14.30 Introduction to Statistical Methods in Economics

Spring 2009

14.30 Introduction to Statistical Methods in Economics
Lecture Notes 22

Konrad Menzel

May 7, 2009

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 $X_1, \ldots, X_n$, the *likelihood ratio*

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

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

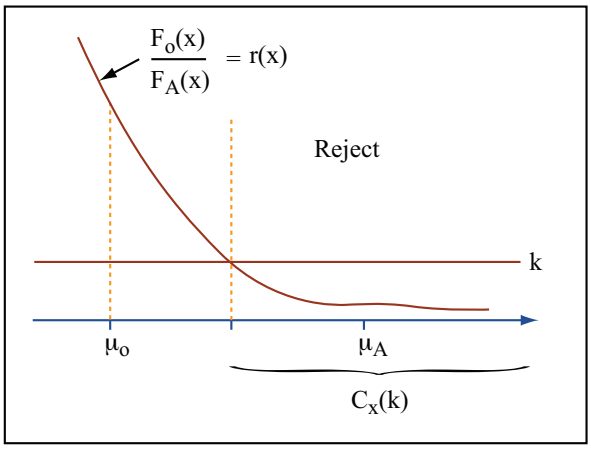

Image by MIT OpenCourseWare.

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 1 *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.*

1 Construction of Tests

In general there is no straightforward answer to how we should construct an optimal test. The Neyman-Pearson Lemma gave us a simple recipe for a most powerful test of one simple hypothesis against another, but in most real-world applications, the alternative hypothesis is composite. The following is a list of recommendations which do not always lead to a uniformly most powerful test (which sometimes does not even exist), but usually yield reasonable procedures:

1. if both H_0 and H_A are simple, the Neyman-Pearson Lemma tells us to construct a statistic
$$T(\mathbf{x}) = \frac{f_0(\mathbf{x})}{f_A(\mathbf{x})}$$
and reject if $T(\mathbf{X}) > k$ for some appropriately chosen value k (typically k is chosen in a way that makes sure that the test has size α). This test is also called the *likelihood ratio test* (LRT).
2. if $H_0 : \theta = \theta_0$ is simple and $H_A : \theta \in \Theta_A$ is composite and 2-sided, we construct a $1-\alpha$ confidence interval $[A(X), B(X)]$ (usually symmetric) using an estimator $\hat{\theta}$. We then reject if $\theta_0 \notin [A(\mathbf{X}), B(\mathbf{X})]$. This gives us a test of size α for H_0.
3. if $H_0 : \theta = \theta_0$ is simple and $H_A : \theta \in \Theta_A$ is composite and one-sided, we construct a symmetric $1-2\alpha$ confidence interval for θ and reject only if the null value is outside the confidence interval *and* in the relevant tail in order to obtain a size α test.
4. either $H_0 : \theta \in \Theta_0$ or $H_A : \theta \in \Theta_A$ composite (or both): define the statistic
$$T(\mathbf{x}) = \frac{\max_{\theta\in\Theta_0} L(\theta)}{\max_{\theta\in\Theta_A\cup\Theta_0} L(\theta)} = \frac{\max_{\theta\in\Theta_0} f(\mathbf{x}|\theta)}{\max_{\theta\in\Theta_A\cup\Theta_0} f(\mathbf{x}|\theta)}$$
and reject if $T(\mathbf{X}) < k$ for some appropriately chosen constant k. This type of test is called the *generalized likelihood ratio test* (GLRT).

Since we haven't discussed the last case yet, some remarks are in order:

- the test makes sense because $T(\mathbf{X})$ will tend to be small if the data don't support H_0
- densities are always positive, so the statistic will be between 0 and 1 (this is because the set over which the density is maximized in the denominator contains the set over which we maximize in the numerator)
- we need to know the exact distribution of the test statistic under the null hypothesis, so that we can find an appropriate critical value k. For most distributions we have that in large samples
$$-2\log T(\mathbf{X}) \sim \chi^2_p$$
where $p = \dim(\Theta_0 \cup \Theta_A) - \dim(\Theta_0)$.
- the GLRT does not necessarily share the optimality properties of the LRT, in fact in this setting with a composite alternative hypothesis a uniformly most powerful test often does not even exist.

2 Examples

Example 2 *Assume that babies' weights (in pounds) at birth are distributed according to* $X \sim N(7,1)$. *Now suppose that if an obstetrician gave expecting mothers poor advice on diet, this would cause babies to be on average 1 pound lighter (but have same variance). For a sample of 10 live births, we observe* $\bar{X}_{10} = 6.2$.

- *How do we construct a* 5% *test of the null that the obstetrician is not giving bad advice against the alternative that he is? We have*

$$H_0 : \mu = 7 \text{ against } H_A : \mu = 6$$

We showed that for the normal distribution, it is optimal to base this simple test only on the sample mean, $\bar{X}_{10}$, *so that* $T(\mathbf{x}) = \bar{x}_{10}$. *Under* H_0, $\bar{X}_{10} \sim N(7, 0.1)$ *and under* H_A, $\bar{X}_{10} \sim N(6, 0.1)$. *The test rejects* H_0 *if* $\bar{X}_{10} < k$. *We therefore have to pick* k *in a way that makes sure that the test has size 5%, i.e.*

$$0.05 = P(\bar{X}_{10} < k | \mu = 7) = \Phi\left(\frac{k-7}{\sqrt{0.1}}\right)$$

where $\Phi(\cdot)$ *is the standard normal c.d.f.. Therefore, we can obtain* k *by inverting this equation*

$$k = 7 + \sqrt{0.01}\Phi^{-1}(0.05) \approx 7 - \frac{1.645}{\sqrt{10}} \approx 6.48$$

Therefore, we reject, since $\bar{X}_{10} = 6.2 < 6.48 = k$.

- *What is the power of this test?*

$$P(\bar{X}_{10} < 6.48 | \mu = 6) = \Phi\left(\frac{6.48-6}{\sqrt{0.1}}\right) \approx \Phi(1.518) \approx 93.55\%$$

- *Suppose we wanted a test with power of at least 99%, what would be the minimum number* n *of newborn babies we'd have to observe? The only thing that changes with* n *is the variance of the sample mean, so from the first part of this example, the critical value is* $k_n = 7 - \frac{1.645}{\sqrt{n}}$, *whereas the power of a test based on* $\bar{X}_n$ *and critical value* k_n *is*

$$1 - \beta = P(\bar{X}_n < k_n | \mu = 6) = \Phi\left(\sqrt{n} - 1.645\right)$$

Setting $1 - \beta \geq 0.99$, *we get the condition*

$$\sqrt{n} - 1.645 \geq \Phi^{-1}(0.99) = 2.326 \Leftrightarrow n \geq 3.971^2 \approx 15.77$$

This type of power calculations is frequently done when planning a statistical experiment or survey - e.g. in order to determine how many patients to include in a drug test in order to be able to detect an effect of a certain size. Often it is very costly to treat or survey a large number of individuals, so we'd like to know beforehand how large the experiment should be so that we will be able to detect any meaningful change with sufficiently high probability.

Example 3 *Suppose we are still in the same setting as in the previous example, but didn't know the variance. Instead, we have an estimate* $S^2 = 1.5$. *How would you perform a test? As we argued earlier, the statistic*

$$T := \frac{\bar{X}_n - \mu_0}{S/\sqrt{n}} \sim t_{n-1}$$

is student-t distributed with $n-1$ degrees of freedom if the true mean is in fact μ_0. Therefore we reject H_0 if

$$T = \frac{\bar{X}_n - 7}{S/\sqrt{10}} < t_9(5\%)$$

Plugging in the values from the problem, $T = -\frac{0.8}{\sqrt{1.5/10}} \approx -2.066$, which is smaller than $t_9(0.05) = -1.83$.

Example 4 *Let $X_i \sim \text{Bernoulli}(p)$, $i = 1, 2, 3$. I.e. we are flipping a bent coin three times independently, and $X_i = 1$ if it comes up heads, otherwise $X_i = 0$. We want to test $H_0 : p = \frac{1}{3}$ against $H_A : p = \frac{2}{3}$. Since both hypotheses are simple, can use likelihood ratio test*

$$T = \frac{f_0(X)}{f_A(X)} = \frac{\prod_{i=1}^3 \left(\frac{1}{3}\right)^{X_i} \left(\frac{2}{3}\right)^{1-X_i}}{\prod_{i=1}^3 \left(\frac{2}{3}\right)^{X_i} \left(\frac{1}{3}\right)^{1-X_i}} = \frac{2^{3-\sum_{i=1}^3 X_i}}{2^{\sum_{i=1}^3 X_i}} = 2^{3-2\sum_{i=1}^3 X_i}$$

Therefore, we reject if

$$2^{3-2\sum_{i=1}^3 X_i} \leq k \Leftrightarrow (3 - 2\sum_{i=1}^3 X_i)\log 2 \leq \log k$$

which is equivalent to $\bar{X}_3 \geq \frac{1}{2} - \frac{\log k}{6 \log 2}$. In order to determine k, let's list the possible values of $\bar{X}_3$ and their probabilities under H_0 and H_A, respectively:

$\bar{X}_3$	Prob. under H_0	Prob. under H_A	cumul. prob. under H_0
1	$\frac{1}{27}$	$\frac{8}{27}$	$\frac{1}{27}$
$\frac{2}{3}$	$\frac{6}{27}$	$\frac{12}{27}$	$\frac{7}{27}$
$\frac{1}{3}$	$\frac{12}{27}$	$\frac{6}{27}$	$\frac{19}{27}$
0	$\frac{8}{27}$	$\frac{1}{27}$	1

So if we want the size of the test equal to $\alpha = \frac{1}{27}$, we could reject if and only if $\bar{X}_3 > \frac{2}{3}$, or equivalently we can pick $k = \frac{1}{2}$. The power of this test is equal to

$$1 - \beta = P(\bar{X}_3 = 1|H_A) = \frac{8}{27} \approx 29.63\%$$

Example 5 *Suppose we have one single observation generated by either*

$$f_0(x) = \begin{cases} 2x & \text{if } 0 \leq x \leq 1 \\ 0 & \text{otherwise} \end{cases} \quad \text{or } f_A(x) = \begin{cases} 2 - 2x & \text{if } 0 \leq x \leq 1 \\ 0 & \text{otherwise} \end{cases}$$

- *Find the testing procedure which minimizes the sum of $\alpha+\beta$ - do we reject if $X = 0.6$? Since we only have one observation X, it's not too complicated the critical region directly in terms of X, and there is nothing to be gained by trying to find some clever statistic (though of course Neyman-Pearson would still work here). By looking at a graph of the densities, we can convince ourselves that the test should reject for small values of X. The probability of type I and type II error is, respectively,*

$$\alpha(k) = P(\text{reject}|H_0) = \int_0^k 2x dx = k^2$$

for $0 \leq k \leq 1$, and

$$\beta(k) = P(\text{don't reject}|H_A) = \int_k^1 (2 - 2x)dx = 2(1 - k) - 1 + k^2 = 1 - k(2 - k)$$

Therefore, minimizing the sum of the error probabilities over k,

$$\min_k\{\alpha(k)+\beta(k)\} = \min_k\{k^2+1-k(2-k)\} = \min_k\{2k^2+1-2k\}$$

Setting the first derivative of the minimand to zero,

$$0 = 4k-2 \Leftrightarrow k = \frac{1}{2}$$

Therefore we should reject if $X < \frac{1}{2}$, and $\alpha = \beta = \frac{1}{4}$. Therefore, we would in particular not reject H_0 for $X = 0.6$.

- *Among all tests such that $\alpha \leq 0.1$, find the test with the smallest β. What is β? Would you reject if $X = 0.4$? - first we'll solve $\alpha(k) = 0.1$ for k. Using the formula from above, $\bar{k} = \sqrt{0.1}$. Therefore,*

$$\beta(\bar{k}) = 1 - 2\bar{k} + \bar{k}^2 = 1.1 - 2\sqrt{0.1} \approx 46.75\%$$

Since $k = \sqrt{0.1} \approx 0.316 < 0.4$, we don't reject H_0 for $X = 0.4$.

Example 6 *Suppose we observe an i.i.d. sample $X_1, \ldots, X_n$, where $X_i \sim U[0,\theta]$, and we want to test*

$$H_0 : \theta = \theta_0 \text{ against } H_A : \theta \neq \theta_0, \theta > 0$$

There are two options: we can either construct a $1-\alpha$ confidence interval for θ and reject if it doesn't cover θ_0. Alternatively, we could construct a GLRT test statistic

$$T = \frac{L(\theta_0)}{\max_{\theta\in\mathbb{R}_+} L(\theta)}$$

The likelihood function is given by

$$L(\theta) = \prod_{i=1}^n f_X(X_i|\theta) = \begin{cases} \left(\frac{1}{\theta}\right)^n & \text{for } 0 \leq X_i \leq \theta, i = 1,\ldots,n \\ 0 & \text{otherwise} \end{cases}$$

The denominator of T is given by the likelihood evaluated at the maximizer, which is the maximum likelihood estimator, $\hat{\theta}_{MLE} = X_{(n)} = \max\{X_1,\ldots,X_n\}$, so that

$$\max_{\theta\in\mathbb{R}_+} L(\theta) = L(\hat{\theta}_{MLE}) = \left(\frac{1}{X_{(n)}}\right)^n$$

Therefore,

$$T = \frac{L(\theta_0)}{\max_{\theta\in\mathbb{R}_+} L(\theta)} = \left(\frac{X_{(n)}}{\theta_0}\right)^n$$

In order to find the critical value k of the statistic which makes the size of the test equal to the desired level, we'd have to figure out the distribution under the null $\theta = \theta_0$ - could look this up in the section on order statistics.
As an aside, even though we said earlier that for large n, the GLRT statistic is χ^2-distributed under the null, this turns out not to be true for this particular example because the density has a discontinuity at the true parameter value.