

Statistics and Data Analysis

Hypothesis Testing

Ling-Chieh Kung

Department of Information Management
National Taiwan University

Introduction

- ▶ How do **scientists** (physicists, chemists, etc.) do research?
 - ▶ Observe phenomena.
 - ▶ Make hypotheses.
 - ▶ Test the hypotheses through experiments (or other methods).
 - ▶ Make conclusions about the hypotheses.
- ▶ In the business world, business researchers do the same thing with **hypothesis testing**.
 - ▶ One of the most important technique of statistical inference.
 - ▶ A technique for (statistically) **proving** things.
 - ▶ Again relies on **sampling distributions**.

Road map

- ▶ **Basic ideas of hypothesis testing.**
- ▶ The first example.
- ▶ The p -value.

People ask questions

- ▶ In the business (or social science) world, people ask questions:
 - ▶ Are older workers more loyal to a company?
 - ▶ Does the newly hired CEO enhance our profitability?
 - ▶ Is one candidate preferred by more than 50% voters?
 - ▶ Do teenagers eat fast food more often than adults?
 - ▶ Is the quality of our products stable enough?
- ▶ How should we answer these questions?
- ▶ Statisticians suggest:
 - ▶ First make a **hypothesis**.
 - ▶ Then **test** it with samples and statistical methods.

Statistical hypotheses

- ▶ A **statistical hypothesis** is a formal way of stating a hypothesis.
 - ▶ Typically it is a mathematical description of parameters to test.
- ▶ It contains two parts:
 - ▶ The **null hypothesis** (denoted as H_0).
 - ▶ The **alternative hypothesis** (denoted as H_a or H_1).
- ▶ The alternative hypothesis is:
 - ▶ The thing that we want (need) to prove.
 - ▶ The conclusion that can be made only if we have **a strong evidence**.
- ▶ The null hypothesis corresponds to a **default** position.
 - ▶ We first **assume** that the null hypothesis is correct.
 - ▶ Then we collect sample data.
 - ▶ If under the null hypothesis it is **quite unlikely** to see our observed result, we claim that the null hypothesis is wrong.

Statistical hypotheses: example 1

- ▶ In our factory, we produce packs of candy whose average weight should be 1 kg.
- ▶ One day, a consumer told us that his pack only weighs 900 g.
- ▶ We need to know whether this is just a rare event or our production system is out of control.
- ▶ If (we believe) the system is out of control, we need to shutdown the machine and spend two days for inspection and maintenance. This will cost us at least \$100,000.
- ▶ So we should not to believe that our system is out of control just because of one complaint. What should we do?

Statistical hypotheses: example 1

- ▶ We first state a hypothesis: “Our production system is under control.”
- ▶ Then we ask: Is there a strong enough evidence showing that the hypothesis is **wrong**, i.e., the system is out of control?
 - ▶ Initially, we assume that our system is under control.
 - ▶ Then we do a survey to see if we have a strong enough evidence.
 - ▶ We shutdown machines **only if** we can “prove” that the system is indeed out of control.
- ▶ Let μ be the average weight, the **statistical hypothesis** is

$$H_0: \mu = 1$$

$$H_a: \mu \neq 1.$$

Statistical hypotheses: example 2

- ▶ In our society, we adopt the presumption of innocence.
 - ▶ One is considered **innocent** until proven **guilty**.
- ▶ So when there is a person who probably stole some money:

H_0 : The person is innocent

H_a : The person is guilty.

- ▶ There are two possible errors:
 - ▶ One is guilty but we think she/he is innocent.
 - ▶ One is innocent but we think she/he is guilty.
- ▶ Which one is more critical?
 - ▶ It is unacceptable that an innocent person is considered guilty.
 - ▶ We will say one is guilty **only if** there is a strong evidence.

Statistical hypotheses: example 3

- ▶ Consider the following hypothesis: “The candidate is preferred by more than 50% voters.”
- ▶ As we need a default position, and the percentage that we care about is 50%, we will choose our null hypothesis as

$$H_0: p = 0.5.$$

- ▶ p is the **population proportion** of voters preferring the candidate.
- ▶ More precisely, let $X_i = 1$ if voter i prefers this candidate and 0 otherwise, $i = 1, \dots, N$, then $p = \frac{\sum_{i=1}^N X_i}{N}$.
- ▶ How about the alternative hypothesis? Should it be

$$H_a: p > 0.5 \quad \text{or} \quad H_a: p < 0.5?$$

Statistical hypotheses: example 3

- ▶ The choice of the alternative hypothesis depends on the related **decisions** or **actions** to make.
- ▶ Suppose one will go for the election only if she thinks she will win (i.e., $p > 0.5$), the alternative hypothesis will be

$$H_a: p > 0.5.$$

- ▶ Suppose one tends to participate in the election and will give up only if the chance is slim, the alternative hypothesis will be

$$H_a: p < 0.5.$$

- ▶ The alternative hypothesis is “the thing we want (need) to prove.”

Remarks

- ▶ For setting up a statistical hypothesis:
 - ▶ Our default position will be put in the null hypothesis.
 - ▶ The thing we want to prove (i.e., the thing that needs a strong evidence) will be put in the alternative hypothesis.
- ▶ For writing the mathematical statement:
 - ▶ The **equal sign** (=) will always be put in the null hypothesis.
 - ▶ The alternative hypothesis contains an **unequal sign** or **strict inequality**: \neq , $>$, or $<$.
- ▶ The direction of the alternative hypothesis, when it is an inequality, depends on the business context.

One-tailed tests and two-tailed tests

- ▶ If the alternative hypothesis contains an unequal sign (\neq), the test is a **two-tailed** test.
- ▶ If it contains a strict inequality ($>$ or $<$), the test is a **one-tailed** test.
- ▶ Suppose we want to test the value of the population mean.
 - ▶ In a two-tailed test, we test whether the population mean significantly deviates from a hypothesized value. We do not care whether it is larger than or smaller than.
 - ▶ In a one-tailed test, we test whether the population mean significantly deviates from a hypothesized value **in a specific direction**.

Road map

- ▶ Basic ideas of hypothesis testing.
- ▶ **The first example.**
 - ▶ **A two-tailed test.**
 - ▶ A one-tailed test.
- ▶ The p -value.

The first example: a two-tailed

- ▶ Now we will demonstrate the process of hypothesis testing.
- ▶ Suppose we test the average weight (in g) of our products.

$$H_0: \mu = 1000$$

$$H_a: \mu \neq 1000.$$

- ▶ The variance of the product weights is $\sigma^2 = 40000 \text{ g}^2$.
 - ▶ The case with unknown σ^2 will be discussed in the next lecture.
- ▶ A random sample has been collected.
 - ▶ Suppose the sample size $n = 100$.
 - ▶ Suppose the sample mean $\bar{x} = 963$.
- ▶ How to make a conclusion?

Controlling the error probability

- ▶ All we can do is to collect a **random** sample and make our conclusion based on the observed sample.
- ▶ It is natural that we may be wrong when we claim $\mu \neq 1000$.
 - ▶ It is possible that $\mu = 1000$ but we unluckily get a sample mean $\bar{x} = 812$.
- ▶ We want to **control the error probability**.
 - ▶ Let α be the maximum probability for us to make this error.
 - ▶ α is called the **significance level**.
 - ▶ $1 - \alpha$ is called the **confidence level**.
 - ▶ Target: If $\mu = 1000$, our sampling and testing process will make us claim that $\mu \neq 1000$ with probability at most α .

Rejection rule

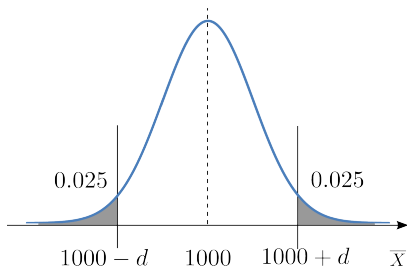
- ▶ Now let's test with the significance level $\alpha = 0.05$.
- ▶ Intuitively, if \bar{X} **deviates** from 1000 **a lot**, we should **reject** the null hypothesis and believe that $\mu \neq 1000$.
 - ▶ If $\mu = 1000$, it is so unlikely to observe such a large deviation.
 - ▶ So such a large deviation provides a **strong evidence**.
- ▶ So we start by sampling and calculating the **sample mean**.
- ▶ We want to construct a **rejection rule**: If $|\bar{X} - 1000| > d$, we reject H_0 . We need to calculate d .

Rejection rule

- ▶ We want a distance d such that **if H_0 is true**, the probability of rejecting H_0 is 5%, i.e.,

$$\Pr(|\bar{X} - 1000| > d \mid \mu = 1000) = 0.05.$$

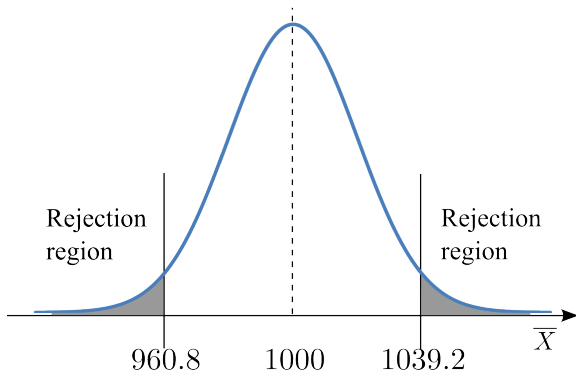
- ▶ People typically hide the condition $\mu = 1000$ and directly write $\Pr(|\bar{X} - 1000| > d)$.
- ▶ Consider \bar{X} :
 - ▶ We know $\sigma = 200$ and $n = 100$.
 - ▶ We **assume** that $\mu = 1000$.
 - ▶ Thanks to the central limit theorem, $\bar{X} \sim \text{ND}(1000, 20)$.



$$\Pr(|\bar{X} - 1000| > d) = 0.05.$$

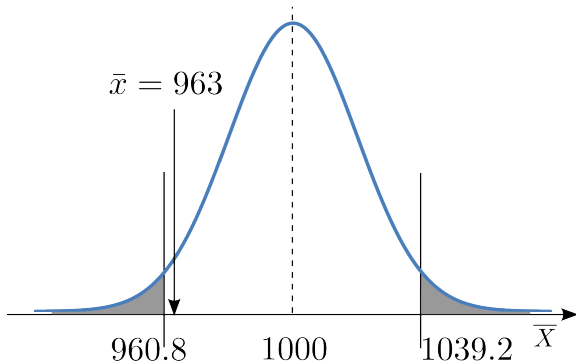
Rejection rule: the critical value

- ▶ According to $\bar{X} \sim \text{ND}(1000, 20)$, $\Pr(|\bar{X} - 1000| > 39.2) = 0.05$. The **rejection region** is $R = (-\infty, 960.8) \cup (1039.2, \infty)$.
- ▶ If \bar{X} falls in the rejection region, we reject H_0 .



Rejection rule: the critical value

- ▶ Because $\bar{x} = 963 \notin R$, we **cannot reject H_0** .
 - ▶ The deviation from 1000 is not large enough.
 - ▶ The evidence is not strong enough.



Rejection rule: the critical value

- ▶ In this example, the two values 960.8 and 1039.2 are the **critical values** for rejection.
 - ▶ If the sample mean is more extreme than one of the critical values, we reject H_0 .
 - ▶ Otherwise, we do not reject H_0 .
- ▶ $\bar{x} = 963$ is not strong enough to support $H_a: \mu \neq 1000$.
- ▶ Concluding statement:
 - ▶ Because the sample mean does not lie in the rejection region, we **cannot reject H_0** .
 - ▶ With a 95% confidence level, there is **no** strong evidence showing that the average weight **is not** 1000 g.
 - ▶ Therefore, we **should not** shutdown machines to do an inspection.

Summary

- ▶ We want to know whether the machine is out of control.
 - ▶ If the machine is actually good, we do not want to reach a conclusion that requires an inspection and maintenance.
 - ▶ We will do the inspection **only if** we have a strong evidence suggesting that $\mu \neq 1000$.
- ▶ We want to know whether H_0 is false, i.e., $\mu \neq 1000$.
- ▶ We control the probability of making a wrong conclusion.
 - ▶ We should not reject H_0 if it is true,
 - ▶ We limit the probability at $\alpha = 5\%$.
- ▶ We will conclude that H_0 is false if \bar{X} falls in the rejection region.
 - ▶ The calculation of the the critical values is based on the normal distribution, which can always be transformed to **the z distribution**.
 - ▶ This is called a **z test**.

Not rejecting vs. accepting

- ▶ We should be careful in writing our conclusions:
 - ▶ **Wrong**: Because the sample mean does not lie in the rejection region, we **accept** H_0 . With a 95% confidence level, there **is** a strong evidence showing that the average weight **is** 1000 g.
 - ▶ **Right**: Because the sample mean does not lie in the rejection region, we **cannot reject H_0** . With a 95% confidence level, there **is no** strong evidence showing that the average weight **is not** 1000 g.
 - ▶ Unable to prove one thing is false does not mean it is true!

Road map

- ▶ Basic ideas of hypothesis testing.
- ▶ **The first example.**
 - ▶ A two-tailed test.
 - ▶ **A one-tailed test.**
- ▶ The p -value.

The first example (part 2)

- ▶ Suppose that we modify the hypothesis into a directional one:¹

$$H_0: \mu = 1000.$$

$$H_a: \mu < 1000.$$

We still have $\sigma^2 = 40000$, $n = 100$, and $\alpha = 0.05$.

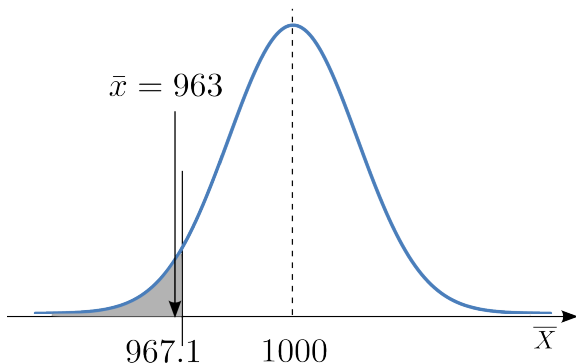
- ▶ This is a **one-tailed test**.
- ▶ Once we have a strong evidence supporting H_a , we will claim that $\mu < 1000$.
- ▶ We need to find a distance d such that

$$\Pr\left(1000 - \bar{X} > d \mid \mu = 1000\right) = 0.05.$$

¹Some researchers write $\mu \geq 1000$ in this case.

Rejection rule: the critical value

- ▶ For $0.05 = \Pr(1000 - \bar{X} > d)$, we have $d = 32.9$.
- ▶ As the observed sample mean $\bar{x} = 963 \in (-\infty, 967.1)$, we **reject H_0** .
 - ▶ The deviation from 1000 is large enough.
 - ▶ The evidence is strong enough.

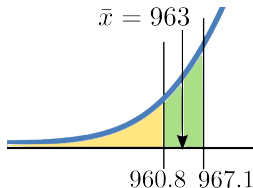


Rejection rule: the critical value

- ▶ In this example, 967.1 is the critical values for rejection.
 - ▶ If the sample mean is more extreme than (in this case, below) the critical value, we reject H_0 .
 - ▶ Otherwise, we do not reject H_0 .
- ▶ There is a strong evidence supporting $H_a: \mu < 1000$.
- ▶ Concluding statement:
 - ▶ Because the sample mean lies in the rejection region, we **reject H_0** .
With a 95% confidence level, there **is a** strong evidence showing that the average weight **is less than** 1000 g.

One-tailed tests vs. two-tailed tests

- ▶ When should we use a two-tailed test?
 - ▶ We use a two-tailed test when we are lack of the direction information.
 - ▶ E.g., we suspect that the population mean **has changed**, but we **have no idea** about whether it becomes larger or smaller.
- ▶ If we know or believe that the change is possible **only in one direction**, we may use a one-tailed test.
- ▶ Having more information (i.e., knowing the direction of change) makes rejection “easier,” i.e., easier to find a strong enough evidence.



Summary

- ▶ Distinguish the following pairs:
 - ▶ One- and two-tailed tests.
 - ▶ No evidence showing H_0 is false and having evidence showing H_0 is true.
 - ▶ Not rejecting H_0 and accepting H_0 .
 - ▶ Using $=$ and using \geq or \leq in the null hypothesis.

Road map

- ▶ Basic ideas of hypothesis testing.
- ▶ The first example.
- ▶ **The p -value.**

The p -value

- ▶ The **p -value** is an important, meaningful, and widely-adopted tool for hypothesis testing.

Definition 1

In a hypothesis testing, for an observed value of the statistic, the p -value is the probability of observing a value that is at least as extreme as the observed value under the assumption that the null hypothesis is true.

- ▶ Calculated based on an **observed** value of the statistic.
- ▶ Is the **tail probability** of the observed value.
- ▶ Assuming that the null hypothesis is true.

The p -value

- ▶ Mathematically:
 - ▶ Suppose we test a population mean μ with a one-tailed test

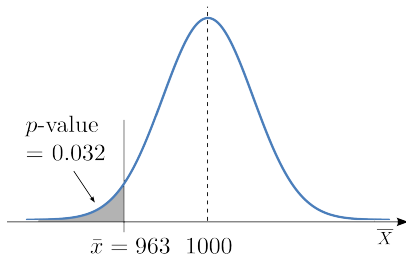
$$H_0: \mu = 1000$$

$$H_a: \mu < 1000.$$

- ▶ Given an observed \bar{x} , the p -value is defined as

$$\Pr(\bar{X} \leq \bar{x}).$$

- ▶ In the previous example, $\sigma = 200$, $n = 100$, $\alpha = 0.05$, and $\bar{x} = 963$.
 - ▶ If H_0 is true, i.e., $\mu = 1000$, we have $\Pr(\bar{X} \leq 963) = 0.032$.
 - ▶ The p -value of \bar{x} is 0.032.



How to use the p -value?

- ▶ The p -value can be used for constructing a **rejection rule**.
- ▶ For a one-tailed test:
 - ▶ If the p -value is **smaller** than α , we **reject** H_0 .
 - ▶ If the p -value is greater than α , we do not reject H_0 .
- ▶ In our example, the one-tailed test is

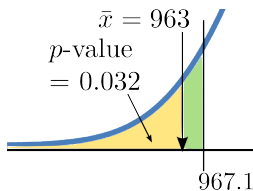
$$H_0: \mu = 1000$$

$$H_a: \mu < 1000.$$

- ▶ We have $\alpha = 0.05$.
- ▶ Because the p -value $0.032 < 0.05$, we reject H_0 .

p -values vs. critical values

- ▶ Using the p -value is **equivalent** to using the critical values.
 - ▶ The rejection-or-not decision we make will be the same based on the two methods.



The benefit of using the p -value

- ▶ In calculating the p -value, we do not need α .
- ▶ After the p -value is calculated, we compare it with α .
- ▶ The p -value, which needs to be calculated **only once**, allows us to know whether the difference is significant under various values of α .
- ▶ In our example:

α	0.1	0.05	0.01
Rejecting H_0 ?	Yes (0.032 < 0.1)	Yes (0.032 < 0.05)	No (0.032 > 0.01)

- ▶ If we use the critical-value method, we need to calculate the critical value for three times, one for each value of α .

The benefit of using the p -value

- ▶ In many studies, researchers do not determine the significance level α **before** a test is conducted.
- ▶ They calculate the p -value and then mark **the significance** of the result with **stars**.
- ▶ One typical way of assigning stars:

p -value	Significant?	Mark
(0, 0.01]	Highly significant	***
(0.01, 0.05]	Moderately significant	**
(0.05, 0.1]	Slightly significant	*
(0.1, 1)	Insignificant	(Empty)

The benefit of using the p -value

- ▶ As an example, suppose one is testing whether people at different ages sleep for at least eight hours per day in average.
 - ▶ Age groups: [10, 15), [15, 20), [20, 35), etc.
 - ▶ For group i , a one-tailed test is conducted. $H_a: \mu_i > 8$.
 - ▶ The result may be presented in a table:

Group	Age group	p -value
1	[10,15)	0.0002***
2	[15,20)	0.2
3	[20,25)	0.06*
4	[25,30)	0.04**
5	[30,35)	0.03**

- ▶ A smaller p -value does NOT mean **a larger deviation!**
 - ▶ We cannot conclude that $\mu_5 > \mu_4$, $\mu_1 > \mu_3$, etc.
 - ▶ There are other tests for the difference between two population means.

The p -value for two-tailed tests

- ▶ How to construct the rejection rule for a **two-tailed** test?
 - ▶ If the p -value is **smaller** than $\frac{\alpha}{2}$, we **reject** H_0 .
 - ▶ If the p -value is greater than $\frac{\alpha}{2}$, we do not reject H_0 .
- ▶ Consider the two-tailed test

$$H_0: \mu = 1000.$$

$$H_a: \mu \neq 1000.$$

- ▶ We have $\alpha = 0.05$.
- ▶ Because the p -value $0.032 > \frac{\alpha}{2} = 0.025$, we do not reject H_0 .

Summary

- ▶ The p -value is the tail probability of the realized value of a statistics assuming the null hypothesis is true.
- ▶ The p -value method is an alternative way of forming the rejection rule.
 - ▶ It is equivalent to the critical-value method.
- ▶ The p -value is related to the probability for H_0 to be false.
- ▶ It does not measure the magnitude of the deviation.