Hypothesis Testing

EECS 126 (UC Berkeley)

Spring 2019

1 Introduction

Thus far in this class, we have been trying to predict behavior for a given model (e.g., how many arrivals do we expect in τ time for a Poisson Process). Hypothesis Testing deals with the reverse direction, of trying to fit a model for a set of observations.[1]

The general setup of the problem is as follows: you are given some observation(s) $x \sim X$, and you are told that there is some $\theta^* \in \Theta$ for which the data was drawn as $X \sim \mathbb{P}_{\theta^*}$, and the task is to determine whether θ^* is in Θ_0 or in Θ_1 (these sets are constructed where $\Theta_0 \cup \Theta_1 = \Theta$ and $\Theta_0 \cap \Theta_1 = \emptyset$). If $\theta^* \in \Theta_0$ we say H_0 is correct, otherwise, H_1 is correct. An example may be more illuminating:

Example 1

Suppose you have 10 fire alarms in your home. k of the fire alarms went off, and you want to determine whether there is actually a fire in your home or its a false alarm. H_0, the null hypothesis, is that there is no fire; H_1, the alternate hypothesis, is that there is a fire in your home. In general, we assume the null hypothesis is correct until we get more information to overturn this assumption. In this example, maybe we decide that if 1 fire alarm went off, this is not enough information to say that the house is on fire, because its more likely that the alarm is faulty. However, if 8 alarms went off, maybe under our knowledge, we would reject the null hypothesis and say the house is on fire. In this example $\Theta_0 = \{0\}$ and $\Theta_1 = \{1\}$. If we decide based on our information that $\theta = 0$, then we say the house is not on fire (H_0). Otherwise we would say $\theta = 1$, that the house is on fire (H_1).

We call H_0 (and H_1) **simple** if Θ_0 and Θ_1 each only contain one item. Notice that in the example, the hypothesis test is simple. Suppose instead of determining whether the house was on fire, we were trying to estimate what percentage of the house was on fire. $\Theta_0 = \{0\}$ and $\Theta_1 = (0, 100]$. In this case the null hypothesis that 0% of the house is on fire is simple, while the alternate hypothesis is not simple. *The focus of this note will be on simple null and simple alternate tests.*

We can design any arbitrary test that uses our observation (later we will find that there is some notion of "optimal" test); in particular our test can be thought of as a **Acceptance Region** A i.e., *the test is: accept* $H_0 \iff x \in A$. More often than not (at least in this class) the problem is to determine what "optimal" A needs to be for a given set of conditions.

[1] the Neyman-Pearson regime of hypothesis testing that we will be doing in this class is a Frequentist method (i.e., H_0 being true is not a random variable). Bayesian hypothesis testing has a different setup and solution which we will not go into in this note.

Example 2: Acceptance Regions

Examples of such arbitrary tests are:

1. Reject H_0 if and only if $x > t$
2. Reject H_0 if and only if $x = t$
3. Reject H_0 with probability γ when $x > t$

where $t \in \mathbb{R}$ is arbitrary and used to define the acceptance region of the test. Note that these tests are arbitrary and are not necessarily "optimal" in the Neyman-Pearson sense.

For any arbitrary test, there is a possibility of rejecting the null when the null is in fact true. This is called a **Type-I Error** or **probability of false alarm (PFA)**. The probability of a type-I error occurring is called the **significance level** which is denoted by α. More formally, this value is in general[2]:

$$\alpha(A) \coloneqq \mathbb{P}_{H_0}(\text{choosing } H_1) = \mathbb{P}_{H_0}(x \notin A)$$

On the flip side, there is a possibility of accepting the null when the alternative is true. This is called a **Type-II Error**. The probability of a type-II error is defined as β and it is in general:

$$\beta(A) \coloneqq \mathbb{P}_{H_1}(\text{choosing } H_0) = \mathbb{P}_{H_1}(x \in A)$$

The probability of the test to correctly reject the null is called the **power** (also called the **probability of correct detection (PCD)**) of the test, which is denoted by $1 - \beta$.

Even though A can be chosen arbitrarily, we want to find the "best" A. It is clear that there are two conflicting interests. We want minimize the probability of a type-I error, while simultaneously maximizing the power of the test. So to formalize this issue, we formulate the problem as "how large of a PCD can we get for a constrained PFA?" In other words, it is the following optimization problem:

$$\begin{aligned} q \coloneqq \max_{A} \; & 1 - \beta(A) \\ \text{s.t. } & \alpha(A) \leq z \end{aligned} \tag{1}$$

Where z is some predefined constant (often times, $z = 0.05$). It turns out for the simple vs simple hypothesis testing regime, we can characterize the "optimal" testing scheme (i.e., the A that maximizes q)

2 "Optimal" Likelihood Ratio test

Definition 1: Likelihood ratio

We define the Likelihood ratio to be:

$$L(x) \coloneqq \frac{\mathbb{P}_{H_1}(x)}{\mathbb{P}_{H_0}(x)} \quad \text{or} \quad L(x) \coloneqq \frac{f_{H_1}(x)}{f_{H_0}(x)}$$

[2] $\mathbb{P}_{H_0}(\ldots)$ denotes the probability of $\ldots$ occurring when H_0 is true. Similarly, $\mathbb{P}_{H_1}(\ldots)$ denotes the probability of $\ldots$ occurring when H_1 is true.

Definition 2: Likelihood ratio test

The Likelihood ratio test (also called the Neyman-Pearson Test): for a critical threshold c and for observation x

1. accept H_0 if $L(x) < c$
2. reject H_0 with probability γ if $L(x) = c$.
3. reject H_0 if $L(x) > c$

Note that this use of $L(x)$ characterizes the acceptance region A

Lemma 1: Neyman-Pearson

Consider a particular choice of c in the Likelihood Ratio test, such that

$$\mathbb{P}_{H_0}(L(x) > c) + \gamma\,\mathbb{P}_{H_0}(L(x) = c) = \alpha_0 \qquad \mathbb{P}_{H_1}(L(x) < c) + (1-\gamma)\,\mathbb{P}_{H_1}(L(x) = c) = \beta_0$$

Suppose there is some other test, with rejection region A, which achieves a smaller or equal false rejection probability $\mathbb{P}_{H_0}(X \in A) \leq \alpha_0$ then $\mathbb{P}_{H_1}(x \notin A) \geq \beta_0$. There is strict inequality $\mathbb{P}_{H_1}(X \notin A) > \beta_0$ when $\mathbb{P}_{H_0}(X \in A) < \alpha_0$

In other words, the Likelihood Ratio Test is the most powerful test.

Proof. In the Bertsekas book page 491. This theorem was paraphrased from the Bertsekas book. □

This lemma characterizes what the optimal acceptance region should look like. As a consequence of the Neyman-Pearson Lemma, (1) is equivalent to:

$$\begin{aligned} q = \max_{c}\ & 1 - \mathbb{P}_{H_1}(L(x) < c) - (1-\gamma)\,\mathbb{P}_{H_1}(L(x) = c) \\ \text{s.t.}\ \ & \mathbb{P}_{H_0}(L(x) > c) + \gamma\,\mathbb{P}_{H_0}(L(x) = c) = z \end{aligned} \tag{2}$$

This has an important consequence as we find the optimal parameters of our decision rule by solving for the value of c that sets $\mathbb{P}_{H_0}(L(x) > c) + \gamma\,\mathbb{P}_{H_0}(L(x) = c) = z$. However, $L(x)$ is often difficult to analyze. So oftentimes in this class, there will be some equivalent condition that is easier to manipulate. i.e., it may be the case that we have a relationship $L(x) > c \iff B$ for an event $B(x)$ which is easier to handle; e.g., in this class, often times $L(x)$ is monotonic in x; this means $L(x) > c \iff x > t$ or $x < t$. We can use these if and only if conditions to rewrite the Likelihood ratio test in an equivalent, more digestible way. This type of trick is elucidated in the next example:

Example 3: Normal RV Hypothesis Test

Suppose we are trying to determine whether X is $\mathcal{N}(0, \sigma^2)$ or $\mathcal{N}(1, \sigma^2)$ by using our observation x for a significance level z via a Likelihood ratio hypothesis test. Our two simple hypotheses are:

- $H_0: \ X \sim \mathcal{N}(0, \sigma^2)$; $\Theta_0 = \{\mu = 0\}$
- $H_1: \ X \sim \mathcal{N}(1, \sigma^2)$; $\Theta_1 = \{\mu = 1\}$

Let's see what the most powerful test is. **i.e., let's conduct a likelihood ratio test.** First, let's see what the likelihood ratio is:

$$L(x) = \frac{f_{\mu=1}(x)}{f_{\mu=0}(x)} = \frac{\frac{1}{\sigma\sqrt{2\pi}} \exp\left(-\frac{(x-1)^2}{2\sigma^2}\right)}{\frac{1}{\sigma\sqrt{2\pi}} \exp\left(-\frac{x^2}{2\sigma^2}\right)} = \exp\left(-\frac{(x-1)^2}{2\sigma^2} + \frac{x^2}{2\sigma^2}\right) = \exp\left(\frac{2x-1}{2\sigma^2}\right)$$

Observe that for some $c \in \mathbb{R}$

$$L(x) = \exp\left(\frac{2x-1}{2\sigma^2}\right) > c \iff x > t$$

for some $t \in \mathbb{R}$ since $L(x)$ is monotonically increasing in t. Using this observation, we can see that the likelihood ratio test reduces to[a]:

1. reject H_0 if $x > t$
2. accept H_0 if $x \leq t$

Notice that this rule intuitively makes sense. Since H_1 corresponds to the hypothesis that the mean is larger, which means that a larger observe value intuitively corresponds to higher likelihood that $\mu = 1$.

With these values defined, we know that $\alpha = \mathbb{P}_{\mu=0}(X > t)$ and $\beta = \mathbb{P}_{\mu=1}(X < t)$. In setting up the optimization problem specified in (2), we see that we need to find that maximizes the following:

$$\max_t \ 1 - \Phi\left((t-1)/\sigma\right)$$
$$\text{s.t. } \ 1 - \Phi\left(t/\sigma\right) = z$$

We can then solve for the t value where $1 - \Phi\left(t/\sigma\right) = z$.

[a] the randomization variable γ is not needed here because $\mathbb{P}(X = t) = 0$ for all t when $X \sim \mathcal{N}(\mu, \sigma^2)$

Oftentimes when the likelihood ratio has non-zero probability on a single value i.e., $\mathbb{P}(L(x) = d) > 0$, γ may need to be calibrated. When X is discrete, the likelihood ratio has a discrete image, so this serves as a sufficient condition indicating when γ may need to be tuned. Below are two examples where randomization is needed, first for a discrete RV and then a continuous one.

Example 4: Discrete RV Hypothesis Test with randomization

Say you are picking a red, green or blue marble out of a jar. You are unsure what the probability distribution of marbles in the jar is, but you know it is one of two hypotheses:

- $H_0 : Y = \begin{cases} \text{red} & \text{w.p. } 0.2 \\ \text{blue} & \text{w.p. } 0.3 \\ \text{green} & \text{w.p. } 0.5 \end{cases}$
- $H_1 : Y = \begin{cases} \text{red} & \text{w.p. } 0.8 \\ \text{blue} & \text{w.p. } 0.1 \\ \text{green} & \text{w.p. } 0.1 \end{cases}$

Let the true jar be X and our prediction be $\hat{X}$. **Let's find a decision rule that maximizes** $P(\hat{X} = 1|X = 1)$ **while ensuring that** PFA $= P(\hat{X} = 1|X = 0) \leq 0.25$

We start by finding the likelihood ratio $L(Y) = \frac{P(Y|X=1)}{P(Y|X=0)}$, which is a piece-wise function of Y.

$$L(y) = \begin{cases} 4 & \text{if y = red} \\ \frac{1}{3} & \text{if y = blue} \\ \frac{1}{5} & \text{if y = green} \end{cases}$$

The Neyman-Pearson lemma tells us that the optimal decision rule will look like

$$\hat{X}(y) = \begin{cases} 1 & \text{if } L(y) > \lambda \\ Bern(\gamma) & \text{if } L(y) = \lambda \\ 0 & \text{L(y)} < \lambda \end{cases}$$

The next step is to choose the appropriate λ. We can choose from $\lambda = 4$, $\lambda = \frac{1}{3}$ or $\lambda = \frac{1}{5}$ because these are the three values that our $L(y)$ function takes. Any other choice of λ for the threshold of the hypothesis test is subsumed by one of these three choices. This is because the randomized decision rule kicks in only for one of these choices of λ. For example, see Figure 1 for $\lambda = 3.9$, which is strictly inferior to $\lambda = 4$ (with γ less than or equal to 1) and equivalent to $\lambda = 1/3$ (with $\gamma = 1$).

If we choose $\lambda = \frac{1}{3}$, with no randomization, $PFA = P(L(y) > \frac{1}{3}|X = 0) = P(Y = \text{red}|X = 0) = 0.2$.

If we choose $\lambda = \frac{1}{5}$, $PFA = P(L(y) > \frac{1}{5}|X = 0) = P(Y = \text{red, blue}|X = 0) = 0.5$.

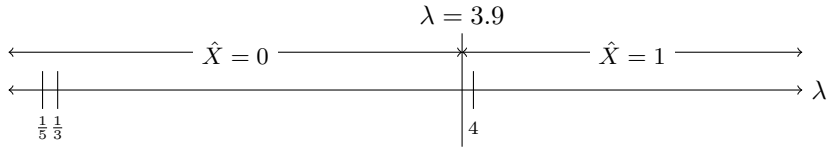
If PFA $= 0.2$ (with $\lambda = \frac{1}{3}$), we can still predict hypothesis 1 a bit more until achieving PFA $= 0.25$. If PFA $= 0.5$ (with $\lambda = \frac{1}{5}$), we are predicting hypothesis 1 too often. So instead, we include $\lambda = \frac{1}{5}$ (i.e. the extra case of $y =$ blue) with some probability γ, giving us a convex combination of the two possible decision rules. This way, we can achieve PFA $= 0.25 = 0.2 + 0.3\gamma \implies \gamma = \frac{1}{6}$.

Equivalently, we have found the following decision rule:

$$\hat{X}(y) = \begin{cases} 1 & \text{if } L(y) > \frac{1}{3} \\ Bern(\frac{1}{6}) & \text{if } L(y) = \frac{1}{3} \\ 0 & \text{else} \end{cases}$$

Stepping back, let's calculate the PCD. Under $\lambda = \frac{1}{3}$ with no randomization, PCD $= P(\hat{X} = 1|X = 1) = P(Y = \text{red}|X = 1) = 0.8$. With randomization, PCD $= P(\hat{X} = 1|X = 1) = P(Y = \text{red}|X = 1) + \gamma P(Y = \text{blue}|X = 1) = 0.8 + \frac{1}{6}0.1 = 0.817$. Moving down to the next threshold $\lambda = \frac{1}{5}$ ((too far for our case), PCD $= P(Y = \text{red, blue}|X = 1) = 0.8 + 0.1 = 0.9$. As we decrease the threshold λ, PCD increases. This is because we predict $X = 1$ more often. The tradeoff is that PFA increases too, limiting how low we can set the threshold λ.

Figure 1: Example decision rule with $\lambda = 3.9$



Example 5: Tuning γ

Suppose we are trying to determine whether X is Uniform$[-1, 1]$ or Uniform$[0, 2]$ for a given observation $x \sim X$ and a significance level z via a hypothesis test. Formally, our hypotheses are:

- $H_0: \ X \sim \text{Uniform}[-1, 1]$
- $H_1: \ X \sim \text{Uniform}[0, 2]$

Let's conduct a likelihood ratio test. The likelihood ratio is:

$$L(x) = \frac{f_{H_1}(x)}{f_{H_0}(x)} = \frac{\mathbb{1}\{0 \le x \le 2\}}{\mathbb{1}\{-1 \le x \le 1\}}$$

Upon using the Neyman-Pearson Lemma, we need to find the solution of (2) which reduces to finding a solution of:

$$\mathbb{P}_{H_0}(L(x) > c) + \gamma\, \mathbb{P}_{H_0}(L(x) = c) = z$$

Since the output of $L(x)$ is $\{0, 1, \infty\}$, we observe that we only need to consider $c = 0, c = 1, c = \infty$. So we can brute force our choice of c directly by evaluating $\mathbb{P}_{H_0}(L(x) > c)$ and $\mathbb{P}_{H_0}(L(x) = c)$.

We make the following observations:

1. Suppose $c = 0$, observe that $\mathbb{P}_{H_0}(L(x) = 0) = 1/2$ and $\mathbb{P}_{H_0}(L(x) > 0) = 1 - \mathbb{P}_{H_0}(L(x) = 0) = 1/2$. Thus, we get the equation $z = 1/2 + \gamma 1/2$

2. Suppose $c = 1$, observe that $\mathbb{P}_{H_0}(L(x) = 1) = 1/2$ and $\mathbb{P}_{H_0}(L(x) > 1) = 0$. Thus, we get the equation $z = 0 + \gamma 1/2$
3. Suppose $c = \infty$, observe that $\mathbb{P}_{H_0}(L(x) = \infty) = 0$ and $\mathbb{P}_{H_0}(L(x) > \infty) = 0$. Thus, we get the (useless) equation $z = 0$, which is not satisfiable for $z \neq 0$.

We observe that we get the following test in accordance to the procedure described in Definition 2:

1. If $0 \leq z \leq 1/2$, (using the second case above) we can set $c = 1$ and $\gamma = 2z$.
2. If $1/2 \leq z \leq 1$, (using the first case above) we can set $c = 0$ and $\gamma = 2(z - 1/2)$

Even though this is an optimal test, we can also disentangle the test from the likelihood ratio altogether by utilizing the following observations:

- $L(x) = 0 \iff x \in [-1, 0)$
- $L(x) = 1 \iff x \in [0, 1]$
- $L(x) = \infty \iff x \in (1, 2]$

Thus in substituting these if and only if conditions into the procedure described in 2, we get the equivalent tests:

1. If $0 \leq z \leq 1/2$:
   (a) accept H_0 if $x \in [-1, 0)$
   (b) reject H_0 with probability $\gamma = 2z$ if $x \in [0, 1]$.
   (c) reject H_0 if $x \in (1, 2]$
2. If $1/2 < z \leq 1$:
   (a) reject H_0 with probability $\gamma = 2(z - 1/2)$ if $x \in [-1, 0)$.
   (b) reject H_0 if $x \in [0, 2]$

Side note: It turns out that the second test for $\gamma > 0$ is "worse"[a] the second test for $\gamma = 0$ because there is no way that H_1 is true if $x \in [-1, 0)$. In other words the PCD of the second test with $\gamma = 0$ is already 1 (the maximum value), so making γ larger only serves to increase the PFA. For this reason, choices of $z > 1/2$ don't really make sense in this problem.

[a] its not worse in the Neyman-Pearson Sense as the PCD is maximized while the PFA is still below a threshold z. i.e., in the Neyman-Pearson notion of optimality, the second test with $\gamma > 0$ is still optimal even though there are clear shortcomings.