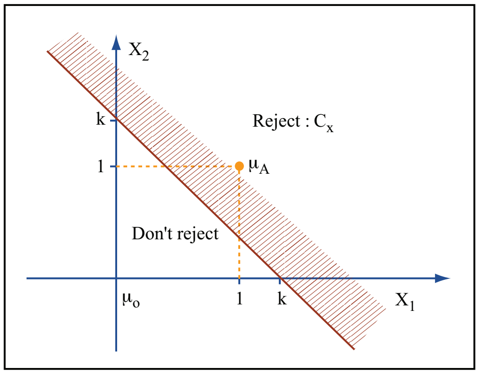

Image by MIT OpenCourseWare.

We could design a test which rejects for values of $X_1 + X_2$ *which are "too large" to be compatible with* $\mu = 0$*. We can also represent this rejection region on a line:*

This representation is much easier to use if n *is large, so it's hard to visualize the rejection region in terms of* $X_1, \dots, X_n$ *directly. However, by condensing the picture from* n *to a single dimension we may loose the ability of specifying really odd-shaped critical regions, but typically those won't be interesting for practical purposes anyway.*
So in this example, we will base our testing procedure on a test statistic $T_n(X_1, \dots, X_n) = \bar{X}_n$ *and reject for large values of* T_n*.*
How do we choose k*? - we'll have to trade off the two types of error. Suppose now that* $n = 25$*, and since* $X_i \sim N(\mu, 4)$*,*

$$T_n := \bar{X}_n \sim \begin{cases} N\left(0, \frac{4}{25}\right) & \text{under } H_0 \\ N\left(1, \frac{4}{25}\right) & \text{under } H_A \end{cases}$$

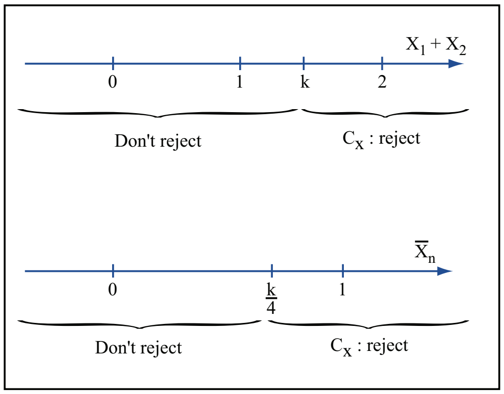

Image by MIT OpenCourseWare.

Now we can calculate the probabilities of type I and type II error

$$\begin{aligned} \alpha &= P(\bar{X} > k|\mu = 0) = 1 - \Phi\left(\frac{k-0}{2/5}\right) = \Phi\left(-\frac{k}{2/5}\right) \\ \beta &= P(\bar{X} \leq k|\mu = 1) = \Phi\left(\frac{k-1}{2/5}\right) \end{aligned}$$

Therefore, fixing any one of α, β, k determines the other two, and that choice involves a specific tradeoff between the probability of type I and type II error - if we increase k, the significance level α goes down, but so does power $1 - \beta$. Specifically, if we choose $k = \frac{3}{5}$, $\alpha \approx 6.7\%$, and $\beta \approx 15.87\%$.
For different sample sizes, we can graph the trade-off between the probability of type I and type II error through the choice of k as follows:

A low value of k would give high power, but also a high significance level, so that increasing k would move us to the left along the frontier.

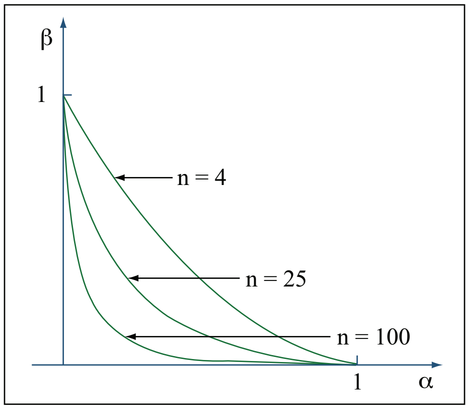

Image by MIT OpenCourseWare.

How should we choose k? Recall that in the usual setting, the first priority is to control the probability α of false rejections, so we'll choose k to keep the probability of a type I error at an acceptably low level, usually 5% or 1%.
Of course, as $n \to \infty$, for fixed α, the power of the test, $1 - \beta$ goes to 1. As a convention, we usually say that a rejection at the significance level $\alpha = 5\%$ is "significant", whereas a rejection at $\alpha = 1\%$ is "highly significant."

Example 2 *In the previous example, the maintained hypothesis was that $\mu \in \{0, 1\}$, but this is a highly artificial assumption, and usually we have no reason to believe that this is the case.*
Suppose that as before $X_1, \dots, X_{25}$ is an i.i.d. sample with $X_i \sim N(\mu, 4)$, but now we want to test

$$H_0 : \mu = 0 \text{ against } H_A : \mu \neq 0$$

Now H_A is a two-sided composite *hypothesis (i.e. under the alternative μ could take several values, some on the left, some on the right of μ_0). Also we'll again look at a test that's only based on the sample mean, $\bar{X}_{25}$ - what should the critical region now look like?*
Intuitively, it makes sense to reject H_0 for both large and small values of $\bar{X}$, i.e. we are unlikely to see values in either tail if the null hypothesis is true, and the alternative hypothesis states that we are interested in evidence that μ is either greater or smaller than 0.

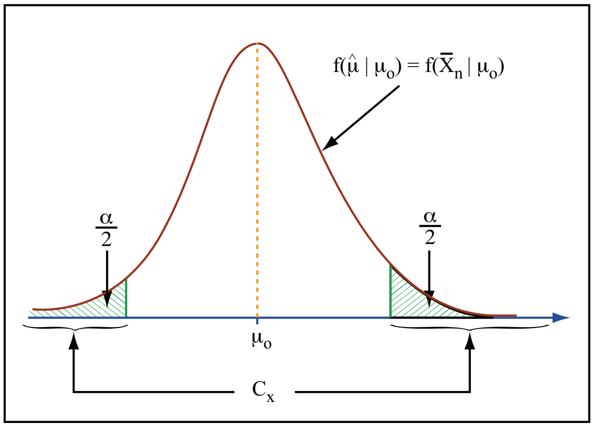

Image by MIT OpenCourseWare.

Therefore, we are going to choose two values k_1 and k_2 such that

$$\alpha = P(\bar{X}_{25} > k_2 | \mu = 0) + P(\bar{X}_{25} < k_1 | \mu = 0) = \left[1 - \Phi\left(\frac{k_2}{2/5}\right)\right] + \Phi\left(\frac{k_1}{2/5}\right)$$

What is β? Since the alternative does not specify a single probability law, but rather a continuum of them, β is not well-defined, i.e. for fixed μ,

$$\beta(\mu) = P(\bar{X}_{25} > k_2 | \mu) + P(\bar{X}_{25} < k_1 | \mu) = \left[1 - \Phi\left(\frac{k_2 - \mu}{2/5}\right)\right] + \Phi\left(\frac{k_1 - \mu}{2/5}\right)$$

Usually for a desired significance level α, we choose k_1, k_2 symmetrically about the value postulated by the null hypothesis (note that since the normal distribution is single-peaked and symmetric, this makes the critical region as large as possible.

The last example should remind you of the way we constructed confidence intervals for μ from a normal population with known variance: the above procedure is in fact identical to the following:

1. construct a $1 - \alpha$ confidence interval $[A(X), B(X)]$ for μ (case 1, see notes for last class)

2. reject H_0 if $\mu_0 = 0 \notin [A(X), B(X)]$

Since we construct the interval $[A(X), B(X)]$ in such a way that $P_\theta(A(X) < \theta < B(X)) = 1 - \alpha$ implicitly under the null assumptions, so that

$$P(\theta_0 \notin [A(X), B(X)] | H_0 : \theta = \theta_0) = \alpha$$

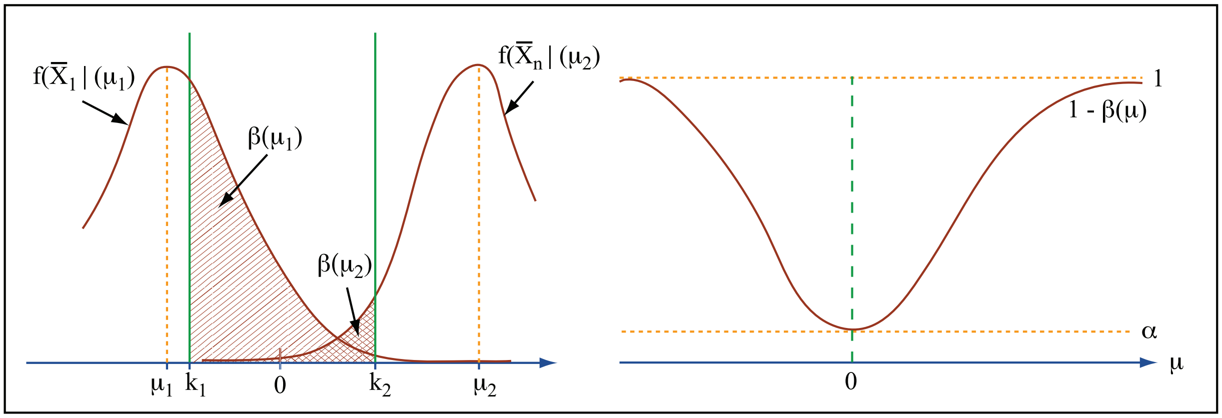

Image by MIT OpenCourseWare.

1 Evaluation and Construction of Tests

As in our discussion of estimation, we first introduced the general idea and saw a few examples. Now we will see how to choose among tests and how to construct them from scratch.

1.1 Properties of Tests

For any test we care about its level of significance $\alpha = P(\text{type I})$ and its power $1 - \beta = 1 - P(\text{type II})$. If H_0 and H_A are both simple hypotheses, α and β are well-defined for given α, and we can simply choose the test with the highest $1 - \beta$, the *most powerful test.*
If H_A is composite, and H_0 is simple, we need a metric for comparing *power functions* $1 - \beta(\theta) = 1 - P(\text{type II}|\theta)$ for a given size α. A test is *uniformly most powerful* (UMP) when it is at least as powerful at every $\theta \in H_A$ as any other test of the same size. In general, a UMP test need not exist.

Example 3 *Sometimes it is possible to find a uniformly most powerful test: Suppose $X_i \sim N(\mu, 4)$, and we are interested in testing*

$$H_0 : \mu = 0 \text{ against } H_A : \mu > 0$$

Recall that the most powerful test for $H_0 : \mu = 0$ against $H_A : \mu = 1$ took the form of reject if $\bar{X} > k$, the general form of the test does not change no matter what μ_A as long as $\mu_A > \mu_0$. Therefore a most powerful test for $H_0 : \mu = 0$ against $H_A : \mu > 0$ will also take the form "reject if $\bar{X}_n > k$."

For the following important result, denote $f_0(\mathbf{x}) = f_0(x_1, \dots, x_n)$ the joint p.d.f. of the sample $X_1, \dots, X_n$ under the simple null hypothesis $H_0 : \mu = \mu_0$, and $f_A(\mathbf{x})$ the joint p.d.f. of the sample under $H_A : \mu = \mu_A$.

Proposition 1 ***(Neyman-Pearson Lemma)*** *In testing f_0 against f_A (where both H_0 and H_A are simple hypotheses), the critical region*

$$C(k) = \left\{ \mathbf{x} : \frac{f_0(\mathbf{x})}{f_A(\mathbf{x})} < k \right\}$$

is most powerful for any choice of $k \geq 0$.

Note that the choice of k depends on the specified significance level α of the test. This means that the most powerful test rejects if for the sample $X_1, \dots, X_n$, the *likelihood ratio*

$$r(X_1, \dots, X_n) = \frac{f_0(X_1, \dots, X_n)}{f_A(X_1, \dots, X_n)}$$

is low, i.e. the data is much more likely to have been generated under H_A.

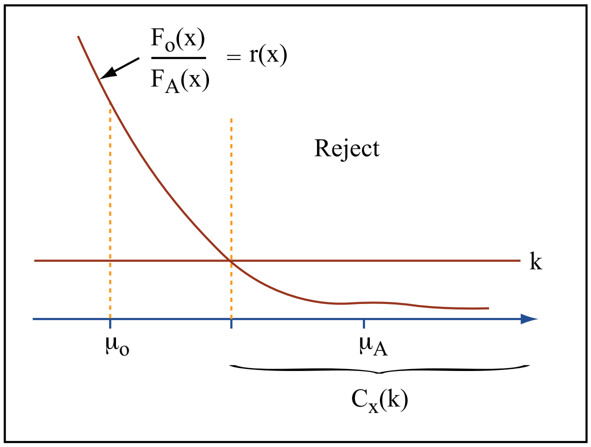

Image by MIT OpenCourseWare.

The most powerful test given in the Neyman-Pearson Lemma explicitly solves the trade-off between size

$$\alpha = P(\text{reject}|H_0) = \int_{C(k)} f_0(\mathbf{x})d\mathbf{x}$$

and power

$$1 - \beta = P(\text{reject}|H_A) = \int_{C(k)} f_A(\mathbf{x})d\mathbf{x}$$

at every point $\mathbf{x}$ in the sample space (where the integrals are over many dimensions, e.g. typically $\mathbf{x} \in \mathbb{R}^n$). From the expressions for α and $1-\beta$ we can see that the likelihood ratio $\frac{f_0(\mathbf{x})}{f_A(\mathbf{x})}$ gives the "price" of including $\mathbf{x}$ with the critical region in terms of how much we "pay" in terms of size α relative to the gain in power from including the point in the critical region C_X.

Therefore, we should start constructing the critical region by including the "cheapest" points $\mathbf{x}$ - i.e. those with a small likelihood ratio. Then we can go down the list of $\mathbf{x}$ ordered according to the likelihood ratio and continue including more points until the size α of the test is down to the desired level.

Example 4 *A criminal defendant (D) is on trial for a purse snatching. In order to convict, the jury must believe that there is a 95% chance that the charge is true.*
There are three potential pieces of evidence the prosecutor may or may not have been able to produce, and in a given case the jury takes a decision to convict based only on which out of the three clues it is presented with. Below are the potential pieces of evidence, assumed to be mutually independent, the probability of observing each piece given the defendant is guilty, and the probability of observing each piece given the defendant is not guilty

		guilty	not guilty	likelihood ratio
1.	D ran when he saw police coming	0.6	0.3	1/2
2.	D has no alibi	0.9	0.3	1/3
3.	Empty purse found near D's home	0.4	0.1	1/4

In the notation of the Neyman-Pearson Lemma, $\mathbf{x}$ *can be any of the* 2^3 *possible combinations of pieces of evidence. Using the assumption of independence, we can therefore list all possible combinations of clues with their respective likelihood under each hypothesis and the likelihood ratio. I already ordered the list by the likelihood ratios in the third column. In the last column, I added*

$$\alpha(k) = \sum_{r(\mathbf{x}) \leq k} f_0(\mathbf{x})$$

the cumulative sum over the ordered list of combinations $\mathbf{x}$*.*

		guilty $f_A(\mathbf{x})$	not guilty $f_0(\mathbf{x})$	likelihood ratio $r(\mathbf{x}) = \frac{f_0(\mathbf{x})}{f_A(\mathbf{x})}$	$\alpha(k)$
1.	all three clues	216/1000	9/1000	0.0417	9/1000
2.	no alibi,found purse	144/1000	21/1000	0.1458	30/1000
3.	ran,no alibi	324/1000	81/1000	0.25	111/1000
4.	no alibi	216/1000	189/1000	0.875	300/1000
5.	ran,found purse	24/1000	21/1000	0.875	321/1000
6.	found purse	16/1000	49/1000	3.0625	370/1000
7.	ran	36/1000	189/1000	5.25	559/1000
8.	none of the clues	24/1000	441/1000	18.375	1

The jury convicting the defendant only if there is at least 95% confidence that the charge is true corresponds to a probability of false conviction (i.e. if the defendant is in fact innocent) of less than 5%. In the terminology of hypothesis test, the sentence corresponds to a rejection of the null hypothesis that the defendant is innocent using the most powerful test of size $\alpha = 5\%$*.*
Looking at the values of $\alpha(k)$ *in the last column of the table, we can read off that including more than the first two combinations of the evidence raises the probability of a false conviction* α *to more than 5%. Therefore, the jury should convict the defendant if he doesn't have an alibi* and *the empty purse was found near his home, regardless whether he ran when he saw the police. In principle, the jury could in addition randomize when the defendant ran, had no alibi, but no purse was found (that is case 3): if in that case, the jury convicted the defendant with probability* $\frac{50-30}{81} \approx \frac{1}{4}$*, the probability of a false conviction would be exactly equal to 5%, but this would probably not be considered an acceptable practice in criminal justice.*

Example 5 *We can now show that a test based on the mean is in fact most powerful in the normal case. Suppose* $X_i \sim N(\mu, 4)$*, and we test* $H_0 : \mu = 0$ *against* $H_A : \mu = 1$*, where we observe an i.i.d. sample* $\mathbf{X} = (X_1, \ldots, X_{25})$ *of 25 observations.*
Since the observations are i.i.d. normal, the likelihood ratio evaluated at the observed sample is given by

$$\begin{aligned} r(\mathbf{X}) &= \frac{f(\mathbf{X}|\mu = 0)}{f(\mathbf{X}|\mu = 1)} = \prod_{i=1}^{25} \frac{\frac{1}{\sqrt{2\pi 2}} e^{-\frac{(X_i - 0)^2}{2 \cdot 4}}}{\frac{1}{\sqrt{2\pi 2}} e^{-\frac{(X_i - 1)^2}{2 \cdot 4}}} \\ &= \prod_{i=1}^{25} e^{\frac{1}{8}\left[(X_i^2 - 2X_i + 1) - (X_i^2)\right]} \\ &= e^{\frac{25}{8} - \frac{1}{4}\sum_{i=1}^{25} X_i} = e^{\frac{25}{8} - \frac{1}{100}\bar{X}_{25}} \end{aligned}$$

We can see that $r(\mathbf{X})$ *depends on the sample only through the sample mean* $\bar{X}_{25}$ *and is strictly decreasing in* $\bar{X}_{25}$*. Therefore, the critical region of a most powerful test takes the form*

$$C_X(k) = \{\mathbf{x} : r(\mathbf{x}) \leq k\}$$