Hypothesis testing (part 1)

Lecture 15a (STAT 24400 F24)

Hypothesis testing (definiton)

Common frame work: suppose our data is drawn from a parametric model,

$$f(\cdot \mid \theta) \quad \text{for some } \theta \in \Theta$$

A **hypothesis test** uses the observed data to choose between two possible statements about θ, e.g.,

- Test $H_0 : \theta = 1$ versus $H_1 : \theta = 2$ $\leftarrow$ H_0 & H_1 are simple
- Test $H_0 : \theta = 1$ versus $H_1 : \theta \neq 1$ $\leftarrow$ H_0 is simple, H_1 is composite
- Test $H_0 : \theta \leq 1$ versus $H_1 : \theta > 1$ $\leftarrow$ H_0 & H_1 are composite

A hypothesis is **simple** if it specifies the distribution exactly
(i.e., a single specific value of θ).

Otherwise it is **composite**.

Hypothesis testing (terminology)

Some terminology:

- H_0 is called the **null hypothesis**
- H_1 is called the **alternative hypothesis** (sometimes written as H_A or H_a)
- A hypothesis test is a function
  mapping observed data to a selection (H_0 or H_1)

- **Type I error** = the probability of selecting H_1, if H_0 is true
  (sometimes called the "**level**" of the test)
- **Type II error** = the probability of selecting H_0, if H_1 is true
- **Power** = prob. of selecting H_1, if H_1 is true = $1 -$ Type II error

Hypothesis testing (type I & type II errors)

Decision	H_0 is true	H_0 is false
"Reject H_0"	Type I error (α)	correct (Power $= 1 - \beta$)
"Do not reject H_0"	correct	Type II error (β)

Common notation:

Type I error $= \alpha$, Type II error $= \beta$, thus power $= 1$ - β.

Hypothesis testing (Poisson example)

Example: The output of an X-ray beam follows a Poisson(λ) distribution.

The intensity parameter λ can be set to 100, 110, 120, or 130.

$$H_0 : \lambda = 100 \qquad vs. \qquad H_1 : \lambda \in \{110, 120, 130\}$$

A possible hypothesis test:

— If X is in the range 84–117 then we *do not reject* the null H_0 (i.e., H_0 is plausible, so we choose H_0)

— If X is not in the range 84–117, then we *reject* the null H_0 (i.e., H_0 is not plausible, so we choose H_1)

Hypothesis testing(Poisson example)

What are the Type I and Type II errors of the test?

assuming H_0, i.e., $X \sim$ Poisson(100)

- Type I error $= \mathbb{P}(\textit{reject } H_0 \mid H_0 \textit{ true}) = \mathbb{P}_{H_0}(X \notin [84, 117])$

  $= 1 - \sum_{k=84}^{117} \frac{100^k e^{-100}}{k!} = 0.089$ ($\neq$ 0.1 due to discreteness)

- Type II error $= \mathbb{P}(\textit{accept } H_0 \mid H_1 \textit{ true}) = \mathbb{P}_{H_1}(X \in [84, 117])$

  Type II error depends on H_1, that is, depends on λ, e.g.,

  — If $\lambda = 110$, Type II error $= \mathbb{P}_{\lambda=110}(X \in [84, 117]) = \sum_{k=84}^{117} \frac{110^k e^{-110}}{k!} = 0.761$

  — If $\lambda = 130$, Type II error $= \mathbb{P}_{\lambda=130}(X \in [84, 117]) = \sum_{k=84}^{117} \frac{130^k e^{-130}}{k!} = 0.136$

Hypothesis testing(binomial example)

Example: flip a coin 100 times. Is the coin fair?

- $X \sim$ Binomial$(100, p)$. H_0: $p = 0.5$, $\quad H_1$: $p \neq 0.5$
- A possible hypothesis test: if $45 \leq X \leq 55$, choose H_0; else, choose H_1

  (in practice, we decide the error rate first then design the test so the decision rule will have the error rate.)

- Type I error $=$

  $$\mathbb{P}_{H_0}(X < 45 \text{ or } X > 55) = \sum_{k=0}^{44} \binom{100}{k} 0.5^k 0.5^{100-k} + \sum_{k=56}^{100} \binom{100}{k} 0.5^k 0.5^{100-k} = 0.271$$

- Type II error — depends on the value of p. For example, if $p = 0.6$,

  $$\mathbb{P}_{H_1}(45 \leq X \leq 55) = \sum_{k=45}^{55} \binom{100}{k} 0.6^k 0.4^{100-k} = 0.178$$

- Power — depends on the value of p. For example, if $p = 0.6$,

  $$\text{Power} = 1 - \text{Type II error} = 0.822$$

  This is the "power against the alternative $p = 0.6$"

Hypothesis testing(binomial example)

How do we choose which hypothesis to label as H_0 / as H_1?

Some conventions:

- If one hypothesis is simple & the other is composite, choose H_0 as the simple one
- If one hypothesis is the one we'd like to prove is likely true, label it as H_1 (because we will try to *reject* the null H_0)
- Possible conclusions based on the evidence from the data:
  - "Reject the H_0" (thus accept H_1)
  - "Do not reject the H_0" (not "accepting the H_0"; subtle importance)

Beyond the parametric setting

In some cases, the hypotheses may not lie in a parametric family.

Examples in the setting where $X_1, \dots, X_n$ are i.i.d. from some distribution:

- H_0: the distrib. is Exponential(λ) for some λ,
  versus H_1: the distrib. is not exponential
  (goodness-of-fit test)
- H_0: the mean of the distribution is 0,
  versus H_1: the mean is $\neq 0$

Another common example—for pairs (X_i, Y_i) i.i.d. from a joint distribution:

- H_0: X & Y are independent,
  versus H_1: X & Y are not independent

Testing two simple hypotheses

Assume the data comes from a parametric family, $f(x \mid \theta)$,
and we are testing

$$H_0 : \ \theta = \theta_0 \qquad \text{vs.} \qquad H_1 : \ \theta = \theta_1$$

How should we decide which is more likely?

One way is to compare their likelihoods.

For a single draw of the data, $X \sim f(\cdot \mid \theta)$:

$$\text{Likelihood of } \theta_0 = f(X \mid \theta_0) \quad \text{vs.} \quad \text{Likelihood of } \theta_1 = f(X \mid \theta_1)$$

For n data points, $X_1, \dots, X_n \stackrel{\text{iid}}{\sim} f(\cdot \mid \theta)$:

$$\underbrace{\text{Likelihood of } \theta_0 = \prod_{i=1}^n f(X_i \mid \theta_0)}_{\text{written as } f_0(\mathbf{X}) \text{ in textbook}} \quad \text{vs.} \quad \underbrace{\text{Likelihood of } \theta_1 = \prod_{i=1}^n f(X_i \mid \theta_1)}_{\text{written as } f_1(\mathbf{X}) \text{ in textbook}}$$

The likelihood ratio test (LRT)

Intuition:

- Higher values of likelihood of $\theta_0 \longleftrightarrow H_0$ seems more plausible
- Higher values of likelihood of $\theta_1 \longleftrightarrow H_1$ seems more plausible

Performing a **likelihood ratio test** (LRT) means that we will
make our decision (H_0 or H_1) based solely on the likelihood ratio

$$\text{LR} = \frac{\text{Likelihood of } \theta_0}{\text{Likelihood of } \theta_1}$$

We will need to set some *threshold* c:

$$\begin{cases} \text{If LR} > c \text{ then choose } H_0 \\ \text{If LR} \leq c \text{ then choose } H_1 \end{cases}$$

(Or use $\geq c$ and $< c$, which may be different in the discrete setting.)

The Neyman–Pearson lemma

For testing H_0 versus H_1 when both hypotheses are simple,
a LR test is the *best* possible test.

Neyman–Pearson lemma:

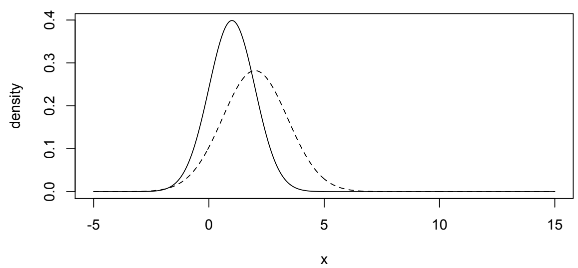
Suppose H_0 and H_1 are simple hypotheses, and fix any $c \geq 0$.
Let $\alpha, \beta =$ Type I error, Type II error for the LR test with threshold c.

Then for any other test of H_0 versus H_1,
if Type I error $= \alpha$ then Type II error $\geq \beta$.

Equivalently:
For any other test of H_0 versus H_1,
if Type II error $= \beta$ then Type I error $\geq \alpha$.

Example: two normal distributions

$$H_0: \; X \sim \mathrm{N}(1,1) \quad \text{versus} \quad H_1: \; X \sim \mathrm{N}(2,2)$$



- Test #1: choose a cutoff $1 < a < 2$ (in between the two means), & if $X < a$ then return H_0, otherwise if $X \geq a$ then return H_1.
- Test #2: LR test — choose a threshold $c > 0$, & if $\mathrm{LR} > c$ then return H_0, otherwise if $\mathrm{LR} \leq c$ then return H_1.

Example: two normal distributions

Implementing the LR test:

$$\mathrm{LR} = \frac{\frac{1}{\sqrt{2\pi\cdot 1}} e^{-(x-1)^2/2\cdot 1}}{\frac{1}{\sqrt{2\pi\cdot 2}} e^{-(x-2)^2/2\cdot 2}} = \sqrt{2} e^{(x-2)^2/4-(x-1)^2/2} = \sqrt{2e}\cdot e^{-x^2/4}$$

If we choose threshold $c = 1.3$:

$$\text{Choose } H_0 \quad \Leftrightarrow \quad \mathrm{LR} > 1.3 \quad \Leftrightarrow \quad |x| < 2\sqrt{\log\left(\frac{\sqrt{2e}}{1.3}\right)} = 1.529$$

$$\text{Type I error} = \mathbb{P}_{\mathrm{N}(1,1)}(\mathrm{LR} \leq 1.3) = \mathbb{P}_{\mathrm{N}(1,1)}(|X| \geq 1.529) = 0.3042$$

$$\text{Type II error} = \mathbb{P}_{\mathrm{N}(2,2)}(\mathrm{LR} > 1.3) = \mathbb{P}_{\mathrm{N}(2,2)}(|X| < 1.529) = 0.3632$$

Example: two normal distributions

Now compare against Test #1.
Suppose we choose cutoff a to get Type I error $= 0.3042$: (to match LR's)

$$0.3042 = \text{Type I error} = \mathbb{P}_{\mathrm{N}(1,1)}(X \geq a) \quad \rightsquigarrow \quad a = 1.5122$$

(same as the LR test with $c = 1.3$)

Then calculate Type II error:

$$\text{Type II error} = \mathbb{P}_{\mathrm{N}(2,2)}(X < 1.5122) = \mathbb{P}_{\mathrm{N}(2,2)}(X < 1.5122) = 0.3651$$

(higher than the Type II error of the LR test with $c = 1.3$)

Illustration of the Neyman–Pearson lemma

The Neyman–Pearson lemma, for this example: