STATS 200: Introduction to Statistical Inference **Autumn 2016**

Lecture 6 — Simple alternatives and the Neyman-Pearson lemma

Last lecture, we discussed a number of ways to construct test statistics for testing a simple null hypothesis, and we showed how to use the null distribution of the statistic to determine the rejection region so as to achieve a desired significance level. The goal of today's lecture is to answer the following question: Which test statistic should we use? The answer depends on the alternative hypothesis that we wish to distinguish from the null.

6.1 The Neyman-Pearson lemma

Let's focus on the problem of testing a simple null hypothesis H_0 against a simple alternative hypothesis H_1. We denote by

$$\beta = \mathbb{P}_{H_1}[\text{accept } H_0]$$

the probability of **type II error**—accepting the null H_0 when in fact the alternative H_1 is true. (Here, $\mathbb{P}_{H_1}[\mathcal{E}]$ denotes probability of an event $\mathcal{E}$ if H_1 is true.) Equivalently,

$$1 - \beta = \mathbb{P}_{H_1}[\text{reject } H_0]$$

is the probability of correctly rejecting H_0 when H_1 is true, which is called the **power** of the test against H_1.[1]

When designing a hypothesis test for testing H_0 versus H_1, we have the following goal:

maximize: the power of the test against H_1
subject to: the significance level of the test under H_0 is at most α

This is an example of a constrained optimization problem, which we can reason about in the following way: Suppose we observe data which are realizations of random variables $X_1, \ldots, X_n$. For notational convenience, let us denote by $\mathbf{X} = (X_1, \ldots, X_n)$ the entire data vector, and by $\mathbf{x} = (x_1, \ldots, x_n)$ a vector of possible values for $\mathbf{X}$. In the discrete case, suppose the hypotheses are

$$\begin{aligned} &H_0 : \mathbf{X} \text{ is distributed with joint PMF } f_0(\mathbf{x}) := f_0(x_1, \ldots, x_n), \\ &H_1 : \mathbf{X} \text{ is distributed with joint PMF } f_1(\mathbf{x}) := f_1(x_1, \ldots, x_n). \end{aligned}$$

Let $\mathcal{X}$ denote the set of all possible values of $\mathbf{X}$ under f_0 and f_1. To define the hypothesis test, for each $\mathbf{x} \in \mathcal{X}$, we must specify whether to accept or reject H_0 if the observed data is $\mathbf{x}$. In other words, we specify a rejection region $\mathcal{R} \subset \mathcal{X}$ such that we reject H_0 if the observed data belongs to $\mathcal{R}$ and we accept H_0 otherwise. Then the probability of rejecting

[1] Caution: Some books/papers use opposite notation and let β denote the power and $1 - \beta$ denote the probability of type II error. Make sure to double-check the meaning of the notation.

H_0 if H_0 were true would be $\sum_{\mathbf{x}\in\mathcal{R}} f_0(\mathbf{x})$, the probability of rejecting H_0 if H_1 were true would be $\sum_{\mathbf{x}\in\mathcal{R}} f_1(\mathbf{x})$, and the above optimization problem is formalized as choosing the rejection region $\mathcal{R} \subset \mathcal{X}$ with the goal

$$\begin{aligned} \text{maximize } & \sum_{\mathbf{x}\in\mathcal{R}} f_1(\mathbf{x}) \\ \text{subject to } & \sum_{\mathbf{x}\in\mathcal{R}} f_0(\mathbf{x}) \leq \alpha. \end{aligned}$$

The continuous case is similar: suppose the hypotheses are

$$\begin{aligned} &H_0 : \mathbf{X} \text{ is distributed with joint PDF } f_0(\mathbf{x}) := f_0(x_1, \ldots, x_n), \\ &H_1 : \mathbf{X} \text{ is distributed with joint PDF } f_1(\mathbf{x}) := f_1(x_1, \ldots, x_n). \end{aligned}$$

We define a hypothesis test by defining the region $\mathcal{R} \subset \mathbb{R}^n$ such that we reject H_0 if and only if the observed data $\mathbf{x}$ belongs to $\mathcal{R}$. The above optimization problem is to choose $\mathcal{R} \subset \mathbb{R}^n$ with the goal

$$\begin{aligned} \text{maximize } & \int_{\mathcal{R}} f_1(\mathbf{x})\, dx_1 \ldots dx_n \\ \text{subject to } & \int_{\mathcal{R}} f_0(\mathbf{x})\, dx_1 \ldots dx_n \leq \alpha. \end{aligned}$$

In either the discrete or continuous case, what are the best points $\mathbf{x}$ to include in this rejection region $\mathcal{R}$? A moment's thought should convince you that $\mathcal{R}$ should consist of those points $\mathbf{x}$ corresponding to the smallest values of $\frac{f_0(\mathbf{x})}{f_1(\mathbf{x})}$, as these give the "smallest increase in type I error per unit increase of power". Another interpretation is that these are the points providing the strongest evidence in favor of H_1 over H_0. The statistic

$$L(\mathbf{X}) = \frac{f_0(\mathbf{X})}{f_1(\mathbf{X})}$$

is called the **likelihood ratio statistic**, and the test that rejects for small values of $L(\mathbf{X})$ is called the **likelihood ratio test**. The Neyman-Pearson lemma shows that the likelihood ratio test is the most powerful test of H_0 against H_1:

Theorem 6.1 (Neyman-Pearson lemma). *Let H_0 and H_1 be simple hypotheses (in which the data distributions are either both discrete or both continuous). For a constant $c > 0$, suppose that the likelihood ratio test which rejects H_0 when $L(\mathbf{x}) < c$ has significance level α. Then for any other test of H_0 with significance level at most α, its power against H_1 is at most the power of this likelihood ratio test.*

Proof. Consider the discrete case, and let $\mathcal{R} = \{\mathbf{x} : L(\mathbf{x}) < c\}$ be the rejection region of the likelihood ratio test. Note that among all subsets of $\mathcal{X}$, $\mathcal{R}$ maximizes the quantity

$$\sum_{x\in\mathcal{R}} (c f_1(\mathbf{x}) - f_0(\mathbf{x})),$$

because $cf_1(\mathbf{x}) - f_0(\mathbf{x}) > 0$ for $\mathbf{x} \in \mathcal{R}$ and $cf_1(\mathbf{x}) - f_0(\mathbf{x}) \leq 0$ for $\mathbf{x} \notin \mathcal{R}$. Hence for any other test with significance level at most α, say with rejection region $\mathcal{R}'$,

$$\sum_{x \in \mathcal{R}} (cf_1(\mathbf{x}) - f_0(\mathbf{x})) \geq \sum_{x \in \mathcal{R}'} (cf_1(\mathbf{x}) - f_0(\mathbf{x})) \,.$$

Rearranging the above, this implies

$$c \left(\sum_{x \in \mathcal{R}} f_1(\mathbf{x}) - \sum_{x \in \mathcal{R}'} f_1(\mathbf{x}) \right) \geq \sum_{x \in \mathcal{R}} f_0(\mathbf{x}) - \sum_{x \in \mathcal{R}'} f_0(\mathbf{x}) = \alpha - \sum_{x \in \mathcal{R}'} f_0(\mathbf{x}) \geq 0,$$

where the last inequality follows because $\sum_{x \in \mathcal{R}'} f_0(\mathbf{x})$ is the significance level of the test that rejects for $\mathbf{x} \in \mathcal{R}'$. Then $\sum_{x \in \mathcal{R}} f_1(\mathbf{x}) \geq \sum_{x \in \mathcal{R}'} f_1(\mathbf{x})$, i.e. the likelihood ratio test has power at least that of this other test. The proof in the continuous case is exactly the same, with all sums above replaced by integrals over $\mathcal{R}$ and $\mathcal{R}'$. □

6.2 Examples

Let's work out what the likelihood ratio test actually is for two simple examples.

Example 6.2. Consider data $X_1, \ldots, X_n$ and the following null and alternative hypotheses:

$$H_0 : X_1, \ldots, X_n \overset{IID}{\sim} \mathcal{N}(0,1)$$
$$H_1 : X_1, \ldots, X_n \overset{IID}{\sim} \mathcal{N}(\mu,1)$$

Here we assume μ is a known, specified value (not equal to 0), so that H_1 is a simple alternative hypothesis. The joint PDF of $(X_1, \ldots, X_n)$ under H_0 is

$$f_0(x_1, \ldots, x_n) = \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi}} e^{-\frac{x_i^2}{2}} = \left(\frac{1}{\sqrt{2\pi}} \right)^n \exp\left(-\frac{x_1^2 + \ldots + x_n^2}{2} \right).$$

The joint PDF under H_1 is

$$f_1(x_1, \ldots, x_n) = \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi}} e^{-\frac{(x_i - \mu)^2}{2}} = \left(\frac{1}{\sqrt{2\pi}} \right)^n \exp\left(-\frac{(x_1 - \mu)^2 + \ldots + (x_n - \mu)^2}{2} \right).$$

Thus, the likelihood ratio statistic is

$$L(X_1, \ldots, X_n) = \frac{f_0(X_1, \ldots, X_n)}{f_1(X_1, \ldots, X_n)} = \exp\left(-\frac{X_1^2 + \ldots + X_n^2}{2} + \frac{(X_1 - \mu)^2 + \ldots + (X_n - \mu)^2}{2} \right).$$

Expanding the squares and simplifying, we obtain

$$L(X_1, \ldots, X_n) = \exp\left(\frac{-2\mu(X_1 + \ldots + X_n) + n\mu^2}{2} \right).$$

Suppose first that $\mu > 0$. Then $L(X_1, \ldots, X_n)$ is a strictly *decreasing* function of the sample mean $\bar{X} = \frac{1}{n}(X_1 + \ldots + X_n)$. Hence, rejecting for small values of $L(X_1, \ldots, X_n)$ is the same as rejecting for large values of $\bar{X}$. So the Neyman-Pearson lemma tells us that the most powerful test should reject when $\bar{X} > c$, for some threshold c. We pick c to ensure that the significance level is α under H_0: Since the null distribution of $\bar{X}$ is $\bar{X} \sim \mathcal{N}(0, \frac{1}{n})$, c should be the $\frac{1}{\sqrt{n}} z(\alpha)$ where $z(\alpha)$ is the "upper α point" of the standard normal distribution.

Now suppose that $\mu < 0$. Then $L(X_1, \ldots, X_n)$ is strictly *increasing* in $\bar{X}$, so rejecting for small $L(X_1, \ldots, X_n)$ is the same as rejecting for small $\bar{X}$. By the same argument as above, to ensure significance level α, the most powerful test rejects when $\bar{X} < -\frac{1}{\sqrt{n}} z(\alpha)$.

Remark 6.3. The most powerful test against the alternative $H_1 : X_1, \ldots, X_n \sim \mathcal{N}(\mu, 1)$ is the same for any $\mu > 0$ (rejecting when $\bar{X} > \frac{1}{\sqrt{n}} z(\alpha)$), and neither the test statistic nor the rejection region depend on the specific value of μ. This means that, in fact, this test is **uniformly most powerful** against the (one-sided) composite alternative

$$H_1 : X_1, \ldots, X_n \overset{IID}{\sim} \mathcal{N}(\mu, 1) \text{ for some } \mu > 0.$$

On the other hand, the most powerful test is different for $\mu > 0$ versus for $\mu < 0$: one test rejects for large positive values of $\bar{X}$, and the other rejects for large negative values of $\bar{X}$. This implies that there does not exist a single most powerful test for the (two-sided) composite alternative

$$H_1 : X_1, \ldots, X_n \overset{IID}{\sim} \mathcal{N}(\mu, 1) \text{ for some } \mu \neq 0.$$

Example 6.4. Let $X_1, \ldots, X_n \in \{0, 1\}$ be the results of n flips of a coin, and consider the following null and alternative hypotheses:

$$H_0 : X_1, \ldots, X_n \overset{IID}{\sim} \text{Bernoulli}\left(\tfrac{1}{2}\right)$$
$$H_1 : X_1, \ldots, X_n \overset{IID}{\sim} \text{Bernoulli}(p).$$

Here we assume that $p \neq \frac{1}{2}$ is a known and specified value, so H_1 is simple. The joint PMF of $(X_1, \ldots, X_n)$ under H_0 and H_1 are, respectively,

$$f_0(x_1, \ldots, x_n) = \prod_{i=1}^{n} \frac{1}{2} = \frac{1}{2^n},$$
$$f_1(x_1, \ldots, x_n) = \prod_{i=1}^{n} p^{x_i}(1-p)^{1-x_i} = p^{x_1+\ldots+x_n}(1-p)^{n-x_1-\ldots-x_n} = (1-p)^n \left(\frac{p}{1-p}\right)^{x_1+\ldots+x_n}.$$

Thus, the likelihood ratio statistic is

$$L(X_1, \ldots, X_n) = \frac{f_0(X_1, \ldots, X_n)}{f_1(X_1, \ldots, X_n)} = \frac{1}{2^n(1-p)^n} \left(\frac{1-p}{p}\right)^{X_1+\ldots+X_n}.$$

First suppose $p > \frac{1}{2}$. Then $L(X_1, \ldots, X_n)$ is a decreasing function of $S = X_1 + \ldots + X_n$, so rejecting for small values of $L(X_1, \ldots, X_n)$ is the same as rejecting for large values of S.

Hence, by the Neyman-Pearson lemma, the most powerful test rejects when $S > c$ for a constant c. We choose c to ensure significance level α: Under H_0, $S \sim \text{Binomial}(n, \frac{1}{2})$, so c should be the $1 - \alpha$ quantile of the $\text{Binomial}(n, \frac{1}{2})$ distribution. This test is the same for all $p > \frac{1}{2}$, so it is in fact uniformly most powerful against the composite alternative

$$H_1 : X_1, \ldots, X_n \overset{IID}{\sim} \text{Bernoulli}(p) \text{ for some } p > \tfrac{1}{2}.$$

For $p < \frac{1}{2}$, $L(X_1, \ldots, X_n)$ is increasing in S, so the most powerful test rejects for $S < c$ and some constant c. To ensure significance level α, c should be the α quantile of the $\text{Binomial}(n, \frac{1}{2})$ distribution. This test is the same for all $p < \frac{1}{2}$, so it is uniformly most powerful against the compositive alternative

$$H_1 : X_1, \ldots, X_n \overset{IID}{\sim} \text{Bernoulli}(p) \text{ for some } p < \tfrac{1}{2}.$$

Remark 6.5. We have glossed over a detail, which is that when the distribution of the likelihood ratio statistic $L(\mathbf{X})$ is discrete under H_0, it might not be possible to choose c so that the significance level is exactly α. For instance, in the previous example, suppose we wish to achieve significance level $\alpha = 0.05$, and $n = 20$. For $S \sim \text{Binomial}(20, \frac{1}{2})$, we have $\mathbb{P}[S \geq 15] = 0.021$ and $\mathbb{P}[S \geq 14] = 0.058$. So if we reject H_0 when $S \geq 14$, we do not achieve significance level $\leq \alpha$, and if we reject H_0 when $S \geq 15$, then we are too conservative.

The theoretically correct solution is to perform a randomized test: Always reject H_0 when $S \geq 15$, always accept H_0 when $S \leq 13$, and reject H_0 with a certain probability when $S = 14$, where this probability is chosen to make the significance level exactly α. A more complete statement of the Neyman-Pearson lemma shows that this type of (possibly randomized) likelihood ratio test is most powerful among all randomized tests.

In practice, it might not be acceptable to use a randomized test. (We found the effects of this drug to be statistically significant because our statistical procedure told us to flip a coin, and our coin landed heads...) So we might take the more conservative option of just rejecting H_0 when $S \geq 15$.