p-Values, Power and the Neyman-Pearson Lemma

Guy Lebanon

October 15, 2006

We have seen in a previous note that a specific test may lead to acceptance or rejection of H_0, and has two types of errors associated with it: type 1 error α and type 2 error β. For most tests, the rejection region may be enlarged or shrunk trading off α for β and causing it to more often reject or accept. This leads to the following concept of p-value. Intuitively, p-value is the type 1 error associated with the a test whose rejection region is just barely small enough to accept H_A (or equivalently reject H_0).

Definition 1. *The p-value, or attained significance level, of a test is the smallest level of α for which the observed data indicates acceptance of H_A.*

The smaller the p-value - the more compelling the evidence that H_A should be accepted. Given some experimental evidence, reporting the p-value contains more information than reporting the specific α of a particular test. We know not only that a particular test was accepted or rejected - but the entire relationship between modifying the rejection region and the test result. We know that for rejection regions leading to $\alpha \geq p$ the test will reject H_0 and for rejection regions leading to $\alpha < p$, the test will accept H_0.

Example: Consider the t-test with $H_A : \mu > \mu_0, H_0 : \mu = \mu_0$, with $RR = [c, \infty)$, and the test statistic $\bar{X}$. We have $\alpha = P(\bar{X} \geq c|\mu_0) = P(T \geq \sqrt{n}\frac{c-\mu_0}{S})$ where T follows a t distribution (see previous note for details). Solving for the empirical mean of the observed data just rejecting $c = \bar{x} = \frac{1}{n}\sum_{i=1}^n x_i$ results in the critical value of the t-distribution $t_p = \sqrt{n}\frac{\bar{x}-\mu_0}{S}$ which can be solved for p using the critical values table of the t-distribution.

In scientific research, researchers often publish alternative hypotheses H_A that are accepted with p smaller than some value (often 0.05). Such practice has its advantage since it could prevents the publication of non-significant research results. On the other hand, strict adherence to this rule may prevent some important discoveries from being made public.

Definition 2. *Consider a hypothesis test with a test statistic T concerning the value of the parameter θ. The power function is*

$$\textit{power}(\theta) = P(T \in RR|\theta) \in [0, 1].$$

For $\theta \in H_0$ we have $\textit{power}(\theta) = \alpha$ and for $\theta \in H_A$ we have $\textit{power}(\theta) = 1 - \beta$. We therefore would like the power function to be small at $\theta \in H_0$ and large for $\theta \in H_A$. In fact, an ideal test would have the power function be 0 on H_0 and 1 on H_A. As mentioned in a previous not, there is in general a tradeoff between α and β and it is not possible to minimize both. A standard way to choose an effective test is to select the one that minimizes β among all tests whose α is fixed at some pre-determined level. In other words, we select a significance level α that we deem acceptable, and among all tests with this α choose the test that minimizes β (or maximize the power function power(θ) for $\theta \in H_A$).

Definition 3. *If a hypothesis contains a single parameter value, it is said to be a simple hypothesis. Otherwise, it is said to be a composite hypothesis.*

We say that a test T_1 is more powerful at $\theta_A \in H_A$ than a test T_2 if $\text{power}_{T_1}(\theta_A) \geq \text{power}_{T_2}(\theta_A)$ (sometime strict inequality is used in the definition). If $\text{power}_{T_1}(\theta) \geq \text{power}_{T_2}(\theta)$ for all $\theta \in H_A$, we say that T_1 is uniformly more powerful than T_2. If a test with significance level α is more powerful than all other tests with the same significance α, it is called the uniformly most powerful (UMP) test. In general, the UMP may not exist or if it exists it may be difficult to find. In the case that both H_0 and H_A are simple, the Neyman Pearson lemma below characterizes the UMP.

Definition 4. *For a simple $H_0 = \{\theta_0\}$, if there exists an α-level test for which power$(\theta), \theta \neq \theta_0$ is larger than the power curve of any other α-level test ($\forall \theta \neq \theta_0$), it is said to be the uniformly most powerful test (UMP). In other words, the power function should be above the power curves of all other α-level tests for $\theta \neq \theta_0$.*

Proposition 1 (Neyman-Pearson)**.** *Consider a test between two simple hypothesis $H_0 = \{\theta_0\}$ and $H_A = \{\theta_A\}$. The test whose rejection region corresponds to*

$$\left\{x_1, \ldots, x_n : \frac{L(x_1, \ldots, x_n|\theta_0)}{L(x_1, \ldots, x_n|\theta_A)} < k\right\},$$

where $L(x_1, \ldots, x_n|\theta)$ is the likelihood, is UMP. Typically, k is chosen to correspond to a specified α.

Proof. We use the notation $1_{\{T \in A\}} = 1$ if $T \in A$ and 0 otherwise. Let T_1 be the Neyman Pearson test and T_2 another test with the same α level. The following inequality

$$1_{\{T_2 \in RR\}}(kL(x_1, \ldots, x_n|\theta_A) - L(x_1, \ldots, x_n)|\theta_0) \leq 1_{\{T_1 \in RR\}}(kL(x_1, \ldots, x_n|\theta_A) - L(x_1, \ldots, x_n|\theta_0))$$

holds for all x (verify by examining both sides in the cases $T_1 \in RR$ and $T_1 \notin RR$. As a result, integrating both sides of the inequality with respect to x would result in a summation of several valid inequalities which gives another valid inequality that proves the lemma

$$k(1 - \beta_{T_2}) - \alpha \leq k(1 - \beta_{T_1}) - \alpha.$$

□

Example: Consider a single sample y from a distribution with pdf $f_\theta(y) = 1_{\{0<y<1\}}\theta y^{\theta-1}$. By the Neyman Pearson lemma, the UMP test for $H_0 : \theta = 2$ vs. $H_A : \theta = 1$ has test statistic $T(y) = y$ and rejection region $k > \frac{L(y|2)}{L(y|1)} = \frac{2y}{1} = 2y$ or $RR = [0, k/2)$. By selecting a specific α, the appropriate k is determined, e.g. $\alpha = P(Y < k|\theta = 2) = \int_0^k 2y\,dy = k^2$ or $k = \sqrt{\alpha}$.

We can use the Neyman Pearson lemma to obtain the UMP for a simple alternative hypothesis. For composite H_A we can still examine the rejection region obtained by the Neyman Pearson lemma. If that region is not a function of θ_A, then the test is UMP for every single simple alternative $\{\theta_A\}$ and is also true for a composite H_A. In other words, if the rejection region computed above does not depend on θ_A the Neyman Pearson lemma can be used to characterize the UMP test for a composite H_A.