1 Hypothesis Testing: the Basic Concepts

Hypothesis testing involves deciding between two statements about the unknown parameter θ:

$$\begin{array}{ccc} H_0 : \theta \in \Theta_0 & \text{vs.} & H_1 : \theta \in \Theta_1 \\ \uparrow & & \uparrow \\ \text{Null Hypothesis} & & \text{Alternative Hypothesis} \end{array}$$

where Θ_0 and Θ_1 partition the parameter space Θ. Although situation seems symmetric, we will focus on the null hypothesis, H_0. The two possible decisions are to *accept* H_0 or to *reject* H_0. We will generally specify the null hypothesis, and then Θ_1 is the complement of Θ_0.

Example 1: A coin will be flipped $n = 10$ times, and we want to test the null hypothesis that the coin is fair or biased against heads, i.e.

$$H_0 : p \leq 1/2,$$

where p (the parameter) is the probability of heads on a single flip. Clearly, there is an obvious *test statistic*, namely X, the total number of heads. We will reject H_0 if X is too large. One might suppose that it suffices to reject H_0 if $X > n/2 = 5$. But in fact, we want to *give the benefit of the doubt to* H_0, and reject H_0 only if there is strong evidence against it. H_0 is sort of like the defendant in a trial – we assume H_0 is "innocent" (i.e., that it is true) unless there is strong evidence it is "guilty" (i.e., false). In fact, we will control the maximum probability of a *Type I Error*, which is rejecting H_0 when true. This maximum Type I Error probability is called the *size* of the test, and we set a bound on this called the *level of significance*, which is small (typically 0.05 or less). So, for this coin flipping problem, we will generally reject H_0 if $X > C$ where C is somewhat greater than 5. The main issue is, how much greater?

To compute the size of the test, we would select a value of C and maximize $P[X > C|p]$ for values of $p \leq 1/2$. It is intuitively clear that the maximum will be achieved at $p = 1/2$.

For instance if the heads probability is 1/4, the probability of seeing more than (say) 8 heads is is smaller than when the heads probability is 1/2. We will discuss this issue later in a more rigorous fashion when we derive Uniformly Most Powerful (UMP) tests. So, we can simply take $p = 1/2$ when we compute the type I error probability.

We have agreed that a test will reject when $X > C$, where C is some constant to be determined, namely to achieve the level of significance. Here, C is called the *critical value*, and the set $x : x > C$ is called the *critical region* or *rejection region*. Since X only takes on integer values between 0 and 10, we can restrict C to integer values as well. We will use the "standard" level of significance $\alpha = 0.05$. Thus, we need to find a value of C such that $P[X > C|p = 1/2]$ is as close as possible to α but doesn't exceed α. Of course, we are assuming i.i.d. flips, so X has a $binomial(10, p)$ distribution. Figure 1 shows a plot of $P[X > C|p = 1/2]$ as a function of C. In this figure, we see that $P[X > 8|p = 1/2] < 0.05$, but $P[X > 7|p = 1/2] > 0.05$. (However, $P[X > 7|p = 1/2]$ is very close to 0.05.) Hence, we should choose the rejection region $X > 8$, i.e. reject the null hypothesis if the number of heads in 10 flips is > 8. The size of this test is $1 - F(8|p = 0.5) = 0.0107$ where $F(x|p)$ is the CDF of $X \sim binomial(10, p)$. The calculation was done in the statistics package R. Note: if we had chosen the critical value $C = 7$, we would have had the size 0.0547.

For now, we are only considering tests that definitely reject or accept H_0, i.e. nonrandomized tests. We will discuss randomized tests shortly. Our *test (function)* is just the indicator of the rejection region:

$$\phi(x) = \begin{cases} 1 & \text{if } x > 8, \\ 0 & \text{if } x \leq 8. \end{cases}$$

Note that $E[\phi(X)|p] = P[X > 8|p]$. The probability of rejection as a function of p is known as the *power function*, which we denote $\beta(p)$. The power function depends on the test, so we are assuming that rejection region has already been determined. In our case, we took $C = 8$, and so we can plot the power function, as is shown if Figure 2. Note that the power function is strictly increasing, an "obvious" property we already alluded to when we said that we could use the value $p = 1/2$ to compute the size of the test.

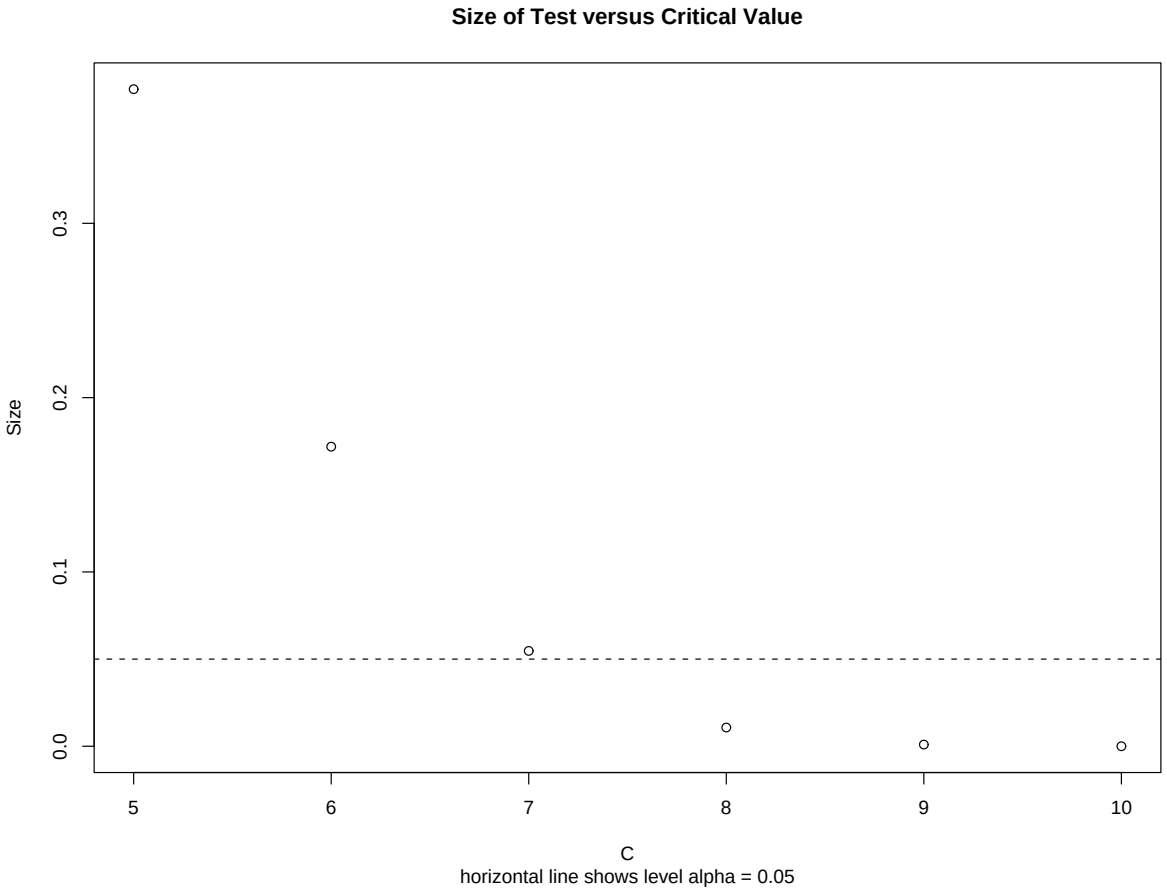

Figure 1: The size of the test which rejects when $X > C$ as a function of C.

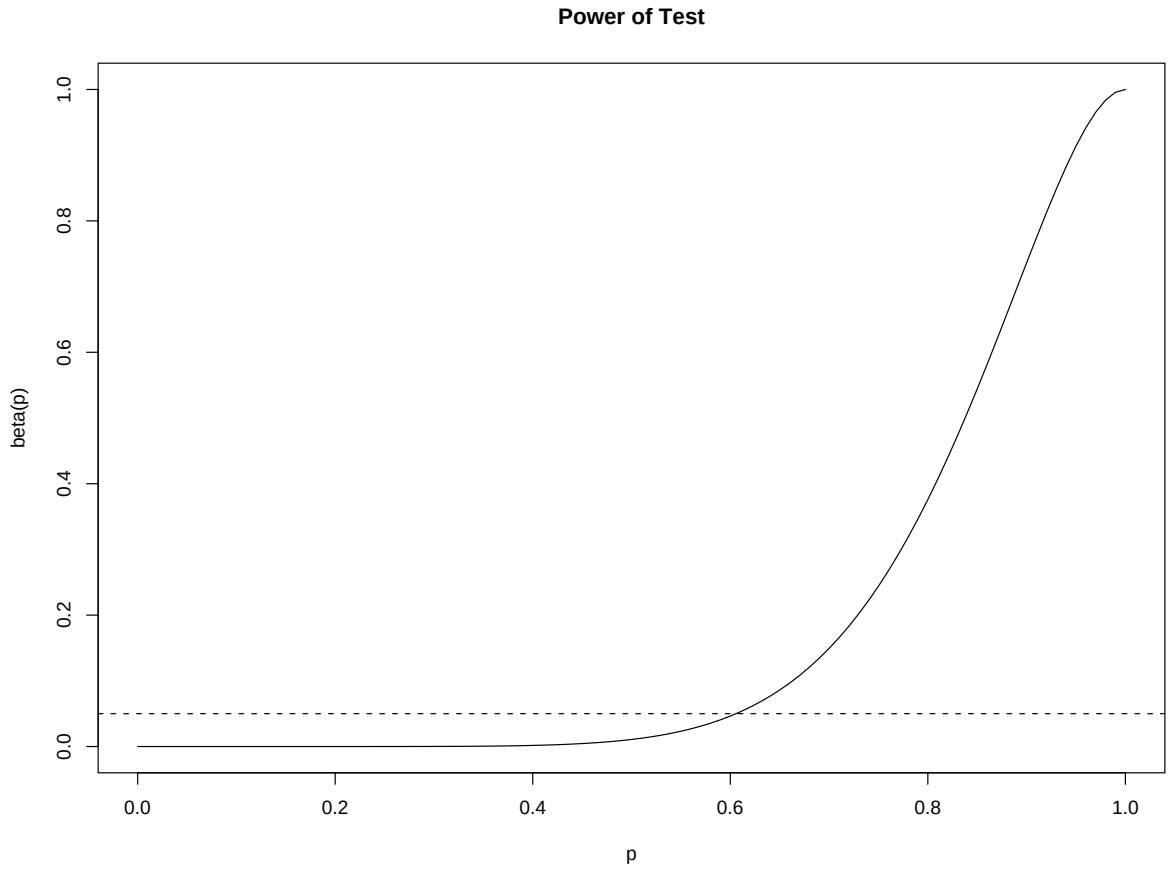

Figure 2: The power of the test which rejects when $X > 8$ as a function of p. We show horizontal line at $\alpha = 0.05$ simply for comparison; the actual size of the test is somewhat smaller.

2 Randomized Tests.

In the above, we only considered tests that definitely accepted or rejected, depending on the value of the observation. For such non-randomized tests, the test function, which is the indicator of the rejection region, takes on only the values 0 and 1. It will be useful to allow tests to take on intermediate values, i.e. test functions $\phi(X)$ such that $0 \leq \phi(X) \leq 1$. The idea is that we will perform an "auxilliary" experiment that results in a binary (0 or 1) outcome, and that decides whether to accept or reject H_0. While one would generally not want to use a randomized test in practice, it is useful from a theoretical perspective for many reasons.

There is a "standard" way to randomize: let $U \sim uniform(0,1)$ be independent of X. Then, for a given test $\phi(X)$, we reject H_0 if $U < \phi(X)$. Thus, if $\phi(X) = 1$, there is no point in generating the random number U; we already know $U < 1$ so we will reject H_0. Similarly, if $\phi(X) = 0$ we will accept H_0 since we already know $U > 0$ with probability 1.

Now when we compute the probability of rejecting H_0, we mean the probability with respect to the joint distribution of U and X. Assuming X has a PMF $f(x|\theta)$ (we actually will typically only need randomization for "good tests" when X is discrete), then

$$\begin{aligned} P[\text{reject } H_0|\theta] &= P[U < \phi(X)|\theta] \\ &= \sum_x P[U < \phi(x)] f(x|\theta) \\ &= \sum_x \phi(x) f(x|\theta) \\ &= E[\phi(X)|\theta]. \end{aligned}$$

The second equality, follows from the independence of U and X: given $X = x$ (where x is a constant), the probability of rejection is simply $P[U < \phi(x)]$. The third equality follows from the uniform distribution of U.

Example 1, Revisited: Recall that we had to take the rejection region to be $X > 8$ in order to achieve the level of significance $\alpha = 0.05$. However, if we allow a randomized

test, we can achieve the level $\alpha = 0.05$. Thus, consider a test:

$$\phi(x) \;=\; \begin{cases} 1 & \text{if } x > 8 \\ \gamma & \text{if } x = 8 \\ 0 & \text{if } x < 8, \end{cases}$$

where $\gamma \in (0, 1)$ is to be determined. It is pretty easy to solve for the value of γ that achieves the desired level of significance $\alpha = 0.05$.

$$\begin{aligned} 0.05 &= E[\phi(X)|p = 1/2] \\ &= P[X > 8|p = 1/2] \;+\; \gamma P[X = 8|p = 1/2] \end{aligned}$$

which gives

$$\gamma \;=\; \frac{0.05 - P[X > 8|p = 1/2]}{P[X = 8|p = 1/2]} \;=\; (0.05 - 0.0107)/0.0439 \;=\; 0.8952.$$

In summary, this test rejects for sure when $X > 8$, and if $X = 8$, rejects with probability 0.8952, and if $X < 8$, accepts for sure. Recall "reject" and "accept" always refer to the null hypothesis.

Of course, all of this would be unnecessary if we simply raised our level of significance to $\alpha = 0.055$, then the test that rejects when $X > 7$, which has size 0.0547. However, most scientists are so locked into $\alpha = 0.05$ that they would not be willing to do this without a lot of convincing.

3 Neyman-Pearson Lemma

First, there are a couple of simple facts about PDFs and PMFs we need. Suppose X has PDF $f(x)$. Then of course $0 \le f(x) < \infty$, i.e. we will not allow infinite values for $f(x)$. Also, $P[f(X) = 0] = 0$. To see this, let $A = \{x : f(x) = 0\}$. Then

$$P[X \in A] \;=\; \int_A f(x)\,dx \;=\; \int_A 0\,dx \;=\; 0,$$

since for $x \in A$, $f(x) = 0$! The same holds true if X has a PMF.

We suppose the parameter space has only two elements: $\Theta = \{\theta_0, \theta_1\}$. We also suppose the observation has a PDF $f(x|\theta)$, for both $\theta = \theta_0$ and $\theta = \theta_1$, or has a PMF in both cases. Here, the size of a test $\phi(X)$ is simply $E[\phi(X)|\theta_0]$. We say a test $\phi^\star$ is *Most Powerful (MP) level* α if it is level α (i.e., $E[\phi^\star(X)|\theta_0] \leq \alpha$) and has maximal power at the alternative among all level α tests, i.e., $E[\phi^\star(X)|\theta_1] \geq E[\phi(X)|\theta_1]$ for any test ϕ satisfying $E[\phi(X)|\theta_0] \leq \alpha$. We shall need the convention that

$$\infty \cdot 0 \;=\; 0.$$

Of course, $\infty \cdot a = \infty$ if $a > 0$. Also, we need the convention that

$$\inf \emptyset \;=\; \infty.$$

This makes sense in the following way: if $A \subset B$ are any sets of real numbers, then $\inf A \geq \inf B$. Since $\emptyset \subset A$ for any set of real numbers A, we should have $\inf \emptyset \geq \inf A$ for any A and this can only happen if the above holds.

Neyman-Pearson Lemma: *Consider testing*

$$H_0 : \theta = \theta_0 \quad \textit{vs.} \quad H_1 : \theta = \theta_1. \tag{1}$$

Let the level of significance $\alpha \in [0, 1]$ *be given.*

(a) Existence: *There exists a test of the form*

$$\phi^\star(x) \;=\; \begin{cases} 1 & \textit{if } f(x|\theta_1) > kf(x|\theta_0) \\ \gamma(x) & \textit{if } f(x|\theta_1) = kf(x|\theta_0) \\ 0 & \textit{if } f(x|\theta_1) < kf(x|\theta_0). \end{cases} \tag{2}$$

where $0 \leq k \leq \infty$ *and* $0 \leq \gamma(X) \leq 1$ *are such that*

$$E[\phi^\star(X)|\theta_0] \;=\; \alpha. \tag{3}$$

If $\alpha > 0$*, then* $k < \infty$*.*

(b) Optimality: *If* $k < \infty$*, then a test of the form (2) satisfying (3) is an MP level* α *test of the problem in (1). If* $k = \infty$ *and* $\gamma(x) \equiv 1$*, then the test is MP level* $\alpha = 0$*.*

PROOF: For part (a), consider the statistic

$$Y \;=\; f_1(X|\theta_1)/f_0(X|\theta_0).$$

If the denominator is 0 and the numerator is positive, $Y = \infty$. If both are 0, we don't care since this happens with probability 0 under each θ_i by our remark above. If $\theta = \theta_0$, then the denominator is positive (except on a set of probability 0), so under θ_0, Y is a well defined RV, and in particular, $P[0 \leq Y < \infty|\theta = \theta_0] = 1$. Hence, Y has a CDF when $\theta = \theta_0$, which we denote $F(y)$.

For the test of the form (2) satisfying (3) that we construct to prove part (a), we will take $\gamma(X)$ to be a constant, which we denote γ. It is useful to allow nonconstant γ in parts (b) and (c).

If $\alpha = 0$, then take $k = 0$ and $\gamma = 0$, and (3) is satisfied. If $0 < \alpha \leq 1$, then let

$$k \;=\; \inf\{y : F(y) \geq 1 - \alpha\}.$$

If $P[Y = k] = 0$, then define γ arbitrarily (in $[0, 1]$). If $P[Y = k] > 0$, then define

$$\gamma \;=\; \frac{F(k) - (1 - \alpha)}{P[Y = k]}.$$

Since F is right continuous, $F(k)$ is equal to $\inf\{y : F(y) \geq 1 - \alpha\}$, so $\gamma \geq 0$. Since F is nondecreasing we have

$$F(k - 0) \;=\; \lim_{y \uparrow k} F(y) \leq 1 - \alpha \leq F(k),$$

and of course

$$P[Y = k] \;=\; F(k) - F(k - 0).$$

Combining these shows $0 \leq \gamma \leq 1$.

Note that if $\alpha > 0$, so $1 - \alpha < 1$, since $\lim_{y \to \infty} F(y) = 1$, we have that there exist (finite) y such that $F(y) > 1 - \alpha$, and hence $k < \infty$ if $\alpha > 0$.

To finish the proof of part (a), we need only verify that (3) holds. We have

$$E[\phi^\star(X)|\theta = \theta_0] \quad = \quad P[Y > k|\theta = \theta_0] + \gamma P[Y = k|\theta = \theta_0]$$

$$
\begin{aligned}
&= (1 - F(k)) + \gamma P[Y = k] \\
&= (1 - F(k)) + F(k) - (1 - \alpha) \\
&= \alpha.
\end{aligned}
$$

Turning to part (b), we apply a Bayesian argument. If we assume $0 - 1$ loss and prior probability

$$P[\theta = \theta_0] \;=\; \pi,$$

then the optimal Bayesian test is of the form (2) where $k = \pi/(1-\pi)$, the prior odds for H_0. For the optimality of the Bayesian test, the value of $\gamma(x)$ is immaterial, since there is equal posterior expected loss if $f(x|\theta_1) = kf(x|\theta_0)$. Remember what the optimality property of the Bayes test is: it minimizes the Bayes risk, i.e. the average of the frequentist risk over the prior. Let ϕ be a test and denote its power function (probability of rejection as a function of θ) by $\beta(\theta) = E[\phi(X)|\theta]$. Now the frequentist risk under $0 - 1$ loss is

$$R(\theta, \phi) \;=\; \begin{cases} \beta(\theta_0) & \text{if } \theta = \theta_0, \\ 1 - \beta(\theta_1) & \text{if } \theta = \theta_1. \end{cases}$$

To explain this, consider the first line. If $\theta = \theta_0$, we lose 1 if we reject H_0, so the expected loss is the probability of rejecting. If $\theta = \theta_1$, we lose 1 if we accept H_0, so the expected loss is the probability of accepting H_0 in this case. The Bayes risk under our prior is

$$r(\phi) \;=\; \beta(\theta_0)\pi + (1 - \beta(\theta_1))(1 - \pi).$$

Let $\beta^\star(\theta)$ denote the power function of the test given in (2) and (3), which we know to be an optimal Bayes test. Assume ϕ is any other level α test of (1), i.e. $\beta(\theta_0) \leq \alpha$. Then $r(\phi) \geq r(\phi^\star)$, so

$$
\begin{aligned}
\alpha\pi + (1 - \beta(\theta_1))(1 - \pi) &\geq \beta(\theta_0)\pi + (1 - \beta(\theta_1))(1 - \pi) \\
&= r(\phi) \\
&\geq r(\phi^\star) \\
&= \beta^\star(\theta_0)\pi + (1 - \beta^\star(\theta_1))(1 - \pi) \\
&= \alpha\pi + (1 - \beta^\star(\theta_1))(1 - \pi).
\end{aligned}
$$

As long as $\pi < 1$, then comparing the beginning and final expressions above, we obtain

$$\beta(\theta_1) \ \leq \ \beta^\star(\theta_1),$$

i.e., $\phi^\star$ is more powerful.

If $\pi = 1$, then $k = \pi/(1-\pi) = \infty$, which implies by part (a) that $\alpha = 0$. For a level 0 test, we should accept H_0 whenever $f(x|\theta_0)$ is positive. This is clear in the PMF case (assume $\phi(x_0) > 0$ and $f(x_0|\theta_0) > 0$ for some x_0; then $E[\phi(X)|\theta = \theta_0] = \sum_x \phi(x) f(x|\theta_0) \geq \phi(x_0) f(x_0|\theta_0) > 0$, but the level $\alpha = 0$ requirement is that $E[\phi(X)|\theta = \theta_0] \leq 0$). For the PDF case, the argument is a little more subtle and left to the reader. However, given that we should accept H_0 whenever $f(x|\theta_0) > 0$, by similar reasoning, a MP test would reject H_0 whenever $f(x|\theta_0) = 0$ but $f(x|\theta_1) > 0$. That is, if $k = \infty$, then $\gamma = 1$.

This completes the proof of the Lemma.

Remarks: There is also a uniqueness part one can prove. This is essentially the "Necessity" part of the Neyman-Pearson Lemma given in the book (p. 388). Note that in practice we would only consider $\alpha < 1$ (in fact, I don't personally recall use of $\alpha > 0.1$ in practice), so worrying about the case $\alpha = 1$ is more for mathematical completeness than practical import. The same is true for considering randomized tests – they allow us to find a UMP test for any allowed value of α, but one wouldn't use them in practice.