ECO 391 Economics and Business Statistics

Lecture 6: Statistical Inference Concerning Two Populations

Xiaozhou Ding

February 19, 2019

Overview

1. Inference Concerning the Difference Between Two Means

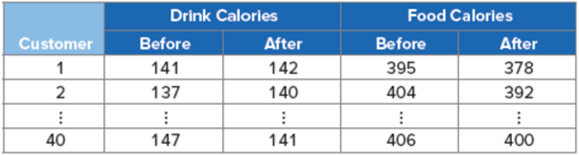
2. Make Inferences about the Mean Difference Based on Matched-Pairs Sampling

Introductory Case: Effectiveness of Mandatory Caloric Postings

- In March 2010, federal health-care law required chain restaurants with 20 locations or more to post caloric information on their menus.
- This would make it easier for consumers to select healthier food options.
- Nutritionist Molly Hosler would like to study the effects of a recent local menu ordinance requiring caloric postings in San Mateo, California.
- Molly obtains transaction data for 40 Starbucks cardholders around the time that San Mateo implemented the ordinance.

Introductory Case: Effectiveness of Mandatory Caloric Postings

- Here is some of the sample data.



Customer	Drink Calories		Food Calories	
	Before	After	Before	After
1	141	142	395	378
2	137	140	404	392
⋮	⋮	⋮	⋮	⋮
40	147	141	406	400

- Molly wants to use the sample information to:
  1. Determine whether average calories of purchased drinks declined after the passage of the ordinance.
  2. Determine whether average calories of purchased food declined after the passage of the ordinance.
  3. Assess the implications of caloric postings for Starbucks and other chains.

Inference Concerning the Difference Between Two Means

Independent Random Samples

- Two (or more) random samples are considered independent if the process that generates one sample is completely separate from the process that generates the other sample.
- The samples are clearly delineated.
- Let μ_1 be the mean of the first population, and μ_2 be the mean of the second population.

Confidence Interval for $\mu_1 - \mu_2$

- $\overline{X}_1 - \overline{X}_2$ is a point estimator for $\mu_1 - \mu_2$.
    - The values of the sample means $\bar{x}_1$ and $\bar{x}_2$ are computed from two independent random samples with n_1 and n_2 observations.
- Sampling distribution of $\overline{X}_1 - \overline{X}_2$ is assumed to be normally distributed.
    - A linear combination of normally distributed random variables is also normally distributed.
    - If underlying distribution is not normal, then by the central limit theorem, the sampling distribution of $\overline{X}_1$ and $\overline{X}_2$ is approximately normal only if both $n_1 > 30$ and $n_2 > 30$.

Three Cases in terms of Variances

- Known population variance σ_1^2 and σ_2^2. We can use z-distribution.
- Unknown population variance, but sample variance s_1^2 and s_2^2 are known and can assume they are equal. t-distribution.
- Unknown population variance, but sample variance s_1^2 and s_2^2 cannot be assumed equal. t-distribution.

Case I

If σ_1^2 and σ_2^2 are known, a $100(1-\alpha)\%$ confidence interval of the difference between two population means $\mu_1 - \mu_2$ is given by

$$\left(\overline{X}_1 - \overline{X}_2\right) \pm z_{\alpha/2}\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}.$$

Case II

If σ_1^2 and σ_2^2 are unknown but assumed equal, a $100(1-\alpha)\%$ confidence interval of the difference between two population means $\mu_1 - \mu_2$ is given by

$$\left(\overline{X}_1 - \overline{X}_2\right) \pm t_{\alpha/2,df}\sqrt{s_p^2\left(\frac{1}{n_1}+\frac{1}{n_2}\right)},$$

where s_p^2 is the polled estimate of the common variance:

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

The degrees of freedom $df = n_1 + n_2 - 2$.

Case III

If σ_1^2 and σ_2^2 are unknown and cannot be assumed to be equal, a $100(1-\alpha)\%$ confidence interval of the difference between two population means $\mu_1 - \mu_2$ is given by

$$\left(\overline{X}_1 - \overline{X}_2\right) \pm t_{\alpha/2,df}\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}},$$

where

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

is rounded down to the nearest integer.

Example

A sample of 20 cigarettes of Brand A has an average nicotine content of 1.68 milligrams with a standard deviation of 0.22 milligram; 25 cigarettes of Brand B has an average of 1.95 milligrams of nicotine with a standard deviation of 0.24 milligram. Construct the 95% confidence interval for the difference between the two population means. Nicotine content is assumed to be normally distributed. In addition, the population variances are unknown but assumed equal.

- Calculate the pooled estimate of the population variance:

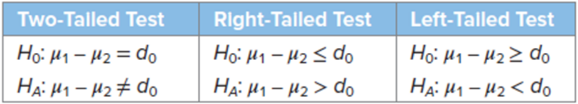
$$s_p^2 = \frac{(n_1-1)s_1^2 + (n_2-1)s_2^2}{n_1+n_2-2} = \frac{(20-1)(0.22)^2 + (25-1)(0.24)^2}{20+25-2} = 0.0535$$

- Find confidence interval:

$$(\bar{x}_1 - \bar{x}_2) \pm t_{0.025,43}\sqrt{s_p^2\left(\frac{1}{n_1}+\frac{1}{n_2}\right)} = -0.27 \pm 2.017\sqrt{0.0535\left(\frac{1}{20}+\frac{1}{25}\right)} = [-0.41, -0.13]$$

Hypothesis Test for $\mu_1 - \mu_2$

- When conducting hypothesis tests concerning $\mu_1 - \mu_2$, the competing hypotheses will take one of the following forms:

Two-Tailed Test	Right-Tailed Test	Left-Tailed Test
H_0: $\mu_1 - \mu_2 = d_0$	H_0: $\mu_1 - \mu_2 \leq d_0$	H_0: $\mu_1 - \mu_2 \geq d_0$
H_A: $\mu_1 - \mu_2 \neq d_0$	H_A: $\mu_1 - \mu_2 > d_0$	H_A: $\mu_1 - \mu_2 < d_0$

  where d_0 is the hypothesized difference between μ_1 and μ_2.
- The formulas for the test statistics are valid only if $\overline{X}_1 - \overline{X}_2$ (approximately) follows a normal distribution.

Test Statistic for the Hypothesis Tests about the Difference Between $\mu_1 - \mu_2$

1. If σ_1^2 and σ_2^2 are known, then the test statistic is assumed to follow the z-distribution and its value is calculated as

$$z = \frac{(\bar{x}_1 - \bar{x}_2) - d_0}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}}$$

Test Statistic for the Hypothesis Tests about the Difference Between $\mu_1 - \mu_2$

2. If σ_1^2 and σ_2^2 are unknown but assumed eqwual, then the test statistic is assumed to follow the t-distribution and its value is calculated as

$$t_{df} = \frac{(\bar{x}_1 - \bar{x}_2) - d_0}{\sqrt{s_p^2\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}}$$

where

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

and $df = n_1 + n_2 - 2$.

Test Statistic for the Hypothesis Tests about the Difference Between $\mu_1 - \mu_2$

2. If σ_1^2 and σ_2^2 are known, then the test statistic is assumed to follow the z-distribution and its value is calculated as

$$t_{df} = \frac{(\bar{x}_1 - \bar{x}_2) - d_0}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}}$$

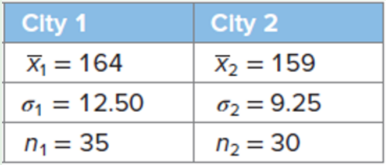
where

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

is rounded down to the nearest integer.

Example

- Claim: Average weekly food expenditure of households in City 1 is more than that of households in City 2.
- A survey of 35 households in City 1 and 30 households in City 2 obtains the following data:



City 1	City 2
$\overline{X}_1 = 164$	$\overline{X}_2 = 159$
$\sigma_1 = 12.50$	$\sigma_2 = 9.25$
$n_1 = 35$	$n_2 = 30$

- Let μ_1 be the mean weekly expenditure for City 1 and μ_2 be the mean weekly expenditure for City 2.
- The economist wishes to determine if the mean weekly food expenditure in City 1 is more than that of City 2, or $\mu_1 > \mu_2$.

This is an example of a right-tailed test:

$$H_0 : \mu_1 \leq \mu_2, \qquad H_A : \mu_1 > \mu_2$$

Since the population standard deviations are known, we compute the z-statistic:

$$z = \frac{(\bar{x}_1 - \bar{x}_2) - d_0}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}} = \frac{(164 - 159) - 0}{\sqrt{\frac{12.50^2}{35} + \frac{9.25^2}{30}}} = \frac{5}{2.70} = 1.85$$

- The p-value of the right-tailed test is computed as $p = P(Z \geq 1.85) = 1 - P(Z \leq 1.85) = 1 - 0.9678 = 0.0322$.
- Since the p-value of $0.0322 < \alpha = 0.05$ 0.05 (the significance level) we reject the null hypothesis.
- Therefore, at the 5% significance level, the economist concludes that average weekly food expenditure in City 1 is more than that of City 2.

Make Inferences about the Mean Difference Based on Matched-Pairs Sampling

Matched-Pairs Sampling

- Parameter of interest is the mean difference μ_D where $D = X_1 - X_2$, and the random variables X_1 and X_2 are matched in a pair.
- Requires that $X_1 - X_2$ is normally distributed or $n \geq 30$.
- For example, assess the benefits of a new medical treatment by evaluating the same patients before (X_1) and after (X_2) the treatment.

Recognizing a Matched-Pairs Experiment

- "Before" and "after" studies characterized by a measurement, some type of intervention, and then another measurement.
  - Example: Measuring the weight of clients before and after a diet plan.
- A pairing of observations, where it is not the same individual who gets sampled twice.
  - Example: Matching 20 adjacent plots of land using a nonorganic fertilizer on one half of the plot and an organic fertilizer on the other.

Confidence Interval for μ_D

- A $100(1-\alpha)\%$ confidence interval of the mean difference μ_D is given by

$$\bar{d} \pm t_{\alpha/2,df}\frac{SD}{\sqrt{n}}$$

  where $\bar{d}$ and SD are the mean and the standard deviation, respectively, of the n sample differences, and $df = n-1$.
- This formula is valid only if (approximately) follows a normal distribution.

Example

Manager is interested in improving productivity at a plant by changing the layout of the workstation. She measures the productivity of 10 workers before and after the change. Given the following sample statistics, construct a 95% confidence interval for the mean difference: $\bar{d} = 8.5, SD = 11.38, n = 10$.

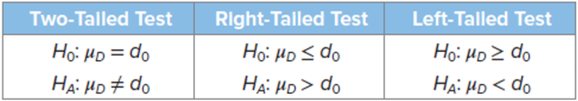
$df = 10 - 1 = 9$ and $\alpha = 0.05$.
Also, $t_{\alpha/2,df} = t_{0.025,9} = 2.262$. Confidence interval is then:

$$8.5 \pm 2.262 \times \frac{11.38}{\sqrt{10}} = 8.5 \pm 8.14.$$

Thus, the 95% confidence interval for the mean difference ranges from 0.36 to 16.64.

Hypothesis Test for μ_D

When conducting hypothesis tests concerning μ_D, the competing hypotheses will take one of the following forms:



Two-Tailed Test	Right-Tailed Test	Left-Tailed Test
H_0: $\mu_D = d_0$ H_A: $\mu_D \neq d_0$	H_0: $\mu_D \leq d_0$ H_A: $\mu_D > d_0$	H_0: $\mu_D \geq d_0$ H_A: $\mu_D < d_0$

where d_0 typically is equal to 0.

Test Statistic for Hypothesis Tests about μ_D

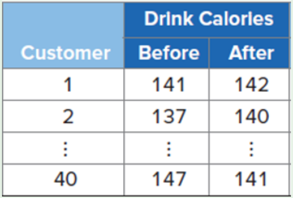
- The test statistic for hypothesis tests about μ_D is computed as

$$t_{df} = \frac{\bar{d} - d_0}{S_D/\sqrt{n}}$$

  where $df = n - 1$, $\bar{d}$ and SD are the mean and standard deviation, respectively, of the n sample differences, and d_0 is a given hypothesized mean difference.
- This formula is valid only if (approximately) follows a normal distribution.

Example

Local ordinance requires chain restaurants to post caloric information on their menus. A nutritionist wants to examine whether average drink calories declined at Starbucks after the passage of the ordinance. The nutritionist obtains transaction data for 40 Starbucks cardholders before and after the ordinance.



Customer	Drink Calories	
	Before	After
1	141	142
2	137	140
⋮	⋮	⋮
40	147	141

Can she conclude at the 5% significance level that the ordinance reduced average drink calories?

- This is a matched-pairs experiment
- Let X_1 be drink calories before and X_2 be drink calories after the ordinance.
- $D = X_1 - X_2$
- $H_0 : \mu_D \leq 0; \qquad H_A : \mu_A > 0.$

We will continue this example after the first midterm using Excel. This section will not be covered in the first midterm.