MA 2B RECITATION 02/16/12

1. Introduction

Homework will be due Tuesday due to the holiday.

Some notes from when the course was taught a couple years ago are available on the official course homepage, and they are certainly worth looking at, especially because they contain additional worked examples..

2. Hypothesis Testing

We're still trying to approach the same goal as last time. We want to approximate the true value of some unknown quantity that we're unable to figure out exactly because of lack of resources, like finding the average number of steps a random person takes in a day.

Now that we know about estimators, we're going to move onto the next level of sophistication. Here, we want to pick two hypotheses that encompass the whole parameter space. For instance, suppose we have a discrete random walk on a line.

Suppose that we know that at time t, you land at 0 or 1. Then we can pick hypotheses "I'm at 0" and "I'm at 1." We can also do this if we are not sure where we will land, but we are interested in knowing whether or not we will be at 0. Then we can divide the parameter space into "We're at 0" and "We're not at 0."

Be careful to observe that the situation is not symmetric. We focus our test on a particular event called the **null hypothesis**. Our goal with this procedure is to either accept or reject the null hypothesis.

2.1. Terminology.

There are a lot of new definitions to introduce here. It can be a bit overwhelming at first, but it will become clear once you work through a couple of examples.

Definition 2.1. Suppose that you are given a null hypothesis H_0.

The probability of rejecting H_0 when it is true ("false negative") is called the **significance level** and is usually denoted by α.

Let β be the probability of accepting H_0 when it is false ("false positive"). Then the quantity $1-\beta$ is called the **power** of the test, and represents the probability of correctly rejecting the hypothesis when it is false.

We often ask about the power "at" a value of the parameter we are trying to estimate. This means that assuming that the true value of the parameter is given (and that it is not in the null hypothesis), we calculate the probability β of rejecting the null hypothesis, and then take $1-\beta$, which corresponds to the probability that it doesn't happen.

The test **statistic** is the *number* that you compute from your data, like the sample mean. In this language, the "test" is determining whether or not this number lies within a certain range.

Date: February 16, 2011.

To "determine critical values" for a test with a given test statistic T and significance α means to find the values of T which should indicate acceptance or rejection of the null hypothesis in such a way that the probability of rejecting the null hypothesis when true is α.

To choose which hypothesis α gets to be the null hypothesis, a good rule of thumb is to pick the one that you would want to favor when you are unsure. For example, typically we set α to be small. Therefore, for statistics near the null hypothesis, we will default to the null hypothesis. That is, it is the one you assume unless it is proven wrong.

Remark 2.2. We can never accept the null hypothesis, we can only fail to reject it. These are *not* the same. Much like how you can never prove something by just giving examples of it, you can never prove that something is true, just that it not being true is false.

It is often the case that you want to choose the "bad" thing as the null hypothesis because the consequences of incorrectly assuming something is good are worse. (A mnemonic for this principle: "You want H_0 to have probability 0 of occurring.") For example, if we are testing a new drug to see if it is safe, the null hypothesis should be "it's not safe." This way, we can control (and make small) the probability of incorrectly rejecting the null (saying it's safe incorrectly). This is clearly better than choosing the alternative, because we want to be really sure that the new drug does not result in deaths, lawsuits, etc.

2.2. **The Algorithm.**

(a) Pick a parameter that you are supposed to draw conclusions about.
(b) Choose a hypothesis and significance level.
(c) Choose a test statistic. (Often the likelihood ratio. We'll get to this later.)
(d) Choose a rejection rule.
(e) Determine the constants in the rejection rule such that the significance level is as given.

Now that we've built all that up, let's look at the simplest example. This will be our running example.

Example 1. Suppose that our test statistic is a single random variable with $(\mu, 1)$-normal distribution. Let's set the null hypothesis as "$\mu = 0$", so the alternate is "$\mu \neq 0$." What are the critical of X if we want $\alpha = 0.01$?

So what do we need to do? We want to find C and D such that the probability that X lies outside the range (C, D) is 0.01, given that the null hypothesis is true. In other words, assume $\mu = 0$. Then we can see that by setting $C = -2.57$ and $D = 2.57$ gives us what we want, because $P(|X| \geq 2.57) \approx 0.01$.

For examples, involving more than a single trial, refer to the course notes.

3. Power

When we set the hypothesis, we can control the probability α, which is the probability of a Type I error, that is, the probability of rejecting the null hypothesis when it true. (A "False negative.") However, in many cases you might also want to reject the null hypothesis with reasonable probability when it is actually false.

Later, we will talk about α-β specification, which gives us a method of showing how to reject when the null is false. Here we will talk about how to determine the power of a test.

Example 2. Suppose that the *true* value of μ is *not* 0, so ideally we would reject the null hypothesis. Q: For a given true value of μ, what is the probability that we do indeed reject? This is the power at μ.

We can even write down a formula for this probability. Assume that the true mean in μ, and let $\rho(\mu)$ be the probability of rejecting the null hypothesis. Then

$$\begin{aligned}\rho(\mu) &= P(|X| \geq 2.57 \mid X \text{ is } (\mu, 1)) \\ &= P(X < -2.57 \mid X \text{ is } (\mu, 1)) + P(X > 2.57 \mid X \text{ is } (\mu, 1)) \\ &= \Phi(-2.57 - \mu) + (1 - \Phi(2.57 - \mu)).\end{aligned}$$

In particular, note that the height of the power curve at 0 is α.

4. Likelihood Ratio

The question of what statistic to use is often difficult to answer. It turns out that a very good statistic to use is the **generalized likelihood ratio**, which compares how the probability under the null hypothesis and the probability under the alternate of our observation. Namely, suppose that we have samples $\{X_i\}_{i=1}^n$ that follow a distribution with density $f(x;\theta)$ and parameter θ. If $L(\theta)$ is the likelihood function, then the likelihood ratio λ is

$$\lambda := \frac{\max_{\theta \in \omega} L(\theta)}{\max_{\theta \in \Omega} L(\theta)},$$

where ω is the set of θ values in the null hypothesis and Ω is the full set of θ values.

The generalized likelihood ratio **test** (GLR test) is the test which rejects the null if λ is too small, that is, it represents "reject if $\lambda < \lambda^*$" where λ^* is chosen so that $P(\lambda < \lambda^* \mid \text{null}) = \alpha$.

Example 3. Let $\{y_i\}_{i=1}^n$ be random samples from a normal pdf with unknown mean μ and variance 1. Find the form of the generalized likelihood ratio test for $H_0 : \mu = \mu_0$ vs. $H_1 : \mu \neq \mu_0$.

Solution. Our statistic will be the likelihood ratio. We first need to maximize the likelihood function over the null parameter space ω. This is simple since $\omega = \{\mu_0\}$, just a single point.

Next, we maximize the likelihood function over all possible values. However, we have already done this, and the maximum occurs at the MLE! Therefore, the maximum likelihood function value occurs when we plug in the MLE into the likelihood function. Recall that the MLE for the mean of a normal is the sample average $\overline{y_i}$. The maximum likelihood function value occurs when we plug the MLE into the

likelihood function:

$$\begin{aligned}\lambda &= \frac{\max_{\mu\in\omega} L(\mu)}{\max_{\mu\in\Omega} L(\mu)} \\ &= \frac{\prod_{i=1}^{n} \frac{1}{\sqrt{2\pi}} \exp(-\frac{1}{2}(y_i-\mu_0)^2)}{\prod_{i=1}^{n} \frac{1}{\sqrt{2\pi}} \exp(-\frac{1}{2}(y_i-\overline{y})^2)} \\ &= \exp\left(\frac{1}{2}\sum_{i=1}^{n}((y_i-\overline{y})^2-(y_i-\mu_0)^2)\right)\end{aligned}$$

Therefore, a good test would be to reject the null hypothesis if λ is small. How should you think about this? To get a small value out of this, the expression inside the exponential must be negative, which means that that many of the y_i must be far from μ_0, which in turn implies that the μ_0 is probably not the mean. □

5. (α, β) Specification

We want to control the significance α, but sometimes we want to set a value of the parameter (colloquially called a "kill value" in some machine learning circles) in the alternative hypothesis space and say that if the true value is close to this chosen value, you want to make sure to reject the null hypothesis. This is surprisingly useful in many different situations. Let's see an example of this.

Example 4. Suppose that we have the distribution

$$f_\theta(x) = \begin{cases}\theta x^{\theta-1}, & \text{if } 0<x<1 \\ 0, & \text{otherwise.}\end{cases}$$

We want to find θ, or at least draw some conclusions about it. Our null hypothesis H_0 will be $\theta < \theta_0$, so H_1 is $\theta \geq \theta_0$. Suppose that we want a significance level of $\alpha = 0.05$ (this is a standard convention), but we also want the probability of accepting the null (incorrectly) when $\theta = \theta_1 > \theta_0$ to be small (say $\beta = 0.05$ as well). This is our (α, β) specification.

There are three questions that we need to answer before we can complete our analysis. How many samples do we need to take? What should our θ_0 be? What should we choose as our statistic?

In this case, the "likelihood ratio" is

$$\prod_i \frac{\theta_0 X_i^{\theta_0-1}}{X_i^{\theta_1-1}\cdot\theta_1} = \left(\frac{\theta_0}{\theta_1}\right)^n \left(\prod_i X_i\right)^{\theta_0-\theta_1}.$$

Remark 5.1. WARNING. This is not the GLR defined above. To simplify things, we've just put the alternate hypothesis on the bottom (not the max over all possible parameters). We are using this statistic because it is essentially the same as the GLR, but makes for a slightly simpler calculation and hence a shorter explanation. I recommend that you try and use the standard GLR as a default option for test statistic if you don't have any more information about the problem.

Note that we can form a bijection between our modified likelihood ratio and $\prod_i X_i$, so we will choose this product as our test statistic. However, since $\theta_1 > \theta_0$, this is an order-reversing bijection. In other words, we will reject the null if our statistic is *larger* than a constant C.

Now, the distribution of $\prod_i X_i$ is not easy to analyze. However, let's take the log and define our test function

$$T(X_i) = \sum_i \log(X_i).$$

We haven't made analyzing the distribution inherently easier, but now we can use the normal approximation, assuming n is fairly large, by the Central Limit Theorem. We need to find the mean and variance of $\log(X_i)$. By integration by parts, we find that the mean is

$$E(\log(X_i)) = \int_0^1 \log(x)\theta x^{\theta-1}\,dx = -\frac{1}{\theta}.$$

Similarly, we obtain

$$E(\log(X_i)^2) = \int_0^1 \log(x)^2\theta x^{\theta-1}\,dx = \frac{2}{\theta^2}.$$

Therefore,

$$\mathrm{Var}(\log(X_i)) = E(\log(X_i)^2) - E(\log(X_i))^2 = \frac{2}{\theta^2} - \frac{1}{\theta^2} = \frac{1}{\theta^2}.$$

We can then approximate the distribution of $\sum_i \log(X_i)$ by a normal $(n(-\frac{1}{\theta}), \frac{n}{\theta^2})$ distribution. This holds for any θ, so we now have two equations: for α we have

$$P(reject|null) = P(\sum_i \log(X_i) > C \mid null) < 0.05$$

and from β we have

$$P(\sum_i \log(X_i) < C|\theta = \theta_1) < 0.05.$$

Our two equations have the unknowns n and C, which we can solve for any two particular θ values.

For instance, suppose that $\theta_0 = 1/4$ and $\theta_1 = 3/4$. Then the null distribution is (approximately) $N(-4n, 16n)$, and the distribution for θ_1 is $N(-(4/3)n, (16/9)n)$. Therefore,

$$P(\sum_i \log(X_i) > C|null) \approx 1 - \Phi\left(\frac{C+4n}{4\sqrt{n}}\right)$$

and

$$P(\sum_i \log(X_i) < C|\theta = \theta_1) \approx \Phi\left(\frac{C+(4/3)n}{(4/3)\sqrt{n}}\right).$$

Note that we need to find Φ^{-1}. We can try to just look this up, but a better way is to use Φ^{-1} to get two equalities for C, and then solve for n. That is, from the first and second equation respectively, we obtain

$$C = (4\sqrt{n})\Phi^{-1}(1-0.05) - 4n$$

and

$$C = (4/3)\sqrt{n}\Phi^{-1}(0.05) - (4/3)n.$$

Setting these equal, we get

$$(4+(4/3))\Phi^{-1}(0.95) = (4-(4/3))\sqrt{n},$$

which gives us $n = (2\Phi^{-1}(0.95))^2 = 10.82$. In other words, we need about 11 trials to be able to satisfy this (C turns out to be about -21.6).