STAT 611-600

Theory of Inference
Lecture 10: Hypothesis Testing

Tiandong Wang

March 11, 2021

Hypothesis Testing

Statistical hypothesis: A conjecture about the distribution of a random variable or random vector.

- Often can be phrased as a conjecture about the *true* population distribution.
- Often involves comparisons:
  - New manufacturing method reduces variability relative to old method.
  - New species of seed increases yield relative to old species.
  - New drug is more effective than old one or more effective than a placebo.
  - Product A is superior to product B.

Example of hypotheses:

- MANUFACTURING: Before modification of the production process the production produces bolts with diameter

$$\text{diam} = \frac{3}{8} + \epsilon_{\text{before}}$$

where

$$\epsilon_{\text{before}} \sim N(0, \sigma^2_{\text{before}}).$$

After modifying the process the diameters of produced bolts has

$$\text{diam} = \frac{3}{8} + \epsilon_{\text{after}}$$

where

$$\epsilon_{\text{after}} \sim N(0, \sigma^2_{\text{after}}).$$

We hope the variance has been reduced so we test the hypothesis that

$$\frac{\sigma^2_{\text{after}}}{\sigma^2_{\text{before}}} < 1.$$

More formal setup

Formulate a hypothesis in statistical terms. The usual framework is that there is a population with a population density (or pmf) $f(x;\theta)$ and we seek to decide if the true θ falls in one subset or another. Thus we have a statistical family

$$\big(S, \mathcal{A}, \{\mathbb{P}_\theta(\cdot), \theta \in \Theta\}\big).$$

Usually the probabilities $\mathbb{P}_\theta(\cdot)$ are specified by a density or pmf

$$f(\boldsymbol{x};\theta), \theta \in \Theta$$

and we seek to decide between two hypotheses, the *null* hypothesis and the *alternative* hypothesis

$$H_0 : \theta \in \Theta_0 \quad vs \quad H_a : \theta \in \Theta_a$$

where

$$\Theta_0 \bigcup \Theta_a \subset \Theta.$$

Different Types of Hypotheses

- (1) *simple* hypothesis: both H_0 and H_1 consist of only one probability distribution.

$$H_0 : \theta = \theta_0 \quad vs \quad H_1 : \theta = \theta_1$$

- (2) *composite* hypotheses: either H_0 or H_1 has more than one distribution:
  - $$H_0 : \theta \geq \theta_0 \quad vs \quad H_1 : \theta < \theta_0$$
  - $$H_0 : \theta \leq \theta_0 \quad vs \quad H_1 : \theta > \theta_0$$
  - $$H_0 : \theta = \theta_0 \quad vs \quad H_1 : \theta \neq \theta_0$$

Examples.

- Based on a random sample $X_1, \ldots, X_n \sim N(\mu, 25)$ decide if

$$H_0 : \mu = 0 \text{ or } \mu > 0.$$

Here

$$\Theta = \mathbb{R}, \quad \Theta_0 = \{0\}, \quad \Theta_a = (0, \infty).$$

- In a binomial population we observe $X_1, \ldots, X_n$, a random sample of successes (have the characteristic) or failures (characteristic absent). Test

$$H_0 : p = \frac{1}{2} \quad vs \quad H_a : p > \frac{1}{2}.$$

Note $p = \frac{1}{2}$ might correspond to guessing a result completely at random as opposed to having skill or efficacy.

Consumer Reports blind taste tests: Every few years, Consumers Reports tests whether New York City residents can distinguish between (free) NYC tap water and (not free) bottled water. Subjects are randomly given a glass of each without identifier. They are then asked to identify which they prefer.

A About 50% preferred NYC tap and 50% preferred bottled water. (What does *about* mean?)

B More than 50% preferred NYC tap water. (What does *more than* mean?)

C More than 50% preferred bottled water.

If [B] or [C] is true, how much more than 50% is required in order to be able to assert the superiority of either tap water or bottled water?

Alternate experiment: Can consumers distinguish between Coke and Pepsi?

Types of possible errors in Hypothesis Testing:

		Types of Errors	
		H_0 true	H_0 false
Decision	Accept H_0	No error	Type II error
	Reject H_0	Type I error	No error

Note:

- The type I error (false rejection of H_0) is regarded as the more serious and we control for this error as a priority.
- This means the wise and experienced experimenter must frame the hypotheses so that rejecting H_0 falsely is more serious than accepting H_a incorrectly (type II) error.
- Typically the hypothesis test is designed so we control for type I error as a priority irrespective of sample size; then type II error is controlled by picking a sufficiently large sample size.
- Easier to remember? We want:

$$P_\theta[\text{reject } H_0] = \begin{cases} \text{small as possible} & \theta \in \Theta_0 \quad (\text{ i.e. } H_0 \text{ true}), \\ \text{big as possible} & \theta \in \Theta_1 \quad (\text{ i.e. } H_0 \text{ false}). \end{cases}$$

		Types of Errors	
		H_0 true	H_0 false
Decision	Accept H_0	No error	Type II error
	Reject H_0	Type I error	No error

Illustrations:

1. Jurisprudence: Consider two hypotheses:

$$H_0: \text{ defendent innocent vs } H_a: \text{ defendent guilty.}$$

The 2-error probabilities are not symmetric in importance:

type I error: False rejection of H_0 means we reject innocence falsely; an innocent person is convicted.

type II error: false acceptance of H_0 means a guilty person is declared innocent.

Civilization regards the type I error as much more serious.

		Types of Errors	
		H_0 true	H_0 false
Decision	Accept H_0	No error	Type II error
	Reject H_0	Type I error	No error

2. Clinical trials:

H_0 : Drug **not** [safe and effective] vs H_a : Drug **is** [safe and effective].

type I error: Reject falsely: Say a drug [safe and effective] when this was false;
either claim efficacy or claim safety falsely.

type II error: false acceptance: Say drug has no effect when it has.
(Conventional wisdom pre-aids: Nobody dies; the only harm is to the drug company.)

Characteristics of H_0 and H_a

Associate with the two hypotheses the characteristics

H_0: status quo, no change, no effect, no skill (guessing)
drug not [safe and effective] (Drug company hopes to reject H_0)

H_a: change in status quo;
dramatic new research findings;
continue to believe H_0 unless
there is strong evidence in favor of H_a.

Steps to formulating a hypothesis test.

Based on observing $X_1, \ldots, X_n$, a random sample from

$$f(x;\theta), \theta \in \Theta,$$

1. Formulate hypotheses

   $$H_0 : \theta \in \Theta_0 \text{ vs } H_a : \theta \in \Theta_a,$$

   where

   $$\Theta_0 \bigcap \Theta_a, \qquad \Theta_0 \bigcup \Theta_a \subset \Theta.$$

   EXAMPLE:

   $$H_0 : p = \frac{1}{2} \text{ vs } H_a : p > \frac{1}{2},$$
   $$H_0 : \mu = 3 \text{ vs } H_a : \mu \neq 3.$$

2. Conventional historical approach: Decide on a *level of significance* α (frequently choose $\alpha = .05$ or 0.01) such that

   $$\sup_{\theta \in \Theta_0} \mathbb{P}_\theta[\text{ reject }] = \mathbb{P}[\text{ reject} \mid \text{should accept }] = \mathbb{P}[\text{ type I error }] \leq \alpha.$$

3. Find a *test statistic*

$$T = T(X_1, \dots, X_n),$$

a function of $X_1, \dots, X_n$ and a rejection region R (a subset of the range of T) and agree that rejection means

$$T \in R.$$

Usually R depends on the significance level α.

3. Find a *test statistic*

$$T = T(X_1, \ldots, X_n),$$

a function of $X_1, \ldots, X_n$ and a rejection region R (a subset of the range of T) and agree that rejection means

$$T \in R.$$

Usually R depends on the significance level α.

EXAMPLE: Test $H_0 : p = \frac{1}{2}$ by agreeing to reject if

$$\hat{p} > .7$$

So $T = \hat{p}$ and $R = (.7, 1]$ are chosen so that $T \in R$ gives strong evidence against H_0 and for H_a. □

The rejection region R must be selected so that if the test statistic falls in R, this is strong evidence against H_0. In more detail:

$$\begin{aligned}\mathbb{P}_\theta[\text{ reject } H_0 \text{]} =& \mathbb{P}_\theta[T \in R] = \mathbb{P}_\theta[\text{ false rejection]} \\ \leq& \alpha, \quad \text{ for all } \theta \in \Theta_0,\end{aligned}$$

and

$$\begin{aligned}\mathbb{P}_\theta[\text{ reject } H_0 \text{]} =& \mathbb{P}_\theta[T \in R] = \mathbb{P}_\theta[\text{ correct rejection]} \\ & \text{ is as large as possible, for all } \theta \in \Theta_a;\end{aligned}$$

or equivalently

$$\begin{aligned}\mathbb{P}_\theta[\text{ accept } H_0] =& \mathbb{P}_\theta[T \in R^c] \\ =& \mathbb{P}[\text{ false acceptance of } H_0 \text{]} \\ & \text{ is as small as possible, for all } \theta \in \Theta_a.\end{aligned}$$

4. Collect data $x_1, \ldots, x_n$, evaluate $T(x_1, \ldots, x_n)$. If

$$T(x_1, \ldots, x_n) \in R$$

reject H_0. Otherwise, *fail to reject at level* α or announce *there is no evidence against the null hypothesis*.

4. Collect data $x_1, \ldots, x_n$, evaluate $T(x_1, \ldots, x_n)$. If

$$T(x_1, \ldots, x_n) \in R$$

reject H_0. Otherwise, *fail to reject at level* α or announce *there is no evidence against the null hypothesis*.

Summary of the traditional setup For testing H_0 vs H_a,

- Choose a level of significance α; say $\alpha = .05$ or 0.01.
- Find a test statistic T and a rejection region R such that

$$\begin{aligned} &\text{for all } \theta \in \Theta_0: \quad && \pi(\theta) := \mathbb{P}_\theta[T \in R] = \mathbb{P}_\theta[\text{ reject }] \leq \alpha; \\ &\text{for all } \theta \in \Theta_a: \quad && \pi(\theta) := 1 - \beta(\theta) := \mathbb{P}_\theta[T \in R] = \mathbb{P}_\theta[\text{ reject }] \\ & && \qquad\qquad \text{is as big as possible .} \end{aligned}$$

$\pi(\theta)$ is called the power function.

Definitions:

1. For $\alpha \in [0,1]$, a test with power function $\pi(\theta)$ is a **size** α test if
$$\sup_{\theta \in \Theta_0} \pi(\theta) = \alpha.$$
2. For $\alpha \in [0,1]$, a test with power function $\pi(\theta)$ is a **level** α test if
$$\sup_{\theta \in \Theta_0} \pi(\theta) \leq \alpha.$$

Examples

1. Let

$$p = \text{proportion of the population in a small town}$$
$$\text{that has an advanced degree.}$$

Setup:

- Sample size $n = 15$.
- X = number of degree holders in sample.
- Test

$$H_0 : p = .3 \quad \text{vs} \quad H_a : p \neq .3.$$

- Set the rejection region

$$R = [X > 7 \text{ or } X < 2] = \{8, 9, \ldots, 14, 15, 0, 1\}.$$

So we have the *size* of the test α

$$\alpha = \mathbb{P}_{0.3}\Big\{[X \leq 1] \cup [X \geq 8]\Big\}$$

and since under H_0, $X \sim b(k; n = 15, p = 0.3)$, we have

$$= .035 + (1 - .950) = .085.$$

In R:

```
pbinom(1,15,.3)+1-pbinom(7,15,.3)
[1] 0.08528014
```

In R:

```
pbinom(1,15,.3)+1-pbinom(7,15,.3)
[1] 0.08528014
```

To get an idea of the power on Θ_a, $\pi(\theta), \theta \in \Theta_a$.

$$\begin{aligned}\pi(.2) =& \mathbb{P}_{0.2}[\text{ reject }] = 1 - \mathbb{P}_{0.2}[2 \le X \le 7] \\ =& 1 - .829 = 0.171 = 1 - \text{sum}(\text{dbinom}(2:7, 15, .2)); \\ \pi(.4) =& 1 - \mathbb{P}_{0.4}[2 \le X \le 7] = 1 - .782 = .218.\end{aligned}$$

Not great power on Θ_a. However, the sample size is small and tests will always have trouble near the boundary of Θ_0 and Θ_a. Summary:

p	$\pi(p)$
0.2	0.171
0.4	0.218
0.6	0.787
0.8	0.996
0.9	0.999

2. For a new curing method for cement, it is claimed that the compression strength is 5000kg/sq. cm. with a standard deviation of 120. Assuming that compression strength is $N(\mu, \sigma^2)$, we test the hypothesis

$$H_0 : \mu = 5000 \text{ vs } H_a : \mu < 5000.$$

We test by examining 50 specimens and decide to reject H_0 if

$$\bar{X} < 4970.$$

- What is the level of significance?
- What is the power at $\mu = 4970, 4960$?

2. For a new curing method for cement, it is claimed that the compression strength is 5000kg/sq. cm. with a standard deviation of 120. Assuming that compression strength is $N(\mu, \sigma^2)$, we test the hypothesis

$$H_0 : \mu = 5000 \text{ vs } H_a : \mu < 5000.$$

We test by examining 50 specimens and decide to reject H_0 if

$$\bar{X} < 4970.$$

- What is the level of significance?
- What is the power at $\mu = 4970, 4960$?

Level of significance: Recall

$$\bar{X} \sim N(\mu, \frac{\sigma^2}{50}) = N(\mu, \frac{120^2}{50}) = N(\mu, 288) = N(\mu, 16.970^2)$$

and so

$$\begin{aligned} \alpha =& \mathbb{P}_{\mu=5000}[\bar{X} < 4970] \\ =& pnorm(4970, mean = 5000, sd = sqrt(288)) = .0385, \end{aligned}$$

Reminder:

$$\bar{X} \sim N(\mu, \frac{\sigma^2}{50}) = N(\mu, \frac{120^2}{50}) = N(\mu, 16.97056^2).$$

Power: Calculate $\pi(4970), \pi(4960)$. Using R, we evaluate

$$\mathbb{P}_\mu[\bar{X} \leq 4970], \quad \mu = 4970, 4960,$$

to get

mu	sd	x	pnorm(x, mean=mu, sd=sd)
4970	16.97056	4970	pnorm(4970,4970,16.97056) $= 0.5$
4960	16.97056	4970	pnorm(4970,4960,16.97056)= 0.7221551

3. Hypothesis test from a normal distribution with unknown μ and known $\sigma = 40$. For $n = 30$, observed $\bar{x} = 788$, test at level 0.05

$$H_0 : \mu = 800, \qquad \text{(simple hypothesis)}$$

VS

$$H_a : \mu \neq 800 \qquad \text{(compound and 2-sided alternative)} .$$

Elements of test:

- We use the test statistic

$$Z = \frac{\bar{X} - 800}{40/\sqrt{n}},$$

  which under H_0 is $N(0, 1)$.

- Rejection region:

$$|Z| > z_{\alpha/2} = z_{.025} = 1.96.$$

  Recall

$$\mathbb{P}[Z > z_\alpha] = \alpha,$$

  and

$$\mathbb{P}[|Z| > z_{\alpha/2}] = \alpha.$$

- Therefore, significance level of the test when H_0 is true is

$$\alpha = \mathbb{P}_{\mu=800}[\text{ Reject }] = \alpha.$$

- Compute value of the test statistic

$$z = \frac{788 - 800}{40/\sqrt{30}} = -1.6431$$

and therefore

$$|z| = 1.6431 < 1.96 = z_{\alpha/2},$$

when $\alpha = 0.05$.

- Conclude: The test fails to reject H_0 at level 0.05 since z is not in the rejection region.

- Therefore, significance level of the test when H_0 is true is

$$\alpha = \mathbb{P}_{\mu=800}[\text{ Reject }] = \alpha.$$

- Compute value of the test statistic

$$z = \frac{788 - 800}{40/\sqrt{30}} = -1.6431$$

and therefore

$$|z| = 1.6431 < 1.96 = z_{\alpha/2},$$

when $\alpha = 0.05$.

- Conclude: The test fails to reject H_0 at level 0.05 since z is not in the rejection region.
- What if we change the level of significance to $\alpha = 0.10$? Then

$$z_{\alpha/2} = z_{.05} = 1.6452$$

and we still fail to reject.

- What if we change the level of significance to $\alpha = .11$? Then

$$z_{\alpha/2} = z_{.055} = 1.598.$$

Then, the observed value IS IN THE REJECTION REGION:

$$|-1.6431| > 1.598.$$

So test rejects at level $\alpha = .11$.

- What if we change the level of significance to $\alpha = .11$? Then

$$z_{\alpha/2} = z_{.055} = 1.598.$$

Then, the observed value IS IN THE REJECTION REGION:

$$|-1.6431| > 1.598.$$

So test rejects at level $\alpha = .11$.

Summary

For this example:

α	$\lvert z \rvert$	$z_{\alpha/2}$	Reject?
0.05	1.6431	1.96	No
0.10	1.6431	1.6452	No
0.11	1.6431	1.598	Yes

P-value

The p-value of the test is the smallest level α at which H_0 is rejected.

For the above example the rejection region was

$$|z| > z_{\alpha/2}.$$

As a function of α,

$$z_\alpha = \Phi^{\leftarrow}(1-\alpha)$$

is decreasing in α. So

- Increase α, (say from .05 to .1 to .11) and you decrease $z_{\alpha/2}$ and hence
- Increase α, increase the size of the rejection region. If you increase the size of rejection region enough, eventually the test statistic falls inside.

Note

- The smaller the α, the more confidence we have that H_0 is false.
- When the data rejects at level α, we say
  the data is significant at level α.

Exact computation of the p-value in the previous example: Take the observed value of Z

$$z = z_{\text{obs}} = -1.643$$

and compute

$$\mathbb{P}[|Z| > |z_{\text{obs}}|] = 2(1 - \Phi(1.643)) = .10034.$$