

Lecture Note 16: Hypothesis Testing and the Neyman-Pearson Lemma, CB 8.1, 8.3.1-8.3.2

We now turn to a different type of statistical problem. Rather than try to provide a guess about the true parameter value (point estimation), we want to decide whether a particular parameter value, or more generally a particular theory about the process underlying the data, is consistent with the data.

For example, we may wish to know whether a given coin is fair, that is, does the probability of heads $p = 1/2$? We observe $X \sim Bin(1, p)$ if we toss the coin once, or $X \sim Bin(N, p)$ if we toss the coin N times, and we need to decide whether it is fair or not fair. In another example we may wish to know whether on average people from one group in society get paid more than people from another group, on the basis of a random sample of individuals. In both cases we are concerned with the question whether a specific value of the parameter is consistent with the evidence provided by the data. This type of question is referred to as hypothesis testing.

The general setup is as follows. We have a random variable X , with a distribution $f_X(x; \theta)$ for some θ about which we only know that it is in some set Θ . We have a hypothesis regarding the value of θ . The hypothesis is that $\theta \in \Theta_0$, where Θ_0 is some subset of the overall parameter space Θ . We refer to this as the null hypothesis, or H_0 . If the null hypothesis is not true, then $\theta \in \Theta_0^c$, where Θ_0^c is the complement of Θ_0 in the parameter space Θ . In that case the null hypothesis is false and the alternative hypothesis H_a is true. Formally:

$$\begin{aligned} H_0 : & \quad \theta \in \Theta_0, \\ H_a : & \quad \theta \in \Theta_0^c. \end{aligned}$$

(In some texts, H_1 is used to denote the alternative, rather than H_a .) We do not entertain the possibility that the true value of θ is not consistent with either the null hypothesis or the alternative hypothesis, for example because the entire model is incorrect.

Given the null and alternative hypothesis and given a realization of the random variable X we are faced with making a decision, or taking an action. We decide either to reject the null hypothesis, or not to reject the null hypothesis, that is to accept the null hypothesis. (Some statistics books argue that it is always wrong to say that an hypothesis is accepted, and that one should only say that the hypothesis did not get rejected. Here we are not so strict.) We make the decision on the basis of the value of the realization of the random variable X . We can therefore characterize our decision rule by the set of values for which we reject the null hypothesis. We refer to this set as the critical region, denoted by C_X . If $x \in C_X$ then we reject the null hypothesis, and if $x \notin C_X$, we accept the null hypothesis. As a result there are four possible outcomes: the null hypothesis can be true or false, and we can accept or reject the null.

Some of these decisions are wrong: if the null hypothesis is true and we reject it, or if the null hypothesis is false and we accept it. These are known as type I and type II errors respectively. Table 1 illustrates the four possibilities.

Table 1: HYPOTHESIS TESTING

State of Nature ↓	Decision	
	Accept H_0 (Do Not Reject H_0)	Reject H_0
H_0	Correct Decision	Type I Error
H_a	Type II Error	Correct Decision

Example

Suppose a college accepts applicants at an overall rate of 25%. There is a particular sub-population from which 20 applicants applied and only 2 were accepted. Is this evidence that the college discriminates against this group? First let us consider the model. It may be reasonable to model the number of accepted applicants from this group as a binomial $Bin(N, p)$ random variable with parameters $N = 20$ and probability p . Recall that the binomial distribution has PMF

$$f_X(x; N, p) = \binom{N}{x} p^x (1-p)^{N-x}.$$

The parameter space is $p \in P = [0, 0.25]$. The null hypothesis is

$$H_0 : p = 0.25,$$

and the alternative hypothesis is

$$H_a : p < 0.25.$$

(We can argue a little about the model or the formulation of the hypotheses. For example, if the college attempts to get a diverse student body, they may not treat applications as independent, and instead try to achieve a particular mix. That may lead the number of acceptances from a particular group to still have mean $N \cdot p$, but variance lower than $N \cdot p(1-p)$, so the Binomial model might not be appropriate. One might also argue that the parameter space is $P = [0, 1]$ and that the null hypothesis should be $p \geq 0.25$ rather than $p = 0.25$. Here we stick for the time being with our initial formulation.)

Let us now consider the shape of the critical region. Suppose we reject the null hypothesis of $p = 0.25$ for some value of x , or $x \in C_X$. What would we decide if we observed $x - 1$? This would appear to be even stronger evidence against the null hypothesis than x (given that p cannot be larger than 0.25), and so it would appear reasonable to have $x - 1 \in C_X$

as well. This implies that the critical region should be of the form $C_X = \{0, 1, \dots, k\}$ for some integer k . (it could be that the critical region is empty, but ignore this for the time being.)

Now suppose we use the critical region $C_X = \{0, 1, \dots, k\}$. What is the probability that we make a type I error?

$$\begin{aligned} Pr(\text{type I error}) &= Pr(X \in C_X | N = 20, p = .25) \\ &= \sum_{i=0}^k \binom{20}{i} \cdot 0.25^i \cdot 0.75^{20-i}. \end{aligned}$$

What is the probability of a type II error? This depends on the value of the probability under the alternative.

$$\begin{aligned} Pr(\text{type II error}) &= Pr(X \notin C_X | N = 20, p) \\ &= \sum_{i=k+1}^{20} \binom{20}{i} \cdot p^i \cdot (1-p)^{20-i}. \end{aligned}$$

Now suppose we expand the critical region from $C_X = \{0, 1, \dots, k+1\}$. The probability of a type I error increases by

$$\binom{20}{k+1} \cdot 0.25^{k+1} \cdot 0.75^{19-k}.$$

The probability of a type II error decreases by

$$\binom{20}{k+1} \cdot p^{k+1} \cdot (1-p)^{19-k}.$$

There is a fundamental tradeoff between type I and type II errors. By expanding the critical region we increase the probability of type I errors and decrease the probability of type II errors. If we contract the critical region, we do the reverse.

How should we decide on the value of k that determines the critical region? Type I and type II errors have very different consequences and costs. In the example a type I error is costly to the college that will be forced to pay a fine and change its practices even though these are perfectly acceptable. A type II error is costly to society at large as well as to the specific individuals denied admission. The cost of this error depends on the actual value of p under the alternative hypothesis. If p is very close to 0.25, presumably the cost is very small. How do we trade these things off? One natural way would be to specify the costs under all the possible outcomes, and then try to minimize the expected costs. This decision-theoretic approach has a lot of merit, but can require a lot of work to specify the costs completely.

The classical approach to testing in statistics, initiated in the work of Neyman and Pearson, avoids specifying these costs explicitly. Instead, we start by requiring that the probability of a type I error is less than some preset number, the level of the test. The level of the

test is conventionally chosen to be either 0.05, or 0.01 or 0.10. For example, if the level is $\alpha = 0.05$, we require that for all $\theta \in \Theta_0$,

$$Pr_{\theta}(X \in C_x) \leq \alpha.$$

(Here Pr_{θ} means we calculate the probability assuming that X is distributed according to $f_X(x; \theta)$.)

Next, define the power function $\beta(\theta)$ as the probability of rejecting the null hypothesis if the true value of the parameter is θ :

$$\beta(\theta) = Pr_{\theta}(X \in C_x).$$

Note that for $\theta \in \Theta_0^c$, the power function is one minus the probability of a type II error:

$$\beta(\theta) = 1 - Pr_{\theta}(\text{type II error}).$$

Intuitively, we want the power function to be as large as possible for values of $\theta \in \Theta_0^c$, subject to the restriction that the probability of a Type I error is not greater than the level α . In other words, we want to maximize $\beta(\theta)$ for $\theta \in \Theta_0^c$ subject to

$$\beta(\theta) \leq \alpha, \quad \forall \theta \in \Theta_0.$$

This is the key idea of the Neyman-Pearson approach to hypothesis testing.

Example continued

Suppose we decide we want the probability of a type I error to be less than 0.01. The probability of a type II error goes down if we increase k , so this means we should set k equal to the largest integer such that

$$Pr(\text{type I}) = \sum_{i=0}^k \binom{20}{i} \cdot 0.25^i \cdot 0.75^{20-i} \leq 0.01.$$

Suppose we set $k = 0$. Then we have

$$Pr(\text{type I}) = \sum_{i=0}^0 \binom{20}{i} \cdot 0.25^i \cdot 0.75^{20-i} = \binom{20}{0} \cdot 0.25^0 \cdot 0.75^{20} = 0.0032.$$

Suppose we choose $k = 1$. Then

$$Pr(\text{type I}) = \sum_{i=0}^1 \binom{20}{i} \cdot 0.25^i \cdot 0.75^{20-i} = 0.0032 + 0.0211 = 0.0243.$$

Hence, $k = 1$ is too large and we conclude that $C_X = \{0\}$ is the optimal critical region given the (self-imposed) restriction that we want the probability that the probability of a type I error should be less than or equal to 0.01.

What does the power function look like?

$$\begin{aligned}\beta(p) &= Pr(X \in \{0\}; N = 20, p) \\ &= (1 - p)^{20}.\end{aligned}$$

It is clearly decreasing in p . At 0.25 it is the same as the probability of a type I error, 0.0032. At other values of p consistent with the alternative hypothesis ($p < 0.25$) it is larger than the probability of a type I error.

□

So how do we look for tests in general? The first issue is the shape of the critical region. Here it was pretty clear that the critical region should be of the form $C_X = \{0, 1, \dots, k\}$. In general that is a more complicated issue. Second, we decide, rather arbitrarily, on the level of the test. Initially we will look at some exact results for these questions. Then we consider large sample approximations similar to those considered in point estimation.

Now we define a criterion that will measure optimality of a test. It requires that the probability of a type II error is minimized for all values of the parameter consistent with the alternative hypothesis.

Definition 1 Consider all tests of level α for the null hypothesis $\theta \in \Theta_0$ against the alternative $\theta \in \Theta_0^c$. A test with power function $\beta(\theta)$ is uniformly most powerful if, for all alternative tests with level α and power function $\beta'(\theta)$, $\beta(\theta) \geq \beta'(\theta)$ for all $\theta \in \Theta_0^c$.

There is no guarantee that uniformly most powerful tests actually exist. We first study a simple case where such tests are easy to find. We focus on the case where both the null hypothesis and the alternative hypothesis are simple, that is, where the sets Θ_0 and Θ_0^c contain a single element each:

$$\begin{aligned}H_0 : & \quad \theta = \theta_0, \\ H_a : & \quad \theta = \theta_1.\end{aligned}$$

(If a hypothesis contains more than a single point, we say that it is a composite hypothesis.)

Result 1 (*Neyman–Pearson lemma*)

Consider testing the null hypothesis $H_0 : \theta = \theta_0$ against the alternative $H_a : \theta = \theta_1$ using a critical region of the form

$$C_X = \{x : f_X(x; \theta_1) \geq k \cdot f_X(x; \theta_0)\}.$$

Let

$$\alpha = \int_{C_X} f_X(x; \theta_0) dx.$$

This test is the uniformly most powerful test of level α .

Proof: Let $\beta(\theta)$ denote the power function of the test proposed. Consider any other test with a critical region C'_X and power function $\beta'(\theta)$. Define

$$\phi(x) = 1\{x \in C_X\},$$

and

$$\phi'(x) = 1\{x \in C'_X\}.$$

Consider

$$(\phi(x) - \phi'(x)) \cdot (f_X(x; \theta_1) - k \cdot f_X(x; \theta_0)).$$

If this expression differs from zero, we must either have $\phi(x) - \phi'(x) = 1$ or $\phi(x) - \phi'(x) = -1$.

If $\phi(x) - \phi'(x) = 1$, $(f_X(x; \theta_1) - k \cdot f_X(x; \theta_0))$ must be nonnegative by the form of the critical region C_X , so the entire expression is nonnegative.

If $\phi(x) - \phi'(x) = -1$, the second factor must be ≤ 0 , and the product again is nonnegative.

Hence,

$$(\phi(x) - \phi'(x)) \cdot (f_X(x; \theta_1) - k \cdot f_X(x; \theta_0)) \geq 0,$$

and therefore

$$\begin{aligned} & \int_x (\phi(x) - \phi'(x)) \cdot (f_X(x; \theta_1) - k \cdot f_X(x; \theta_0)) dx \\ & \int_x (\phi(x) - \phi'(x)) \cdot f_X(x; \theta_1) dx - k \cdot \int_x (\phi(x) - \phi'(x)) \cdot f_X(x; \theta_0) dx \\ & = \beta(\theta_1) - \beta'(\theta_1) - k \cdot (\beta(\theta_0) - \beta'(\theta_0)) \geq 0. \end{aligned}$$

If both tests are level α tests, $\beta(\theta_0) = \beta'(\theta_0) = \alpha$, and so it must be the case that

$$\beta(\theta_1) - \beta'(\theta_1) \geq 0,$$

and the second test cannot be the most powerful test. \square