Lecture 26

Hypothesis Testing

Manju M. Johny

STAT 330 - Iowa State University

Hypothesis Testing

Definition:
A statistical *hypothesis* is a statement about a parameter θ

There are 2 competing hypotheses in a testing problem:

- *Null Hypothesis (H_0):* the default/pre-data view about the parameter. (What we already believe - never prove this)
- *Alternative Hypothesis (H_A):* usually what you want your data/study to show. (what you're trying to prove)

Note: H_0 and H_A have to be disjoint. There can not be any outcomes in common between the null and alternative hypotheses.

Motivating Example

Example 1: I have a coin and I'm interested in the probability of flipping a "head". I flip a coin 100 times and record the number of heads obtained.

$$X = \# \text{ of heads}$$
$$X \sim Bin(n = 100, p)$$

where $p = P(\text{"}heads\text{"})$ is unknown

By default, we assume coin is fair $p = 0.5$ (null hypothesis).

Alternative hypothesis should contradict the null hypothesis.

Hypotheses:

- $H_0 : p = 0.5$ (coin is fair)
- $H_A : p \neq 0.5$ (coin is unfair)

Motivating Example Continued

Data: Out of 100 flips, I get 71 heads. $\hat{p} = 0.71$

Idea of Hypothesis Testing:

- Assume H_0 (our default belief) is true until our *data* tells us otherwise.
- Ask ourselves "what is the probability of getting 71 heads if the null hypothesis is true (coin is fair)?"

  $\rightarrow$ probability $= 0.000032$ (called the "$p - value$")
- There is a 0.000032 probability that we observed our data if the null hypothesis that the coin is fair is true.

  $\rightarrow$ Now we have evidence against the null hypothesis (that coin is fair), and in favor of the alternative hypothesis (that coin is unfair).

General Hypothesis Testing Procedure

Hypothesis Tests

We will look at 4 different hypothesis testing scenarios.
Their null hypotheses are given below:

- $H_0 : \mu = \#$
- $H_0 : p = \#$

test whether the population parameter is equal to some value(#).

- $H_0 : \mu_1 - \mu_2 = \#$
- $H_0 : p_1 - p_2 = \#$

usually 0

test whether the difference in parameters for 2 groups is equal to some value (#). This value (#) is almost always 0.

The above all follow the same general hypothesis testing procedure.

Testing Procedure

General Hypothesis Testing Procedure

Note: θ is just stand in symbol for parameter
θ can be μ, p, $\mu_1 - \mu_2$, $p_1 - p_2$

1. Determine the Null and Alternative Hypotheses:

   always "=" → $H_0 : \theta = \#$

   sign depends on problem → $H_A : \theta < \#$, $>$, $\neq$

   # should be same in Ho and HA

2. Gather data and calculate a *test statistic* under the assumption that H_0 is true. Test statistic has general form:

$$Z = \frac{\hat{\theta} - \#}{SE(\hat{\theta})}$$

3. Calculate the *p-value*. Use p−value to determine whether you have enough evidence to reject the null hypothesis.
   - small p−value → H_0 unlikely → Reject H_0
   - large p−value → No evidence against H_0 → Do not reject Ho
     Fail to reject Ho

Calculating p-values

Calculating $p-$value

Definition: $p-$value

The *$p-$value* is the probability of observing your test statistic or *more extreme* if the null hypothesis (H_0) is true.

"*more extreme*" can be bigger, smaller or both depending on the the sign in the alternative hypothesis (H_A)

- Small $p-$ *value* indicates a small probability of seeing your data if H_0 is true. The data is evidence against H_0 (Reject H_0)
- Large $p-$ *value* indicates a large probability of seeing your data if H_0 is true. No evidence against H_0 (Do Not Reject H_0)
- $P-$ *value* is often *wrongly* interpreted as the probability of the null hypothesis. (Don't make this mistake)

Calculating the $p-$ *value*

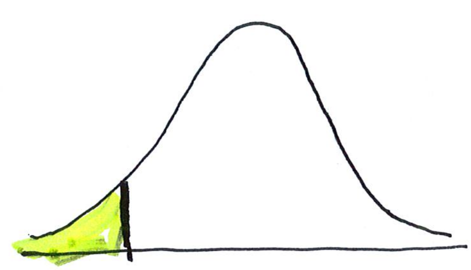
- By central limit theorem, the estimator follows a normal distribution. Standardizing the estimator gives us the test statistic Z, which follows $N(0,1)$ distribution
- Obtain $p-$ *value* from the $\underline{z-\text{table}}$ as left-hand area, right-hand area or both (depending on sign in H_A)

Left-sided Hypothesis Test

$H_0 : \theta = \#$

$H_A : \theta < \#$

$Z = \frac{\hat{\theta} - \#}{SE(\hat{\theta})}$



z

p-value = P(Z ≤ z)

test statistic value

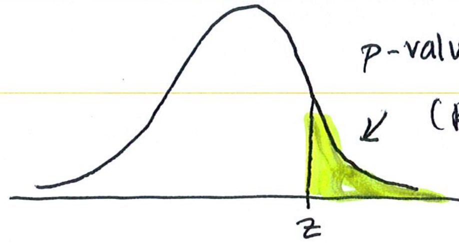
Calculating $p-$value Cont.

Right-sided Hypothesis Test

$H_0 : \theta = \#$

$H_A : \theta > \#$

$Z = \frac{\hat{\theta} - \#}{SE(\hat{\theta})}$



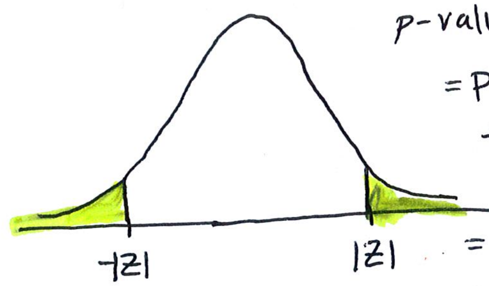
p-value $= P(Z > z)$
(Right-hand area)

2-sided Hypothesis Test

$H_0 : \theta = \#$

$H_A : \theta \neq \#$

$Z = \frac{\hat{\theta} - \#}{SE(\hat{\theta})}$



p-value $= P(|Z| > |z|)$
$= P(Z < -|z|) + P(Z > +|z|)$
$= 2P(Z < -|z|)$
$= 2P(Z > +|z|)$

Get one side & double it.

Easiest way: make your z a negative value, and find the left-hand area, then double it.

Types of Errors

Not on Final!

In the testing framework, it is possible to make errors that are inherent to the testing procedure (not calculation mistakes).

Types of errors

- Type I Error (wrongly reject H_0)
  $\rightarrow$ P(Type I error) $= \alpha$
- Type II Error (wrongly fail to reject H_0)
  $\rightarrow$ P(Type II error) $= \beta$

Decision

Truth	Reject Ho	Don't Reject Ho
Ho True	Type I error	No error
Ho False	No Error	Type II error

Note:

- α (significance level) can be viewed as a cut-off for how small the $p-$value needs to be to reject H_0. Reject H_0 if $p - value < \alpha$. (α set before conducting the test).
- In this class, we use a strength of evidence argument without a "cut-off" for $p - value$.

Hypothesis Testing Examples

Tax Fraud Example

Example: Tax Fraud

Default belief
Historically, IRS taxpayer compliance audits have revealed that about 5% of individuals do things on their tax returns that invite criminal prosecution.

A sample of $n = 1000$ tax returns produces $\hat{p} = 0.061$ as an estimate of the fraudulent returns.

Does this provide a clear signal of change in the tax payer behavior?

no specific direction. ~~the~~ can be greater or less
use "$\neq$" in H_A

1. State the Hypotheses

$$H_0 : p = 0.05$$
$$H_A : p \neq 0.05$$

same value
(never use $\hat{p}$)

Tax Fraud Example

2. The *test statistic* will be obtained from

$$Z = \frac{\hat{p} - \#}{\sqrt{\frac{\#(1-\#)}{n}}} = \frac{\hat{p} - 0.05}{\sqrt{\frac{0.05(0.95)}{n}}}$$

Under the null hypothesis, Z follows a N(0,1) distribution.

data

$\hat{p} = 0.061$

$n = 1000$

Plugging in our data values, we get the test statistic

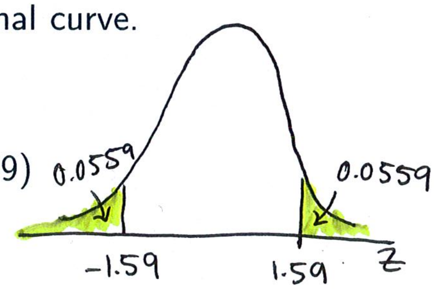
$$z = \frac{0.061 - 0.05}{\sqrt{\frac{0.05(0.95)}{1000}}} = 1.59$$

Tax Fraud Cont.

3. Since we have a "$\neq$" in the H_A, the p−value is obtained from both the left-hand and right-hand area of the normal curve.

$$\begin{aligned} p - value &= P(|Z| \geq 1.59) \\ &= P(Z < -1.59) + P(Z > 1.59) \\ &= 2 \cdot P(Z < -1.59) \\ &= 2 * 0.0559 \\ &= 0.1118 \end{aligned}$$

$\approx$ 11.%



⇒ H_0: $p = 0.05$

X H_A: $p \neq 0.05$

This is not a very small p−value. We therefore only have very weak evidence against H_0. Thus, we *do not* reject the null hypothesis in favor of the alternative hypothesis.

There is not much evidence of change in tax payer behavior.

Disk Drive Example

Example: Disk Drive

$n_1 = 30$ and $n_2 = 40$ disk drives of 2 different designs were tested under conditions of "accelerated" stress and times to failure recorded:

Group 1 Standard Design	Group 2 New Design
$n_1 = 30$	$n_2 = 40$
$\bar{x}_1 = 1205$ hr	$\bar{x}_2 = 1400$ hr
$s_1 = 1000$ hr	$s_2 = 900$ hr

Does the new design have a larger mean time to failure under "accelerated" stress? In other word, is the new design better?

1. State the Hypotheses

$$H_0: \mu_1 = \mu_2 \Rightarrow \mu_1 - \mu_2 = 0$$

$$H_A: \mu_1 < \mu_2 \Rightarrow \mu_1 - \mu_2 < 0$$

Disk Drive Cont.

2. The *test statistic* will be obtained from

$$Z = \frac{(\bar{X}_1 - \bar{X}_2) - 0}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}}$$

Under the null hypothesis, Z follows a N(0,1) distribution.

Plugging in our data values, we get the test statistic

$$z = \frac{(1205 - 1400) - 0}{\sqrt{\frac{1000^2}{30} + \frac{900^2}{40}}} = -0.84$$

observed test statistic

Disk Drive Cont.

3. Since we have a "<" in the H_A, the $p-$value is obtained from the left-hand area of the normal curve.

$$p - value = P(Z < -0.84)$$
$$= 0.2005 \quad \approx 20\%$$

$\Rightarrow H_0 : \mu_1 = \mu_2$

$\times \; H_A : \mu_1 < \mu_2$

This is not a small $p-$value. We therefore only have very weak evidence against H_0. Thus, we do not reject the null hypothesis in favor of the alternative hypothesis.

There is not significant evidence that the new design is better.

Queuing System Example

Example: Queuing System

Suppose we have 2 queuing systems A and B. We'd like to know whether system A has a higher probability of having an available server in the long run than system B. The simulation data for the 2 servers is shown below:

Group 1 System A	Group 2 System B
$n_1 = 500$ runs	$n_2 = 1000$ runs
$\hat{p}_1 = \frac{303}{500} = 0.606$	$\hat{p}_2 = \frac{551}{1000} = 0.551$

$\hat{p}$ = sample proportion of available server in runs

where $\hat{p}$ is the proportion runs with available servers at $t = 2000$.

1. State the Hypotheses

$$H_0 : p_1 = p_2 \quad \Rightarrow \quad p_1 - p_2 = 0$$

$$H_A : p_1 > p_2 \quad \Rightarrow \quad p_1 - p_2 > 0$$

Queuing System Cont.

2. The *test statistic* will be obtained from

$$Z = \frac{(\hat{p}_1 - \hat{p}_2) - 0}{\sqrt{\hat{p}_{pool}(1 - \hat{p}_{pool})}\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}$$

Under the null hypothesis, Z follows a N(0,1) distribution.

Next, calculate $\hat{p}_{pool}$ to plug into the denominator of the test statistic.

$$\hat{p}_{pool} = \frac{n_1\hat{p}_1 + n_2\hat{p}_2}{n_1 + n_2} = \frac{303 + 551}{500 + 1000} = 0.569$$

Plugging in our data values, we get the test statistic

$$z = \frac{(0.606 - 0.551) - 0}{\sqrt{0.569(1 - 0.569)}\sqrt{\frac{1}{500} + \frac{1}{1000}}} = 2.03$$

observed test statistic value

Queuing System Cont.

3. Since we have a ">" in the H_A, the $p-$value is obtained from the right-hand area of the normal curve.

$H_0: P_1 = P_2$

$H_A: P_1 > P_2$

$$\begin{aligned} p - value &= P(Z > 2.03) \\ &= 1 - 0.9788 \\ &= 0.0212 \quad \approx 2\% \end{aligned}$$

This is a small $p-$value. We therefore have strong evidence against H_0. Thus, we reject the null hypothesis in favor of the alternative hypothesis.

There is strong evidence that system A has a higher probability of having an available server than system B.

Hypothesis Testing Summary

Null Hypothesis	Test-Statistic	Reference Dist.
$H_0 : \mu = \#$	$Z = \frac{\bar{X} - \#}{s/\sqrt{n}}$	$Z \sim N(0,1)$
$H_0 : p = \#$	$Z = \frac{\hat{p} - \#}{\sqrt{\frac{\#(1-\#)}{n}}}$	$Z \sim N(0,1)$
$H_0 : \mu_1 - \mu_2 = \#$	$Z = \frac{(\bar{X}_1 - \bar{X}_2) - \#}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}}$	$Z \sim N(0,1)$
$H_0 : p_1 - p_2 = \#$	$Z = \frac{(\hat{p}_1 - \hat{p}_2) - \#}{\sqrt{\hat{p}_{pool}(1-\hat{p}_{pool})}\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}$ where $\hat{p}_{pool} = \frac{n_1\hat{p}_1 + n_2\hat{p}_2}{n_1 + n_2}$	$Z \sim N(0,1)$

"#" is usually 0