Notes on hypothesis testing

November 21, 2010

1 Null and alternate hypotheses

In scientific research one most often plays off some hypotheses against certain others, and then one performs an experiment to decide whether to reject or not reject certain of these hypotheses. Notice that I said "reject" or "not reject"; that is, I didn't say, "reject" or "accept". By saying that I "reject" or "not reject", I am actually saying *less* than if I said "reject" or "accept"; and in saying less, the conclusion has a greater chance of actually being *true*, though often more daring individuals will actually say "accept" instead of "not reject".

Scientific expierments could potentially involve testing many hypotheses at once, but typically one only works with two of them, called the *null hypothesis*, denoted H_0, and the *alternate hypothesis*, denoted H_a.

The null hypothesis is so named because it represents the "default position" or "prior belief". An example would be the hypothesis "the drug has no effect" in testing a drug for efficacy against some disease, and another example would be that "all electrons have almost exactly the same rest mass".

It is actually slightly inaccurate to call the null hypothesis a "default position", because prior to setting up the experiment it may be that the hypothesis hasn't even been considered before. It is only "default" in the sense that if the hypothesis were true it would have little obvious significance to our expanding body of scientific knowledge (though maybe on deeper reflection it *could* have significance). This is perhaps why sometimes one hears the phrase "hypothesis of no consequence" when defining H_0.

The alternate hypothesis H_a basically represents what we would like to be true, since it *would* have some obvious significance if it were true. For this

reason we hope that the outcome of a scientific experiment indicates that we should reject H_0 (or even that we should accept H_a). By negating the above two examples of null hypotheses, we arrive at two good examples of alternate hypotheses: H_a might be the claim that "the drug *does* have an effect" against some disease; or, it might be the claim that "not all electrons have the same rest mass".

Often, null hypotheses are written in terms of parameters related to the experiment; for example, it might be $H_0 : \mu = \mu_0$. This may seem a little strange since, of course, in many cases we wouldn't expect to be able to measure some parameter μ accurately enough to say that it has value *exactly* equal to μ_0; however, within the limits of the test we may not be able to exclude this possibility, so we would continue to accept (or not reject) that $\mu = \mu_0$.

In deciding whether or not to reject the null hypothesis, we must decide upon what *test statistic* to use. A *test statistic* is just some function $f(X_1, ..., X_n)$ of a given data sample $X_1, ..., X_n$, which are random variables, and therefore, numbers. In many scenarios there are standard ones that are used, like the "χ^2 statistic" or the "student-t statistic", which we will discuss below.

Once we have chosen the test statistic, we then choose a region of the real line called the *rejection region* (abbreviated RR) so that we "reject H_0" if $f(X_1, ..., X_n) \in RR$; and otherwise, we don't reject H_0 if $f(x_1, ..., X_n) \notin RR$. We will see some examples of test statistics and rejection regions below.

1.1 Type I and Type II errors

A type I error occurs when we *reject* the null hypothesis when it happens to be true; and a type II error occurs when we *fail to reject* the null hypothesis (or accept the alternate hypothesis) when it happens to be false. I like to think of a type I error as being the sort that credulous people make – they reject conventional wisdom in favor of any new thing that comes along. And, I like to think of a type II error as being the sort that overly conservative people make – they become too set in their ways, and fail to discard old ideas that turn out to be wrong.

It is standard to use the Greek letters α and β to indicate the probabilities of making type I and type II errors, respectively, subject to certain

assumptions. It often helps to write these in probability language as follows:

$$\alpha \ = \ \mathbb{P}(\text{rejecting } H_0 \mid H_0 \text{ is true}),$$

and

$$\beta \ = \ \mathbb{P}(\text{not rejecting } H_0 \mid H_0 \text{ is false}).$$

What assumptions? Typically, it is that the data $X_1, ..., X_n$ fits some particular type of distribution, like maybe that they are sampled from a normal distribution with some unknown mean and variance. This may seem a little upsetting at first, since we tend to think of the methods of science are precise, quantitative, and foolproof – that the only source of error there could be is in the data itself (e.g. measurement error); this isn't so, unfortunately. But *all* empirical subjects must begin with at least *some* assumptions.

1.2 An uncertainty principle

Science tends to be conservative. And as such, it tends to focus much more on trying to keep α small, within practical limits; it prefers to keep the probability of accepting newfangled false claims low, through keeping the probability of making a type I error low.

There is a downside, however, in the form of an uncertainty principle: you can only make α smaller at the expense of making β larger; and you can only make β smaller at the expense of making α larger. This is what I mean by an "uncertainty principle. Let's see why it is true: in order to make α smaller, you basically must shrink the size of the rejection region, in order that it is less likely that the test statistic $f(X_1, ..., X_n)$ falls inside it. In so doing, however, you increase the probability of failing to reject H_0, which opens you up to making more type II errors.

2 Some examples

2.1 The voter example from earlier in the semester

Suppose you have a population that is divided into k different categories. Further, you hypothesize that the percent of individuals in the jth category is p_j. Note that $p_1 + \cdots + p_k = 1$. You now wish to test this hypothesis by picking a large number N of individuals with replacement (i.e. you may pick

the same person more than once), and checking to see which category they fall into. Suppose that the number observed in category j is X_j. Note that $\mathbb{E}(X_j) = p_j N$ and that $X_1 + \cdots + X_k = N$.

Define the parameter

$$E \;=\; \sum_{j=1}^{k} \frac{(X_j - p_j N)^2}{p_j N}. \tag{1}$$

Then, $E \geq 0$, and will be "large" if too many of the classes contain a number of individuals that are far away from the expected number.

In order to be able to check our hypothesis that "p_j percent of the population belongs to class j, for all $j = 1, 2, ..., k$", we need to know the probability distribution of E, and the following result gives us this needed information:

Theorem. For large values of N, the random variable E given in (1) has approximately a chi-squared distribution with $k - 1$ degrees of freedom.

And now our example problem:

Example: Suppose we read in a newspaper that likely voters in Florida break down according to the following distribution: 40% will vote Republican, 40% will vote Democrat, 10% will vote Independent, 5% will vote Libertarian, and 5% will vote "other". We decide to test this, and so we let our null and alternate hypotheses be:

$$H_0 \;:\; \text{The newspaper is correct, and } H_a \;:\; \text{The newspaper is incorrect.}$$

To test H_0 we use the standard χ^2 test as follows: suppose that we ask $N = 10,000$ (polled with replacement) likely Florida voters which group they will vote for. We let X_1, X_2, X_3, X_4, and X_5 denote the number that answered Republican, Democrat, Independent, Libertarian and Other, respectively. We let

$$p_1 = 0.4,\; p_2 = 0.4,\; p_3 = 0.1,\; p_4 = 0.05,\; \text{and } p_5 = 0.05.$$

We will then use the function E above as our test statistic; and let us suppose that we use the following upper-tailed rejection region:

$$RR \;:\; [\chi^2_{0.05,4},\; \infty).$$

That is, if $E \in RR$, then we reject H_0.

In this case, $\alpha = 0.05$; but, we cannot actually compute a value for β, since to do so we would need to know the true population percentages of people who vote Republican, Democrat, Independent, Libertarian, and Other.

Let us suppose that the following is the result of our poll:

4,200 will vote Republican;
3,900 will vote Democrat;
1,000 will vote Independent;
700 will vote Libertarian; and,
200 will vote "other".

So, we have that

$$\begin{aligned} E &= \frac{(4200-4000)^2}{4000} + \frac{(3900-4000)^2}{4000} + 0 + \frac{(700-500)^2}{500} + \frac{(200-500)^2}{500} \\ &= 10 + 2.5 + 8 + 180 \ = 200.5. \end{aligned}$$

And one can check that this certainly lies inside the rejection region; so, we reject the null hypothesis H_0.

2.2 An example involving normal random variables with unknown mean and variance

Recall the following theorem.

Theorem. Suppose that $X_1, ..., X_k$ are i.i.d. $N(\mu, \sigma^2)$ random variables. Let $\overline{X}$ represent the sample mean, and let $\hat{\sigma}$ represent the sample standard deviation. Then, we have that

$$t \ = \ \frac{(\overline{X} - \mu)\sqrt{k}}{\hat{\sigma}}$$

has a Student-t distribution with $k-1$ degrees of freedom.

We will use this in addressing the following problem.

Example: You want to test the theory that the average resistivity of Atlantic Ocean seawater is 0.2 ohm-meters. Suppose you know in advance

that resistivity of ocean water is normally distributed (a BIG assumption, but what can you do?), and let μ and σ^2 denote the corresponding mean and variance of this distribution. In this case, we have that

$$H_0 \ : \ \mu = 0.2, \text{ and } H_a \ : \ \mu \neq 0.2.$$

Let us suppose that we do an experiment by randomly selecting 6 assays of Atlantic Ocean water. We will use the t function in the above theorem with $k = 6$ and $\mu = 0.2$ as our test statistic, and we will use the following two-tailed rejection region:

$$RR \ : \ (-\infty, \ t_{0.05,5}] \ \cup \ [-t_{0.05,5}, \infty).$$

where here we use the notation $t_{\theta,5}$ to denote the θ percentile of a student-t distribution with $5 = 6 - 1$ degrees of freedom; note that it is *percentile* and note *upper percentile* as we used in the χ^2 test.

The percentile value $t_{0.05,5}$ can be computed using the following Maple commands:

```
>with(Statistics):

> Percentile(StudentT(5), 5, numeric);
                                  -2.015042560
```

So,

$$RR \ : \ (-\infty, \ -2.015042560] \ \cup \ [2.015042560, \infty).$$

It is obvious from the way we set things up that $\alpha = 0.1$; and, again, we cannot compute a particular value for β unless we know something about μ. We can, however, determine values for $\beta(\mu)$ – that is, we can determine the probability of making a type II error, given that we know some particular value for μ. An exercise for YOU: determine $\beta(0.18)$ and $\beta(0.22)$ using $\sigma = 1$.

Suppose that the following are the resistivities of our 6 assays in ohm-meters

$$0.26, \ 0.15, \ 0.25, \ 0.22, \ 0.18, \ 0.20.$$

Then, we will have that

$$t \ = \ \frac{(0.21 - 0.20)\sqrt{6}}{0.04335896677...} \ = \ 0.564932...$$

Clearly, since this does not lie in the RR, we *do not* reject H_0.

2.3 The p-value of a test

It is common nowadays to report in a scientific paper not only whether H_0 was rejected or not, but also the p-value associated with the outcome. The p-value is often called the "observed significance level", and is the smallest value that α could be in order that H_0 is rejected. That is, if $p \leq \alpha$, then we reject H_0; and if $p > \alpha$, then we do not reject H_0.

Two more ways of thinking about the p-value: 1) it is the value that α would need to be in order that we are right on the boundary between rejecting and not rejecting H_0; and 2) intuitively, it is measuring the probability that H_0 could be true, subject to the usual underlying assumptions (like that the X_i are sampled from a particular distribution).

Typically we want to report p-values for when H_0 was *rejected*; so, let us take a look at the elections example: there, we found that $E = 200.5$, and so the rejection region would have to be $[200.5, \infty)$ for H_0 to just barely be rejected. The probability associated to this, assuming $E \sim \chi^2_4$ can be found using Maple again:

```
> with(Statistics);
> 1- CDF(ChiSquare(4),200.5);
```

```
2.93341306*10^(-42)
```

So, the p-value is incredibly close to 0. It is unusual for it to be this small; typically, p-values are of size 0.05 or so. Sometimes, when p-values are as small as we just found, one will just write "$p < 0.01$".