S&DS 242/542: Theory of Statistics

Lecture 8: Statistical power and the Neyman-Pearson lemma

Midterm exam logistics

Our midterm exam will take place on

Monday Feb 24, 7-9PM, YSB Marsh Auditorium

- It is a closed-book exam. You are allowed to bring 1 page of notes (front-and-back, standard letter or A4 size paper).
- The exam will cover material up to the end of lecture on Wed Feb 19, with a focus on Units 0 and 1 of our course.

If you have a conflict with the exam time or need alternative exam arrangements, please email our course manager Bella Bao:

`bella.bao@yale.edu`

Type I error and power

For testing a null hypothesis H_0 against an alternative H_1, recall

$$\mathbb{P}[\text{Type I error}] = \mathbb{P}_{H_0}[\text{reject } H_0]$$

A test with significance level α guarantees that

$$\mathbb{P}[\text{Type I error}] \leq \alpha$$

Among several different level-α tests of the same hypotheses, we may prefer the test that maximizes

$$\text{Power} = \mathbb{P}_{H_1}[\text{reject } H_0]$$

Q: Given two arbitrary hypotheses H_0 and H_1, is there an optimal test that maximizes power, among all possible level-α tests?

Simple and composite hypotheses

We will see that the answer to this question is generally "yes" if both hypotheses H_0 and H_1 are simple.

H_0 or H_1 is **simple** if it describes a *single* distribution for the data — there are no unknown parameters or other missing information about the distribution. Otherwise, the hypothesis is **composite**.

A simple hypothesis provides all the information that would be needed to *simulate* the data. A composite hypothesis requires some further specification of the data distribution in order to perform a simulation.

Simple and composite hypotheses

Example: The null and alternative hypotheses

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

are both simple. The null hypotheses

$$H_0 : X_1, \ldots, X_n \overset{IID}{\sim} \mathcal{N}(0, \sigma^2) \text{ for some (unknown) } \sigma^2 > 0$$
$$H_0 : X_1, \ldots, X_n \text{ are IID from a distribution with mean } 0$$

are both composite. The alternative hypothesis

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

is also composite.

A simple vs. simple testing example

We observe a single value $X \in \{1, \ldots, 5\}$, sampled from one of two discrete distributions:

x	1	2	3	4	5
$f_0(x)$	0.2	0.2	0.2	0.2	0.2
$f_1(x)$	0.0	0.1	0.2	0.3	0.4

We wish to test

$$H_0 : X \sim f_0 \quad \text{vs.} \quad H_1 : X \sim f_1$$

at the significance level $\alpha = 0.4$. What is the test based on the observation X that would maximize power against H_1?

A simple vs. simple testing example

x	1	2	3	4	5
$f_0(x)$	0.2	0.2	0.2	0.2	0.2
$f_1(x)$	0.0	0.1	0.2	0.3	0.4

To ensure

$$\mathbb{P}[\text{Type I error}] \leq \alpha = 0.4$$

we are allowed to reject H_0 for two possible values of X, because each value has probability 0.2 under H_0.

To maximize the power against H_1, we want to pick the two values that have maximum probability under H_1: These are 4 and 5. So the most powerful test at level $\alpha = 0.4$ would reject H_0 if $X \in \{4, 5\}$ and accept H_0 if $X \in \{1, 2, 3\}$.

Testing as constrained optimization

When designing an optimal test of H_0 vs. H_1, we have the following goal:

maximize: power of the test against H_1

subject to: probability of Type I error under H_0 is $\leq \alpha$

This is a constrained optimization problem.

Suppose we observe random data $\mathbf{X} = (X_1, \ldots, X_n)$, taking possible values denoted $\mathbf{x} = (x_1, \ldots, x_n)$. To define a test, we must decide, for each possible value $\mathbf{x}$, whether to accept or reject H_0 if we observe $\mathbf{X} = \mathbf{x}$.

I.e., we must define the set of values $\mathbf{x}$ that belong to the *acceptance* and *rejection* regions of the test.

The likelihood ratio test

Suppose the distribution of $\mathbf{X}$ is discrete, and the hypotheses are

$$H_0 : \mathbf{X} \text{ is distributed with (joint) PMF } f_0(\mathbf{x})$$
$$H_1 : \mathbf{X} \text{ is distributed with (joint) PMF } f_1(\mathbf{x})$$

Which values $\mathbf{x}$ should we include in the rejection region?

Intuition suggests to reject H_0 for those points $\mathbf{x}$ with largest values of

$$\frac{f_1(\mathbf{x})}{f_0(\mathbf{x})}$$

because these give the "largest increase in power per unit increase of Type I error". Alternatively, these provide the "strongest evidence" in favor of H_1 over H_0.

The likelihood ratio test

The case of continuous $\mathbf{X}$ is similar: Suppose the hypotheses are

$$H_0 : \mathbf{X} \text{ is distributed with (joint) PDF } f_0(\mathbf{x})$$
$$H_1 : \mathbf{X} \text{ is distributed with (joint) PDF } f_1(\mathbf{x})$$

Intuition suggests to reject H_0 for those points $\mathbf{x}$ with largest values of

$$\frac{f_1(\mathbf{x})}{f_0(\mathbf{x})}$$

In both the discrete and continuous settings, the test statistic

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

is called the **likelihood ratio statistic**. The test that rejects H_0 in favor of H_1 for large ~~$T(\mathbf{X})$~~ is the **likelihood ratio test**.

$L(x)$

The Neyman-Pearson lemma

For testing a simple null hypothesis versus a simple alternative, the Neyman-Pearson lemma guarantees that the likelihood ratio test is the *most powerful test*.

Theorem (Neyman-Pearson lemma)

Let H_0 and H_1 be simple hypotheses, and fix a significance level $\alpha \in (0,1)$. Suppose there exists a value $c > 0$ such that the likelihood ratio test which

$$\begin{cases} \textit{rejects } H_0 \textit{ if } L(\mathbf{X}) > c \\ \textit{accepts } H_0 \textit{ if } L(\mathbf{X}) \leq c \end{cases}$$

has Type I error probability exactly equal to α.

Then for any other test with probability of Type I error $\leq \alpha$, its power against H_1 is at most the power of this likelihood ratio test.

Proof of the Neyman-Pearson lemma

Consider the discrete case. Let

$$\mathcal{R} = \{\mathbf{x} : L(\mathbf{x}) > c\} = \{\mathbf{x} : f_1(\mathbf{x}) > cf_0(\mathbf{x})\}$$

be the rejection region of the likelihood ratio test.

Among all possible rejection regions, this set $\mathcal{R}$ maximizes

$$\sum_{\mathbf{x}\in\mathcal{R}} (f_1(\mathbf{x}) - cf_0(\mathbf{x}))$$

because $f_1(\mathbf{x}) - cf_0(\mathbf{x}) > 0$ for $\mathbf{x} \in \mathcal{R}$ and $f_1(\mathbf{x}) - cf_0(\mathbf{x}) \leq 0$ for $\mathbf{x} \notin \mathcal{R}$. Then for any test, say with rejection region $\mathcal{R}'$,

$$\sum_{\mathbf{x}\in\mathcal{R}} (f_1(\mathbf{x}) - cf_0(\mathbf{x})) \geq \sum_{\mathbf{x}\in\mathcal{R}'} (f_1(\mathbf{x}) - cf_0(\mathbf{x})) \,.$$

Proof of the Neyman-Pearson lemma

Rearranging this inequality,

$$\underbrace{\sum_{\mathbf{x}\in\mathcal{R}} f_1(\mathbf{x}) - \sum_{\mathbf{x}\in\mathcal{R}'} f_1(\mathbf{x})}_{\text{difference in power}} \geq c \underbrace{\left(\sum_{\mathbf{x}\in\mathcal{R}} f_0(\mathbf{x}) - \sum_{\mathbf{x}\in\mathcal{R}'} f_0(\mathbf{x})\right)}_{\text{difference in probability of Type I error}}$$

If the likelihood ratio test (with rejection region $\mathcal{R}$) has Type I error probability α, and the other test (with rejection region $\mathcal{R}'$) has Type I error probability $\leq \alpha$, then

$$\text{difference in probability of Type I error} \geq 0$$

So this implies

$$\text{difference in power} \geq 0$$

i.e. power of likelihood ratio test $\geq$ power of the other test. The continuous case is the same, with all sums replaced by integrals.

Testing a normal mean

Example: Consider data $\mathbf{X} = (X_1, \ldots, X_n)$, and a test of

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

Assume that $\mu > 0$ is a *known and pre-specified* value, so both H_0 and H_1 are simple hypotheses. Let's derive the form of $L(\mathbf{X})$:

Under H_0: $f_0(\mathbf{x}) = \prod_{i=1}^{n} f_0(x_i) = \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi}} e^{-\frac{x_i^2}{2}} = \left(\frac{1}{\sqrt{2\pi}}\right)^n e^{-\frac{1}{2}(x_1^2 + \ldots + x_n^2)}$

Under H_1: $f_1(\mathbf{x}) = \prod_{i=1}^{n} f_1(x_i) = \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi}} e^{-\frac{(x_i-\mu)^2}{2}}$

$= \left(\frac{1}{\sqrt{2\pi}}\right)^n e^{-\frac{1}{2}\left[(x_1-\mu)^2 + \ldots + (x_n-\mu)^2\right]}$

Testing a normal mean

$$L(\mathbf{x}) = \frac{f_1(\mathbf{x})}{f_0(\mathbf{x})} = \frac{e^{-\frac{1}{2}[(x_1-\mu)^2+\dots+(x_n-\mu)^2]}}{e^{-\frac{1}{2}[x_1^2+\dots+x_n^2]}}$$

$$= \exp\left(-\frac{1}{2}[(x_1-\mu)^2+\dots+(x_n-\mu)^2]+\frac{1}{2}[x_1^2+\dots+x_n^2]\right)$$

$$= \exp\Big(-\frac{1}{2}[x_1^2-2\mu x_1+\mu^2+\dots+x_n^2-2\mu x_n+\mu^2]$$
$$+\frac{1}{2}[x_1^2+\dots+x_n^2]\Big)$$

$$= \exp\Big(\mu x_1+\mu x_2+\dots+\mu x_n \underbrace{-\frac{\mu^2}{2}-\dots-\frac{\mu^2}{2}}_{n}\Big)$$

$$= \exp\left(\mu(x_1+\dots+x_n)-\frac{n}{2}\mu^2\right)$$

Testing a normal mean

The Neyman-Pearson lemma ensures that the most powerful test is the test which rejects H_0 when $L(\mathbf{X}) > c$, where c is chosen so that

$$\mathbb{P}[\text{Type I error}] = \mathbb{P}_{H_0}[L(\mathbf{X}) > c] = \alpha$$

Thus c is the upper-α point of the distribution of $L(\mathbf{X})$ under H_0.

Observe that, for $\mu > 0$, the statistic

$$L(\mathbf{X}) = e^{\mu(X_1+\ldots+X_n)-\frac{n\mu^2}{2}}$$

depends on $\mathbf{X}$ only via the sample mean $\bar{X} = \frac{1}{n}(X_1 + \ldots + X_n)$. Furthermore, $L(\mathbf{X})$ is an *increasing* function of $\bar{X}$.

Testing a normal mean

Because $L(\mathbf{X})$ is increasing in $\bar{X}$, the rejection event

$$L(\mathbf{X}) > \text{upper-}\alpha \text{ point of the null distribution of } L(\mathbf{X})$$

is exactly the same as the rejection event

$$\sqrt{n}\,\bar{X} > \text{upper-}\alpha \text{ point of the null distribution of } \sqrt{n}\,\bar{X}$$

Under H_0, recall $Z = \sqrt{n}\,\bar{X} \sim \mathcal{N}(0,1)$, with upper-$\alpha$ point $z^{(\alpha)}$.

Thus the Neyman-Pearson lemma implies that the most powerful test is exactly the z-test, which rejects H_0 when $Z > z^{(\alpha)}$.

Testing a normal mean

- The form of this most powerful test is the same against any simple alternative with known and pre-specified mean $\mu > 0$. Thus this z-test is *uniformly most powerful* against the compositive alternative $H_1 : \mu > 0$ when μ is unknown.
- If we specify an alternative $\mu < 0$, then $L(\mathbf{X}) = e^{\mu(X_1+\ldots+X_n)-\frac{n\mu^2}{2}}$ is *decreasing* in $\bar{X}$. So

$$L(\mathbf{X}) > \text{upper-}\alpha \text{ point of the null distribution of } L(\mathbf{X})$$
$$\Updownarrow$$
$$\sqrt{n}\,\bar{X} < \text{lower-}\alpha \text{ point of the null distribution of } \sqrt{n}\,\bar{X}$$

  The most powerful test would reject H_0 for *small* values of Z.
- There is no single test that is uniformly most powerful against both positive and negative alternatives, because the most powerful test in each case rejects H_0 for different values of Z.

Testing if a coin is fair

Example: Let $X_1, \ldots, X_n \in \{0, 1\}$, and consider testing

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

Assume $p > \frac{1}{2}$ is a known and pre-specified value, so both hypotheses are simple. Let's derive the form of $L(\mathbf{X})$:

Under H_0: $f_0(\mathbf{x}) = \prod_{i=1}^{n} f_0(x_i) = \prod_{i=1}^{n} \frac{1}{2} = \frac{1}{2^n}$

Under H_1: $f_1(\mathbf{x}) = \prod_{i=1}^{n} f_1(x_i) = \prod_{i=1}^{n} p^{x_i}(1-p)^{1-x_i}$

$= p^{x_1 + \ldots + x_n}(1-p)^{1-x_1+1-x_2+\ldots+1-x_n}$

$= (1-p)^n \left(\frac{p}{1-p}\right)^{x_1+\ldots+x_n}$

$\Rightarrow L(\mathbf{x}) = \frac{f_1(\mathbf{x})}{f_0(\mathbf{x})} = 2^n (1-p)^n \left(\frac{p}{1-p}\right)^{x_1+\ldots+x_n}$

Testing if a coin is fair

The Neyman-Pearson lemma ensures that the most powerful test is the test which rejects H_0 when $L(\mathbf{X}) > c$, where c is chosen so that

$$\mathbb{P}[\text{Type I error}] = \mathbb{P}_{H_0}[L(\mathbf{X}) > c] = \alpha$$

Thus c is the upper-α point of the distribution of $L(\mathbf{X})$ under H_0. Here, for any fixed $p > \frac{1}{2}$,

$$L(\mathbf{X}) = 2^n(1-p)^n(\tfrac{p}{1-p})^{X_1+\ldots+X_n}$$

is *increasing* in $S = X_1 + \ldots + X_n$. Under H_0, $S \sim \text{Binomial}(n, \frac{1}{2})$. So equivalently, the most powerful test rejects H_0 when

$$S > b_n^{(\alpha)} \quad \text{the "upper-}\alpha\text{ point" of Binomial}(n, \tfrac{1}{2})$$

Test statistics with discrete distributions

In this case, both $S = X_1 + \ldots + X_n$ and $L(\mathbf{X})$ have discrete null distributions. There may not exist a value of c for which

$$\mathbb{P}_{H_0}[L(\mathbf{X}) > c] = \alpha$$

exactly, i.e. there may not exist a value $b_n^{(\alpha)}$ for which

$$\mathbb{P}_{H_0}[S > b_n^{(\alpha)}] = \alpha$$

Example: Suppose $n = 20$. For $S \sim \text{Binomial}(20, \frac{1}{2})$, we have $\mathbb{P}[S > 14] = 0.021$ and $\mathbb{P}[S > 13] = 0.058$. We cannot perform this test to attain Type I error probability exactly $\alpha = 0.05$.

A level-α test would be conservative and reject H_0 when $S > 14$. The Neyman-Pearson lemma would not guarantee that this is most powerful, although we would usually go with this test in practice.

Beyond the Neyman-Pearson lemma

If there is a single test that maximizes power, why do we still have so many different testing procedures?

- Alternative hypotheses H_1 in practice are oftentimes not simple, and we may wish to balance power against different types of alternatives.
- Null hypotheses H_0 in practice are sometimes not simple, and we may wish to restrict to test statistics that are pivotal under broad specifications of H_0.
- We may be unsure about a specific data model for H_0 and prefer to sacrifice some power to achieve greater robustness against misspecification of the null model.