Biomedical Engineering Degree

3. Hypothesis Testing: Two-Sample Inference

Felipe Alonso Atienza
✉felipe.alonso@urjc.es
@FelipeURJC

Escuela Técnica Superior de Ingeniería de Telecomunicación
Universidad Rey Juan Carlos

References

1. R. Bernard. *Fundamentals of Biostatistics*. Ed.: Thompson. Chapter 8
2. B. Caffo. *Statistical Inference for Data Science*. Leanpub. Chapters $9 - 11$
3. D. Díez, M Cetinkaya-Rundel and CD Barr. *OpenIntro Statistics*. Chapter 7.

Outline

1. Introduction
2. Testing for the equality of two means
   - Paired samples
   - Independent Samples
3. Testing for the equality of two variances

Introduction

We want to compare **two populations**

$$X_1 \sim \mathcal{N}(\mu_1, \sigma_1^2) \quad \text{vs} \quad X_2 \sim \mathcal{N}(\mu_2, \sigma_2^2)$$

by using **two random samples** of n_1 and n_2 **observations** from X_1 and X_2, respectively. We are interested in:

1. Testing for the equality of two **population means**:
   - Paired or matched samples
   - Independent samples
     - Equal vs unequal variances
2. Testing for the equality of two **population variances**
   - Independent samples

Paired vs Independent samples

- Two samples are said to be **paired** when each data point in the first sample is matched and is related to a unique data point in the second sample.
  - Prices of books in *Casa del libro* vs *Amazon*
  - Longitudinal or follow-up studies, where the same group of people is followed over time.

- Two samples are said to be **independent** when the data points in one sample are unrelated to the data points in the second sample
  - Cross-sectional studies, where the participants are seen at only one point in time.

Outline

Paired t Test

- Let X_1 be a population with mean μ_1, and X_2 be a population with mean μ_2
- Suppose we have a random sample of n paired observations from these two populations and let

$$\begin{aligned} d_1 &= x_{1,1} - x_{1,2} \\ d_2 &= x_{2,1} - x_{2,2} \\ &\vdots \\ d_n &= x_{n,1} - x_{n,2} \end{aligned}$$

represent n differences with

- Sample mean: $\overline{d}$
- Quasi-standard deviation[1]: s_d

Let assume that the population of differences is normal $D \sim \mathcal{N}(\mu_d, \sigma_d)$

[1] We will use the notation s_d to refer to $s_{*,d}$

Example

i	SBP level while not using OCs (x_{i1})	SBP level while using OCs (x_{i2})	d_i^*
1	115	128	13
2	112	115	3
3	107	106	–1
4	119	128	9
5	115	122	7
6	138	145	7
7	126	132	6
8	105	109	4
9	104	102	–2
10	115	117	2

Where:

- $\overline{d} = \frac{1}{10}\sum_{i=1}^{10} d_i = 4.8$
- $s_d = \frac{1}{n-1}\sum_{i=1}^{10}(d_i - \overline{d})^2 = 5.56$

Paired t Test

- To test the hypothesis

$$H_0 : \mu_d = 0 \quad \text{vs} \quad H_1 : \mu_d \neq 0$$

with unknown σ_d with a significance level of α

Paired $t-$Test

Compute

$$t = \frac{\bar{d}}{s_d/\sqrt{n}} \sim t_{n-1}$$

- if $|t| > t_{n-1,1-\alpha/2}$, then we reject H_0
- if $|t| \leq t_{n-1,1-\alpha/2}$, then we accept H_0

Paired t Test. Confidence interval

- To test the hypothesis

$$H_0 : \mu_d = 0 \quad \text{vs} \quad H_1 : \mu_d \neq 0$$

with unknown σ_d with a significance level of α

Paired $t-$Test

Compute the confidence interval

$$\bar{d} \pm t_{n-1,1-\alpha/2} \frac{s_d}{\sqrt{n}}$$

- if $100\,\%(1-\alpha)$ CI **does not contain** 0, then we reject H_0
- if $100\,\%(1-\alpha)$ CI **does contain** 0, then we accept H_0

Example

Paired $t-$Test

Assess the statistical significance of the example data shown in the previous table. Use:

- The critical-value method
- $p-$value method

Example solution

- Using the critical-value approach, first we compute the statistic

$$t = \frac{\bar{d}}{s_d/\sqrt{n}} = \frac{4.8}{4.56/\sqrt{10}} = 3.32$$

  and then: $t_{n-1,1-\alpha/2} = $ `t(df=9).ppf(0.975)` $= 2.262$. Thus, $t \geq t_{9,0.975}$
- On the other hand, using the $p-$value approach, we calculate

$$\begin{aligned} p = 2 * P(t_9 > 3.32) &= 2 \cdot (1 - P(t_9 \leq 3.32)) = \\ &= \texttt{2*(1-t(df=9).cdf(3.32))} = 0.0089 \end{aligned}$$

  then, the results are statistically significant to reject H_0

$$\boxed{\text{We REJECT the null hypothesis } H_0 \text{ at a significance level of } 0.05}$$

$t-$Test for independent samples with equal variances

- Let X_1 be a population with mean μ_1 and variance σ_1^2
- Let X_2 be a population with mean μ_2 and variance σ_2^2

Let assume that both population are **normally distributed** with unknown but equal variances

$$\sigma^2 = \sigma_1^2 = \sigma_2^2$$

- Suppose we have a random sample of n_1 observations from X_1 and an **independent** random sample of n_2 observations from X_2
- Thus, we have access to:
  - $\bar{x}_1, s_1$, for X_1 population
  - $\bar{x}_2, s_2$, for X_2 population

Variance estimation

- Since both populations have equal variances, we can estimate σ from either n_1 or n_2 observations. Thus,
  - On the one hand, $\hat{\sigma} = s_{*,1}$
  - On the other hand, $\hat{\sigma} = s_{*,2}$
- ... or we could use both of them (weighted average):

Pooled estimate of the variance

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

$t-$Test for independent samples with equal variances

- We want to test the hypothesis

$$H_0 : \mu_1 = \mu_2 \quad \text{vs} \quad H_1 : \mu_1 \neq \mu_2$$

with unknown σ with a significance level of α

$t-$Test for independent samples with equal variances

Compute

$$t = \frac{\bar{x}_1 - \bar{x}_2}{s\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} \sim t_{n_1+n_2-2}$$

- if $|t| > t_{n_1+n_2-2,1-\alpha/2}$, then we reject H_0
- if $|t| \leq t_{n_1+n_2-2,1-\alpha/2}$, then we accept H_0

$t-$Test for independent samples with equal variances

Example

A study attempted to assess the effect of the presence of a moderator on the number of ideas generated by a group. Groups of four members, with or without moderator, were observed. For a random sample of four groups with a moderator, the mean number of ideas generated per group was 78.0, and the sample quasi-standard deviation was 24.4. For an independent sample of four groups without a moderator, the mean number of ideas generated was 63.5, and the sample quasi-standard deviation was 20.2. Assuming that the populations distributions are normal with equal variances, test the null hypothesis $(\alpha = 0.1)$ that the population means are equal against the alternative that the true mean is higher for groups with a moderator.

Example solution

- We know that $\bar{x}_1 = 78.0$, $s_1 = 24.4$, and $n_1 = 4$
- We also know that $\bar{x}_2 = 63.5$, $s_1 = 20.2$, and $n_2 = 4$
- Then,

$$s^2 = \frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{n_1 + n_2 - 2} = \frac{(4-1)24.4^2 + (4-1)20.2^2}{4+4-2} = 501.7$$

- Therefore $\boxed{s = \sqrt{501.7} = 22.4}$. Using this value, we can calculate t

$$t = \frac{\bar{x}_1 - \bar{x}_2}{s\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} = \frac{78.0 - 63.5}{22.4\sqrt{\frac{1}{4} + \frac{1}{4}}} = 0.915$$

  and then $t_{n_1+n_2-2,1-\alpha/2} = t_{4+4-2,1-0.1/2} =$ `t(df=6).ppf(0.95)` $= 1.94$
- Since $t < t_{n_1+n_2-2,1-\alpha/2}$, then

$$\boxed{\text{We ACCEPT the null hypothesis } H_0 \text{ at a significance level of } 0.1}$$

$t-$Test for independent samples with equal variances

- We want to test the hypothesis

$$H_0 : \mu_1 = \mu_2 \quad \text{vs} \quad H_1 : \mu_1 \neq \mu_2$$

with unknown σ with a significance level of α

Compute the confidence interval

$$\bar{x}_1 - \bar{x}_2 \pm t_{n_1+n_2-2,1-\alpha/2} \cdot s\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}$$

- if $100\,\%(1-\alpha)$ CI **does not contain** 0, then we reject H_0
- if $100\,\%(1-\alpha)$ CI **does contain** 0, then we accept H_0

Test for independent samples with **unequal and known** variances

- Let X_1 be a population with mean μ_1 and variance σ_1^2
- Let X_2 be a population with mean μ_2 and variance σ_2^2

Let assume that both population are **normally distributed with known** variances

- Suppose we have a random sample of n_1 observations from X_1 and an **independent** random sample of n_2 observations from X_2
- Thus, we have access to:
  - $\bar{x}_1, \sigma_1$, for X_1 population
  - $\bar{x}_2, \sigma_2$, for X_2 population

Homework!

If we want to test the hypothesis

$$H_0 : \mu_1 = \mu_2 \quad \text{vs} \quad H_1 : \mu_1 \neq \mu_2$$

with a significance level of α, **what would be test we should apply?**[a]

[a]Hint: if $\overline{X}_i \sim \mathcal{N}(\mu_i, \frac{\sigma_i}{n_i})$, $i = 1, 2$, how is distributed $\overline{X}_1 - \overline{X}_2$?

Test for independent samples with **unequal and unknown** variances

- Let X_1 be a population with mean μ_1 and variance σ_1^2
- Let X_2 be a population with mean μ_2 and variance σ_2^2

Let assume that both population are **normally distributed with unknown** variances

- Suppose we have a random sample of n_1 observations from X_1 and an **independent** random sample of n_2 observations from X_2
- Thus, we have access to:
  - $\bar{x}_1, s_1$, for X_1 population
  - $\bar{x}_2, s_2$, for X_2 population

$t-$Test for independent samples with unknown and unequal variances

- We want to test the hypothesis

$$H_0 : \mu_1 = \mu_2 \quad \text{vs} \quad H_1 : \mu_1 \neq \mu_2$$

with unknown σ_1, σ_2 and a significance level of α

$t-$Test for independent samples with unknown and unequal variances

Compute

$$t = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}} \sim t_d \text{ (Satterthwaite approximation)}$$

- if $|t| > t_{d,1-\alpha/2}$, then we reject H_0
- if $|t| \leq t_{d,1-\alpha/2}$, then we accept H_0

where

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

Outline

Testing for the equality of two variances

- Let X_1 be a population with mean μ_1 and variance σ_1^2
- Let X_2 be a population with mean μ_2 and variance σ_2^2

Let assume that both population are **normally distributed**

- Suppose we have a random sample of n_1 observations from X_1 and an **independent** random sample of n_2 observations from X_2
- Thus, we have access to:
  - $\bar{x}_1, s_1$, for X_1 population
  - $\bar{x}_2, s_2$, for X_2 population

Testing for the equality of two variances

- We want to test the hypothesis

$$H_0 : \sigma_1 = \sigma_2 \quad \text{vs} \quad H_1 : \sigma_1 \neq \sigma_2$$

with a significance level of α

Testing for the equality of two variances

Compute the statistic

$$f = \frac{s_1^2}{s_2^2} \sim F_{n_1-1,n_2-1} \text{ (F distribution)}$$

- if $f > F_{n_1-1,n_2-1,1-\alpha/2}$, or $f < F_{n_1-1,n_2-1,\alpha/2}$ then we reject H_0
- Otherwise we accept H_0

F-distribution

- if $x_1, x_2, \ldots, x_n$ and $y_1, y_2, \ldots, y_m$ denote r.v.'s following a $\mathcal{N}(0,1)$ distribution. Then, the r.v.

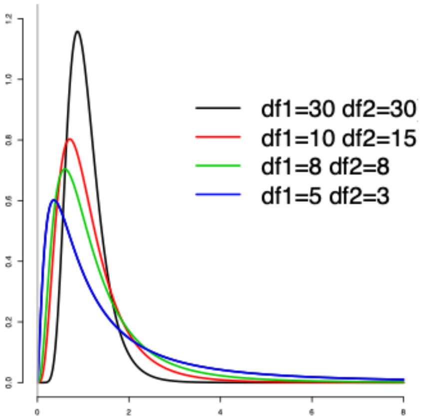
$$F = \frac{\frac{1}{n}\sum_{i=1}^{n} x_i^2}{\frac{1}{m}\sum_{i=1}^{m} y_i^2}$$

  follows a $\mathbf{F_{n,m}}$ **with** $\mathbf{n}$ **and** $\mathbf{m}$ **degrees of freedom**

- We can view it as a ratio of two normalized chi-square r.v.'s

$$\frac{s_1^2}{s_2^2} = \frac{\frac{1}{n_1-1}\overbrace{\frac{(n_1-1)s_1^2}{\sigma^2}}^{\chi^2_{n_1-1}}}{\frac{1}{n_2-1}\underbrace{\frac{(n_2-1)s_2^2}{\sigma^2}}_{\chi^2_{n_2-1}}} \sim F_{n_1-1,n_2-1}$$

F-distribution



Testing for the equality of two variances

Example

For a random sample of 17 newly issued AAA-rated industrial bonds, the quasi-variance of maturities (in years squared) was 123.35. For an independent random sample of 11 issued CCC-rated industrial bonds, the quasi-variance of maturities was 8.02. If the respective population variances are denoted σ_1 and σ_2, perform a two-sided test at a $5\,\%$ level.

Example solution

- Calculate the f statistic

$$f = \frac{s_1^2}{s_2^2} = \frac{123.35}{8.02} = 15.38 \sim F_{16,10}$$

where

- $F_{16,10,1-\alpha/2} = \texttt{f(16,10).ppf(0.975)} = 3.496$
- $F_{16,10,\alpha/2} = \texttt{f(16,10).ppf(0.025)} = 0.335$

and since $f > F_{16,10,1-\alpha/2}$ then

$$\boxed{\text{We REJECT the null hypothesis } H_0 \text{ at a significance level of } 0.05}$$