Lecture 10

STATISTICAL INFERENCE FOR TWO SAMPLES

FPT Education

FPT UNIVERSITY

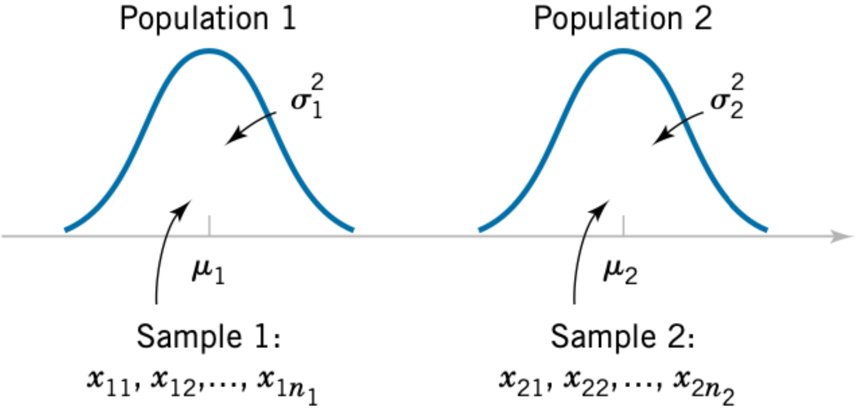
Department of Mathematics

Võ Văn Nam

Contents

1. Two independent populations

Two independent populations



Assumptions

- $X_{11}, X_{12}, ..., X_{1n_1}$ is a random sample from population 1.
- $X_{21}, X_{22}, ..., X_{2n_2}$ is a random sample from population 2.
- The two populations represented by X_1 and X_2 are independent.
- Both populations are normal.

Based on the assumptions, we may state the following.

Theorem

The quantitive

$$Z = \frac{\bar{X}_1 - \bar{X}_2 - (\mu_1 - \mu_2)}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}}$$

has a $\mathcal{N}(0, 1)$ distribution.

2. Inference on the Difference in Means of Two Normal Distributions

Confidence interval on the difference in means (variances known)

Theorem (σ_1, σ_2 known)

A $100(1-\alpha)\%$ C.I. on the difference in means $\mu_1 - \mu_2$

$$\bar{x}_1 - \bar{x}_2 - z_{\alpha/2} \cdot \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}} \leq \mu_1 - \mu_2 \leq \bar{x}_1 - \bar{x}_2 + z_{\alpha/2} \cdot \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}$$

FPT Education
FPT UNIVERSITY

One-sided Confidence interval on the difference in means (variances known)

Theorem (σ_1, σ_2 known)

- *A* $100(1-\alpha)\%$ *upper-confidence bound for* μ *is*

$$\mu_1 - \mu_2 \leq \bar{x}_1 - \bar{x}_2 + z_\alpha \cdot \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}$$

- *A* $100(1-\alpha)\%$ *lower-confidence bound for* μ *is*

$$\mu_1 - \mu_2 \geq \bar{x}_1 - \bar{x}_2 - z_\alpha \cdot \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}$$

Example

A product developer aims to reduce the drying time of a primer paint and tests two formulations: formulation 1, the standard chemistry, and formulation 2, which includes a new drying ingredient expected to decrease drying time.

It is known from prior experience that the standard deviation of drying time is 8 minutes, and this inherent variability is assumed to be unaffected by the addition of the new ingredient. Ten specimens are painted with formulation 1, and another 10 specimens with formulation 2; the 20 specimens are painted in random order. The sample average drying times are $\bar{x}_1 = 121$ minutes and $\bar{x}_2 = 112$ minutes, respectively.

Construct a 95% confidence interval based on the difference in means.

Answer. $2 \leq \mu_1 - \mu_2 \leq 16$

Hypothesis Test for Difference in Means (variances known)

Theorem (Traditional Method)

- *Step 1. Construct the two hypotheses*

$$H_0 : \mu_1 - \mu_2 = \Delta_0 \text{ vs } H_1 : \mu_1 - \mu_2 \neq \Delta_0.$$

- *Step 2. Find the test statistic* $z_0 = \dfrac{\bar{x}_1 - \bar{x}_2 - \Delta_0}{\sqrt{\dfrac{\sigma_1^2}{n_1} + \dfrac{\sigma_2^2}{n_2}}}$
- *Step 3. Identify acceptance region (use $\mathcal{N}(0,1)$).*
- *Step 4. Make a decision:*
  If z_0 is in critical region, then reject H_0.
  If z_0 is in acceptance region, then we fail to reject H_0

FPT UNIVERSITY

Hypothesis Test for Difference in Means (variances known)

Theorem (P-value Method)

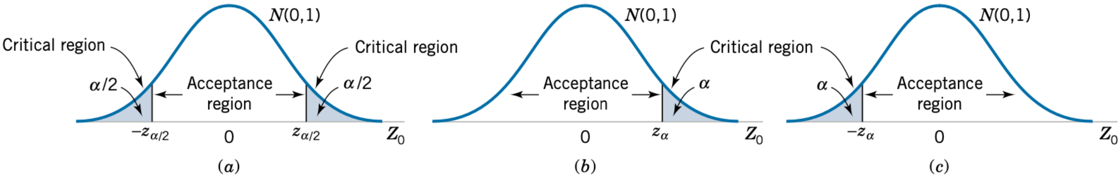
- *Step 1. Construct the two hypotheses*

$$H_0 : \mu_1 - \mu_2 = \Delta_0 \text{ vs } H_1 : \mu_1 - \mu_2 \neq \Delta_0.$$

- *Step 2. Find the test statistic* $z_0 = \dfrac{\bar{x}_1 - \bar{x}_2 - \Delta_0}{\sqrt{\dfrac{\sigma_1^2}{n_1} + \dfrac{\sigma_2^2}{n_2}}}$
- *Step 3. Find the P-value.*
- *Step 4. Make a decision:*
  If P-value $\leq \alpha$, *then reject* H_0.
  If P-value $> \alpha$, *then fail to reject* H_0

Critical regions and P-values for one-tailed tests (σ known)

- Critical regions in two-tailed, upper-tailed, and lower-tailed tests (left to right)



-

$$\text{P-value} = \begin{cases} 2\Big(1 - \Phi(|z_0|)\Big) & \text{for case} H_0 : \mu_1 - \mu_2 = \Delta_0 \text{ vs } H_1 : \mu_1 - \mu_2 \neq \Delta_0 \\ 1 - \Phi(z_0) & \text{for case } H_0 : \mu_1 - \mu_2 = \Delta_0 \text{ vs } H_1 : \mu_1 - \mu_2 > \Delta_0 \\ \Phi(z_0) & \text{for case} H_0 : \mu_1 - \mu_2 = \Delta_0 \text{ vs } H_1 : \mu_1 - \mu_2 < \Delta_0 \end{cases}$$

Example (cont')

What conclusions can the product developer draw about the effectiveness of the new ingredient, using $\alpha = 0.05$?

Hint. The hypotheses

$$H_0 : \mu_1 - \mu_2 = 0 \text{ vs } H_1 : \mu_1 - \mu_2 > 0.$$

The test statistic

$$z_0 = \frac{121 - 112 - 0}{\sqrt{\frac{8^2}{10} + \frac{8^2}{10}}} \approx 2.52.$$

- Traditional Method: acceptance region is $(-\infty, 1.645]$
- P-value Method: P-value $= 1 - P(Z < 2.52)$

FPT Education
FPT UNIVERSITY

Inference on the difference in means of two normal distributions, variances unknown (assume equal variances)

Question. What if we do NOT know population variances? (Assume equal variances)

- We need to replace population variances by **pooled variances**

$$s_p^2 = \frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{n_1 + n_2 - 2}$$

- Use t-distribution with degree of freedom

$$df = n_1 + n_2 - 2$$

FPT Education
FPT UNIVERSITY

Confidence interval on the difference in means (variances unknown)

Theorem (σ_1, σ_2 unknown)

A $100(1-\alpha)\%$ *C.I. on the difference in means* $\mu_1 - \mu_2$

$$\bar{x}_1 - \bar{x}_2 - t_{\alpha/2,df} \cdot \sqrt{\frac{s_p^2}{n_1} + \frac{s_p^2}{n_2}} \leq \mu_1 - \mu_2 \leq \bar{x}_1 - \bar{x}_2 + t_{\alpha/2,df} \cdot \sqrt{\frac{s_p^2}{n_1} + \frac{s_p^2}{n_2}}$$

One-sided Confidence interval on the difference in means (variances unknown)

Theorem (σ_1, σ_2 unknown)

- *A* $100(1-\alpha)\%$ *upper-confidence bound for* μ *is*

$$\mu_1 - \mu_2 \leq \bar{x}_1 - \bar{x}_2 + t_{\alpha,df} \cdot \sqrt{\frac{s_p^2}{n_1} + \frac{s_p^2}{n_2}}$$

- *A* $100(1-\alpha)\%$ *lower-confidence bound for* μ *is*

$$\mu_1 - \mu_2 \geq \bar{x}_1 - \bar{x}_2 - t_{\alpha,df} \cdot \sqrt{\frac{s_p^2}{n_1} + \frac{s_p^2}{n_2}}$$

Example

Two catalysts are being analyzed to determine how they affect the mean yield of a chemical. Construct a 95% confidence interval for the difference in means.

No.	Catalyst 1	Catalyst 2
1	91.50	89.19
2	94.18	90.95
3	92.18	90.46
4	95.39	93.21
5	91.79	97.19
6	89.07	97.04
7	94.72	91.07
8	89.21	92.75
$\bar{x}$	92.255	92.733
s	2.39	2.98

Hypothesis Test for Difference in Means (variances unknown)

Theorem (Traditional Method)

- *Step 1. Construct the two hypotheses*

$$H_0 : \mu_1 - \mu_2 = \Delta_0 \text{ vs } H_1 : \mu_1 - \mu_2 \neq \Delta_0.$$

- *Step 2. Find the test statistic* $t_0 = \dfrac{\bar{x}_1 - \bar{x}_2 - \Delta_0}{\sqrt{\dfrac{s_p^2}{n_1} + \dfrac{s_p^2}{n_2}}}$
- *Step 3. Identify acceptance region (use t-distribution with* $df = n_1 + n_2 - 2$*).*
- *Step 4. Make a decision:*
  If z_0 *is in critical region, then reject* H_0*.*
  If z_0 *is in acceptance region, then we fail to reject* H_0

FPT UNIVERSITY

Hypothesis Test for Difference in Means (variances unknown)

Theorem (P-value Method)

- *Step 1. Construct the two hypotheses*

$$H_0 : \mu_1 - \mu_2 = \Delta_0 \text{ vs } H_1 : \mu_1 - \mu_2 \neq \Delta_0.$$

- *Step 2. Find the test statistic* $t_0 = \dfrac{\bar{x}_1 - \bar{x}_2 - \Delta_0}{\sqrt{\dfrac{s_p^2}{n_1} + \dfrac{s_p^2}{n_2}}}$
- *Step 3. Find the P-value (use t-distribution with* $df = n_1 + n_2 - 2$*).*
- *Step 4. Make a decision:*
  If P-value $\leq \alpha$, *then reject* H_0.
  If P-value $> \alpha$, *then fail to reject* H_0

Example (cont')

Use significant level 0.05 and assume equal variances, is there any difference in the mean yields.

Hint. Hypothesis test

$$H_0 : \mu_1 - \mu_2 = 0 \text{ vs } H_1 : \mu_1 - \mu_2 \neq 0.$$

The test statistic

$$t_0 \approx -0.354$$

- Traditional Method: acceptance region $[-2.145, 2.145]$
- P-value Method: P-value 0.728

3. Inference on the Difference of Two Population Proportions

Assumption for two sample inference

Two independent random samples of size n_1 and n_2 (large enough).

Remarks.

- Sample proportion: $\hat{p}_1 = \dfrac{x_1}{n_1}$ and $\hat{p}_2 = \dfrac{x_2}{n_2}$
- $\hat{p}_1 - \hat{p}_2$ is a point estimator of $p_1 - p_2$
- If n_1 and n_2 are large enough, we have

$$\hat{p}_1 - \hat{p}_2 \sim \mathcal{N}\left(p_1 - p_2, \frac{p_1(1-p_1)}{n_1} + \frac{p_2(1-p_2)}{n_2}\right)$$

- Pooled proportion: $\bar{p} = \dfrac{x_1 + x_2}{n_1 + n_2}$

Confidence interval on the difference of two proportions

Theorem (two-sided C.I.)

A $100(1-\alpha)\%$ *confidence interval for* $(p_1 - p_2)$ *is*

$$\hat{p}_1 - \hat{p}_2 - z_{\alpha/2} \cdot \sqrt{\frac{\hat{p}_1(1-\hat{p}_1)}{n_1} + \frac{\hat{p}_2(1-\hat{p}_2)}{n_2}} \le p_1 - p_2$$
$$\le \hat{p}_1 - \hat{p}_2 + z_{\alpha/2} \cdot \sqrt{\frac{\hat{p}_1(1-\hat{p}_1)}{n_1} + \frac{\hat{p}_2(1-\hat{p}_2)}{n_2}}$$

FPT Education
FPT UNIVERSITY

Example

Extracts of St. Johns Wort are widely used to treat depression. An article in the April 18, 2001, issue of the *Journal of the American Medical Association* ("Effectiveness of St. Johns Wort on Major Depression: A Randomized Controlled Trial") compared the efficacy of a standard extract of St. Johns Wort with a placebo in 200 outpatients diagnosed with major depression. Patients were randomly assigned to two groups; one group received St. Johns Wort, and the other received a placebo. After eight weeks, 19 of the placebo-treated patients showed improvement, and 27 of those treated with St. Johns Wort improved.

Construct a 95% confidence interval for the difference between these two proportions.

Hypothesis Test for Difference in Population proportions

Theorem (Traditional Method)

- *Step 1. Form the two hypotheses*

$$H_0 : p_1 - p_2 = 0 \text{ vs } H_1 : p_1 - p_2 \neq 0.$$

- *Step 2. Find the test statistic* $z_0 = \dfrac{\hat{p}_1 - \hat{p}_2}{\sqrt{\dfrac{\bar{p}(1-\bar{p})}{n_1} + \dfrac{\bar{p}(1-\bar{p})}{n_2}}}$
- *Step 3. Identify acceptance region (use $Z = \mathcal{N}(0,1)$).*
- *Step 4. Make a decision:*
  If z_0 is in critical region, then reject H_0.
  If z_0 is in acceptance region, then we fail to reject H_0

Example (cont')

Is there any reason to believe that St. Johns Wort is effective in treating major depression? Use $\alpha = 0.05$.

Remark. We can also use P-value method to solve this problem.

Thank you!

namvv14@fe.edu.vn