MSF100
Statistical Inference Principles - Lecture 8

Rebecka Jörnsten
Mathematical Statistics
University of Gothenburg/Chalmers University of Technology

February 12, 2012

1 Testing

We can think of testing as a particular kind of estimation problem. Let $\delta(\tilde{X})$ be a statistic that takes on values 0 or 1. A value of 0 means you don't reject a hypothesis (to be specified later) and a value of 1 means you reject. Then $E_{\theta_0}[\delta(\tilde{X})] = P_{\theta_0}(\delta(\tilde{X}) =)$ relates to the test outcome under a particular true value for the parameter θ_0. The outcome can be correct or incorrect.

1.1 Setup

We have $\tilde{X}$ iid $f_\theta(x)$ with $\theta \in \Theta$ indexing a known model family. We consider two hypotheses: $H_0 : \theta \in \Theta_0$ is called the null hypothesis and $H_1 : \theta \in \Theta_0^c$ the alternative. Together they comprise the whole parameter space.

We construct a rule $\delta(\tilde{X})$ s.t. $\delta(\tilde{X}) = 1$ means we reject H_0. Usually this rule is a function of a statistic $T(\tilde{X})$ and not $\tilde{X}$ directly - just like when we did estimation of a function $q(\theta)$. To go from a statistic $T(\tilde{X})$ to a rule δ we define a *rejection region* such that

$$\delta(\tilde{X}) = 1\{T(\tilde{X}) \in R\} = 1\{\tilde{X} \in R'\}$$

Example:

- We play a coin tossing game. The probability of a win in each game is θ
- We want to test the hypothesis $H_0 : \theta \geq 1/2$ against $H_1 : \theta < 1/2$.
- We play n times and get X wins with $T(X) = X/n$ the proportion of wins in the game.
- Consider two different rejection regions: $R_1 = \{x : T(x) < 1/2\}$ and $R_2 = \{x : T(x) < 0.35\}$
- What if $\theta_0 = 1/2$? With rejection region R_1 we reject this hypothesis with probability

  $$P_{\theta_0}(T(x) < 1/2) = P(Bin(n, 1/2) < n/2)$$

  which is 0.54 for $n = 100$ and 0.63 for $n = 10$
- Using R_3 these probabilities are 0.001 and 0.17 respectively
- So, using R_3 means you do not falsely reject the hypothesis $\theta_0 = 1/2$ as frequently
- What if the true $\theta_0 = 1/4$ meaning H_0 is false?
- Using R_1 we reject the null 100% of the times when $n = 100$ and 98% of the time when $n = 10$
- Using R_2 this is reduced to 99% and 78% respectively.
- The example illustrates that it is difficult to construct a test that behaves as desired in the region of the null AND in the region of the alternative

1.2 Type errors

The above can be summarized by two kinds of errors:

- Type I error: rejecting H_0 when H_0 is true
- Type II error: failing to reject H_0 when H_0 is false
- Correct decision: rejection H_0 when H_0 is false, failing to reject H_0 when H_0 is true

Note that I always say "fail to reject" never "accept". The reason for this is that all calculations done to construct tests are done under the assumption that the null is true. "Accept" sounds like you can prove the null to be true, but that would be a circular argument.

We look more closely at the two errors:

$$P(type\ I) = P(\delta(X) = 1|H_0) = P(X \in R|H_0)$$

This error rate is usually fairly easy to retrieve - we only need to evaluate it for $\theta \in \Theta_0$ and often only for values on the boundary.

$$P(type\ II) = P(\delta(X) = 0|H_1) = P(X \notin R|H_1)$$

This can be more tricky. This error rate depends on the actual value $\theta \in \Theta_0^c$ and its distance from the Θ_0. It is more common to convert the type II error rate into a *power function*:

$$\beta(\theta) = P_\theta(\delta(X) = 1)$$

We want this power function to be as small as possible for $\theta \in \Theta_0$ (to control the type I error rate) and as large as possible for $\theta \in \Theta_0^c$ (to enable detection of the null not being true). Since you can't optimize these two goals at the same time, one working against the other, we usually make controlling the type I error rate our priority and look for tests with maximum power from those that control the type I error. We say that a test has level α if $P(typeI) \leq \alpha$.

1.3 Hypothesis testing

A simple hypothesis test compares two specific values of θ, e.g. θ_0 vs θ_1. A complex or composite hypothesis test is one that compares sets of parameter values, e.g. $\theta \leq \theta_0$ vs $\theta > \theta_0$.

We can say much more about the behaviour of simple hypothesis tests. For a simple test we can find R such that

$$P_{\theta_0}(\delta(X) = 1) = P_{\theta_0}(X \in R') = \alpha$$

(perhaps not exact if we have discreet distributions, but then $\leq \alpha$). We can also compute

$$\beta(\theta_1) = P_{\theta_1}(\delta(X) = 1)$$

In the complex case we look at

$$P_{H_0}(\delta(X) = 1) = P_{H_0}(X \in R') \leq \alpha, \quad for\ all\ \theta \in \Theta_0$$

and we need the inequality to be an equality for at least one point or R is too small. We then compute

$$\beta(\theta) = P(X \in R'|\theta), \quad for\ all\ \theta \in \Theta_0^c$$

The power function is especially interesting to look at for values of θ near the null. The sharpness of the power curve tells you how easy/difficult it is to distinguish between the null and the alternative. For values of θ far from the null almost any test will do the right thing and reject the null, but it's the behaviour near the boundary that tells you how useful the test will be in practise. As n gets larger, the power curve $\beta(\theta)$ gets sharper - meaning testing, like estimation, behaves better for large sample sizes.

We want to maximize the power outside the null parameter region while constraining the type I error rate at level α.
Example:

- Let's look at a multinomial example: outcomes A and B occur with probability θ and outcome C with probability $1-2\theta$.
- We want to test $H_0 : \theta \leq \theta_0 = 1/4$ against $H_1 : \theta > \theta_0$
- In n games we get X_1 A's, X_2 B's and $n-(X_1+X_2)$ C's.
- Test 1: Reject H_0 if X_1 is too large: $R_1 = \{X : X_1 > x\}$
- Test 2: Reject H_0 if $X_1 + X_2$ is too large: $R_2 : \{X : X_1 + X_2 > y\}$
- Is one test better than the other?
- We have $L(\theta|\tilde{X}) = \theta^{X_1}\theta^{X_2}(1-2\theta)^{n-X_1-X_2}const$
- That is, $L(\theta|\tilde{X}) = \theta^{X_1+X_2}(1-2\theta)^{n-X_1-X_2}const$ and so $T(X) = X_1 + X_2$ is a sufficient statistic.
- Test 1: $P(X_1 > x|\theta \leq 1/4) = P(Bin(n,\theta) > x|\theta \leq 1/4)$. We want this to be less than α
- Test 2: $P(T(X) > y|\theta \leq 1/4) = P(Bin(n, 2\theta) > y|\theta \leq 1/4)$
- Test 1: using a large sample approximation that $Bin(n,p) \sim N(np, np(1-p)$ we can compute

$$P(\frac{X-np}{\sqrt{np(1-p)}} = Z > \frac{x-np}{\sqrt{np(1-p)}}|p \leq 1/4) = \alpha$$

  at $p = 1/4$. This gives us that

$$x = np + z_{1-\alpha}\sqrt{np(1-p)}$$

  where z_q are quantiles of the standard normal. For $n = 100$ and $\alpha = .05$, $x = 32$.
- Test 2: via the same kind of calculation we find that $y = 58$ results in a level $\alpha = .05$ test.
- Let us now say we observe the following data set: $n = 100, X_1 = 30, X_2 = 25$. Then we fail to reject the null in either case.
- What if $\theta = 1/3$ in actuality?
- Test 1:

$$\beta(\theta = 1/3) = P(X_1 > 32|\theta = 1/3) = P(Bin(n, 1/3) > 32) = 0.56$$

- Test 2:

$$\beta(\theta = 1/3) = P(X_1 + X_2 > 58|\theta = 1/3) = P(Bin(n, 2/3) > 58) = 0.96$$

- Test 2 is clearly better (more powerful), at least for $\theta = 1/3$.
- Turns out that if you compare the power for other values of θ in the alternative, Test 2 is always more powerful than Test 1.
- Test 2 is thus *uniformly more powerful* than Test 1.

UMP test: If a test is uniformly more powerful than any other test of size α, the we call it the UMP (uniformly most powerful) test.

Usually, such a test does not exist except for special cases such as simple-vs-simple tests or one-sided tests for one-parameter exponential families. We can still try to construct tests that are as powerful as possible in principle, and we may not even care about all values of θ. As I said above, it's more important that you have a locally powerful test (near the null).

2 Neyman-Pearson lemma, LRT

NP lemma: We test a simple-vs-simple hypothesis: $H_0 : \theta = \theta_0$ vs $H_1 : \theta = \theta_1$. If we construct a test from the rejection region

$$R = \{\tilde{x} : \frac{f_{\theta_1}(\tilde{x})}{f_{\theta_0}(\tilde{x})} > k_\alpha\}$$

where k_α such that

$$P_{\theta_0}(\frac{f_{\theta_1}(\tilde{x})}{f_{\theta_0}(\tilde{x})} > k_\alpha) = \alpha$$

then it is the most powerful test.

If we go back to our example above:

- $H_0 : \theta = \theta_0$ and $H_1 : \theta = \theta_1$, note that

  $$\frac{f_{\theta_1}}{f_{\theta_0}} = (\frac{\theta_1}{\theta_0})^t(\frac{1-2\theta_1}{1-2\theta_0})^{n-t}$$

  where $t = x_1 + x_2$.
- Both terms of this ratio (assuming $\theta_1 > \theta_0$) increases with t and so setting a rejection on the ratio is equivalent to setting a rejection on t.
- Thus Test 2 is in fact a test of the form of the NP lemma.

Things get more complicated when we don't have simple-vs-simple hypotheses. The *likelihood ratio test* is defined as follows:

LRT, likelihood ratio test:

$$\lambda(\tilde{x}) = \frac{sup_{\Theta_0} L(\theta|\tilde{x})}{sup_{\Theta} L(\theta|\tilde{x})} = \frac{L(\hat{\theta}_0|\tilde{x})}{L(\hat{\theta}|\tilde{x})}$$

where $\hat{\theta}$ is the MLE and $\hat{\theta}_0$ is the maximizer of the likelihood in the restricted parameter space Θ_0. The rejection region is of the form

$$R = \{\tilde{x} : \lambda(\tilde{x}) < c_\alpha\}$$

where c_α is such that $P(\lambda(\tilde{x}) < c_\alpha | \theta \in \Theta_0) \leq \alpha$

Actually, this doesn't quite look like just a generalizing of the NP lemma because the denominator is not the sup over θ in the alternative. However, if the null is very small, these are almost the same, but foremost this is the chosen form for the LRT because it makes it easie to work out the sampling distribution for this statistic (the LRT) in practice.

Example

- 2 independent samples: $\tilde{x}$ n samples from $exp(\theta)$ and $\tilde{y}$ m samples from $exp(\mu)$
- Test $H_0 : \theta = \mu$
- The LRT is

  $$\lambda(\tilde{x}, \tilde{y}) = \frac{sup_\theta \frac{1}{\theta^n} e^{-\sum x/\theta} \frac{1}{\theta^m} e^{-\sum y/\theta}}{\sup_{\theta,\mu} \frac{1}{\theta^n} e^{-sumx/\theta} \frac{1}{\mu^m} e^{-\sum y/\mu}}$$

- The numerator is maximized at $\theta_0 = \frac{\sum x + \sum y}{n+m}$ and the denominator at $\hat{\theta} = \sum x/n$ and $\hat{\mu} = \sum y/m$.
- Plug this into the LRT to obtain

  $$\lambda(\tilde{x}, \tilde{y}) = \frac{(n+m)^{n+m}(\sum x)^n(\sum y)^m}{n^n m^m (\sum x + \sum y)^{n+m}}$$

- We reject the null when the LRT is less than some c. How can we figure out what c should be to give us a level α test?
- Denote $t = T(\tilde{x}, \tilde{y}) = \frac{\sum x}{\sum x + \sum y}$. Then we can write the LRT as

$$\lambda(\tilde{x}, \tilde{y}) = \frac{(n+m)^{n+m}}{n^n m^m} t^n (1-t)^m$$

- The LRT is maximized at $t = \frac{n}{n+m}$ and is unimodal. So the LRT rejects if $\lambda \leq c$ corresponds to rejecting when $T \leq a$ or $T \geq b$ for some a and b with $a^n(1-a)^m = b^n(1-b)^m$ (to make rejection as easy in either direction away from the max value).
- Now, sine $\sum x$ and $\sum y$ are sums of exponentials they are independent gamma-distributed. A ratio of the form of t follows a Beta distribution. We can now use the Beta distribution to figure out what a and b should be (looking at the upper and lower $\alpha/2$ quantiles of the Beta(n,m).

The LRT is a function of the sufficient statistic (follows from the factorization theorem). Now, the LRT is not a UMP in general but can be for certain special cases. We can also arrive at a likelihood based test that is *locally* most powerful.

We want to find a test with maximum power for θ near θ_0, e.g. for $\theta_1 = \theta_0 + \delta$. We look at the rate of change of the power function at θ_0 and want to maximize it (sharpness of the power function). Now, locally, comparing just these two values we can arrive at a test using the NP lemma as

$$L(\theta_0 + \delta)/L(\theta_0) > k_\alpha$$

We linearize the numerator (Taylor expansion) and write

$$\frac{L(\theta_0 + \delta)}{L(\theta_0)} \simeq \frac{L(\theta_0) + \delta L'(\theta_0)}{L(\theta_0)} > k_\alpha$$

We can write this as

$$\delta \frac{L'(\theta_0)}{L(\theta_0)} > c_\alpha$$

or (since $\delta > 0$)

$$\frac{L'(\theta_0)}{L(\theta_0)} > b_\alpha.$$

Using what we know about the derivative of logs we have

$$d \log L(\theta)|_{\theta_0} > b_\alpha$$

That is, we reject the null when the score $S(\theta_0)$ is too large. This is called the *score test*.