PROBABILITY AND STATISTICS

CHAPTER 6: INFERENCES BASED ON TWO SAMPLES

Dr. Phan Thi Huong

HoChiMinh City University of Technology
Faculty of Applied Science, Department of Applied Mathematics
Email: huongphan@hcmut.edu.vn

HCM city — 2021.

Outline

Outline

Example

Example 1

Trace metals in drinking water affect the flavor, and unusually high concentrations can pose a health hazard. An article reports on a study in which six river locations were selected (six experimental objects) and the zinc concentration (mg/L) determined for both surface water and bottom water at each location. The six pairs of observations are displayed in the accompanying table. Does the data suggest that true average concentration in bottom water exceeds that of surface water? ($\alpha = 0.05$)

Zinc concentration	1	2	3	4	5	6
in bottom water (x)	0.430	0.266	0.567	0.531	0.707	0.716
in surface water (y)	0.415	0.238	0.390	0.410	0.605	0.609

Paired data

Paired data usually given as this form:

Observation	1	2	3	4	...	n
Sample/Property 1 (X)	x_1	x_2	x_3	x_4	...	x_n
Sample/Property 2 (Y)	y_1	y_2	y_3	y_4	...	y_n

PAIRED DATA

Paired data usually given as this form:

Observation	1	2	3	4	...	n
Sample/Property 1 (X)	x_1	x_2	x_3	x_4	...	x_n
Sample/Property 2 (Y)	y_1	y_2	y_3	y_4	...	y_n

$\Rightarrow$ Problem: Compare the true mean difference between property 1 and property 2, e.i $E(X) - E(Y) = E(X - Y)$.

PAIRED DATA AND HYPOTHESIS TESTING

Let $D_i = X_i - Y_i$, we have:

Observation	1	2	3	4	...	n
Sample/Property 1 (X)	x_1	x_2	x_3	x_4	...	x_n
Sample/Property 2 (Y)	y_1	y_2	y_3	y_4	...	y_n
$D = X - Y$	d_1	d_2	d_3	d_4	...	d_n

PAIRED DATA AND HYPOTHESIS TESTING

Let $D_i = X_i - Y_i$, we have:

Observation	1	2	3	4	...	n
Sample/Property 1 (X)	x_1	x_2	x_3	x_4	...	x_n
Sample/Property 2 (Y)	y_1	y_2	y_3	y_4	...	y_n
$D = X - Y$	d_1	d_2	d_3	d_4	...	d_n

$\Rightarrow$ Let $\mu_D = \mathbb{E}(X - Y)$, consider the test of hypothesis for one sample d_i:

H0: $\mu_D = \Delta_0$
H1: $\mu_D \neq \Delta_0$

H0: $\mu_D = \Delta_0$
H1: $\mu_D > \Delta_0$

H0: $\mu_D = \Delta_0$
H1: $\mu_D < \Delta_0$

Distribution of the Sample Differences

Assumptions

The data consists of n independently selected pairs (X_1, Y_1), $(X_2, Y_2), \ldots, (X_n, Y_n)$, with $E(X_i) = \mu_1$ and $E(Y_i) = \mu_2$. Let

$$D_1 = X_1 - Y_1, D_2 = X_2 - Y_2, \ldots, D_n = X_n - Y_n$$

So the D_i's are the differences within pairs. Then the D_i's are assumed to be normally distributed with mean μ_D and variance σ_D^2 (unknown).

Note:

$$T = \frac{\overline{D} - \mu_D}{s_D/\sqrt{n}} \sim t_{n-1}$$

HYPOTHESIS TESTING ON THE DIFFERENCE MEANS

Consider the null hypothesis: $H0: \mu_D = \Delta_0$
The test statistic:

$$T = \frac{\bar{D} - \Delta_0}{S_D / \sqrt{n}}$$

HYPOTHESIS TESTING ON THE DIFFERENCE MEANS

Consider the null hypothesis: $H0: \mu_D = \Delta_0$
The test statistic:

$$T = \frac{\bar{D} - \Delta_0}{S_D/\sqrt{n}}$$

H_1	Rejection region
$\mu_D \neq \Delta_0$	$\lvert T\rvert > t_{n-1,\alpha/2}$
$\mu_D > \Delta_0$	$T > t_{n-1,\alpha}$
$\mu_D < \Delta_0$	$T < -t_{n-1,\alpha}$

$\Rightarrow$ use t-test.

CONFIDENCE INTERVALS

The paired t CI for μ_D is

$$\overline{D} \pm t_{n-1,\alpha/2}\frac{S_D}{\sqrt{n}}$$

A one-sided confidence bound results from retaining the relevant sign and replacing $t_{\alpha/2}$ by t_α.

Normal Population + Known variances (Hypothesis Tests on the Difference in Means)

Example 2

A consumer-research organization routinely selects several car models each year and evaluates their fuel efficiency. In this year's study of two similar subcompact models from two different automakers, the average gas mileage for twelve cars of brand A was 27.2 miles per gallon. The nine brand B cars that were tested averaged 32.1 mpg. At $\alpha = 0.01$ should it conclude that brand B cars have higher average gas mileage than brand A cars do? Suppose that two populations have normal distribution with standard deviations 3.8 mpg and 4.3 mpg respectively.

Normal Population + Known variances (Hypothesis Tests on the Difference in Means)

Assumptions

1. $X_1, X_2, \ldots, X_m$ is a random sample from the normal population with mean μ_1 and variance σ_1^2.
2. $Y_1, Y_2, \ldots, Y_n$ is a random sample from the the normal population with mean μ_2 and variance σ_2^2.
3. The X and Y samples are independent of one another.
4. The variances σ_1^2 and σ_2^2 are given.

Normal Population + Known variances

Note: If both X and Y are normal then

$$Z = \frac{(\overline{X} - \overline{Y}) - (\mu_1 - \mu_2)}{\sqrt{\frac{\sigma_1^2}{m} + \frac{\sigma_2^2}{n}}} \sim N(0,1)$$

Consider the null hypothesis $H_0 : \mu_1 - \mu_2 = \Delta_0$.
The test statistic:

$$z = \frac{(\overline{x} - \overline{y}) - \Delta_0}{\sqrt{\frac{\sigma_1^2}{m} + \frac{\sigma_2^2}{n}}}$$

Then apply the following decision rule (z-test)

Normal Population + Known variances

Note: If both X and Y are normal then

$$Z = \frac{(\overline{X} - \overline{Y}) - (\mu_1 - \mu_2)}{\sqrt{\frac{\sigma_1^2}{m} + \frac{\sigma_2^2}{n}}} \sim N(0,1)$$

Consider the null hypothesis $H_0 : \mu_1 - \mu_2 = \Delta_0$.
The test statistic:

$$z = \frac{(\overline{x} - \overline{y}) - \Delta_0}{\sqrt{\frac{\sigma_1^2}{m} + \frac{\sigma_2^2}{n}}}$$

Then apply the following decision rule (z-test)

H_1	Rejection Region
$\mu_1 - \mu_2 \neq \Delta_0$	$\lvert z\rvert > z_{\alpha/2}$
$\mu_1 - \mu_2 < \Delta_0$	$z < -z_\alpha$
$\mu_1 - \mu_2 > \Delta_0$	$t > z_\alpha$

NORMAL POPULATION + UNKNOWN VARIANCES

ASSUMPTIONS

1. $X_1, X_2, \ldots, X_m$ is a random sample from the normal population with mean μ_1 and variance σ_1^2.
2. $Y_1, Y_2, \ldots, Y_n$ is a random sample from the normal population with mean μ_2 and variance σ_2^2.
3. The X and Y samples are independent of one another.
4. The variances σ_1^2 and σ_2^2 **are unknown**.

THE RULE OF THUMB

- If $\frac{s_1}{s_2} \in [0.5, 2]$, we assume that $\sigma_1^2 = \sigma_2^2$.
- If $\frac{s_1}{s_2} \notin [0.5, 2]$, we assume that $\sigma_1^2 \neq \sigma_2^2$.

Normal Population + Unknown σ + Equal Variances

- **In case** $\sigma_1^2 = \sigma_2^2 = \sigma^2$, we use the same sample variance to estimate for σ_1^2 and σ_2^2 that is S_p^2 called the pooled sample variance.

$$S_p^2 = \frac{(n-1)S_1^2 + (m-1)S_2^2}{n+m-2} \tag{1}$$

- Then, the statistic

$$T_0 = \frac{\bar{X} - \bar{Y} - (\mu_1 - \mu_2)}{S_p\sqrt{\frac{1}{n} + \frac{1}{m}}} \tag{2}$$

follows Student distribution with $n+m-2$ degrees of freedom.

Normal Population + Unknown σ + Equal Variances

Consider the null hypothesis $H_0 : \mu_1 - \mu_2 = \Delta_0$.
The test statistic:

$$t = \frac{(\overline{x} - \overline{y}) - \Delta_0}{S_p \sqrt{\frac{1}{n} + \frac{1}{m}}}$$

Then apply the following decision rule (t-test)

Normal Population + Unknown σ + Equal Variances

Consider the null hypothesis $H_0 : \mu_1 - \mu_2 = \Delta_0$.
The test statistic:

$$t = \frac{(\overline{x} - \overline{y}) - \Delta_0}{S_p \sqrt{\frac{1}{n} + \frac{1}{m}}}$$

Then apply the following decision rule (t-test)

H_1	Rejection Region
$\mu_1 - \mu_2 \neq \Delta_0$	$\lvert t \rvert > t_{\alpha/2, m+n-2}$
$\mu_1 - \mu_2 < \Delta_0$	$t < -t_{\alpha, m+n-2}$
$\mu_1 - \mu_2 > \Delta_0$	$t > t_{\alpha, m+n-2}$

Normal Population + Unknown σ + Equal Variances

Example 3

The course coordinator wants to determine if two ways of taking the course resulted in a significant difference in achievement as measured by the final exam for the course. The following table gives the scores on an examination with 45 possible points for two groups.

Online	32	37	35	28	41	44	35	31	34
Classroom	35	31	29	25	34	40	27	32	31

Do these data present sufficient evidence to indicate that the average grade for students who take the course online is significantly higher than for those who attend a conventional class? Assume that the population are both normal and the significance level $\alpha = 0.01$.

Normal Population + Unknown σ + Equal Variances σ (Confidence Interval on a Difference in Means)

If both X and Y are normal then

$$\frac{(\overline{X}-\overline{Y})-(\mu_1-\mu_2)}{se} \sim t_{m+n-2}$$

where

$$se=\sqrt{s_p^2\left(\frac{1}{m}+\frac{1}{n}\right)} \text{ and } s_p^2=\frac{(m-1)s_1^2+(n-1)s_2^2}{m+n-2}.$$

A $100(1-\alpha)\%$ confidence interval for $\mu_1-\mu_2$ is

$$(\overline{x}-\overline{y})-t_{m+n-2,\alpha/2}\times se \le \mu_1-\mu_2 \le (\overline{x}-\overline{y})+t_{m+n-2,\alpha/2}\times se$$

Normal Population + Unknown σ + Equal Variances (Hypothesis Tests on the Difference in Means)

Example 4

Ten samples of standard cement had an average weight percent calcium of $\overline{x} = 90.0$ with a sample standard deviation of $s_1 = 5.0$, and 15 samples of the lead-doped cement had an average weight percent calcium of $\overline{y} = 87.0$ with a sample standard deviation of $s_2 = 4.0$. Assume that weight percent calcium is normally distributed with same standard deviation. Find a 95% confidence interval on the difference in means, $\mu_1 - \mu_2$, for the two types of cement.

Normal Population + Unknown σ + Unequal Variances

- **when** $\sigma_1^2 \neq \sigma_2^2$, we use the test statistic

$$T_0 = \frac{\bar{X} - \bar{Y} - \Delta_0}{\sqrt{\frac{S_1^2}{n} + \frac{S_2^2}{m}}} \tag{3}$$

- T_0 follows the Student distribution with df degrees of freedom which is the **the closest integer** of

$$\nu = \frac{\left[(s_1^2/n) + (s_2^2/m)\right]^2}{\frac{(s_1^2/n)^2}{n-1} + \frac{(s_2^2/m)^2}{m-1}} \tag{4}$$

Normal Population + Unknown σ + Unequal Variances

The Two-Sample t test for testing $H_0 : \mu_1 - \mu_2 = \Delta_0$

We can test hypotheses about this difference based on the statistic

$$T = \frac{(\overline{x} - \overline{y}) - \Delta_0}{\sqrt{\frac{s_1^2}{m} + \frac{s_2^2}{n}}} \xrightarrow{t-test}$$

H_1	Rejection Region
$\mu_1 - \mu_2 \neq \Delta_0$	$\lvert t \rvert > t_{df,\alpha/2}$
$\mu_1 - \mu_2 > \Delta_0$	$t > t_{df,\alpha}$
$\mu_1 - \mu_2 < \Delta_0$	$t < -t_{df,\alpha}$

where df is the closest integer of

$$\nu = \frac{\left(\frac{s_1^2}{m} + \frac{s_2^2}{n}\right)^2}{\frac{1}{m-1}\left(\frac{s_1^2}{m}\right)^2 + \frac{1}{n-1}\left(\frac{s_2^2}{n}\right)^2}.$$

Normal Population + Unknown σ + Unequal Variances

Example 5

Arsenic concentration in public drinking water supplies is a potential health risk. The below table reported drinking water arsenic concentrations (in ppb) for 10 metropolitan Phoenix communities and 10 communities in rural Arizona. Determine if there is any difference in mean arsenic concentrations between metropolitan Phoenix communities and communities in rural Arizona. Assumed the arsenic concentration is following normal distributions and the significant level $\alpha = 0.05$.

Metro Phoenix		Rural Arizona	
Phoenix	3	Rimrock	48
Chandler	7	Goodyear	44
Gilbert	25	New River	40
Glendale	10	Apache Junction	38
Mesa	15	Buckeye	33
Paradise Valley	6	Nogales	21
Peoria	12	Black Canyon City	20
Scottsdale	25	Sedona	12
Tempe	15	Payson	1
Sun City	7	Casa Grande	18

Normal Population + Unknown σ + Unequal Variances σ (Confidence Interval on a Difference in Means)

The Two-Sample t Confidence Interval for $\mu_1 - \mu_2$

$$\overline{X} - \overline{Y} \pm t_{df,\alpha/2}\sqrt{\frac{s_1^2}{m} + \frac{s_2^2}{n}}$$

here df is the closest integer of

$$\nu = \frac{\left(\frac{s_1^2}{m} + \frac{s_2^2}{n}\right)^2}{\frac{1}{m-1}\left(\frac{s_1^2}{m}\right)^2 + \frac{1}{n-1}\left(\frac{s_2^2}{n}\right)^2}$$

A one-sided CI can be calculated as described earlier.

EXAMPLE 6

The void volume within a textile fabric affects comfort, flammability, and insulation properties. Permeability of a fabric refers to the accessibility of void space to the flow of a gas or liquid. An article gave summary information on air permeability (cm3/cm2/sec) for a number of different fabric types. Consider the following data on two different types of plain-weave fabric:

Fabric Type	Sample Size	Sample Mean	Sample Std
Cotton	10	51.71	0.79
Triacetate	10	136.14	3.59

Assuming that the porosity distributions for both types of fabric are normal, let's calculate a confidence interval for the difference between true average porosity for the cotton fabric and that for the acetate fabric, using $\gamma = 95\%$.

Large Sample Size

Assumptions

1. $X_1, X_2, \ldots, X_m$ is a random sample from the any distribution with mean μ_1 and variance σ_1^2.
2. $Y_1, Y_2, \ldots, Y_n$ is a random sample from the any distribution with mean μ_2 and variance σ_2^2.
3. The X and Y samples are independent of one another.
4. The sample sizes are enough large.

Large Sample Size

If m and n are large then

$$z = \frac{(\overline{X} - \overline{Y}) - (\mu_1 - \mu_2)}{\sqrt{\frac{S_1^2}{m} + \frac{S_2^2}{n}}} \simeq N(0,1)$$

Consider the null hypothesis: $H0 : \mu_1 - \mu_2 = \Delta_0$.
The test statistic (if the variances are known, then use σ_1 and σ_2 instead of s_1 and s_2):

$$z = \frac{(\overline{x} - \overline{y}) - \Delta_0}{\sqrt{\frac{s_1^2}{m} + \frac{s_2^2}{n}}}$$

Then apply the following decision rule (z-test)

Large Sample Size

If m and n are large then

$$z = \frac{(\overline{X} - \overline{Y}) - (\mu_1 - \mu_2)}{\sqrt{\frac{S_1^2}{m} + \frac{S_2^2}{n}}} \simeq N(0,1)$$

Consider the null hypothesis: $H0 : \mu_1 - \mu_2 = \Delta_0$.
The test statistic (if the variances are known, then use σ_1 and σ_2 instead of s_1 and s_2):

$$z = \frac{(\overline{x} - \overline{y}) - \Delta_0}{\sqrt{\frac{s_1^2}{m} + \frac{s_2^2}{n}}}$$

Then apply the following decision rule (z-test)

H_1	Rejection Region
$\mu_1 - \mu_2 \neq \Delta_0$	$\lvert z \rvert > z_{\alpha/2}$
$\mu_1 - \mu_2 < \Delta_0$	$z < -z_\alpha$
$\mu_1 - \mu_2 > \Delta_0$	$z > z_\alpha$

Large Sample Size

Example 7

To compare the average life of two brands of 9-volt batteries, a sample of 100 batteries from each brand is tested. The sample selected from the first brand shows an average life of 47 hours and a standard deviation of 4 hours. A mean life of 48 hours and a standard deviation of 3 hours are recorded for the sample from the second brand. Is the observed difference between the means of the two samples significant at the 0.01 level?

Distribution of the Difference in Proportions

Proposition 2.1

Let $\hat{p}_1 = X/m$ and $\hat{p}_2 = Y/n$, where $X \sim B(m, p_1)$ and $Y \sim B(n, p_2)$ with $X \perp Y$. Then

$$\mathbb{E}(\hat{p}_1 - \hat{p}_2) = p_1 - p_2$$

So $(\hat{p}_1 - \hat{p}_2)$ is an unbiased estimator of $(p_1 - p_2)$, and

$$\mathbb{V}ar(\hat{p}_1 - \hat{p}_2) = \frac{p_1 q_1}{m} + \frac{p_2 q_2}{n}$$

The following test statistic is distributed approximately as standard normal and is the basis of the test:

$$Z = \frac{(\hat{p}_1 - \hat{p}) - (p_1 - p_2)}{\sqrt{\frac{p_1 q_1}{m} + \frac{p_2 q_2}{n}}}$$

Large-Sample (Hypothesis Tests on the Difference in Proportion)

A Large-Sample z Test $H_0 : p_1 - p_2 = 0$

Test statistic

$$Z = \frac{(\hat{p}_1 - \hat{p}_2) - \Delta_0}{\sqrt{\overline{p}\,\overline{q}\left(\frac{1}{m} + \frac{1}{n}\right)}},$$

$$\overline{p} = \frac{m\hat{p}_1 + n\hat{p}_2}{m+n} = \frac{X+Y}{m+n}$$

H_1	Rejection
$\hat{p}_1 - \hat{p}_2 \neq 0$	$\lvert Z\rvert > z_{\alpha/2}$
$\hat{p}_1 - \hat{p}_2 > 0$	$Z > z_\alpha$
$\hat{p}_1 - \hat{p}_2 < 0$	$Z < -z_\alpha$

The test can safely be used as long as $m\hat{p}_1, m\hat{q}_1, n\hat{p}_2$, and $n\hat{q}_2$ are all at least 10.

A Large-Sample z Test $H_0 : \hat{p}_1 - \hat{p}_2 = 0$

Example 8

Extracts of St. John's Wort are widely used to treat depression. An article in the April 18, 2001, issue of the Journal of the American Medical Association compared the efficacy of a standard extract of St. John's Wort with a placebo in 200 outpatients diagnosed with major depression. Patients were randomly assigned to two groups; one group received the St. John's Wort, and the other received the placebo. After eight weeks, 19 of the placebo- treated patients showed improvement, and 27 of those treated with St. John's Wort improved. Is there any reason to believe that St. John's Wort is effective in treating major depression? Use $\alpha = 0.05$.

Large-Sample (Confidence Interval on the Difference in Proportion)

A CI for $p_1 - p_2$ is

$$(\hat{p}_1-\hat{p}_2)-z_{\alpha/2}\sqrt{\frac{\hat{p}_1\hat{q}_1}{m}+\frac{\hat{p}_2\hat{q}_2}{n}} \le p_1-p_2 \le (\hat{p}_1-\hat{p}_2)+z_{\alpha/2}\sqrt{\frac{\hat{p}_1\hat{q}_1}{m}+\frac{\hat{p}_2\hat{q}_2}{n}}$$

Large-Sample (Confidence Interval on the Difference in Proportion)

A CI for $p_1 - p_2$ is

$$(\hat{p}_1-\hat{p}_2)-z_{\alpha/2}\sqrt{\frac{\hat{p}_1\hat{q}_1}{m}+\frac{\hat{p}_2\hat{q}_2}{n}} \le p_1-p_2 \le (\hat{p}_1-\hat{p}_2)+z_{\alpha/2}\sqrt{\frac{\hat{p}_1\hat{q}_1}{m}+\frac{\hat{p}_2\hat{q}_2}{n}}$$

- This interval can safely be used as long as $m\hat{p}_1, m\hat{q}_1, n\hat{q}_2$, and $n\hat{q}_2$ are all at least 10.
- A one-sided confidence bound results from retaining the relevant sign and replacing $z_{\alpha/2}$ by z_α.
- The estimated standard deviation of $(\hat{p}_1 - \hat{p}_2)$ is different here from what it was for hypothesis testing when $\Delta_0 = 0$.

Large-Sample (Confidence Interval on the Difference in Proportion)

Example 9

Consider the process of manufacturing crankshaft bearings. Suppose that a modification is made in the surface finishing process and that, subsequently, a second random sample of 85 bearings is obtained. The number of defective bearings in this second sample is 8. Suppose that

$$m = 85, \hat{p}_1 = 10/85 = 0.1176, n = 85, \hat{p}_2 = 8/85 = 0.0941$$

Obtain an approximate 95% confidence interval on the difference in the proportion of defective bearings produced under the two processes.