Chapter 6: Introduction to hypothesis testing

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

Duzhe Wang

https://dzwang91.github.io/stat371/

WISCONSIN
UNIVERSITY OF WISCONSIN–MADISON

What do we study

Key concepts: null hypothesis, alternative hypothesis, test statistic, rejection region, Type I error, Type II error, power, p-value, significance level

A joke: “dependent love” story

- Question: Do you love me?

A joke: “dependent love” story

- Question: Do you love me?
- Claim: You love me.

A joke: “dependent love” story

- Question: Do you love me?
- Claim: You love me.
- Reasoning: If you love me, you would take the trash out every week and put your socks away.

A joke: “dependent love” story

- Question: Do you love me?
- Claim: You love me.
- Reasoning: If you love me, you would take the trash out every week and put your socks away.
- Data: Some weeks you don’t take the trash out or leave your socks where they fall.

A joke: “dependent love” story

- Question: Do you love me?
- Claim: You love me.
- Reasoning: If you love me, you would take the trash out every week and put your socks away.
- Data: Some weeks you don’t take the trash out or leave your socks where they fall.
- Conclusion: I don’t believe you love me(reject the claim).

A joke: “dependent love” story

- Question: Do you love me?
- Claim: You love me.
- Reasoning: If you love me, you would take the trash out every week and put your socks away.
- Data: Some weeks you don’t take the trash out or leave your socks where they fall.
- Conclusion: I don’t believe you love me(reject the claim).
- Philosophy: disprove(reject) a claim by contradiction

What is hypothesis testing?

- To prove that a hypothesis is true or false with absolute certainty, we would need absolute knowledge, that is, we would have to examine the entire population.
- Instead, 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.

  For example,

  $$H_0 = \text{" you love me"},\ H_A = \text{" you don't love me"}.$$

- The hypothesis we want to test is if H_A is "likely" true.

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. Data $X_1, ..., X_n$ are gathered, 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.

Steps of a hypothesis test

1. Data $X_1, ..., X_n$ are gathered, 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.
2. We 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. The rejection region consists of values that comprise evidence against H_0.

Steps of a hypothesis test

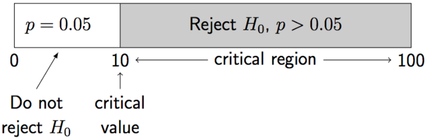
1. Data $X_1, ..., X_n$ are gathered, choose a test statistic $T_n = T_n(X_1, ..., X_n)$. The test statistic is an RV. Based on data, we can calculate the realization of the test statistic.
2. We specify a set of values of the test statistic such that, if it realizes to one of these values, we reject H_0. This region is called the rejection region. The rejection region consists of values that comprise evidence against H_0.
   If the test statistic falls outside of the rejection region, there is insufficient evidence against the null, and we say we fail to reject the null.

An 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:

- $H_0 : p = 0.05$
- $H_A : p > 0.05$

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

We are making a decision on a finite sample, so there is a possibility that we will make mistakes. The possible outcomes are:

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 α	
	H_0 False		Type II Error β

- 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 α.
- α=P(reject $H_0 \mid H_0$ is true). Smaller α is better. Typically, 0.05 or smaller.
- Use the distribution of the test statistic to determine a rejection region that limits the type I error at significance level α.

An example continued

- $$\alpha = P(X \geq 10 \text{ when } p = 0.05) = \sum_{n=10}^{100} \binom{100}{n} 0.05^n (1-0.05)^{100-n}$$
  $$= 0.0282$$
- So the level of significance is $\alpha = 0.0282$.
- Why can we calculate α in this way for the example?

Types of errors continued

- 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 β.
- β=P(not reject H_0 | H_0 is false). Smaller β is better.
- Note it is impossible to compute β unless we have a specific alternative hypothesis.
- Suppose we have $H_A : p = 0.1$, then

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

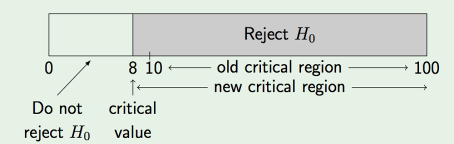
$$= 0.4513$$

Trade-off between α and β

Moving the critical value provides a trade-off between α and β. Given a fixed sample size, a reduction in β is always possible by increasing the size of the rejection region, but this increases α. Likewise, reducing α is possible by decreasing the rejection region.

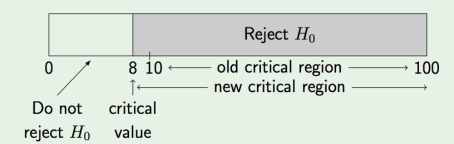
Trade-off between α and β

Moving the critical value provides a trade-off between α and β. Given a fixed sample size, a reduction in β is always possible by increasing the size of the rejection region, but this increases α. Likewise, reducing α is possible by decreasing the rejection region.



Trade-off between α and β

Moving the critical value provides a trade-off between α and β. Given a fixed sample size, a reduction in β is always possible by increasing the size of the rejection region, but this increases α. Likewise, reducing α is possible by decreasing the rejection region.



- The new significance level is $\alpha = \sum_{n=8}^{100} \binom{100}{n} 0.05^n 0.95^{100-n} = 0.128$, larger than before.
- The new β is $\beta = \sum_{n=0}^{7} \binom{100}{n} 0.1^n 0.9^{100-n} = 0.206$, lower than before.

Effect of the sample size

- Both α and β can be reduced simultaneously by increasing the sample size.

Effect of the sample size

- Both α and β can be reduced simultaneously by increasing the sample size.
- For the example, consider that the sample size is $n = 150$ and the critical value is 12. Then, reject H_0 if $X \geq 12$, where X is the number of defectives in the sample of 150 chips.
  - The significance level 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 a specific alternative hypothesis is true. That is, $Power = 1 - \beta = P(\text{reject } \mathrm{H}_0 \text{ when } H_0 \text{ is false})$.

P-value

- 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 when 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 α, we would reject the null, otherwise we would not reject the null.
- In most situations, reporting the p-values so that it may be used as the degree of evidence against the null is better than only stating the reject or not-reject decision.

Summary

- In hypothesis testing, we need to choose the test statistic and the rejection region so that the test has good statistical properties(for example, small errors).
- α and β are related, decreasing one generally increases the other.
- α can be set to a desired value by adjusting the critical value. Typically, α is set at 0.05.
- Increasing sample size decreases both α and β.
- Two methods of making a conclusion in hypothesis testing: one is using rejecting region and the other is using p-value.

What's the next?

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