Hypothesis Testing

STAT 330 - Iowa State University

Outline

In this lecture students will learn about Hypothesis Testing. We define a hypothesis test, look at a motivating example, and introduce the steps to the hypothesis testing procedure. We will look at tests for:

1. μ
2. p
3. $\mu_1 - \mu_2$
4. $p_1 - p_2$

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. (Assumed value for θ)
- *Alternative Hypothesis (H_A):* usually what you want your data/study to show.

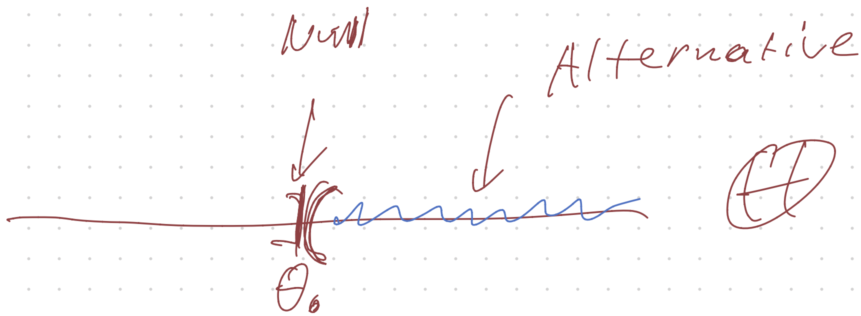
Note: H_0 and H_A have to be disjoint. The value of the parameter is either in the "null space" or "alternative space".

θ = true Response to treatment

"gold standard" value is θ_0

$H_0: \theta = \theta_0$

$\rightarrow H_A: \theta > \theta_0$



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:

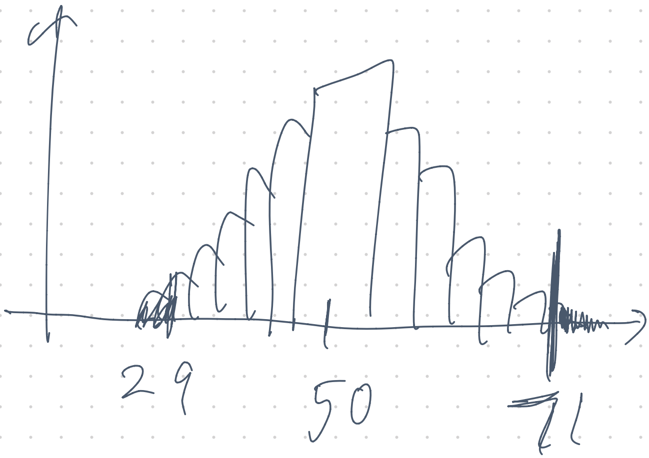
- Assume H_0 (our default belief) is true until our *data* tells us otherwise.
- Ask ourselves "what is the probability of getting 71 heads if the null hypothesis is true (coin is fair)?"

  $\rightarrow$ 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).

$H_0: p = .5$

$X \sim \text{Bin}(100, .5)$



$\mathbb{P}(X \geq 71) + \mathbb{P}(X \leq 29)$

(Assuming $p = .5$)

General Hypothesis Testing Procedure

Hypothesis Tests

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

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

The above all follow the same general hypothesis testing procedure.

Testing Procedure

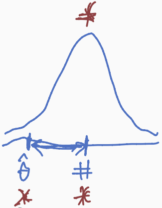
General Hypothesis Testing Procedure

1. Determine the Null and Alternative Hypotheses: $(\hat{\theta})$

   $H_0 : \theta = \#$

   $H_A : \theta \begin{matrix} < \\ > \\ \neq \end{matrix} \#$

   $\hat{\theta}$ estimate θ



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

   $\hat{\theta} \approx N(\#, [se(\hat{\theta})]^2)$

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

3. Calculate the *p-value*. Use $p-$value to determine whether you have enough evidence to reject the null hypothesis.
   - small $p-$value $\rightarrow$ H_0 unlikely $\rightarrow$ Reject H_0
   - large $p-$value $\rightarrow$ No evidence against H_0

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

$\hat{\theta} \simeq N(\#,$

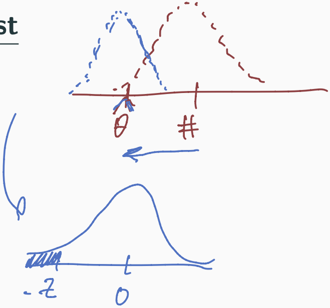
- Obtain $p-value$ from the $z-$table as left-hand area, right-hand area or both (depending on sign in H_A)

Left-sided Hypothesis Test

$$H_0 : \theta = \#$$

$$H_A : \theta < \#$$

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



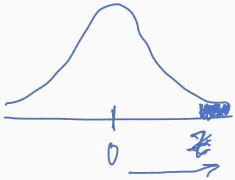
Calculating $p-$value Cont.

Right-sided Hypothesis Test

$$H_0 : \theta = \#$$

$$H_A : \theta > \#$$

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

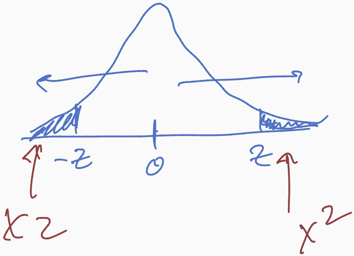


2-sided Hypothesis Test

$$H_0 : \theta = \#$$

$$H_A : \theta \neq \#$$

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



Types of Errors

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

H_0:

		True	False
Evidence	Against	✗	✓
	For	✓	✗

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) $= \beta$

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 Summary

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

Examples

Tax Fraud Example

Example 2: Tax Fraud

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 fraction of fraudulent returns.

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

1. State the Hypotheses

$H_0: p = .05$

$H_1: p \neq .05$

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.

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.

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

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 3: 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:

Standard Design	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

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

$H_A: \mu_1 < \mu_2$

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

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

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.

$$\begin{aligned} p - value &= \underline{P(Z < -0.84)} \\ &= \underline{\underline{0.2005}} \end{aligned}$$

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 4: 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:

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

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

1. State the Hypotheses

$H_0: p_1 - p_2 = 0$

$H_A: p_1 - p_2 > 0$

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

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

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.

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

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.

CI's vs Hyp. tests

μ:

$\bar{X}$ estimate μ

$se(\bar{X})$

$(1-\alpha)100\%$ CI for μ

$$\left(\bar{X} - z_{\alpha/2}\, se(\bar{X}),\ \bar{X} + z_{\alpha/2}\, se(\bar{X})\right)$$

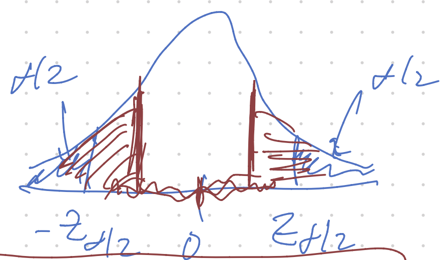
$H_0: \mu = \#$ vs $H_a: \mu \neq \#$

suppose # is in our Interval ...

$$|\bar{X} - \#| < z_{\alpha/2}\, se(\bar{X})$$

$$\Rightarrow \left|\frac{\bar{X} - \#}{se(\bar{X})}\right| < z_{\alpha/2}$$

$$\Rightarrow |z| < z_{\alpha/2}$$



$\boxed{\text{pvalue} > \alpha}$

95% CI for μ

(35, 47)

$H_0: \mu = 40$

$H_A: \mu \neq 40$

pvalue > .05

$H_0: \mu = 50$

$H_A: \mu \neq 50$

pvalue < .05

Recap

Students should now be familiar with the idea of Hypothesis Testing in Statistics. They should be able to set up appropriate hypotheses for a parameter (or difference of parameters) and carry out the testing procedure. They should be aware of the logic and types of conclusions we reach in testing.