Probability and Statistics

Nels Grevstad

Metropolitan State University of Denver

ngrevsta@msudenver.edu

May 6, 2019

Notes

Topics

Notes

Objectives

Objectives:

- Carry out:
  - Two-sample z test for the difference between two normal means $\mu_1 - \mu_2$ when σ_1 and σ_2 are known.
  - Two-sample z test for the difference between two general population means $\mu_1 - \mu_2$ when m and n are large.
  - Two-sample t test for the difference between two normal means $\mu_1 - \mu_2$ when σ_1 and σ_2 are unknown.

Notes

Test for the Difference Between Two Normal Means $\mu_1 - \mu_2$ when σ_1 and σ_2 are Known (9.1)

- Suppose
  1. $X_1, X_2, \ldots, X_m$ are a random sample from a **N**(μ_1, σ_1) population.
  2. $Y_1, Y_2, \ldots, Y_n$ are a random sample from a **N**(μ_2, σ_2) population.
  3. The population meansμ_1 and μ_2 **are unknown** but the standard deviations σ_1 and σ_2 **are known**.
  4. The X and Y samples are **independent** of each other.
- We'll see how to test whether μ_1 and μ_2 are different from each other.

Notes

- When σ_1 and σ_2 are known, the appropriate test is called the ***two-sample z test***
- The difference $\bar{X} - \bar{Y}$ between the two sample means is an **estimator** of the (unknown) difference between the population means $\mu_1 - \mu_2$.
- $\bar{X} - \bar{Y}$ is a difference between two **normal** random variables, so it too follows a **normal distribution**.

Notes

Normality of $\bar{X} - \bar{Y}$: If $X_1, X_2, \ldots, X_m$ is a random sample from a N(μ_1, σ_1) distribution, and $Y_1, Y_2, \ldots, Y_n$ is a random sample from a N(μ_2, σ_2) distribution, and the two samples are **independent** of each other, then

$$\bar{X} - \bar{Y} \sim \mathrm{N}\left(\mu_1 - \mu_2, \sqrt{\frac{\sigma_1^2}{m} + \frac{\sigma_2^2}{n}}\right).$$

It follows that

$$\frac{\bar{X} - \bar{Y} - (\mu_1 - \mu_2)}{\sqrt{\sigma_1^2/m + \sigma_2^2/n}} \sim \mathrm{N}(0, 1).$$

Notes

Proof: From Slides 14,

$$\bar{X} \sim \mathrm{N}\left(\mu_1, \frac{\sigma_1}{\sqrt{m}}\right) \qquad \text{and} \qquad \bar{Y} \sim \mathrm{N}\left(\mu_2, \frac{\sigma_2}{\sqrt{n}}\right).$$

Furthermore, $\bar{X} - \bar{Y}$ is a linear combination of $\bar{X}$ and $\bar{Y}$ (which are independent because the two samples are), so (also from Slides 14)

$$\bar{X} - \bar{Y} \sim \mathrm{N}\left(\mu_1 - \mu_2, \sqrt{\frac{\sigma_1^2}{m} + \frac{\sigma_2^2}{n}}\right).$$

Notes

- We're seeking to "disprove" the claim that μ_1 is equal to μ_2, so the **null hypothesis** is that they *are* equal.

Null Hypothesis:

$$H_0 : \mu_1 - \mu_2 = 0$$

(H_0 could also be written as $H_0 : \mu_1 = \mu_2$.)

Notes

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

Alternative Hypothesis: The alternative hypothesis will be one of

1. $H_a : \mu_1 - \mu_2 > 0$ (**one-sided, upper-tailed**)
2. $H_a : \mu_1 - \mu_2 < 0$ (**one-sided, lower-tailed**)
3. $H_a : \mu_1 - \mu_2 \neq 0$ (**two-sided, two-tailed**)

depending on what we're trying to verify using the data.

Notes

- The **test statistic** for the ***two-sample z test for $\mu_1 - \mu_2$*** is

Two-Sample Z Test Statistic:

$$Z = \frac{\bar{X} - \bar{Y} - 0}{\sqrt{\sigma_1^2/m + \sigma_2^2/n}}$$

When H_0 is true, $Z \sim \mathrm{N}(0, 1)$.

- Z measures how many standard errors $\bar{X} - \bar{Y}$ is away from **zero**.

Notes

- $\boldsymbol{\bar{X} - \bar{Y}}$ is an estimator of the unknown difference between population means $\boldsymbol{\mu_1 - \mu_2}$, so ...
  1. Z will be approximately **zero** if $\mu_1 - \mu_2 = 0$.
  2. It will be **positive** if $\mu_1 - \mu_2 > 0$.
  3. It will be **negative** if $\mu_1 - \mu_2 < 0$.

Notes

1. **Large positive** values of Z provide **evidence against H_0 in favor of $\boldsymbol{H_a : \mu_1 - \mu_2 > 0}$.**
2. **Large negative** values of Z provide **evidence against H_0 in favor of $\boldsymbol{H_a : \mu_1 - \mu_2 < 0}$.**
3. **Large positive *and* large negative** values of Z provide **evidence against H_0 in favor of $\boldsymbol{H_a : \mu_1 - \mu_2 \neq 0}$.**

Notes

- Recall that
$$\frac{\bar{X} - \bar{Y} - (\mu_1 - \mu_2)\boldsymbol{(\mu_1 - \mu_2)}}{\sqrt{\sigma_1^2/m + \sigma_2^2/n}} \sim \text{N}(0, 1).$$
- It follows that **if $\boldsymbol{H_0}$ is true** (so $\mu_1 - \mu_2 = 0$),
$$\frac{\bar{X} - \bar{Y} - 0\mathbf{0}}{\sqrt{\sigma_1^2/m + \sigma_2^2/n}} \sim \text{N}(0, 1).$$

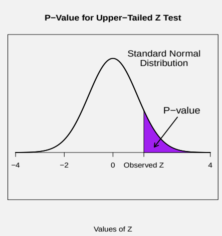
Notes

Sampling Distribution of the Test Statistic Under H_0:
If Z is the two-sample Z test statistic, then when
$$H_0: \ \mu_1 - \mu_2 \ = \ 0$$
is true,
$$Z \sim \text{N}(0, 1).$$

- 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.

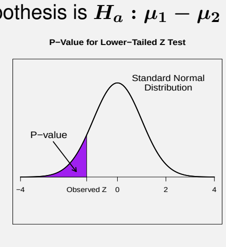
Notes

1. **P-value** = Area to the **right** of the observed z if the alternative hypothesis is $\boldsymbol{H_a : \mu_1 - \mu_2 > 0}$.



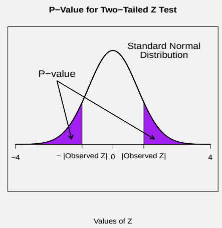
Notes

1. **P-value** = Area to the **left** of the observed z if the alternative hypothesis is $\boldsymbol{H_a : \mu_1 - \mu_2 < 0}$.



Notes

1. **P-value** = Area to the **left** of $-|z|$ **and right** of $|z|$ if the alternative hypothesis is $\boldsymbol{H_a : \mu_1 - \mu_2 \neq 0}$.



Notes

Two-Sample Z Test for $\mu_1 - \mu_2$ when σ_1 and σ_2 are Known

Assumptions: The data $x_1, x_2, \ldots, x_m$ are a random sample from a N(μ_1, σ_1) distribution and $y_1, y_2, \ldots, y_n$ are a random sample from a N(μ_2, σ_2) distribution, where σ_1 and σ_2 are known. Also, the two samples are independent.

Null hypothesis: $H_0 : \mu_1 - \mu_2 = 0$.

Test statistic value: $z = \frac{\bar{x} - \bar{y} - 0}{\sqrt{\sigma_1^2/m + \sigma_2^2/n}}$.

Decision rule: Reject H_0 if p-value $< \alpha$.

Alternative hypothesis	**P-value** = area under N$(0,1)$ distribution:
$H_a : \mu_1 - \mu_2 > 0$	to the right of z
$H_a : \mu_1 - \mu_2 < 0$	to the left of z
$H_a : \mu_1 - \mu_2 \neq 0$	to the left of $-|z|$ and right of $|z|$

Notes

Large Sample Test for the Difference Between Two General Population Means $\mu_1 - \mu_2$ (9.1)

- When the **sample sizes m and n are both large**, two things are true:
  1. Regardless of the shape of the populations, by the Central Limit Theorem,
  $$Z \;=\; \frac{\bar{X} - \bar{Y} - (\mu_1 - \mu_2)}{\sqrt{\sigma_1^2/m + \sigma_2^2/n}} \;\sim\; \text{N}(0,1) \qquad \text{(approximately)}.$$
  2. The **sample standard deviations S_1** and S_2 will remain fairly **constant** from one sample to the next, and **approximately equal to σ_1 and σ_2**.

Notes

- As a consequence, **when n is large**,
$$Z \;=\; \frac{\bar{X} - \bar{Y} - (\mu_1 - \mu_2)}{\sqrt{S_1^2/m + S_2^2/n}} \;\sim\; \text{N}(0,1) \qquad \text{(approximately)}.$$
  even if the samples are from **non-normal** populations.

  (Note that $\boldsymbol{\sigma_1}$ and $\boldsymbol{\sigma_2}$ were replaced by $\boldsymbol{S_1}$ and $\boldsymbol{S_2}$ above.)
- It follows that **if H_0 is true** (so $\mu_1 - \mu_2 = 0$),
$$\frac{\bar{X} - \bar{Y} - 0}{\sqrt{S_1^2/m + S_2^2/n}} \;\sim\; \text{N}(0,\,1) \qquad \text{(approximately)}.$$
- Thus we can use this as our **test statistic** in a ***two-sample z test for $\mu_1 - \mu_2$***.

Notes

Two-Sample Z Test Statistic:

$$Z \;=\; \frac{\bar{X} - \bar{Y} \,-\, 0}{\sqrt{S_1^2/m + S_2^2/n}}$$

When H_0 is true, $Z \sim \text{N}(0, 1)$.

Notes

Sampling Distribution of the Test Statistic Under H_0:
If Z is the two-sample Z test statistic (from the previous slide), then when

$$H_0: \;\mu_1 - \mu_2 \;=\; 0$$

is true,

$$Z \;\sim\; \text{N}(0, 1).$$

- The ***p-value*** is the appropriate tail area under the $\text{N}(0,\,1)$ curve.
- In practice, m and n are **large enough** if $m > 40$ and $n > 40$.

Notes

Two-Sample Z Test for $\mu_1 - \mu_2$ when m and n and n are Large

Assumptions: The data $x_1, x_2, \ldots, x_m$ are a random sample from *any* distribution whose mean and standard deviation are μ_1 and σ_1 and $y_1, y_2, \ldots, y_n$ are a random sample from *any* distribution whose mean and standard deviation are μ_2 and σ_2. Also, the two samples are independent of each other.

Null hypothesis: $H_0: \mu_1 - \mu_2 = 0$.

Test statistic value: $z = \frac{\bar{x} - \bar{y} - 0}{\sqrt{s_1^2/m + s_2^2/n}}$.

Decision rule: Reject H_0 if p-value $< \alpha$.

Alternative hypothesis	**P-value** = area under $\text{N}(0, 1)$ distribution:
$H_a: \mu_1 - \mu_2 > 0$	to the right of z
$H_a: \mu_1 - \mu_2 < 0$	to the left of z

Notes

Test for the Difference Between Two Normal Means $\mu_1 - \mu_2$ When σ_1 and σ_2 are Unknown (9.2)

- It can be shown that if $X_1, X_2, \ldots, X_n$ and $Y_1, Y_2, \ldots, Y_n$ are independent random samples from two **normal** populations, the random variable

$$\frac{\bar{X} - \bar{Y} - (\mu_1 - \mu_2)}{\sqrt{S_1^2/m + S_2^2/n}} \;\sim\; t(\nu),$$

a t ***distribution*** with ν ***degrees of freedom***, where

$$\nu \;=\; \frac{\left(\frac{s_1^2}{m} + \frac{s_2^2}{n}\right)^2}{\frac{(s_1^2/m)^2}{m-1} + \frac{(s_2^2/n)^2}{n-1}},$$

which should be truncated **down** to the nearest integer.

Notes

- It follows that **if H_0 is true** (so $\mu_1 - \mu_2 = 0$),

$$\frac{\bar{X} - \bar{Y} - 0}{\sqrt{S_1^2/m + S_2^2/n}} \sim t(\nu),$$

- Thus we can use this as our **test statistic** in a ***two-sample t test for $\mu_1 - \mu_2$***.

Notes

- The **test statistic** for the ***two-sample t test for $\mu_1 - \mu_2$*** is

Two-Sample t Test Statistic:

$$T = \frac{\bar{X} - \bar{Y} - 0}{\sqrt{S_1^2/m + S_2^2/n}}.$$

When H_0 is true, $T \sim t(\nu)$.

Notes

Sampling Distribution of the Test Statistic Under H_0:
If T is the two-sample t test statistic, then when

$$H_0: \ \mu_1 - \mu_2 = 0$$

is true,

$$T \sim \text{t}(\nu).$$

- The ***p-value*** is the appropriate tail area under the t(ν) curve.

Notes

Two-Sample t Test for $\mu_1 - \mu_2$

Assumptions: The data $x_1, x_2, \ldots, x_m$ are a random sample from a N(μ_1, σ_1) distribution and $y_1, y_2, \ldots, y_n$ are a random sample from a N(μ_2, σ_2) distribution. Also, the two samples are independent of each other.

Null hypothesis: $H_0 : \mu_1 - \mu_2 = 0$.

Test statistic value: $t = \frac{\bar{x} - \bar{y} - 0}{\sqrt{s_1^2/m + s_2^2/n}}$.

Decision rule: Reject H_0 if p-value $< \alpha$.

Alternative hypothesis	**P-value** = area under $t(\nu)$ distribution*:
$H_a : \mu_1 - \mu_2 > 0$	to the right of t
$H_a : \mu_1 - \mu_2 < 0$	to the left of t
$H_a : \mu_1 - \mu_2 \neq 0$	to the left of $-\lvert t \rvert$ and right of $\lvert t \rvert$

* $t(\nu)$ is the t distribution with d.f. ν given a few slides back.

Notes

Example

An engineer in a garment factory must compare two different work sequences for measuring the strength of polyester fibers **to decide if one sequence 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 **Work Sequence 1** and the second measures it using **Work Sequence 2**.

Notes

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

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

Notes

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

Work Sequence 1	Work Sequence 2
$m = 6$	$n = 6$
$\bar{x} = 214.3$	$\bar{y} = 224.5$
$s_1 = 12.9$	$s_2 = 14.6$

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

Notes

The **hypotheses** are

$$\begin{aligned} H_0: \ \mu_1 - \mu_2 &= 0 \\ H_a: \ \mu_1 - \mu_2 &\neq 0 \end{aligned}$$

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

Notes

The observed **test statistic** is

$$\begin{aligned} t &= \frac{\bar{x}-\bar{y}-0}{\sqrt{s_1^2/m+s_2^2/n}} \\ &= \frac{214.3-224.5-0}{\sqrt{12.9^2/6+14.6^2/6}} \\ &= \mathbf{-1.28}. \end{aligned}$$

Thus the observed difference between **sample mean** completion times, $\bar{x}-\bar{y} = -10.2$, 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** in the **population means** μ_1 and μ_2.

Notes

Under H_0, the test statistic would follow a $t(\nu)$ **distribution** with **degrees of freedom**

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

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

From the **two tail** areas of the $t(9)$ **distribution**, to the **left** of **-1.28** and **right** of **1.28**,

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

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 $\mathbf{0.232 \geq 0.05}$, we **fail to reject H_0**.

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

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

Notes

Notes