Hypothesis testing and likelihood ratio tests

We will adopt the following model for observed data. The distribution of $\mathbf{Y} = (Y_1, ..., Y_n)$ is considered known except for some **parameter** θ, which may be a vector $\theta = (\theta_1, ..., \theta_k)$; $\theta \in \Theta$, the **parameter space**. The parameter space will usually be an open set. If $\mathbf{Y}$ is a continuous random variable, its **probability density function** (pdf) will de denoted $f(\mathbf{y};\theta)$. If $\mathbf{Y}$ is discrete then $f(\mathbf{y};\theta)$ represents the **probability mass function** (pmf); $f(\mathbf{y};\theta) = P_\theta(\mathbf{Y}=\mathbf{y})$.

A **statistical hypothesis** is a statement about the value of θ. We are interested in testing the null hypothesis H_0: $\theta \in \Theta_0$ versus the alternative hypothesis H_1: $\theta \in \Theta_1$. Where Θ_0 and $\Theta_1 \subset \Theta$. Naturally $\Theta_0 \cap \Theta_1 = \varnothing$, but we need not have $\Theta_0 \cup \Theta_1 = \Theta$. A **hypothesis test** is a procedure for deciding between H_0 and H_1 based on the sample data. It is equivalent to a **critical region**: a critical region is a set $C \subset \mathbb{R}^n$ such that if $\mathbf{y} = (y_1, ..., y_n) \in C$, H_0 is rejected. Typically C is expressed in terms of the value of some **test statistic**, a function of the sample data. For example, we might have $C = \{(y_1, ..., y_n)$: $\frac{\bar{y} - \mu_0}{s / \sqrt{n}} \geq 3.324\}$. The number 3.324 here is called a **critical value** of the test statistic $\frac{\bar{Y} - \mu_0}{S / \sqrt{n}}$.

If $\mathbf{y} \in C$ but $\theta \in \Theta_0$, we have committed a Type I error. If $y \notin C$ but $\theta \in \Theta_1$, we have committed a Type II error. The ideal hypothesis test would simultaneously minimize the probabilities of both types of error, but this turns out to be impossible in principle. So what we do is to select a **significance level** $\alpha = \max_{\theta \in \Theta_0} P_\theta(Y \in C)$ to be a small number; $\alpha = 0.05$ and $\alpha = 0.01$ are traditional. Then a "good" test is one with a small probability of Type II error, among all possible tests with significance level α. When $\mathbf{y} \in C$ for a test of level α (that is, H_0 is rejected), we say the results are **statistically significant** at level α.

For a particular set of data, the smallest significance level that leads to the rejection of H_0 is called the **p–value**. H_0 is rejected if and only if $p \leq \alpha$.

$P_\theta(Y \in C)$ can be viewed as a function of θ. If $\theta \in \Theta_1$, we refer to this quantity as the **power** of the test C. For any good test, we will have $P_\theta(Y \in C) \uparrow 1$ as $n \to \infty$ for each $\theta \in \Theta_1$. This provides a way to choose sample size. For a fixed value $\theta \in \Theta_1$ that is of scientific interest, choose n large enough so that the probability of rejecting H_0 is acceptably high.

How do we construct good hypothesis tests? Usually it is hard to beat the **likelihood ratio tests**, which are defined as follows. $C = \{\mathbf{y}: \lambda = \frac{\underset{\theta \in \Theta_0}{\text{Max}} L(\theta;\mathbf{y})}{\underset{\theta \in \Theta}{\text{Max}} L(\theta;\mathbf{y})} \leq k\}$. The value of k (0<k<1) varies from problem to problem. It is chosen so that the test will have significance level ("size") α.

Notice that the denominator of λ is just the likelihood function evaluated at the MLE. The numerator is the likelihood function evaluated at a sort of restricted MLE, in which θ is forced to stay in the set Θ_0. Also notice that when $\lambda=0$, H_0 is always rejected, and when $\lambda=1$, H_0 is never rejected.

There are two main versions of the likelihood ratio approach. One version leads to exact tests, and the other leads to large–sample approximations. The **exact likelihood ratio tests** are obtained by working on the critical region C, and re–expressing it in terms of the value of some test statistic whose distribution (given that H_0 is true) is known exactly. This is where we get most of the standard statistical tests, including t–tests and F–tests in regression and the analysis of variance.

Sometimes, after the critical region C has been re–expressed in terms of some seemingly convenient test statistic, nobody can figure out its distribution under H_0. This means we are unable to choose $k \in (0,1)$ so that the test has size α. And if the MLE has to be approximated numerically, there is little hope of an exact test. In such cases we resort to **large–sample likelihood ratio tests**; we will use the following result.

Let $\theta = (\theta_1, ..., \theta_p)$; we want to test H_0: $\theta \in \Theta_0$ versus H_1: $\theta \in \Theta_1$, where $\Theta_0 \cup \Theta_1 = \Theta$. Let $r \le p$ be the number of parameters θ_j whose values are restricted by H_0. Then under some smoothness conditions, $G = -2 \log(\lambda)$ has (for large n) an approximate χ^2 distribution with r degrees of freedom. We reject H_0 when G is greater than the critical value of the χ^2 distribution. If g is the observed value of the statistic G calculated from the sample data, the p−value is $1-\gamma(g)$, where γ is the cumulative distribution function of a χ^2 distribution with r degrees of freedom.