# Chapter 6: Introduction to hypothesis testing

(Ott & Longnecker Sections: 5.1, 5.4, 5.6)

https://dzwang91.github.io/stat324/

WISCONSIN
UNIVERSITY OF WISCONSIN–MADISON

# Key concepts

- null hypothesis
- alternative hypothesis
- test statistic
- rejection region
- type I error
- type II error
- power
- p-value
- significance level

# "Dependent love"

A girl: "Do you love me? "

# "Dependent love"

A girl: "Do you love me? "

- Hypotheses:

$$H_0 : \text{You love me.} \quad v.s.\ H_A : \text{You don't love me.}$$

# “Dependent love”

A girl: “Do you love me? ”

- Hypotheses:

$$H_0 : \text{You love me.} \quad v.s. \; H_A : \text{You don't love me.}$$

- Reasoning: If you love me (if $H_0$ is true), you would take the trash out every week and put your socks away.

# "Dependent love"

A girl: "Do you love me? "

- Hypotheses:

$$H_0 : \text{You love me.} \quad v.s. \ H_A : \text{You don't love me.}$$

- Reasoning: If you love me (if $H_0$ is true), you would take the trash out every week and put your socks away.
- Data: Some weeks the girl's boyfriend doesn't take the trash out or leave the socks where they fall.

# “Dependent love”

A girl: “Do you love me? ”

- Hypotheses:

$$H_0 : \text{You love me.} \quad v.s. \; H_A : \text{You don't love me.}$$

- Reasoning: If you love me (if $H_0$ is true), you would take the trash out every week and put your socks away.
- Data: Some weeks the girl's boyfriend doesn't take the trash out or leave the socks where they fall.
- Conclusion: I don't believe you love me (reject the $H_0$).

# “Dependent love”

A girl: “Do you love me? ”

- Hypotheses:

$$H_0 : \text{You love me.} \quad v.s. \; H_A : \text{You don't love me.}$$

- Reasoning: If you love me (if $H_0$ is true), you would take the trash out every week and put your socks away.
- Data: Some weeks the girl's boyfriend doesn't take the trash out or leave the socks where they fall.
- Conclusion: I don't believe you love me (reject the $H_0$).
- Hypothesis testing concerns on how to use a random sample to judge if it is evidence that supports or not the hypothesis.

# Ingredients of a hypothesis test

- In hypothesis testing, there are two competing hypotheses:
  - $H_0$: the null hypothesis;
  - $H_A$: the alternative hypothesis.
- The hypothesis we want to test is if $H_A$ is "likely" true.
- One-tailed tests:

$$H_0 : \theta = \theta_0 \ v.s. \ H_A : \theta > \theta_0.$$

  or

$$H_0 : \theta = \theta_0 \ v.s. \ H_A : \theta < \theta_0.$$

- Two-tailed test:

$$H_0 : \theta = \theta_0 \ v.s. \ H_A : \theta \neq \theta_0.$$

# Ingredients of a hypothesis test

- There are two possible outcomes:
  - Reject $H_0$ because of sufficient evidence in the sample in favor of $H_A$.
  - Do not reject $H_0$ because of insufficient evidence to support $H_A$.
- Note that failure to reject $H_0$ does not mean the null hypothesis is true. It only means that we do not have sufficient evidence to support $H_A$.

# Steps of a hypothesis test

1. State the $H_0$ and $H_A$ hypotheses
2. Devise a test statistic: collect data $(X_1, ...., X_n)$ and choose a test statistic $T_n = T_n(X_1, ..., X_n)$. The test statistic is an RV. Based on data, we can calculate the realization of the test statistic.
3. Decide the rejection region: specify a set of values of the test statistic such that, if it realizes to one of these values, we reject $H_0$. This region is called the rejection region. (how do we specify the rejection region?)
4. Conclusion: if the test statistic falls outside of the rejection region, then there is insufficient evidence against $H_0$, and we say we fail to reject the null. Otherwise we say reject the null.

# Example

A company manufacturing RAM chips claims the defective rate of the population is 5%. Let p denote the true defective probability. We want to test:

# Example

A company manufacturing RAM chips claims the defective rate of the population is 5%. Let p denote the true defective probability. We want to test:

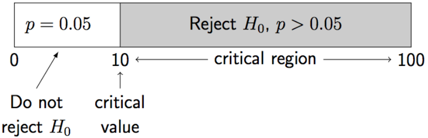
$$H_0 : p = 0.05, \;\; v.s. \; H_A : p > 0.05$$

# Example

A company manufacturing RAM chips claims the defective rate of the population is 5%. Let p denote the true defective probability. We want to test:

$$H_0 : p = 0.05, \ \ v.s. \ H_A : p > 0.05$$

We are going to use a sample of 100 chips from the production to test. Let $X$ denote the number of defective in the sample of 100. Reject $H_0$ if $X \geq 10$. Then $X$ is a test statistic.



# Types of errors

## Type I Error vs Type II Error

| | | Decision (based on sample) | |
|---|---|---|---|
| | | Reject $H_0$ | Not Reject $H_0$ |
| Truth (for population studied) | $H_0$ True | Type I Error $\alpha$ | |
| | $H_0$ False | | Type II Error $\beta$ |

- The acceptance of $H_A$ when $H_0$ is true is called a Type I error. The probability of committing a type I error is called the level of significance and is denoted by $\alpha$:

$$\alpha = P(\text{reject } H_0 \mid H_0 \text{ is true}) = P(T_n \in \text{rejection region} \mid H_0 \text{ is true}).$$

- Smaller $\alpha$ is better. Typically, 0.05 or smaller.

# Types of errors continued

- Use the distribution of the test statistic at $H_0$ to determine a rejection region: control the type I error at significance level $\alpha$.
- Failure to reject $H_0$ when $H_A$ is true is called a Type II error. The probability of committing a type II error is denoted by $\beta$.

$$\beta = P(\text{not reject } H_0 \mid H_A \text{ is true}).$$

- Smaller $\beta$ is better.
- It is usually impossible to compute $\beta$ unless we have a specific alternative hypothesis.

# Example

- Type I error:

$$\alpha = P(X \geq 10 \text{ when } p = 0.05) = \sum_{n=10}^{100} \binom{100}{n} 0.05^n (1 - 0.05)^{100-n}$$

$$= 0.0282$$

- Suppose we have $H_A : p = 0.1$, then

$$\beta = P(X < 10 \text{ when } p = 0.1) = \sum_{n=0}^{9} \binom{100}{n} 0.1^n (1 - 0.1)^{100-n}$$

$$= 0.4513$$

# Trade-off between $\alpha$ and $\beta$ for fixed sample size

$$H_0 : p = 0.05 \ v.s. \ H_A : p = 0.01.$$

- When rejection region is $X \geq 10$:

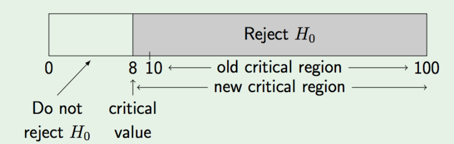
$$\alpha = 0.0282, \beta = 0.4513$$

# Trade-off between $\alpha$ and $\beta$ for fixed sample size

$$H_0 : p = 0.05 \ v.s. \ H_A : p = 0.01.$$

- When rejection region is $X \geq 10$:

$$\alpha = 0.0282, \beta = 0.4513$$

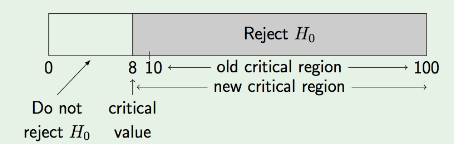


# Trade-off between $\alpha$ and $\beta$ for fixed sample size

$$H_0 : p = 0.05 \; v.s. \; H_A : p = 0.01.$$

- When rejection region is $X \geq 10$:

$$\alpha = 0.0282, \beta = 0.4513$$




- When the rejection region is $X \geq 8$:

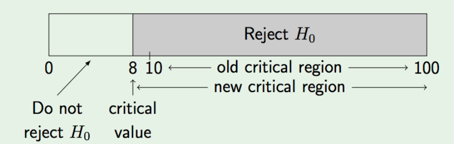
$$\alpha = 0.128, \beta = 0.206.$$

# Trade-off between $\alpha$ and $\beta$ for fixed sample size

$$H_0 : p = 0.05 \; v.s. \; H_A : p = 0.01.$$

- When rejection region is $X \geq 10$:

$$\alpha = 0.0282, \beta = 0.4513$$




- When the rejection region is $X \geq 8$:

$$\alpha = 0.128, \beta = 0.206.$$

How can we reduce $\alpha$ and $\beta$ simultaneously?

# Effect of the sample size

- Both $\alpha$ and $\beta$ can be reduced simultaneously by increasing the sample size.
- Consider that the sample size is $n = 150$ and the critical value is 12. Then, rejection region is $X \geq 12$, where X is the number of defectives in the sample of 150 chips.
  - The type I error is

$$\alpha = \sum_{n=12}^{150} \binom{150}{n} 0.05^n 0.95^{150-n} = 0.074,$$

    lower than 0.128 for n=100 and critical value of 8.
  - The type II error is

$$\beta = \sum_{n=0}^{11} 0.1^n 0.9^{150-n} = 0.171,$$

    lower than 0.206 for n=100 and critical value of 8.

# Power of a test

- The power of a test is the probability of rejecting $H_0$ given that the alternative hypothesis is true. That is,

$$Power = 1 - \beta = P(\text{reject } H_0 \text{ when } H_0 \text{ is false}).$$

# Hypothesis testing using p-value

- What is the drawback of hypothesis testing using rejection region?

# Hypothesis testing using p-value

- What is the drawback of hypothesis testing using rejection region?
- Different significance levels require recomputing the rejection region.

# Hypothesis testing using p-value

- What is the drawback of hypothesis testing using rejection region?
- Different significance levels require recomputing the rejection region.
- The p-value is defined to be the probability of a test statistic realizing to a value that is as or more extreme than the one actually observed given the null hypothesis is true.
- Smaller p-values indicate relatively more evidence against the null hypothesis.
- If the p-value is smaller than the given significance level $\alpha$, we would reject the null, otherwise we would not reject the null.

# Example

- Suppose that, for a given hypothesis test, the p-value is 0.09. Can $H_0$ be rejected?

# Example

- Suppose that, for a given hypothesis test, the p-value is 0.09. Can $H_0$ be rejected?
- Depends! At a significance level of 0.05, we cannot reject $H_0$ because $p = 0.09 > 0.05$. However, for significance levels greater than or equal to 0.09, we can reject $H_0$.

# What's the next?

We'll introduce some specific tests based on samples from one population in the next lecture.