Inference about two population means

MA 116

June 2025

1. Determine H_0, H_1, test type.
2. Assume H_0 is true. Check if the distribution of $\hat{p}_1 - \hat{p}_2$ is approximately normal. ($n_i < 0.05N_i$; $n_i\hat{p}_i(1-\hat{p}_i) \geq 10$ for $i = 1$ and 2.)
3. If normal, change to a standard normal variable z using the approximation formula of $\sigma_{\hat{p}_1-\hat{p}_2}$ (Keep assuming H_0 is true so that the $\sigma_{\hat{p}_1-\hat{p}_2}$ formula is

$$\frac{\hat{p}_1 - \hat{p}_2}{\sqrt{\hat{p}(1-\hat{p})}\sqrt{\frac{1}{n_1}+\frac{1}{n_2}}}, \text{ where } \hat{p} = \frac{x_1 + x_2}{n_1 + n_2}.$$

4. Determine the critical value(s) and the critical region on the standard normal curve diagram. (Keep assuming H_0 is true.)
5. Calculate the test statistic z_0 by plugging in $\hat{p}_1$ and $\hat{p}_2$ values of our particular sample into z formula. If z_0 falls into the critical region, reject H_0.

Outline

1. Two sample data sets that are dependent, matched-pairs
2. Inference about two means: matched-pairs design
   - distribution of μ_d
   - test hypothesis
   - Confidence interval/interval estimator
3. Inference about two means: independent samples
   - population mean difference $\mu_1 - \mu_2$, distribution of variable $\overline{x_1} - \overline{x_2}$.
   - test hypothesis
   - Confidence interval/interval estimator

Independent data sets vs. Dependent data sets

Example: independent sample data sets, proportion

Investigate whether the proportion of Boston residents who drink coffee every day is approximately equal to the proportion who say they like coffee. Although both groups come from the same population, the two samples must be independently and randomly drawn. Be careful to ensure that selecting one group does not influence the selection of the other.

Example: independent sample data sets, mean

Investigate whether the average number of cups of coffee Boston residents consumed a day per person is about the same as the average number of cups of coffee New York residents consumed a day per person.

Example: dependent (matched-pair) sample data sets, mean

Investigate whether, on average, a Boston resident consumes more cups of coffee than cups of milk per day.

Matched-pair data

Require population to be quantitative.

Definition. (population mean difference μ_d)

This is a population parameter associated to the mean of difference. (Not the difference of the two means!)

In this example, we investigate whether, on average, a Boston resident consumes more cups of coffee than cups of milk per day. Our population parameter μ_d is **the population mean of (# of cups of coffee a Boston resident consumes per day-# of cups of milk a Boston resident consumes per day)**. μ_d is NOT defined to be μ_x(the population mean of # of cups of coffee a Boston resident consumes per day)-μ_y(the population mean of # of cups of milk a Boston resident consumes per day)! Even though, in this particular scenario we do have

$$\mu_d = \mu_x - \mu_y.$$

population parameter vs. sample statistic

When doing inference about two means in a matched-pair setting we only use population parameter μ_d, not μ_x or μ_y. We analyze the distribution of a new variable d, instead of x or y.

The sample statistic associated to this population parameter μ_d is $\overline{d}$.

$$\overline{d} = \frac{\sum_i (x_i - y_i)}{n} = \overline{x} - \overline{y}$$

We can talk about distributions of d and (wrt some fixed n) $\overline{d}$.

Sampling distribution of $\overline{d}$

Note. s_d is a sample statistic.

We say $t = \dfrac{\overline{d} - \mu_d}{s_d/\sqrt{n}}$ can be approximated by Student's t-distribution of $df = n - 1$ if either

1. d is approximately normally distributed, for example, when both x and y are normally distributed.
2. $n > 30$, so the distribution of t is approximately **standard** normal.

Hypothesis test of two means: matched-pairs design

$$H_0 : \mu_d = 0,\ H_1 :\ \mu_d >, <,\ \text{or}\ \neq 0.$$

Assume H_0 is true, we may verify whether the variable $t = \dfrac{\overline{d} - \mu_d}{s_d/\sqrt{n}}$ follows Student's t-distribution. If it does, we may determine the critical value(s) and the critical region on the Student's t-distribution graph with the appropriate df. Finally we calculate the test statistic

$$t_0 = \frac{\overline{d} - \mu_d}{s_d/\sqrt{n}} = \frac{\overline{d} - 0}{s_d/\sqrt{n}}.$$

If the test statistic lies in the critical region, reject H_0.

Example 11.2

Question: Matched-Pair Hypothesis Test

A company wants to test whether a new training program improves employee productivity. To investigate this, a random sample of 12 employees is selected. For each employee, the number of units produced per day is recorded **before** the training and again **after** the training. Let x_i be the number of units produced **before** training, and y_i be the number of units produced **after** training for employee i.

Assume the differences d_i are normally distributed. The sample of differences yields a sample mean $\overline{d} = 5.2$ units and a sample standard deviation $s_d = 4.8$ units. Conduct a hypothesis test at the $\alpha = 0.05$ significance level.

Solution

Hypotheses:

$$H_0\colon \mu_d = 0 \quad \text{(no improvement)}$$
$$H_1\colon \mu_d > 0 \quad \text{(productivity improves after training)}$$

Given:

$$\overline{d} = 5.2, \quad s_d = 4.8, \quad n = 12$$
$$E = \frac{s_d}{\sqrt{n}} = \frac{4.8}{\sqrt{12}} \approx 1.3856$$
$$\text{Test statistic: } t = \frac{\overline{d} - 0}{E} = \frac{5.2}{1.3856} \approx 3.75$$

Degrees of freedom: $df = n - 1 = 11$

Using a Student's t-distribution table, the critical value for a one-sided test at $\alpha = 0.05$ and $df = 11$ is approximately $t_{0.05} = 1.796$

Since $t = 3.75 > 1.796$, we **reject the null hypothesis**.

There is statistically significant evidence at the 5% level to conclude that the training program improves employee productivity.

Interval estimator

We want to construct an interval estimator for the population parameter μ_d. If the variable $t = \dfrac{\overline{d} - \mu_d}{s_d/\sqrt{n}}$ follows Student's t-distribution (there are 2 situations in which this happend), then we can use the following formula.

$$E = t_{\alpha/2}\frac{s_d}{\sqrt{n}}.$$

We claim that given a particular random sample of matched-pair data of size n, a (1-α)100% confidence interval estimator of d is $\overline{d} \pm E$, where $\overline{d}$ and s_d are sample statistics of this particumar sample.

Matched-Pair data

A new training program might improve employee productivity. Before giving this training program to every employee, a random sample of 12 employees is selected for testing purpose. For each employee, the number of units produced per day is recorded **before** the training and again **after** the training. Let x_i be the number of units produced **before** training, and y_i be the number of units produced **after** training for employee i. Assume the differences d_i are normally distributed. The sample of differences yields a sample mean $\overline{d} = 5.2$ units and a sample standard deviation $s_d = 4.8$ units. Conduct a hypothesis test at the $\alpha = 0.05$ significance level.

Suppose we want to estimate the average increasement of number of units produced, if all the employees in this company would receive this training program.

- point estimator $\overline{d} = 5.2$
- interval estimator $\overline{d} \pm E$

11.3 Inference about two means: Independent samples

Population parameter. $\mu_1 - \mu_2$.
Variable. $(\overline{x}_1 - \overline{x}_2) \mapsto t$

- Mean of this random variable is $\mu_1 - \mu_2$
- Standard deviation of this random variable is ?

Sample statistics. $\overline{x}_1 - \overline{x}_2$ (numerical value associated to a particular random sample)
Distribution. Under certain circumstance, the new variable t roughly follows the Student's t-distribution of some df.

Example: Independent samples

We want to know if an experimental drug relieves symptoms attributable to the common cold. Let μ_1 be the mean time until cold go away for anyone who (hypothetically) take the drug. Let μ_2 be the mean time until cold go away for anyone who is not taking this drug.

Want to investigate $\mu_1 - \mu_2$. To do hypothesis test, we need to know some information about the distribution of the variable $\overline{x}_1 - \overline{x}_2$. But there's an issue: population standard deviation is unknown.

Recall we've encountered the same issue in 9.2 in constructing an interval estimator for μ.

Solution

We'd need to approximate the population standard deviation of this variable in order to describe the distribution of this variable.

Welch's approximate t

Suppose a simple random sample of size n_1 is taken from a population with unknown mean μ_1. In addition, a simple random sample of size n_2 is taken from a population with unknown mean μ_2 independent to the first sample. If the two populations are normally distributed or the sample sizes are sufficiently large (n_1, $n_2 > 30$) then the new variable

$$t = \frac{(\overline{x}_1 - \overline{x}_2) - (\mu_1 - \mu_2)}{\sqrt{\dfrac{s_1^2}{n_1} + \dfrac{s_2^2}{n_2}}}$$

approximately follows Student's t-distribution with the smaller of $n_1 - 1$ or $n_2 - 1$ degrees of freedom.

Hypothesis test 11.3

$H_0 : \mu_1 = \mu_2$, $H_1 : \mu_1 >, <, = \mu_2$.

Assume that H_0 is true. Verify if

$$t = \frac{(\overline{x}_1 - \overline{x}_2) - (\mu_1 - \mu_2)}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}} = \frac{\overline{x}_1 - \overline{x}_2}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}}$$

approximately follows Student's t-distribution with the smaller of $n_1 - 1$ or $n_2 - 1$ degrees of freedom. I.e. check whether at least one of the following two conditions are met:

1. Both populations are approximately normal.
2. $n_1 > 30$, $n_2 > 30$.

Once we know that the new variable t approximately follows the Student's t-distribution with certain df, we may draw the curve of that Student's t-distribution. Based on the level of significance α and the type of the test we find the t_α in the Student's t-distribution table with the right df and find the critical value(s). On the diagram, we find the correct critical region.

Calculate the test statistic t_0 from our particular sample

Our particular sample provides us with its $\overline{x}_1$, $\overline{x}_2$, s_1, s_2–all numerical values. Plug those values into

$$t = \frac{(\overline{x}_1 - \overline{x}_2) - (\mu_1 - \mu_2)}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}} = \frac{\overline{x}_1 - \overline{x}_2}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}}$$

to get our test statistic t_0.

If t_0 lies in the critical region, **reject** H_0.

Example

We want to know if an experimental drug relieves symptoms attributable to the common cold. Let μ_1 be the mean time until cold go away for anyone who (hypothetically) take the drug. Let μ_2 be the mean time until cold go away for anyone who is not taking this drug. Do a hypothesis test at $\alpha = 0.05$.

We conduct a randomized experiment:

- 14 individuals with colds are randomly assigned to take the drug (Group 1).
- Another 14 individuals are randomly assigned to take a placebo (Group 2).

The data are assumed to come from independent normal populations. I.e. we assume the time until cold go away for people who take/not take the drug are both normally distributed; in other words, we assume x_1 and x_2 are approximately normal variables.

Sample data (given)

Group 1 (Drug): sample mean $\overline{x}_1 = 5.9$ days, sample standard deviation $s_1 = 1.2$ days.
Group 2 (Placebo): sample mean $\overline{x}_2 = 7.1$ days, sample standard deviation $s_2 = 1.5$ days.

$$H_0\colon \mu_1 = \mu_2 \quad \text{(no difference)}$$
$$H_1\colon \mu_1 < \mu_2 \quad \text{(drug reduces the cold duration)}$$

Assume H_0 is true. By our assmuption of normality of the two populations, the variable $t = \frac{\overline{x}_1 - \overline{x}_2}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}}$ approximately follows Student's t-distribution with $df = 13$. The next step is to determine the critical value(s) and the critical region.

After that, we calculate $t_0 = \frac{5.9-7.1}{0.5134} \approx \frac{-1.2}{0.5134} \approx -2.34$.

Conclusion. Because t_0 lies in the critical region, there is sufficient evidence to conclude that the new experimental drug reduces the cold duration.

Interval estimator

Recall that in Welch's t formula, the denominator is an approximation of the standard deviation of the variable $\overline{x}_1 - \overline{x}_2$. Then, it makes sense to say when Welch's t follows the Student's t-distribution of some df, a margin of error of $\mu_1 - \mu_2$ could be

$$E = t_{\alpha/2}\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}.$$

A confidence interval (or interval estimator of $\mu_1 - \mu_2$) of confidence level $(1-\alpha)100\%$ can then be given by $[\overline{x}_1 - \overline{x}_2 - E, \overline{x}_1 - \overline{x}_2 + E]$, where $\overline{x}_1 - \overline{x}_2$ is a numerical sample statistic of a particular sample.