HCMC University of Technology

Dung Nguyen

Probability and Statistics

Inferences Based on Two Samples

Outline I

1. Two Independent samples
   - Introduction
   - Inferences for Two Population Means

Basic Assumptions

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

If both X and Y are normal then

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

$$\mathsf{E}(\overline{X} - \overline{Y}) = \mathsf{E}(\overline{X}) - \mathsf{E}(\overline{Y}) = \mu_1 - \mu_2$$

$$\mathsf{V}(\overline{X} - \overline{Y}) = \mathsf{V}(\overline{X}) + \mathsf{V}(\overline{Y}) = \frac{\sigma_1^2}{m} + \frac{\sigma_2^2}{n}$$

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

First of all, compute a statistic

$$z = \frac{(\overline{x} - \overline{y}) - \Delta}{\mathsf{se}}, \quad \mathsf{se} = \sqrt{\frac{\sigma_1^2}{m} + \frac{\sigma_2^2}{n}}$$

Then apply the following decision rule

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

First of all, compute a statistic

$$z = \frac{(\overline{x} - \overline{y}) - \Delta}{\text{se}}, \quad \text{se} = \sqrt{\frac{\sigma_1^2}{m} + \frac{\sigma_2^2}{n}}$$

Then apply the following decision rule

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

Example 1 – Gas Mileage

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.

Solution

We test the following hypotheses

$$H_0 : \mu_1 - \mu_2 \geq 0 \quad vs. \quad H_1 : \mu_1 - \mu_2 < 0 \quad (i.e. \mu_1 < \mu_2)$$

Compute the statistic

$$z = \frac{(27.2 - 32.1) - 0}{\sqrt{\frac{3.8^2}{12} + \frac{4.3^2}{9}}} = -2.715$$

Since $c = -z_{0.01} = -2.326$ and $z < c$, we can reject H_0.

Example 2 – Drying Time

A product developer is interested in reducing the drying time of a primer paint. Two formulations of the paint are tested; formulation 1 is the standard chemistry, and formulation 2 has a new drying ingredient that should reduce the drying time. From experience, it is known that the standard deviation of drying time is 8 minutes, and this inherent variability should be unaffected by the addition of the new ingredient. Ten specimens are painted with formulation 1, and another 10 specimens are painted with formulation 2; the 20 specimens are painted in random order. The two sample average drying times are $\overline{X} = 121$ minutes and $\overline{Y} = 112$ minutes, respectively. What conclusions can the product developer draw about the effectiveness of the new ingredient, using $\alpha = 0.05$?

Solution

We test the following hypotheses

$$H_0 : \mu_1 - \mu_2 \leq 0 \quad vs. \quad H_1 : \mu_1 - \mu_2 > 0 \quad (i.e. \quad \mu_1 > \mu_2).$$

$m = n = 10$. $\overline{X} = 121, \overline{Y} = 112$. Thus $z = 2.52$.

Interpretation: We can conclude that adding the new ingredient to the paint significantly reduces the drying time.

Normal Population + Known σ (Confidence Interval on a Difference in Means)

If both X and Y are normal then

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

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

$$\mu_1 - \mu_2 = (\overline{X} - \overline{Y}) \pm z_{\alpha/2}\,\mathsf{se}$$

where

$$\mathsf{se} = \sqrt{\frac{\sigma_1^2}{m} + \frac{\sigma_2^2}{n}}$$

Example 3 – Wings of aircrafts

Tensile strength tests were performed on two different grades of aluminum spars used in manufacturing the wing of a commercial transport aircraft. From past experience with the spar manufacturing process and the testing procedure, the standard deviations of tensile strengths are assumed to be known. The data obtained are as follows: $m = 10, \overline{x} = 87.6, \sigma_1 = 1, n = 12, \overline{y} = 74.5, \sigma_2 = 1.5$. If μ_1 and μ_2 denote the true mean tensile strengths for the two grades of spars, determine a 90% confidence interval on the difference in mean strength $\mu_1 - \mu_2$.

Example 3 – Wings of aircrafts

Tensile strength tests were performed on two different grades of aluminum spars used in manufacturing the wing of a commercial transport aircraft. From past experience with the spar manufacturing process and the testing procedure, the standard deviations of tensile strengths are assumed to be known. The data obtained are as follows: $m = 10, \overline{x} = 87.6, \sigma_1 = 1, n = 12, \overline{y} = 74.5, \sigma_2 = 1.5$. If μ_1 and μ_2 denote the true mean tensile strengths for the two grades of spars, determine a 90% confidence interval on the difference in mean strength $\mu_1 - \mu_2$.

Solution

$$\mu_1 - \mu_2 = (87.6 - 74.5) + 1.645\sqrt{\frac{1^2}{10} + \frac{1.5^2}{12}} = [12.22, 13.98]$$

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

If both X and Y are normal then

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

where

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

First of all, compute a statistic

$$t=\frac{(\overline{x}-\overline{y})-\Delta}{\mathsf{se}}$$

Then apply the following decision rule

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

If both X and Y are normal then

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

where

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

First of all, compute a statistic

$$t = \frac{(\overline{x}-\overline{y})-\Delta}{\mathsf{se}}$$

Then apply the following decision rule

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

Example 4 – Online vs Classroom

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 sample population are both normal and have the same variances and the significance level $\alpha = 0.01$.

Solution

We test the following hypotheses

$$H_0 : \mu_1 - \mu_2 \leq 0 \quad vs. \quad H_1 : \mu_1 - \mu_2 > 0 \quad (i.e. \ \mu_1 > \mu_2)$$

Compute $m = 9, n = 9, \overline{x} = 35.22, \overline{y} = 31.56, s_1 = 4.94, s_2 = 4.48$. Thus

$$s^2 = \frac{(9-1)4.94^2 + (9-1)4.48^2}{9+9-2} = 22.2361$$

and the statistic

$$t = \frac{(35.22 - 31.56) - 0}{\sqrt{22.2361\left(\frac{1}{9} + \frac{1}{9}\right)}} = 1.6495$$

Since $c = t_{0.01,9+9-2} = t_{0.01,16} = 2.583 > t$, we fail reject H_0.

Example 5 – Yield from a Catalyst

Two catalysts are being analyzed to determine how they affect the mean yield of a chemical process. Specifically, catalyst 1 is currently in use, but catalyst 2 is acceptable. Since catalyst 2 is cheaper, it should be adopted, providing it does not change the process yield. A test is run in the pilot plant and results in the data shown in the table. Is there any difference between the mean yields? Use $\alpha = 0.05$, and assume equal variances.

Obs.	1	2	3	4	5	6	7	8
Cat. 1	91.50	94.18	92.18	95.39	91.79	89.07	94.72	89.21
Cat. 2	89.19	90.95	90.46	93.21	97.19	97.04	91.07	92.75

Solution

We test the following hypotheses

$$H_0 : \mu_1 - \mu_2 = 0 \quad vs. \quad H_1 : \mu_1 - \mu_2 \neq 0.$$

$\overline{x} = 92.255, \overline{y} = 92.733, \sigma_x = 2.39, \sigma_y = 2.98$. Then $s^2 = 7.30$ and $t_0 = -0.35$.
The null hypothesis cannot be rejected.

Interpretation: At 5% level of significance, we do not have strong evidence to conclude that catalyst 2 results in a mean yield that differs from the mean yield when catalyst 1 is used.

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

If both X and Y are normal then

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

where

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

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

$$\mu_1-\mu_2 = (\overline{X}-\overline{Y}) \pm t_{\alpha/2}\cdot \text{se}$$

Example 6 – Cement Hydration

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.

Example 6 – Cement Hydration

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.

Solution

$s^2 = 19.52, \implies \text{se} = \sqrt{19.62\left(\frac{1}{10} + \frac{1}{15}\right)} = 1.796.$ *Thus*

$$\mu_1 - \mu_2 = 90.0 - 87.0 \pm (2.069)(1.796)$$

Normal Population + Unknown σ + Unequal Variances

If both X and Y are normal then

$$\frac{(\overline{X}-\overline{Y})-(\mu_1-\mu_2)}{\sqrt{\frac{s_1^2}{m}+\frac{s_2^2}{n}}} \approx t_\nu$$

where

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

Normal Population + Unknown σ + Unequal Variances

If both X and Y are normal then

$$\frac{(\overline{X}-\overline{Y})-(\mu_1-\mu_2)}{\sqrt{\frac{s_1^2}{m}+\frac{s_2^2}{n}}} \approx t_\nu$$

where

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

Example 1

Determine the number of degrees of freedom for the two-sample t test or CI in each of the following situations:

Normal Population + Unknown σ + Unequal Variances

If both X and Y are normal then

$$\frac{(\overline{X}-\overline{Y})-(\mu_1-\mu_2)}{\sqrt{\frac{s_1^2}{m}+\frac{s_2^2}{n}}} \approx t_\nu$$

where

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

Example 1

Determine the number of degrees of freedom for the two-sample t test or CI in each of the following situations:

1. $m=10, n=10, s_1=5.0, s_2=6.0 \quad \Longrightarrow \nu = 17.433 \approx 17.$

Normal Population + Unknown σ + Unequal Variances

If both X and Y are normal then

$$\frac{(\overline{X}-\overline{Y})-(\mu_1-\mu_2)}{\sqrt{\frac{s_1^2}{m}+\frac{s_2^2}{n}}} \approx t_\nu$$

where

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

Example 1

Determine the number of degrees of freedom for the two-sample t test or CI in each of the following situations:

1. $m = 10, n = 10, s_1 = 5.0, s_2 = 6.0 \implies \nu = 17.433 \approx 17.$
2. $m = 10, n = 15, s_1 = 5.0, s2 = 6.0 \implies \nu = 21.711 \approx 22.$

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

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

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

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

where

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

Example 7 – Arsenic in Drinking Water

Arsenic concentration in public drinking water supplies is a potential health risk. An article in the Arizona Republic (May 27, 2001) reported drinking water arsenic concentrations (in ppb) for 10 metropolitan Phoenix communities and 10 communities in rural Arizona.

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

Determine if there is any difference in mean arsenic concentrations between metropolitan Phoenix communities and communities in rural Arizona.

Solution

We test the following hypotheses

$$H_0 : \mu_1 = \mu_2 \quad vs. \quad H_1 : \mu_1 \neq \mu_2$$

$\overline{x} = 12.5, \quad \overline{y} = 27.5, \quad s_1 = 7.63, \quad s_2 = 15.3$

$\implies t_0 = -2.77, \quad \nu = 13.2 \approx 13, \quad t_{0.025,13} = 2.160$. Thus we reject the null hypothesis.

Interpretation: We can conclude that mean arsenic concentration in the drinking water in rural Arizona is different from the mean arsenic concentration in metropolitan Phoenix drinking water.

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

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

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

where

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

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

Example 8 –

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 $\gamma = 95\%$.

Solution

$$\text{se} = \sqrt{\frac{0.79^2}{10} + \frac{3.59^2}{10}} = 1.162$$

$$df = \frac{\left(\frac{0.79^2}{10} + \frac{3.59^2}{10}\right)^2}{\frac{1}{9}\left(\frac{0.79^2}{10}\right)^2 + \frac{1}{9}\left(\frac{3.59^2}{10}\right)^2} = 9.8696$$

Thus

$$\mu_1 - \mu_2 = (51.71 - 136.14) \pm (2.228)(1.162).$$

Large Sample Size

If m and n are large then

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

First of all, compute a statistic

$$z = \frac{(\overline{x} - \overline{y}) - \Delta}{\mathsf{se}}, \quad \mathsf{se} = \sqrt{\frac{s_1^2}{m} + \frac{s_2^2}{n}}$$

Then apply the following decision rule

Large Sample Size

If m and n are large then

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

First of all, compute a statistic

$$z = \frac{(\overline{x} - \overline{y}) - \Delta}{\text{se}}, \quad \text{se} = \sqrt{\frac{s_1^2}{m} + \frac{s_2^2}{n}}$$

Then apply the following decision rule

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

Example 9 – Battery Lifetime

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?

Example 9 – Battery Lifetime

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?

Solution

We test the following hypotheses

$$H_0 : \mu_1 - \mu_2 = 0 \quad vs. \quad H_1 : \mu_1 - \mu_2 \neq 0 \quad (i.e.\ \ \mu_1 \neq \mu_2)$$

Compute the statistic

$$z = \frac{(47 - 48) - 0}{\sqrt{\frac{4^2}{100} + \frac{3^2}{100}}} = -2.$$

Since $c = z_{0.005} = 2.58$ *and* $|z| < c$, *we cannot reject* H_0.

Summary

Compute

$$z = \frac{(\overline{x} - \overline{y}) - \Delta}{\text{se}} \quad \text{or} \quad t = \frac{(\overline{x} - \overline{y}) - \Delta}{\text{se}}$$

- Test two means
  - Population distribution
    - Normal
      - Known σ^2: $\boxed{\text{se} = \sqrt{\frac{\sigma_1^2}{m} + \frac{\sigma_2^2}{n}}}$ $\boxed{\mathrm{N}(0,1)}$
      - Unknown σ^2 Equal: $\boxed{\text{se} = s\sqrt{\frac{1}{m} + \frac{1}{n}}}$ $\boxed{\mathbf{t}(m+n-2)}$
    - Any, $m, n \gg 1$: $\boxed{\text{se} = \sqrt{\frac{s_1^2}{m} + \frac{s_2^2}{n}}}$ $\boxed{\mathrm{N}(0,1)}$

where

$$s = \sqrt{\frac{(m-1)s_1^2 + (n-1)s_2^2}{m+n-2}}$$

2 Analysis of Paired Data

- Normal distribution
- Inferences for Population Proportions (Large-Sample)

Distribution of the Sample Differences

Assumptions

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

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

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.

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.

Remark

Let $D = X - Y$. Then the expected difference is

$$\mu_D = \mathrm{E}(X - Y) = \mathrm{E}(X) - \mathrm{E}(Y) = \mu_1 - \mu_2.$$

Then D_i's constitute a normal random sample with mean μ_D. Moreover,

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

CI and HT on the Difference in Means

Confidence Intervals

The paired t CI for μ_D is

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

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

CI and HT on the Difference in Means

Confidence Intervals

The paired t CI for μ_D is

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

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

Hypothesis Testing

Test statistic

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

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

Example 10 – Parallel Parking

The journal Human Factors (1962, pp. 375–380) reported a study in which 14 subjects were asked to parallel park two cars having very different wheel bases and turning radii. The time in seconds for each subject was recorded and is given in the table. Find the 90% confidence interval for $\mu_D = \mu_1 - \mu_2$.

Subject	1st car	2nd car	Difference
1	37.0	17.8	19.2
2	25.8	20.2	5.6
3	16.2	16.8	-0.6
4	24.2	41.4	-17.2
5	22.0	21.4	0.6
6	33.4	38.4	-5.0
7	23.8	16.8	7.0
8	58.2	32.2	26.0
9	33.6	27.8	5.8
10	24.4	23.2	1.2
11	23.4	29.6	-6.2
12	21.2	20.6	0.6
13	36.2	32.2	4.0
14	29.8	53.8	-24.0

Solution

From the column of observed differences, we calculate $\overline{D} = 1.21$ and $s_D = 12.68$. Thus

$$\mu_D = 1.21 \pm (1.771)(12.68)/\sqrt{14} = [-4.79, 7.21]$$

Example 11 – Zinc Concentration

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
Difference	0.015	0.028	0.177	0.121	0.102	0.107

Solution

We test the following hypotheses

$$H_0 : \mu_1 - \mu_2 \leq 0 \quad vs. \quad H_1 : \mu_1 - \mu_2 > 0.$$

$\overline{D} = 0.0917, \quad s_D = 0.0607.$ Compute the statistic

$$z = \frac{0.0917 - 0}{0.0607\sqrt{6}} = 3.6998.$$

$t_{0.05} = 2.015 \implies$ Reject H_0.

Distribution of the Difference in Proportions

Proposition

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*

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

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

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

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

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

Example 12 – Crankshaft Bearings

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.

Example 12 – Crankshaft Bearings

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.

Solution

$$\begin{aligned} p_1 - p_2 &= (0.1176 - 0.0941) \pm 1.96(0.0472) \\ &= [-0.0685, 0.1155] \end{aligned}$$

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

Although for population means the case $\Delta_0 = 0$ presented no difficulties, for population proportions $\Delta_0 = 0$ and $\Delta_0 \neq 0$ must be considered separately. Since the vast majority of actual problems of this sort involve $\Delta_0 = 0$, we'll concentrate on this case.

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

Although for population means the case $\Delta_0 = 0$ presented no difficulties, for population proportions $\Delta_0 = 0$ and $\Delta_0 \neq 0$ must be considered separately. Since the vast majority of actual problems of this sort involve $\Delta_0 = 0$, we'll concentrate on this case.

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

Test statistic

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

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.

Example 13 – St. John's Wort

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 $\alpha = 0.05$.

Example 13 – St. John's Wort

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 $\alpha = 0.05$.

Solution

$\hat{p}_1 = 27/100 = 0.27, \hat{p}_2 = 19/100 = 0.19, m = n = 100$. *Then*

$$\overline{p} = \frac{19+27}{100+100} = 0.23 \quad \textit{and} \quad SE = \sqrt{0.23(1-0.23)(1/100+1/100)} = 0.0595.$$

Since $z_0 = 1.34$ *and* $z_{0.025} = 1.96$, *we cannot reject the null hypothesis.*

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

The Paired t Confidence Interval

A CI for $p_1 - p_2$ is

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

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

The Paired t Confidence Interval

A CI for $p_1 - p_2$ is

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

- 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.

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

The Paired t Confidence Interval

A CI for $p_1 - p_2$ is

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

- 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.

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

The Paired t Confidence Interval

A CI for $p_1 - p_2$ is

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

- 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 $\Delta_0 = 0$.