Introduction to Statistics

Nels Grevstad

Metropolitan State University of Denver

ngrevsta@msudenver.edu

October 27, 2019

Notes

Topics

Notes

Objectives

Objectives:

- Carry out a two-mean t test for two population means μ_1 and μ_2 when the population standard deviations σ_1 and σ_2 are unknown.

Notes

Hypothesis Test for μ_1 and μ_2 when σ_1 and σ_2 are Unknown (and Not Assumed Equal) (10.1, 10.3)

Introduction to the Two-Mean t Test

- The ***two-mean t test*** is used to decide if ***two* population means** μ_1 and μ_2 are **different**.

  It's used when we have ***two* random samples**, one from each of ***two* populations**.

  (It can also be used to compare responses to ***two* treatments** in an **experiment**.)

  The **population means** μ_1 and μ_2 and **standard deviations** σ_1 and σ_2 are **all unknown**.

Notes

Example

An engineer in a garment factory must compare two different methods for measuring the strength of polyester fibers **to decide if one method is, on average, faster than the other**.

Twelve workers are randomly assigned to two groups of **six** workers **each**.

The first group measures the strength of the fabric using **Method 1** and the second measures it using **Method 2**.

Notes

Example

The following data are the **completion times** (in **seconds**) for each group:

Method 1	Method 2
220	247
235	223
214	215
197	219
206	207
214	236

In a later example, we'll carry out a **two-mean t test** to decide **which method, if any, is faster**.

Notes

Example

A random sample of **fruits** and another of **vegetables** were taken and the **moisture content** (by percent) measured in each piece of food.

Here are the data:

Fruits		Vegetables	
Apricot	86	Artichoke	85
Banana	75	Bamboo Shoots	91
Avocado	72	Beets	88
Blackberry	88	Broccoli	89
Clementine	87	Cucumber	95
Fig	79	Iceberg Lettuce	96
Pink Grapefruit	92	Mushroom	92
Mango	84	Radish	95
		Tomato	94

Notes

In a later example, we'll carry out a **two-mean t test** to decide **which food type, if any, has the higher mean moisture content**.

Notes

- The difference $\bar{x}_1 - \bar{x}_2$ between the two **sample means** is an **estimator** of the (unknown) difference between the **population means** $\mu_1 - \mu_2$.
- The **sampling error** is:

> **Sampling Error of the Difference Between Two Sample Means**:
>
> $$\text{Sampling Error} = (\bar{x}_1 - \bar{x}_2) - (\mu_1 - \mu_2)$$

- The **sampling distribution** of $\bar{x}_1 - \bar{x}_2$ can be used to gauge how large the **sampling error** of $\bar{x}_1 - \bar{x}_2$ might be when estimating $\mu_1 - \mu_2$.

Notes

- In the slides ahead, we'll see that:
  1. When we sample from **two normal populations**, the **sampling distribution of** $\bar{x}_1 - \bar{x}_2$ will be **normal** too.
  2. Furthermore, even if we sample from **two non-normal populations** (e.g. right skewed ones), as long as the **sample sizes** n_1 **and** n_2 **are large**, the **sampling distribution of** $\bar{x}_1 - \bar{x}_2$ will be **approximately normal**.

Notes

Normality of the Sampling Distribution of $\bar{X}_1 - \bar{X}_2$ When the Samples are from Normal Populations

> **Normality of $\bar{X}_1 - \bar{X}_2$**: If we take **two samples** of sizes n_1 and n_2 independently from **two *normal* populations** whose means are μ_1 and μ_2 and whose and standard deviations are σ_1 and σ_2, then:
>
> The $\bar{x}_1 - \bar{x}_2$ **distribution** will be ***normal*** with mean $\mu_{\bar{x}_1-\bar{x}_2}$ and standard deviation $\sigma_{\bar{x}_1-\bar{x}_2}$, where
>
> $$\mu_{\bar{x}_1-\bar{x}_2} = \mu_1 - \mu_2 \quad \text{and} \quad \sigma_{\bar{x}_1-\bar{x}_2} = \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}.$$

Notes

- **Interpretation of $\mu_{\bar{x}_1-\bar{x}_2}$ and $\sigma_{\bar{x}_1-\bar{x}_2}$**:
  - $\mu_{\bar{x}_1-\bar{x}_2}$ is the value that $\bar{x}_1 - \bar{x}_2$ takes, **on average**. Thus, because $\mu_{\bar{x}_1-\bar{x}_2} = \mu_1 - \mu_2$, **on average** the **difference** between the **sample means** equals the **difference** between the **population means**.
  - $\sigma_{\bar{x}_1-\bar{x}_2}$ represents a **typical deviation** of $\bar{x}_1 - \bar{x}_2$ away from $\mu_1 - \mu_2$, i.e. a typical **sampling error**. Thus, because $\sigma_{\bar{x}_1-\bar{x}_2} = \sqrt{\sigma_1^2/n_1 + \sigma_2^2/n_2}$, the size of a **typical sampling error** is $\sqrt{\sigma_1^2/n_1 + \sigma_2^2/n_2}$.
- $\sqrt{\sigma_1^2/n_1 + \sigma_2^2/n_2}$ is often called the ***standard error*** of $\bar{x}_1 - \bar{x}_2$.

Notes

Normality of $\bar{X}_1 - \bar{X}_2$ When the Populations *Aren't* Normal but n_1 and n_2 are Large

Normality of $\bar{X}_1 - \bar{X}_2$: If we take **two samples** of sizes n_1 and n_2 independently from **two *non-normal* populations** whose means are μ_1 and μ_2 and whose and standard deviations are σ_1 and σ_2, then as long as the **sample sizes** n_1 and n_2 are ***large***:

The $\bar{x}_1 - \bar{x}_2$ **distribution** will be (at least approximately) ***normal*** with mean $\mu_{\bar{x}_1 - \bar{x}_2}$ and standard deviation $\sigma_{\bar{x}_1 - \bar{x}_2}$, where

$$\mu_{\bar{x}_1 - \bar{x}_2} \;=\; \mu_1 - \mu_2 \quad \text{and} \quad \sigma_{\bar{x}_1 - \bar{x}_2} \;=\; \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}.$$

Notes

- The **standardized version of $\bar{x}_1 - \bar{x}_2$**,

$$z \;=\; \frac{(\bar{x}_1 - \bar{x}_2) \;-\; (\mu_1 - \mu_2)}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}},$$

  follows a **standard normal** distribution.

- If we replace the **population standard deviations** σ_1 and σ_2 by the **sample standard deviations** s_1 and s_2, the resulting **standardized version of $\bar{x}_1 - \bar{x}_2$**,

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

  follows a t **distribution** (with **df** given in the slides ahead).

Notes

Carrying Out the (Nonpooled) Two-Mean t Test

- The **null hypothesis** is:

Null Hypothesis:

$$H_0: \; \mu_1 \;=\; \mu_2$$

Notes

- The **alternative hypothesis** will depend on what we're trying to **substantiate**:

Alternative Hypothesis: The alternative hypothesis will be one of

1. $H_a: \; \mu_1 \;>\; \mu_2$ (**one-sided, upper-tailed test**)
2. $H_a: \; \mu_1 \;<\; \mu_2$ (**one-sided, lower-tailed test**)
3. $H_a: \; \mu_1 \;\neq\; \mu_2$ (**two-sided, two-tailed test**)

depending on what we're trying to substantiate in our study.

Notes

- The **test statistic** for the ***(nonpooled) two-mean t test for μ_1 and μ_2*** is

Two-Mean T Test Statistic:

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

When

$$H_0 : \mu_1 = \mu_2$$

is true, the **sampling distribution** of the test statistic t is a t **distribution** with **df** given by Δ (on the next slide):

Notes

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

which we always round **down** to the nearest integer.

(The *two-mean t test* is valid if either the samples are from **normal** populations or the sample sizes n_1 and n_2 are **large** ($n_1 \geq 30$ and $n_2 \geq 30$).)

Notes

- t measures (approximately) how many **standard errors** $\bar{x}_1 - \bar{x}_2$ is **away from zero** (the claimed value for $\mu_1 - \mu_2$) .

Notes

1. Values of t **close to zero** provide almost ***no* evidence against** the **null hypothesis** $H_0 : \mu_1 = \mu_2$.

Notes

2. **Positive** values of t provide **evidence against the null hypothesis in favor of** $H_a : \mu_1 > \mu_2$.

3. **Negative** values of t provide **evidence against the null hypothesis in favor of** $H_a : \mu_1 < \mu_2$.

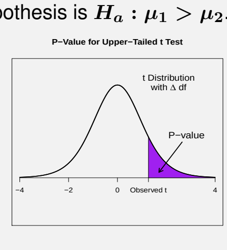
4. **Positive *and* negative** values of t provide **evidence against the null hypothesis in favor of** $H_a : \mu_1 \neq \mu_2$.

Notes

- The ***p-value*** is the probability that just by chance (under H_0) we'd get a test statistic value as far from zero, in the direction predicted by H_a, as the observed value.
- The **p-value** for the **two-mean t test** is obtained from the **t curve** with **df** given by $\boldsymbol{\Delta}$ (see the next slides).

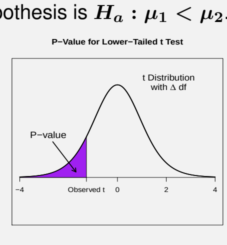
Notes

1. **P-value** = Area to the **right** of the observed t if the alternative hypothesis is $H_a : \mu_1 > \mu_2$.



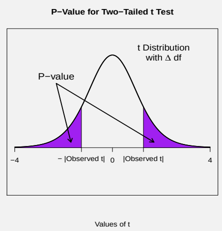
Notes

2. **P-value** = Area to the **left** of the observed t if the alternative hypothesis is $H_a : \mu_1 < \mu_2$.



Notes

3. **P-value** = Area to the **left** of $-|t|$ **and right** of $|t|$ if the alternative hypothesis is $H_a : \mu_1 \neq \mu_2$.



Notes

- As always, after choosing a **level of significance** α, we **reject** H_0 if **p-value** $< \alpha$, otherwise fail reject H_0.
- The next examples both illustrate **two-tailed** t **tests**.

  The first involves an **experiment** in which people were assigned to **two treatment groups** – the *two-mean* t *test* is valid in this context too.

  The second involves **sampling** from **two populations** (fruits and vegetables).

Notes

Example

Here (again) are the data (in seconds) from the study of two methods for measuring strength of polyester fibers.

Method 1	Method 2
220	247
235	223
214	215
197	219
206	207
214	236

Notes

The **summary statistics** for the two groups are:

Method 1	Method 2
$n_1 = 6$	$n_2 = 6$
$\bar{x}_1 = 214.3$	$\bar{x}_2 = 224.5$
$s_1 = 12.9$	$s_2 = 14.6$

We'll carry out a **two-mean** t **test** to decide **which work method, if any, is faster**.

Notes

The **hypotheses** are

$$H_0: \mu_1 = \mu_2$$
$$H_a: \mu_1 \neq \mu_2$$

where μ_1 and μ_2 are the **true (unknown) population mean completion times**.

Notes

The observed value of the **test statistic** is

$$t = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}} = \frac{214.3 - 224.5}{\sqrt{\frac{12.9^2}{6} + \frac{14.6^2}{6}}} = \mathbf{-1.28}.$$

Thus the observed difference between **sample mean** completion times, $\bar{x}_1 - \bar{x}_2 = -10.2$, is about **1.28 standard errors below zero**.

The **p-value** is the **probability** that we'd get a t value this far away from zero (in either direction) by chance **if** there was **no difference** between the **population means** μ_1 and μ_2.

Notes

The **p-value** is obtained from the **two tail areas** under the t **curve** with **df**

$$\Delta = \frac{\left(\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}\right)^2}{\frac{(s_1^2/n_1)^2}{n_1 - 1} + \frac{(s_2^2/n_2)^2}{n_2 - 1}} = \frac{\left(\frac{12.9^2}{6} + \frac{14.6^2}{6}\right)^2}{\frac{(12.9^2/6)^2}{6-1} + \frac{(14.6^2/6)^2}{6-1}} = \mathbf{9.8},$$

which we round **down** to **9**.

From the **two tails** of the t **curve** with **9 df**, to the **left** of **-1.28** and **right** of **1.28**,

$$\textbf{p-value} = 2 \times 0.116 = \mathbf{0.232}.$$

Notes

Thus we'd get a result like the one we got **23.2%** of the time **even if** the **population mean** completion times μ_1 and μ_2 were equal.

Using a **level of significance** $\alpha = 0.05$, the **decision rule** is

Reject H_0 if p-value < 0.05.
Fail to reject H_0 if p-value ≥ 0.05.

Because $0.232 \geq 0.05$, we **fail to reject** H_0.

There's **no statistically significant evidence** for any difference in the mean completion times for the two methods.

The observed difference can be explained by chance variation (sampling error).

Notes

Exercise

Here are the data (again) on **moisture contents** (by percent) measured in a random sample of **fruits** and another of **vegetables**.

Fruits		Vegetables	
Apricot	86	Artichoke	85
Banana	75	Bamboo Shoots	91
Avocado	72	Beets	88
Blackberry	88	Broccoli	89
Clementine	87	Cucumber	95
Fig	79	Iceberg Lettuce	96
Pink Grapefruit	92	Mushroom	92
Mango	84	Radish	95
		Tomato	94

Notes

We want **to decide which food type, if any, has the higher mean moisture content**.

The summary statistics for the two samples are

Fruits	Vegetables
$n_1 = 8$	$n_2 = 9$
$\bar{x}_1 = 82.88$	$\bar{x}_2 = 91.67$
$s_1 = 6.90$	$s_2 = 3.74$

Notes

Do the data provide **statistically significant** evidence, at the $\alpha = 0.05$ **level**, for **any difference** between the population mean moisture contents μ_1 and μ_2?

a) State the **hypotheses** that should be tested.

b) Calculate the **test statistic**.

Hint: You should get **-3.21**.

c) Find the **df**, Δ, and use them to determine the **p-value**.

Hint: You should get $\Delta = 10.51 \approx 10$ and **p-value = 0.010**.

Notes

d) State the **conclusion** using a **level of significance** $\alpha = 0.05$, in which case the **decision criterion** is

Reject H_0 if p-value < 0.05.
Fail to reject H_0 if p-value ≥ 0.05.

e) **Interpret** the result: Is there **statistically significant** evidence that the population mean **moisture contents** μ_1 and μ_2 **differ** for **fruits** and **vegetables**?

If so, **which type of food** has a **higher moisture content**?

Notes