STA 2023

Module 12
Comparing Means

Learning Objectives

Upon completing this module, you should be able to:

1. Perform inferences based on independent simple random samples to compare the means of two populations when the population standard deviations are unknown but are assumed to be equal.
2. Perform inferences based on independent simple random samples to compare the means of two populations when the population standard deviations are unknown but are not assumed to be equal.

Two Population Means

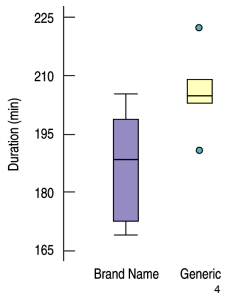
In the previous module, we learned how to obtain confidence intervals and perform hypothesis tests for one population mean.

Sometimes, however, we might want to perform a hypothesis test to decide whether the mean price of single family homes at Orlando is greater than the mean price of single family homes at Cocoa; or, we might want to find a confidence interval for the difference between the two mean prices.

In this module, we are going to learn how to perform inferences for two populations means when the population standard deviations are unknown.

Comparing Two Groups

- The natural display for comparing two groups is boxplots of the data for the two groups, placed side-by-side. For example:



4

Comparing Two Means

- Once we have examined the side-by-side boxplots, we can turn to the comparison of two means.
- The parameter of interest is the difference between the two means, $\mu_1 - \mu_2$.

5

Comparing Two Means

- Remember that, for independent random quantities, variances add.
- So, the standard deviation of the difference between two sample means is

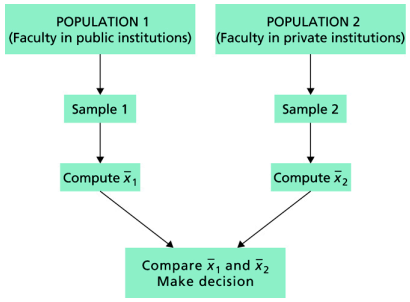
$$SD(\bar{y}_1 - \bar{y}_2) = \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}$$

- We still don't know the true standard deviations (population standard deviations) of the two groups, so we need to estimate and use the standard error.

$$SE(\bar{y}_1 - \bar{y}_2) = \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}$$

6

Comparing Two Means (Cont.)



Comparing Two Means (Cont.)

Because we are working with means and estimating the standard error of their difference using the data, we shouldn't be surprised that the sampling model is a Student's *t*.

- The confidence interval we build is called a two-sample *t*-interval (for the difference in means).
- The corresponding hypothesis test is called a two-sample hypothesis test.

Sampling Distribution for the Difference Between Two Means

- When the conditions are met, the standardized sample difference between the means of two independent groups

$$t = \frac{(\bar{y}_1 - \bar{y}_2) - (\mu_1 - \mu_2)}{SE(\bar{y}_1 - \bar{y}_2)}$$

  can be modeled by a Student's *t*-model with a number of degrees of freedom found with a special formula.
- We estimate the standard error with

$$SE(\bar{y}_1 - \bar{y}_2) = \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}$$

Assumptions and Conditions

- Independence Assumption (Each condition needs to be checked for both groups.):
  - Randomization Condition: Were the data collected with suitable randomization (representative random samples or a randomized experiment)?
  - 10% Condition: We don't usually check this condition for differences of means. We will check it for means only if we have a very small population or an extremely large sample.

Assumptions and Conditions (cont.)

- Normal Population Assumption:
  - Nearly Normal Condition: This must be checked for *both* groups. A violation by either one violates the condition.
- Independent Groups Assumption: The two groups we are comparing must be independent of each other.

Two-Sample *t*-Interval

When the conditions are met, we are ready to find the confidence interval for the difference between means of two independent groups.

The confidence interval is $(\bar{y}_1 - \bar{y}_2) \pm t^*_{df} \times SE(\bar{y}_1 - \bar{y}_2)$

where the standard error of the difference of the means is

$$SE(\bar{y}_1 - \bar{y}_2) = \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}$$

The critical value depends on the particular confidence level, *C*, that you specify and on the number of degrees of freedom, which we get from the sample sizes and a special formula.

Degrees of Freedom

- The special formula for the degrees of freedom for our *t* critical value is a bear:

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

- Because of this, we will let technology calculate degrees of freedom for us!

Testing the Difference Between Two Means

- The hypothesis test we use is the two-sample *t*-test for means.
- The conditions for the two-sample *t*-test for the difference between the means of two independent groups are the same as for the two-sample *t*-interval.

Testing the Difference Between Two Means (cont.)

We test the hypothesis $H_0: \mu_1 - \mu_2 = \mu_0$, where the hypothesized difference, μ_0, is almost always 0, using the statistic

$$t = \frac{(\bar{y}_1 - \bar{y}_2) - \Delta_0}{SE(\bar{y}_1 - \bar{y}_2)}$$

The standard error is $SE(\bar{y}_1 - \bar{y}_2) = \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}$

When the conditions are met and the null hypothesis is true, this statistic can be closely modeled by a Student's *t*-model with a number of degrees of freedom given by a special formula. We use that model to obtain a P-value.

Why Pooling the Data?

- If we are willing to *assume* that the variances of two means are equal, we can pool the data from two groups to estimate the common variance and make the degrees of freedom formula much simpler.
- We are still estimating the pooled standard deviation from the data, so we use Student's *t*-model, and the test is called a pooled *t*-test.

16

The Pooled *t*-Test

- If we assume that the variances are equal, we can estimate the common variance from the numbers we already have:

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

- Substituting into our standard error formula, we get:

$$SE_{pooled}(\bar{y}_1 - \bar{y}_2) = s_{pooled}\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}$$

- Our degrees of freedom are now $df = n_1 + n_2 - 2$.

17

The Pooled *t*-Test and Confidence Interval

- The conditions for the pooled *t*-test and corresponding confidence interval are the same as for our earlier two-sample *t* procedures, with the assumption that the variances of the two groups are the same.
- For the hypothesis test, our test statistic is

$$t = \frac{(\bar{y}_1 - \bar{y}_2) - \Delta_0}{SE_{pooled}(\bar{y}_1 - \bar{y}_2)}$$

  which has $df = n_1 + n_2 - 2$.
- Our confidence interval is

$$(\bar{y}_1 - \bar{y}_2) \pm t^*_{df} \times SE_{pooled}(\bar{y}_1 - \bar{y}_2)$$

18

Is There Ever a Time When Assuming Equal Variances Makes Sense?

- Yes. In a randomized comparative experiment, we start by assigning our experimental units to treatments at random.
- Each treatment group therefore begins with the same population variance.
- In this case assuming the variances are equal is still an assumption, and there are conditions that need to be checked, but at least it's a plausible assumption.

What Can Go Wrong?

- Watch out for paired data.
  - The Independent Groups Assumption deserves special attention.
  - If the samples are not independent, you can't use two-sample methods.
- Look at the plots.
  - Check for outliers and non-normal distributions by making and examining boxplots.

How to Perform a Pooled t-Test?

Pooled *t*-Test

Purpose To perform a hypothesis test to compare two population means, μ_1 and μ_2

Assumptions

1. Simple random samples
2. Independent samples
3. Normal populations or large samples
4. Equal population standard deviations

STEP 1 The null hypothesis is H_0: $\mu_1 = \mu_2$, and the alternative hypothesis is

H_a: $\mu_1 \neq \mu_2$ (Two tailed) or H_a: $\mu_1 < \mu_2$ (Left tailed) or H_a: $\mu_1 > \mu_2$ (Right tailed)

STEP 2 Decide on the significance level, α.

STEP 3 Compute the value of the test statistic

$$t = \frac{\bar{x}_1 - \bar{x}_2}{s_p\sqrt{(1/n_1) + (1/n_2)}}$$

where

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

Denote the value of the test statistic t_0.

How to Perform a Pooled t-Test? (Cont.)

CRITICAL-VALUE APPROACH or **P-VALUE APPROACH**

STEP 4 The critical value(s) are

$\pm t_{\alpha/2}$ (Two tailed) or $-t_\alpha$ (Left tailed) or t_α (Right tailed)

with $df = n_1 + n_2 - 2$. Use Table IV to find the critical value(s).

STEP 5 If the value of the test statistic falls in the rejection region, reject H_0; otherwise, do not reject H_0.

STEP 4 The t-statistic has $df = n_1 + n_2 - 2$. Use Table IV to estimate the P-value, or obtain it exactly by using technology.

STEP 5 If $P \le \alpha$, reject H_0; otherwise, do not reject H_0.

STEP 6 Interpret the results of the hypothesis test.

The hypothesis test is exact for normal populations and is approximately correct for large samples from nonnormal populations.

22

How to Perform a Pooled t-Interval?

Pooled *t*-Interval Procedure

Purpose To find a confidence interval for the difference between two population means, μ_1 and μ_2

Assumptions

1. Simple random samples
2. Independent samples
3. Normal populations or large samples
4. Equal population standard deviations

STEP 1 For a confidence level of $1 - \alpha$, use Table IV to find $t_{\alpha/2}$ with $df = n_1 + n_2 - 2$.

STEP 2 The endpoints of the confidence interval for $\mu_1 - \mu_2$ are

$$(\bar{x}_1 - \bar{x}_2) \pm t_{\alpha/2} \cdot s_p \sqrt{(1/n_1) + (1/n_2)}.$$

STEP 3 Interpret the confidence interval.

The confidence interval is exact for normal populations and is approximately correct for large samples from nonnormal populations.

23

What if the Variances are Not Equal?

- If the variances of two means are not equal, then the standard deviations of two means are not equal.
- In this case, we cannot pool the data from two groups, and we cannot use the pooled *t*-procedures, which require that the standard deviations (or the variances) of the two populations be equal.
- In general, if you are not sure that the populations have nearly equal standard deviations or variances, then it's always safer to use a nonpooled *t*-procedure.

24

How to Perform a Nonpooled t-Test?

Nonpooled *t*-Test

Purpose To perform a hypothesis test to compare two population means, μ_1 and μ_2

Assumptions

1. Simple random samples
2. Independent samples
3. Normal populations or large samples

STEP 1 The null hypothesis is H_0: $\mu_1 = \mu_2$, and the alternative hypothesis is

H_a: $\mu_1 \neq \mu_2$ (Two tailed) or H_a: $\mu_1 < \mu_2$ (Left tailed) or H_a: $\mu_1 > \mu_2$ (Right tailed)

STEP 2 Decide on the significance level, α.

STEP 3 Compute the value of the test statistic

$$t = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{(s_1^2/n_1) + (s_2^2/n_2)}}.$$

Denote the value of the test statistic t_0.

25

How to Perform a Nonpooled t-Test (Cont.)

CRITICAL-VALUE APPROACH or **P-VALUE APPROACH**

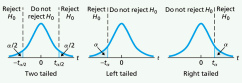
STEP 4 The critical value(s) are

$\pm t_{\alpha/2}$ (Two tailed) or $-t_\alpha$ (Left tailed) or t_α (Right tailed)

with df = Δ, where

$$\Delta = \frac{\left[(s_1^2/n_1) + (s_2^2/n_2)\right]^2}{\frac{(s_1^2/n_1)^2}{n_1 - 1} + \frac{(s_2^2/n_2)^2}{n_2 - 1}}$$

rounded down to the nearest integer. Use Table IV to find the critical value(s).



STEP 5 If the value of the test statistic falls in the rejection region, reject H_0; otherwise, do not reject H_0.

STEP 4 The t-statistic has df = Δ, where

$$\Delta = \frac{\left[(s_1^2/n_1) + (s_2^2/n_2)\right]^2}{\frac{(s_1^2/n_1)^2}{n_1 - 1} + \frac{(s_2^2/n_2)^2}{n_2 - 1}}$$

rounded down to the nearest integer. Use Table IV to estimate the P-value, or obtain it exactly by using technology.

STEP 5 If $P \leq \alpha$, reject H_0; otherwise, do not reject H_0.

STEP 6 Interpret the results of the hypothesis test.

26

How to Perform a Nonpooled t-Interval?

Nonpooled *t*-Interval Procedure

Purpose To find a confidence interval for the difference between two population means, μ_1 and μ_2

Assumptions

1. Simple random samples
2. Independent samples
3. Normal populations or large samples

STEP 1 For a confidence level of $1 - \alpha$, use Table IV to find $t_{\alpha/2}$ with df = Δ, where

$$\Delta = \frac{\left[(s_1^2/n_1) + (s_2^2/n_2)\right]^2}{\frac{(s_1^2/n_1)^2}{n_1 - 1} + \frac{(s_2^2/n_2)^2}{n_2 - 1}}$$

rounded down to the nearest integer.

STEP 2 The endpoints of the confidence interval for $\mu_1 - \mu_2$ are

$$(\bar{x}_1 - \bar{x}_2) \pm t_{\alpha/2} \cdot \sqrt{(s_1^2/n_1) + (s_2^2/n_2)}.$$

STEP 3 Interpret the confidence interval.

27

What have we learned?

We have learned to:

1. Perform inferences based on independent simple random samples to compare the means of two populations when the population standard deviations are unknown but are assumed to be equal.
2. Perform inferences based on independent simple random samples to compare the means of two populations when the population standard deviations are unknown but are not assumed to be equal.

Credit

Some of these slides have been adapted/modified in part/whole from the slides of the following textbooks.

- Weiss, Neil A., Introductory Statistics, 8th Edition
- Bock, David E., Stats: Data and Models, 3rd Edition