Lecture 20. Hypothesis Testing

December 5, 2011

One more thing you can do with data is hypothesis testing, i.e. to evaluate the validity of hypotheses. In general, we have some hypotheses about the distribution of a random variable/vector, X, of interest, and would like to know whether our hypotheses are correct or not based on a random n-sample of X. The hypotheses could come from some kind of theory, or they could be some kind of wild guess. We are not concerned with where they come from. We take them as given and think about how to test them.

The statistical hypothesis testing framework requires us to have two competing hypotheses, a **null hypothesis** (denoted H_0) and an **alternative hypothesis** (denoted H_1), both are some restrictions on the underlying distribution of X. For example, for a scalar random variable X, a pair of hypotheses can be:

$$H_0 : X \sim N(0,1) \; vs. \; H_1 : X \sim N(1,1). \tag{1}$$

The union of the two competing hypotheses usually is not the whole universe. We call the union the "**maintained hypothesis**", which are assumed to be true when doing the hypothesis test. Though people usually do not say it explicitly, they are always doing hypothesis testing under certain maintained hypotheses. In the example above, the maintained hypothesis is "$X \sim N(a,1), \; a = 0 \; or \; 1$.

A simple hypothesis is a hypothesis that restrict the distribution of X to one particular distribution. In the above example, both H_0 and H_1 are simple hypotheses.

The opposite of "**simple hypothesis**" is "**composite hypothesis**". A composite hypothesis restrict the distribution of X to a set composed of more than one distribution. For example, both the H_0 and H_1 below are composite hypotheses:

$$\begin{aligned} H_0 :& X \sim N(0,\sigma^2) \text{ for some } \sigma^2 > 0 \; vs. \\ H_1 :& X \sim N(1,\sigma^2) \text{ for some } \sigma^2 > 0. \end{aligned}$$

The maintained hypothesis is $X \sim N(a, \sigma^2)$ for some $\sigma^2 > 0$and $a \in \{0, 1\}$.

Conventionally, we state the maintained hypothesis as assumptions (or as prior knowledge) before we state the H_0 and H_1 and simplify H_0 and H_1 to only emphasize their additional restrictions on the data relative to the maintained hypothesis. Following this convention, the first example can be rewritten as follows: assume that $X \sim N(\mu, 1)$, the hypotheses we would like to test are:

$$H_0 : \mu = 0 \; vs. \; H_1 : \mu = 1. \tag{2}$$

The second example can be rewritten as follows: assume that $X \sim N(\mu, \sigma^2)$ for some $\sigma^2 > 0$, the hypotheses we would like to test are:

$$H_0 : \mu = 0 \; vs. \; H_1 : \mu = 1. \tag{3}$$

As you can see, because people follow this convention, it is important to be aware of the mainted hypotheses as well as what's after H_0 : and H_1 :. Otherwise, we may wrongly think that the hypotheses in (3) are simple hypotheses.

The two example above are parametric hypotheses, i.e. hypotheses about a finite dimensional parameter μ. We may sometimes be interested in hypotheses that are not about a finite dimensional parameter as well. For example, we may be interested in testing the normality of X:

$$\begin{aligned} &H_0 : X \sim N(\mu, \sigma^2) \text{ for some } (\mu, \sigma^2) \in R \times R_+ \\ &H_1 : E(X^2) < \infty \text{ and } X \text{ is not normally distributed.} \end{aligned}$$

For this course, we focus on parametric hypotheses, where both the null and the alternative hypotheses are parametric. In other words, in the maintained hypothesis, we assume the distribution of X is known upto a finite dimensional parameter θ and the null and the alternative hypotheses imposes different restrictions on θ. To establish notation, assume that X has density $f_X(x, \theta)$, where $\theta \in \Theta \subset R^{d_\theta}$. The null and alternative hypotheses are of the form:

$$H_0 : \theta \in \Theta_0 \text{ vs. } H_1 : \theta \in \Theta_1 := \Theta / \Theta_0. \tag{4}$$

Clearly, when Θ_0 is a singleton, the null hypothesis is a simple hypothesis (but the alternative is a composite hypothesis). When Θ_0 is not a singleton, the null is a composite hypothesis. For reasons that become clear below, we never let the alternative hypothesis be a singleton unless Θ contains only two points.

A **hypothesis test** is a rejection region C_R, which is a set that the random sample $\mathbb{X} = \{X_1, X_2, ..., X_n\}$ may belong to. The test "rejects H_0" if $\mathbb{X} \in C_R$ and does not reject H_0 if $\mathbb{X} \notin C_R$. Sometimes, one may say "accept H_0" when $\mathbb{X} \in C_R$. Because the sample does not reveal all information about the distribution of X, the hypothesis test may make errors. There are two types of errors: the false rejection (type I error) and the false acceptance (type II error). The probability of making the type I error is called the null rejection probability and it potentially depends on which point in Θ_0 is the true parameter value

$$\alpha(\theta) = \Pr{}_\theta(\mathbb{X} \in C_R) \text{ for } \theta \in \Theta_0. \tag{5}$$

The probability of making the type II error almost always depends on which point in Θ_1 is the true parameter value:

$$\beta(\theta) = \Pr{}_\theta(\mathbb{X} \notin C_R) \text{ for } \theta \in \Theta_1. \tag{6}$$

$1 - \beta(\theta)$ is called the "**power**" of the test at $\theta \in \Theta_1$. We sometimes extend the domain of the function $1 - \beta(\theta)$ to the whole parameter space Θ and define the "**power function**" of the test C_R to be:

$$\gamma(\theta) = \Pr{}_\theta(\mathbb{X} \in C_R) \text{ for } \theta \in \Theta. \tag{7}$$

In practice, the costs of making the type-I error and the type-II error may be very different, ideally, one would like to balance the two types of errors to minimize some cost function. One difficulty with specifying this cost function is that it requires some subjective belief about θ and is very sensitive to such a belief. Suppose our prior belief is that $\theta \in \Theta_0$, then we will not care about the type II error and our rejection rule should be always accepting H_0. Suppose our prior belief is that $\theta \in \Theta_1$, then we will not care about the type I error and our rejection rule should be always rejecting H_0. This may be resolved by agreeing upon some nondegenerate prior distribution on Θ. The second difficulty is that the cost function is often problem-specific and it is hard to discuss it in any level of generality. The difficulties do not mean that the cost approach is not useful or important. But conventionally, that is not the approach that people take to balance the type-I and the type-II errors.

The conventional approach, established by Neyman and Pearson, is to control the maximum probability of type-I error at a prespecified level, and then make the probability of type-II error as small as possible. A few concepts are central to this Neyman-Pearson framework. A test is said to have "**significance level**"(or level) α if the maximum probability of making the type-I error is less than or equal to α:

$$\max_{\theta\in\Theta_0} \Pr{}_\theta(\mathbb{X} \in C_R) \leq \alpha. \tag{8}$$

The "**size**" of a test is $\max_{\theta\in\Theta_0} \Pr{}_\theta(\mathbb{X} \in C_R)$: i.e. the maximum null rejection probability, or maximum probability of making the type-I error. The level is the first and foremost requirement. Two tests, C_R^1 and C_R^2, of level α are compared by their power. The test C_R^1 is uniformly more powerful than C_R^2 if

$$\begin{aligned} \beta_1(\theta) &\leq \beta_2(\theta)\ \forall\theta \in \Theta_1 \text{ and} \\ \beta_1(\theta) &< \beta_2(\theta)\ \exists\theta \in \Theta_1, \end{aligned} \tag{9}$$

where $\beta_j(\theta) = \Pr_\theta(\mathbb{X} \notin C_R^j)$.

What does the rejection region C_R usually look like? Typically, it involves a test statistic and a critical value:

$$C_R = \{\mathbb{X} : T(\mathbb{X}) > c(\mathbb{X}, \alpha)\}, \tag{10}$$

where $T(\mathbb{X})$ is the test statistic and $c(\mathbb{X}, \alpha)$ is the critical value and α is a prespecified significance level, typically 1%, 5% or 10%. The test statistic is a statistic, i.e., a known function of the data. The critical value $c(\mathbb{X})$ could depend on the data, but for the cases discussed in this course, it is a constant. The test statistic and the critical value are chosen to (1) control the size (i.e., making $\max_{\theta\in\Theta_0} \Pr{}_\theta(\mathbb{X} \in C_R) \leq \alpha$) and (2) to make the power large.

Exercise 1. Suppose I tossed a coin 10 times and head came up 2 times. Is the coin a fair coin? Clearly, in this problem, the random variable of interest is the Bernoulli random variable X that equals 1 if the coin is head-side up in a toss and equals zero if the coin is tail-side up in the toss. The coin is fair iff the success rate p of the Bernoulli X is 0.5. We observe a random sample of size 10: $\{X_1, ..., X_{10}\}$ and would like to test the following hypotheses:

$$H_0 : p = 0.5 \text{ vs. } H_1 : p \neq 0.5. \tag{11}$$

Let $S(\mathbb{X}) = \sum_{i=1}^{10} X_i$. Then $S(\mathbb{X}) \sim Bin(10, p)$. If H_0 is true, then $S(\mathbb{X})$ has mean 5 and is likely to take values around 5. If H_1 is true, then $S(\mathbb{X})$ can be reasonably believed to be more likely to take a value far from 5. Thus, a reasonable rejection rejection region can be

$$C_R = \{\mathbb{X} : |S(\mathbb{X}) - 5| > c\}.$$

For each $c = 0, 1, 2, ..., 5$, the probability of type-I error can be calculated:

$$\begin{aligned}
\Pr{}_{0.5}(|S(\mathbb{X}) - 5| > 0) &= 1 - \begin{pmatrix} 10 \\ 5 \end{pmatrix} \times 0.5^5 \times 0.5^5 = 0.76 \\
\Pr{}_{0.5}(|S(\mathbb{X}) - 5| > 1) &= 0.76 - 2 \begin{pmatrix} 10 \\ 4 \end{pmatrix} \times 0.5^4 \times 0.5^6 = 0.34 \\
\Pr{}_{0.5}(|S(\mathbb{X}) - 5| > 2) &= 0.34 - 2 \begin{pmatrix} 10 \\ 3 \end{pmatrix} \times 0.5^3 \times 0.5^7 = 0.1 \\
\Pr{}_{0.5}(|S(\mathbb{X}) - 5| > 3) &= 0.1 - 2 \begin{pmatrix} 10 \\ 2 \end{pmatrix} \times 0.5^2 \times 0.5^8 = 0.012 \\
\Pr{}_{0.5}(|S(\mathbb{X}) - 5| > 4) &= 2 \times 0.5^1 \times 0.5^9 = 0.00195
\end{aligned}$$

A test of level $\alpha = 5\%$ can take $c = 3$ or $c = 4$. The test with $c = 3$ has uniformly better power than the test with $c = 4$ because for all $p \neq 0.5$:

$$\Pr{}_p(|S(\mathbb{X}) - 5| > 3) > \Pr{}_p(|S(\mathbb{X}) - 5| > 4). \tag{12}$$