Lecture 14: Hypothesis Testing (I)

Ye Tian

Department of Statistics, Columbia University
Calculus-based Introduction to Statistics (S1201)

August 2-3, 2022

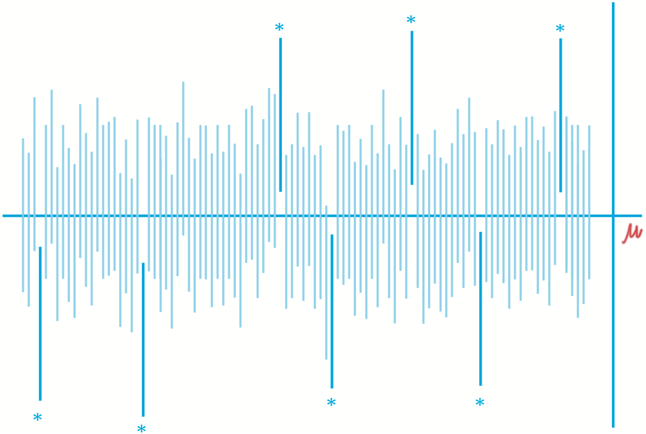
Recap: $100(1-\alpha)\%$ confidence intervals

- Normal distribution $N(\mu, \sigma^2)$
  - σ^2 known, the CI of μ: $\left[\bar{X} - \frac{\sigma}{\sqrt{n}} z_{\alpha/2}, \bar{X} + \frac{\sigma}{\sqrt{n}} z_{\alpha/2}\right]$
  - σ^2 unknown, the CI of μ: $\left[\bar{X} - \frac{S}{\sqrt{n}} t_{n-1,\alpha/2}, \bar{X} + \frac{S}{\sqrt{n}} t_{n-1,\alpha/2}\right]$
  - The CI of σ^2 (or σ): $\left[\sqrt{\frac{(n-1)S^2}{\chi^2_{n-1,\alpha/2}}}, \sqrt{\frac{(n-1)S^2}{\chi^2_{n-1,1-\alpha/2}}}\right]$
- Bernoulli trials Bernoulli(p): The **approximate** CI of p:
  $\left[\hat{p} - z_{0.025}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}, \hat{p} + z_{0.025}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}\right]$

How to understand this 95% coverage probability

$$\mathbb{P}\left(\bar{X} - \frac{z_{0.025}}{\sqrt{n}} \le \mu \le \bar{X} + \frac{z_{0.025}}{\sqrt{n}}\right) = 95\%.$$

We get a **random** interval $\left[\bar{X} - \frac{z_{0.025}}{\sqrt{n}}, \bar{X} + \frac{z_{0.025}}{\sqrt{n}}\right]$ which can cover μ with 95% probability!



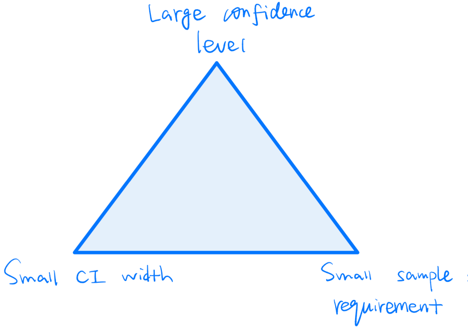
*: The "realization" of the random interval that does NOT cover μ

Confidence level, precision (width), and sample size

The ideal scenario:

- **Large** confidence level
- **Small** CI width
- **Small** sample size requirement

But this is an **impossible trinity**!



$$\mathbb{P}\left(\bar{X} - \frac{z_{\alpha/2}}{\sqrt{n}} \leq \mu \leq \bar{X} + \frac{z_{\alpha/2}}{\sqrt{n}}\right) = 1 - \alpha.$$

Confidence level $= 100(1-\alpha)\%$, CI width $= \frac{2z_{\alpha/2}}{\sqrt{n}}$, sample size n.

- Fixed confidence level: CI width $\searrow$, n needs to be $\nearrow$
- Fixed confidence level: $n \searrow$, CI width $\nearrow$
- Fixed sample size n: CI width $\nearrow$, corresponding confidence level $\nearrow$
- Fixed sample size CI width: $n \nearrow$, corresponding confidence level $\nearrow$

Today's goal

- Know the formulation of hypothesis testings
- Understand the general procedure of hypothesis testing
- Know two types of errors, the power and p-value

Intuitions

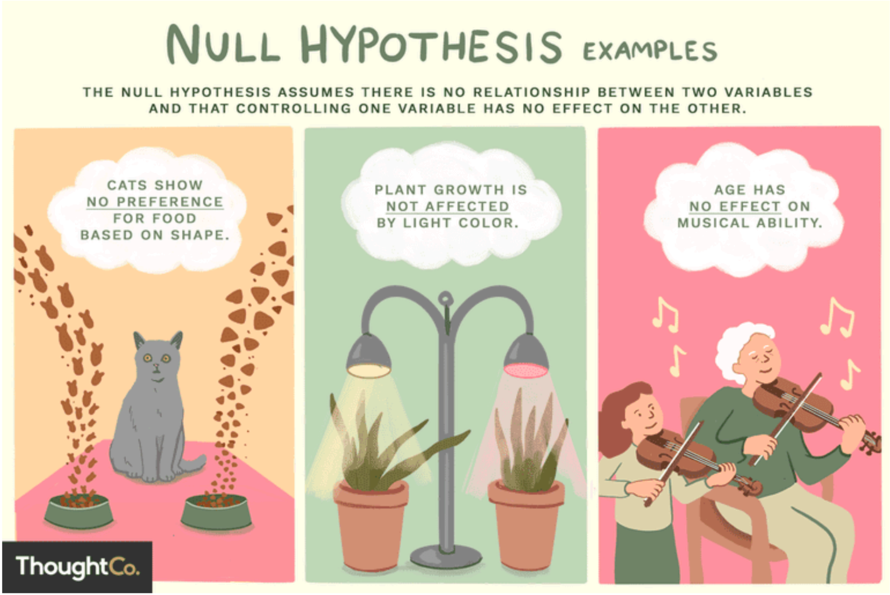
Intuitions

- Point estimation provides us a single number estimate for a parameter of interest
  e.g. estimate the lifetime of light bulbs by a statistic (r.v.) $\Rightarrow$ For a sample of data, it is a number
- Confidence intervals provide an entire interval of plausible values for a parameter of interest
  e.g. estimate the lifetime of light bulbs by a random interval $\Rightarrow$ For a sample of data, it is an deterministic interval
- But in some cases, instead of knowing the value of parameter, **we want to decide which of two contradictory claims about the parameter is correct**
  e.g.: decide whether the lifetime > 500 hours or not.

Intuitions

- A researcher develops a theory (through observation or previous exploratory analysis) about how something works.
- They wish to formally test their theory (TRUE/FALSE?) through the collection and analysis of sample data
- They can conduct a **hypothesis test** to test their theory.
  - Theory is TRUE? or
  - Theory is FALSE
- Let the data tell whether we should believe the theory is TRUE is FALSE
- Two hypotheses/statements are called the **null hypothesis** (the theory is FALSE) and **alternative** hypothesis (the theory is TRUE)
- Some principles:
  - People usually put a statement of the currently accepted belief as the null hypothesis (say, the theory is FALSE)
  - And put the new statement/the claim they want to justify as the alternative
  - Usually we need **very strong evidence** to believe something new!

Null hypothesis examples



Intuition about errors

- Since there are two statements, and we claim either one or the other, we would make mistakes in the decision process.
- For example, we want to claim the lifetime T of the light bulb from a factory is over 500 hours.
  - Null: $T \leq 500$ hrs
  - Alternative: $T > 500$ hrs
- Two types of errors:
  - $T \leq 500$ hrs, but we claim $T > 500$ hrs falsely (Type-I error)
  - $T > 500$ hrs actually, but we claim $T \leq 500$ hrs falsely (Type-II error)

Concepts

Hypothesis testing

Statistical hypothesis: A claim or assertion either about the value of a parameter of population.
E.g.: the lifetime of the light bulb > 500 hrs, population mean height of men is equal to 70 inches etc..

Two hypotheses: Usually people decide between two claims. The **null hypothesis**, denoted by H_0, is the claim that is initially assumed to be true. The **alternative hypothesis**, denoted by H_a or H_1, is the assertion that is contradictory to H_0.

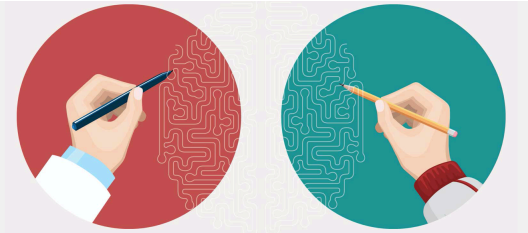
- Usually if we want to claim some new findings, we put them as H_a
- H_0 and H_a are NOT symmetric --- we tend to "**protect**" H_0, and require **very strong evidence** to claim H_a is true

A **test (of hypotheses)** is a method for using **sample data** to decide whether the null hypothesis should be rejected. It is a **function of sample data**, whose value indicates the result of hypothesis testing.

Examples: null and alternative hypotheses

What should the null and alternative hypotheses be?

- A researcher thinks that if knee surgery patients go to physical therapy twice a week (instead of 3 times), their recovery period will be longer.
- The annual return of ABC Limited bonds is assumed to be 7.5%. We want to test if the scenario is true or false.
- About 10% of the human population is left-handed. Suppose a researcher at Columbia speculates that students in Columbia College are more likely to be left-handed than people found in the general population.



Type-I and Type-II errors

There are two competing hypotheses: the null and the alternative. In a hypothesis test, we make a decision about which might be true, but our choice might be incorrect.

Decision

Truth		accept H_0	reject H_0
	H_0 true	✓	Type 1 error
	H_a true	Type 2 error	✓

- A Type 1 Error is rejecting the null hypothesis when H_0 is true.
- A Type 2 Error is failing to reject the null hypothesis when H_a is true.
- We never know if H_0 or H_a is true, but we need to consider all possibilities

Type-I and Type-II errors

Truth \ Decision	accept H_0	reject H_0
H_0 true	✓	Type 1 error
H_a true	Type 2 error	✓

- Usually Type-1 error is more serious than Type-2 error
- This matches our intuition that strong evidence is needed to claim H_a is true (we don't want to make Type-1 error!)

Hypothesis Test as a Trial

If we think of a hypothesis test as a criminal trial:Declaring the defendant innocent when they are actually guilty

$$H_0 : \text{Defendant is innocent}$$
$$H_a : \text{Defendant is guilty}$$

- **Type-I error:** Declaring the defendant innocent when they are actually guilty
- **Type-II error:** Declaring the defendant guilty when they are actually innocent

Which error do you think is the worse error to make? *"better that ten guilty persons escape than that one innocent suffer"* - William Blackstone

Type-I and Type-II error **rates**

We call probability $\mathbb{P}(\text{Type-I error}|H_0 \text{ true}) = \mathbb{P}(\text{reject } H_0|H_0 \text{ true})$ as **Type-I error rate** (of the test), and call $\mathbb{P}(\text{Type-II error}|H_1 \text{ true}) = \mathbb{P}(\text{accept } H_0|H_1 \text{ true})$ as **Type-II error rate** (of the test). And $1-$ **Type-II error** rate is called the **power** (of the test).

- This definition can be problematic. We will discuss more on this.
- Recall that Type-I error is more serious

The ideal scenario: Both Type-I and Type-II error rates are small.

But it's IMPOSSIBLE! (We will see this later)

Neyman-Pearson Criterion: We would like to construct a test that:

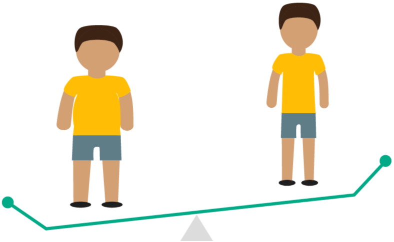
(1) controls Type-I error rates under some small level α

(2) under (1), minimizes the Type-II error rate

In this course, we will focus on the goal (1).

General Procedure of a Hypothesis Test

Example



"A common belief among the lay public is that body weight increases after entry into college, and the phrase 'freshman 15' has been coined to describe the 15 pounds that students presumably gain over their freshman year." Suppose everyone's weight is **normally distributed**.

Let μ denote the true average weight gain of students over the course of their first year in college.

$$H_0 : \mu = 15 \quad \text{v.s.} \quad H_a : \mu \neq 15$$

Example

Let μ denote the true average weight gain of students over the course of their first year in college.

$$H_0 : \mu = 15 \quad \text{v.s.} \quad H_a : \mu \neq 15$$

Suppose we collected weight gain data from n students: $X_1, \ldots, X_n$.

Goal: control Type 1 error under 5%.
Intuition: $\bar{X} \sim N(\mu, \sigma^2/n)$. Under H_0, $\bar{X} \sim N(15, \sigma^2/n)$ and $\frac{\sqrt{n}(\bar{X}-15)}{S} \sim t_{n-1}$.
Therefore, $\bar{X}$ is farther from 15, H_0 is **less** possible to be true.
Thus we can reject H_0 when $|\bar{X} - 15| > c$ for some constant $c > 0$.

Question: How to determine c?
Recall that Type-I error rate $= \mathbb{P}(|\bar{X} - 15| > c | H_0 \text{ true}) =$ $\mathbb{P}(|\bar{X} - 15| > c | \mu = 15)$. Note that $\frac{\sqrt{n}(\bar{X}-15)}{S} \sim t_{n-1}$. Therefore
$\mathbb{P}(|\bar{X} - 15| > c | \mu = 15) = \mathbb{P}(\sqrt{n}(\bar{X} - 15)/S > \sqrt{n}|c - 15|/S)$
$= 2F(-\sqrt{n}|c - 15|/S) \leq 5\%$ (F is the cdf of t_{n-1})

Example

Let μ denote the true average weight gain of students:

$$H_0 : \mu = 15 \quad \text{v.s.} \quad H_a : \mu \neq 15$$

Suppose we collected weight gain data from n students: $X_1, \ldots, X_n$.

Intuition: Under H_0, $\bar{X} \sim N(15, \sigma^2/n)$ and $\frac{\sqrt{n}(\bar{X}-15)}{S} \sim N(0,1)$.
Therefore, $\bar{X}$ is farther from 15, H_0 is **less** possible to be true.
Thus we can reject H_0 when $|\bar{X} - 15| > c$ for some constant $c > 0$.

Question: How to determine c?
Recall that Type-I error rate $= \mathbb{P}(|\bar{X} - 15| > c | H_0 \text{ true}) =$
$\mathbb{P}(|\bar{X} - 15| > c | \mu = 15)$. Note that $\frac{\sqrt{n}(\bar{X}-15)}{S} \sim N(0,1)$. Therefore
$\mathbb{P}(|\bar{X} - 15| > c | \mu = 15) = \mathbb{P}(|\sqrt{n}(\bar{X} - 15)/S| > \sqrt{n}c/S)$
$= 2F(-\sqrt{n}c/S) \leq 5\%$

Hence it suffices to let $-\sqrt{n}c/S = t_{n-1,0.975}$ i.e. $c = \frac{St_{n-1,0.025}}{\sqrt{n}}$.

Therefore, we reject H_0 if $\bar{X} > 15 + \frac{St_{n-1,0.025}}{\sqrt{n}}$ or $\bar{X} < 15 - \frac{St_{n-1,0.025}}{\sqrt{n}}$.

We call $\left(-\infty, 15 - \frac{St_{n-1,0.025}}{\sqrt{n}}\right) \cup \left(15 + \frac{St_{n-1,0.025}}{\sqrt{n}}, +\infty\right)$ the **rejection region**.

General Procedure

(1) Construct the null and alternative hypotheses

(2) Given the level α, develop a test that satisfies Type-1 error rate $\mathbb{P}(\text{reject } H_0 | H_0 \text{ true}) \leq \alpha$

How to do step (2):

- Construct a **test statistic** $T(X_1, \ldots, X_n)$, which is a function of samples $X_1, \ldots, X_n$ and whose distribution under H_0 is **known**
- Find a random region B such that $\mathbb{P}(\text{reject } H_0 | H_0 \text{ true}) = \mathbb{P}(T(X_1, \ldots, X_n) \in B) = \alpha$ (why "=" instead of $\leq$?)
- Then the rejection region is B, and the corresponding test is

$$\begin{cases} \text{reject } H_0, & \text{if } T(X_1, \ldots, X_n) \in B, \\ \text{accept } H_0, & \text{if } T(X_1, \ldots, X_n) \notin B. \end{cases}$$

Example: revisited

Let μ denote the true average weight gain of students:

$$H_0 : \mu = 15 \quad \text{v.s.} \quad H_a : \mu \neq 15$$

Suppose we collected weight gain data from n students: $X_1, \ldots, X_n$.

- Under H_0, $T = \frac{\sqrt{n}(\bar{X}-15)}{S} \sim t_{n-1}$-distribution would be the test statistic.
- We want to find c such that Type-I error rate $= \mathbb{P}(|T| > c) = 5\%$

It suffices to let $c = t_{n-1,0.025}$.
Therefore, we reject H_0 if $\bar{X} > 15 + \frac{St_{n-1,0.025}}{\sqrt{n}}$ or $\bar{X} < 15 - \frac{St_{n-1,0.025}}{\sqrt{n}}$. We call $\left(-\infty, 15 - \frac{St_{n-1,0.025}}{\sqrt{n}}\right) \cup \left(15 + \frac{St_{n-1,0.025}}{\sqrt{n}}, +\infty\right)$ the **rejection region**.

Example: revisited

$$H_0 : \mu = 15 \quad \text{v.s.} \quad H_a : \mu \neq 15$$

- Under H_0, $T = \frac{\sqrt{n}(\bar{X}-15)}{S} \sim t_{n-1}$-distribution would be the test statistic.
- We want to find c such that Type-I error rate $= \mathbb{P}(|T| > c) = 5\%$

It suffices to let $c = t_{n-1,0.025}$.

Therefore, we reject H_0 if $\bar{X} > 15 + \frac{St_{n-1,0.025}}{\sqrt{n}}$ or $\bar{X} < 15 - \frac{St_{n-1,0.025}}{\sqrt{n}}$. We call $\left(-\infty, 15 - \frac{St_{n-1,0.025}}{\sqrt{n}}\right) \cup \left(15 + \frac{St_{n-1,0.025}}{\sqrt{n}}, +\infty\right)$ the **rejection region**.

- Its Type-I error rate $= 5\%$ (by construction of the test!).
- Recall that Type-II error rate $= \mathbb{P}(\text{accept } H_0 | H_1 \text{ true}) = \mathbb{P}(|T| \leq t_{n-1,0.025} | H_1 \text{ true})$
- $\{\mu \neq 15\}$ includes infinite choices of μ. Without knowing the true μ, we cannot calculate the power or Type-II error rate.
- But as long as we are given a value $\mu = \mu_0$, we can calculate $\mathbb{P}(|T| \leq t_{n-1,0.025} | \mu_0)$ (will see more examples later).

Revisit Type-I and Type-II error **rates**

$$H_0 : \mu = 15 \quad \text{v.s.} \quad H_a : \mu \neq 15$$

We call probability $\mathbb{P}(\text{Type-I error}|H_0 \text{ true}) = \mathbb{P}(\text{reject } H_0|H_0 \text{ true})$ as **Type-I error rate**, and call $\mathbb{P}(\text{Type-II error}|H_1 \text{ true}) = \mathbb{P}(\text{accept } H_0|H_1 \text{ true})$ as **Type-II error rate**. And $1-$ **Type-II error rate** is called the **power**.

We mentioned that "This definition can be problematic."

- We need to specify **one single μ value** to calculate Type-I or Type-II error rates.
- Not a problem for Type-I error rate. We will only talk about H_0 that contains a single point (e.g.: $\mu = 15$).
- For Type-II error rate, both the power and Type-II error rate are a **function of underlying μ value**. It means, given a $\mu \neq 15$, we have a power and Type-II error rate.

More Examples

Example

The drying time of a type of paint under specified test conditions is known to be normally distributed with mean value 75 min and standard deviation 9 min. Chemists have proposed a new additive designed to **decrease** average drying time μ. It is believed that drying times with this additive will remain normally distributed with $\sigma = 9$.

- Develop a test that controls Type-I error under 5%.
- We have collected drying time $X_1, \ldots, X_{25}$ with $\bar{X} = 69.5$ min. What's the conclusion?

$$H_0 : \mu = 75 \quad \text{v.s.} \quad H_a : \mu \leq 75$$

- This is called a **one-sided** test.

  Test statistic $T = \frac{\sqrt{n}(\bar{X}-75)}{9} \sim N(0,1)$ when $\mu = 75$ (H_0 true).

  Intuition: When $T <$ some positive constant c, we would reject H_0.

  By Type-I error rate control: Let $\mathbb{P}(T < c|\mu = 75) = 5\%$, i.e. $c = z_{0.95}$, implying a rejection region $\bar{X} < 75 - \frac{9}{\sqrt{n}} z_{0.05} = 72.039$.

Example

The drying time of a type of paint under specified test conditions is known to be normally distributed with mean value 75 min and standard deviation 9 min. Chemists have proposed a new additive designed to **decrease** average drying time μ. It is believed that drying times with this additive will remain normally distributed with $\sigma = 9$.

- Develop a test that controls Type-I error under 5%.
- We have collected drying time $X_1, \ldots, X_{25}$ with $\bar{X} = 69.5$ min. What's the conclusion?

$$H_0 : \mu = 75 \quad \text{v.s.} \quad H_a : \mu \leq 75$$

- A rejection region $\bar{X} < 75 - \frac{9}{\sqrt{n}} z_{0.05} = 72.039$, which contains 69.5. Therefore, we have sufficient evidence to reject H_0 under significance level (or Type-I error rate control level) 5%, i.e. we believe that the new additive can indeed decrease average drying time μ under significance level 5%.

Example

$$H_0 : \mu = 75 \quad \text{v.s.} \quad H_a : \mu \leq 75$$

A rejection region $\bar{X} < 75 - \frac{9}{\sqrt{n}} z_{0.05} = 72.039$, which contains 69.5. Therefore, we have sufficient evidence to reject H_0 under significance level (or Type-I error rate control level) 5%, i.e. we believe that the new additive can indeed decrease average drying time μ under significance level 5%.

Wait... can we say anything more?

- Suppose $T_0 = \frac{\sqrt{n}(\bar{x}-75)}{9} = \frac{5\times(69.5-75)}{9} = -3.06$. Then $\mathbb{P}(T < T_0 | \mu = 75) = \mathbb{P}(N(0,1) < T_0) = \Phi(-3.06) = 0.0011$.
- We call $\mathbb{P}(T < T_0)$ as the **p-value** of the test. It is the probability of obtaining a value of the test statistic at least as contradictory to H_0 * as the value calculated from the available sample data**, when assuming that the null hypothesis is true**.
  *Here it means we see $T <$ the current value of T, which is $T_0 = 3.06$
  **Here it means $T_0 = -3.06$, which is calculated from the current data

More about p-values

- p-value is a **probability**
- We reject the null when p-value $<$ significance level α, and accept the null when p-value $\geq$ significance level α. This is **equivalent** to the test we came up with before by calculating the rejection region.
- For one-sided test like $H_0 : \mu = 15$ v.s. $H_a : \mu > 15$, the p-value usually looks like $\mathbb{P}(T > T_0)$, where T is the test statistic and T_0 is the current value of T calculated from the available sample
- For one-sided test like $H_0 : \mu = 15$ v.s. $H_a : \mu < 15$, the p-value usually looks like $\mathbb{P}(T < T_0)$
- For two-sided test like $H_0 : \mu = 15$ v.s. $H_a : \mu \neq 15$, the p-value usually looks like $\mathbb{P}(|T| > |T_0|)$
- The smaller p-value we have, more evidence we have against H_0
- p-value is **NOT** the probability that H_0 is true!!! (H_0 can only be true or false, so it's deterministic, although we don't know.)

Many thanks to

- Chengliang Tang
- Joyce Robbins
- Yang Feng
- Owen Ward
- Wenda Zhou
- And all my teachers in the past 25 years