Introduction to Statistical Inference

Marta Bofill Roig

March 11, 2025

GRBIO
UNIVERSITAT POLITÈCNICA DE CATALUNYA
BARCELONATECH
UPC

Outline

Warming up

Estimation

Confidence intervals

Statistical Inference and confidence intervals

Warming up

Population and Sample

The aim is to understand a phenomena

Research questions:

- will the new treatment for hypertension work better than the standard one?
- which are risk factors for IC admission due to COVID-19?
- are genetic factors related to the onset of breast cancer?
- ...

Population and Sample (cont'd)

The **target population** is the precise definition of the total group of individuals for whom we want to draw conclusions.

- This is achieved by formulating the **inclusion criteria** for the study

Ideally, we collect data from the whole population (i.e., from all subjects), and proceed to analyze them:

- data are actually the realizations from the random variables of interest,
- e.g., blood pressure measurements

Population and Sample (cont'd)

When all subjects from the population have the same chance to be included in the sample we obtain a **random sample**.

- Such a sample is guaranteed to provide us with valid statements about the target population.

Population and Sample (cont'd)

To be able to make generalizations from our sample, we want it to be sufficiently **representative** of the target population.

A **representative sample** is a group of subjects from the target population that adequately replicates the population according to whatever characteristic or quality is under study.

A representative sample parallels key variables and characteristics of the larger population.

Statistical inference refers to the use of statistics to draw conclusions about an unknown aspect of a population based on a random sample.

Estimation and Sampling Variability

Sampling Error: There will be a difference between the characteristic we measure in the sample and the same characteristic in the population.

Sampling variability is the variability in the analysis results caused by the fact that we work with the sample and not the whole population.

Estimation and Sampling Variability (cont'd)

The **estimand** is the parameter of the target population we wish to estimate from a sample.

Example: the mean of the blood pressure μ.

Estimation and Sampling Variability (cont'd)

Example:

- let's assume that we will obtain a representative sample from this population of size n (i.e., the number of patients in our sample)
- each subject in the sample has a random variable X_i describing his/her blood pressure levels
- we could then estimate μ using the sample mean

$$\bar{X} = \frac{1}{n} \sum_{i=1}^{n} X_i$$

The **estimator** is a rule for estimating the parameter in the population using the data we will collect in a sample.

Estimation and Sampling Variability (cont'd)

Example:

- when we have available specific values x_i from a **realized sample**, we calculate the realized value of the sample mean

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$$

where x_i denotes the blood pressure measurements for patient i

The estimate of a particular population characteristic we obtain from a specific sample using an estimator is called the **point estimate**.

Estimation and Sampling Variability (cont'd)

But what would be the variability of this point estimate in all different samples of size n from this target population?

Standard error of the sample mean

$$se(\bar{X}) = \frac{\sigma}{\sqrt{n}}$$

where σ is the standard deviation of blood pressure values in our population.

Estimation

Estimation, Inference, and Sampling Distributions

Key concepts:

- **Population and Sample:** We aim to learn about a population by analysing random samples drawn from it.
- **Randomness:** The values in the sample are random variables due to chance, arising from sampling and biological variation.
- **Probability:** Statements about the population will involve probabilities due to the inherent randomness.

In essence, statistical inference uses sample data and probability to make conclusions about a population.

Estimation

- Estimation involves assigning likely values to unknown population parameters based on sample data.
- Examples:
  - Sample mean ($\bar{Y}$) as an estimator for population mean (μ).
  - Sample variance (s^2) as an estimator for population variance (σ^2).

Estimation (cont'd)

For a given parameter, there can be **multiple possible estimators**.

- Example: Estimating population variance (σ^2).
  - Standard sample variance:

$$s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(Y_i - \bar{Y})^2$$

  - Alternative estimator:

$$s_*^2 = \frac{1}{n}\sum_{i=1}^{n}(Y_i - \bar{Y})^2$$

  - The alternative uses n instead of $n-1$ as the divisor.

Estimation (cont'd)

If multiple estimators exist for a parameter, the obvious question is: how do we choose the best one?

- Example: Choosing between different estimators for the population variance.

Estimation (cont'd)

- Estimators like $\bar{Y}$ and s^2 have their own underlying populations with probability distributions.
- Consider the population of all possible $\bar{Y}$ values from all possible samples of size n.

Estimation (cont'd)

- The mean of the population of $\bar{Y}$ values is equal to the population mean μ:

$$\text{Mean of population of } \bar{Y} = \mu_{\bar{Y}} = \mu$$

  in other words, $E(\bar{Y}) = \mu$.

- This means $\bar{Y}$ is an **unbiased** estimator of μ.

- Unbiasedness is a desirable property: The estimator's average value matches the true parameter.

Estimation (cont'd)

- If the mean of the estimator's probability distribution equals the parameter being estimated, the estimator is **unbiased**.
- Example: $\bar{Y}$ (sample mean) is an unbiased estimator of μ (population mean).
- Unbiasedness is a key criterion for choosing among competing estimators.

Estimation (cont'd)

Example:

- We have two estimators for population variance σ^2:
  - $s^2 = \frac{1}{n-1} \sum_{i=1}^{n} (Y_i - \bar{Y})^2$
  - $s_*^2 = \frac{1}{n} \sum_{i=1}^{n} (Y_i - \bar{Y})^2$

Estimation (cont'd)

- s_*^2 (dividing by n) has a mean of $\left(\frac{n-1}{n}\right)\sigma^2$, not σ^2.
- s_*^2 is **biased downward**.
- It underestimates σ^2, especially for small n.

Estimation (cont'd)

- s^2 (dividing by $n-1$) is designed to be an **unbiased** estimator.
- Dividing by $n-1$ corrects the bias.
- Therefore, s^2 is the accepted estimator for σ^2 due to its unbiasedness.

Estimation (cont'd)

- If two estimators are both unbiased, how do we choose the better one?
- Unbiasedness is desirable, but not the only criterion.

Estimation (cont'd)

- We want estimators to be "close" to the true value.
- This means we want the estimator's probability distribution to have **small variance**.
- Small variance means the estimator's values vary little across different samples.

Estimation (cont'd)

- If two estimators are unbiased, choose the one with the **smaller variance**.
- This ensures the estimator is more consistent and precise.

Estimators for population mean and variance

It turns out that for normally distributed data Y, the estimators $\bar{Y}$ (for μ) and s (for σ^2) have this desirable property.

We use $\bar{Y}$ and s as estimators for population mean and variance because they have the desirable properties of **unbiasedness and minimum variance**.

Confidence intervals

Confidence interval

Estimators

- Estimators like $\bar{Y}$ and s^2 provide "likely" values for population parameters μ and σ^2.
- Due to chance, they are rarely exactly equal to the true parameters.
- However, "good" estimators tend to be "close" to the true values.

Confidence interval (cont'd)

Interval Estimates

- Instead of reporting a single value, we report an **interval** based on the estimator.
- This interval is likely to contain the true parameter value.
- "Likely" implies that **probability** is involved.

Here, we discuss the notion of such an interval, known as a **confidence interval**.

Confidence interval for μ

Example: we wish to estimate μ by the sample mean $\bar{Y}$.

- Assume $Y \sim N(\mu, \sigma^2)$, where μ and σ^2 are unknown.
- We have a random sample $Y_1, \ldots, Y_n$.
- Our estimator for μ is the sample mean, $\bar{Y}$.

Confidence interval for μ (cont'd)

- To make probability statements about $\bar{Y}$, we need to account for the unknown σ^2.
- We must estimate σ^2 even if it's not our primary interest.
- We use the statistic:

$$\frac{\bar{Y} - \mu}{s_{\bar{Y}}}$$

  where $s_{\bar{Y}}$ is the estimated standard error of $\bar{Y}$.

Confidence interval for μ (cont'd)

- We want to quantify the uncertainty of estimating the fixed value μ using the random variable $\bar{Y}$.
- The randomness arises from the random sample.
- Probability statements reflect the uncertainty of trying to get an understanding of μ using $\bar{Y}$.

Confidence interval for μ (cont'd)

Wait a minute, here's one more detail about probability distributions!

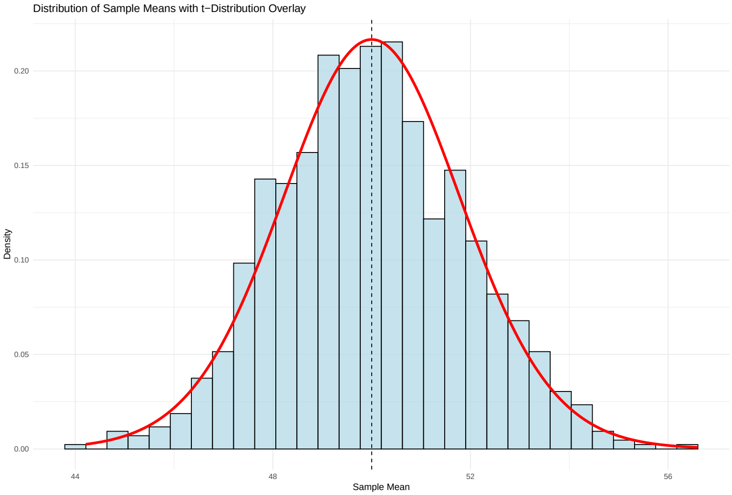
- Let $t_{n-1,\alpha/2}$ be the point such that the shaded region under the Student's t density with $(n-1)$ degrees of freedom has area $\alpha/2$.

- Note that then unshaded region has area $(1-\alpha)/2)$.

- By symmetry, $t_{n-1,\alpha/2}$ is such that each shaded region has area $\alpha/2$, with $1-\alpha$ in the middle.

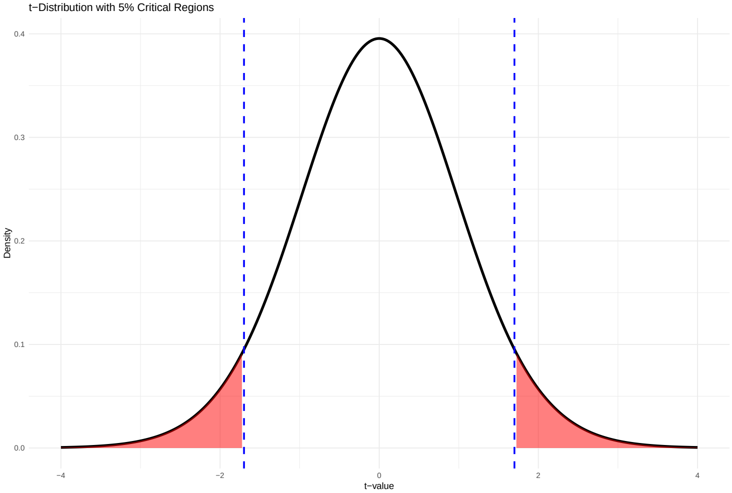
Confidence interval for μ (cont'd)

Wait a minute, here's one more detail about probability distributions!



Confidence interval for μ (cont'd)

Wait a minute, here's one more detail about probability distributions!



Confidence interval for μ (cont'd)

- We aim to create an interval that likely contains the true population mean μ.
- The formula is derived from the t-distribution:

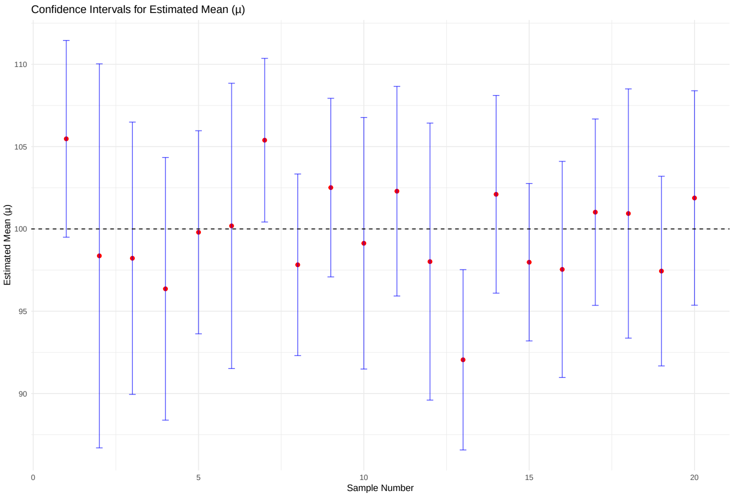
$$P\left(\bar{Y} - t_{n-1,\alpha/2} \cdot s_{\bar{Y}} \leq \mu \leq \bar{Y} + t_{n-1,\alpha/2} \cdot s_{\bar{Y}}\right) = 1 - \alpha$$

- $t_{n-1,\alpha/2}$ is the critical t-value, $s_{\bar{Y}}$ is the standard error of the mean.

Confidence interval for μ (cont'd)

- The probability $(1-\alpha)$ is **not** about μ being in the interval.
- μ is a **fixed constant**.
- The probability refers to the **random interval** $(\bar{Y} \pm t_{n-1,\alpha/2} \cdot s_{\bar{Y}})$.
- It's the probability that this interval will contain the true μ.

Confidence interval for μ (cont'd)



Confidence interval for μ (cont'd)

The interval

$$(\bar{Y} - t_{n-1,\alpha/2} \cdot s_{\bar{Y}}, \bar{Y} + t_{n-1,\alpha/2} \cdot s_{\bar{Y}})$$

is called a $100 \cdot (1-\alpha)\%$ **confidence interval** for μ.

- For example, if $\alpha = 0.05$, then the interval would be called a 95% confidence interval.
- In general the value $1-\alpha$ is called the **confidence coefficient**.

Confidence interval for μ (cont'd)

- The probability associated with a confidence interval is about the **interval's endpoints**, not about the population mean μ.
- It's wrong to say "the probability that μ falls in the interval is 95%".
- μ is a **fixed but unknown** value.

Confidence interval for μ (cont'd)

- The interval is calculated from the sample mean ($\bar{Y}$) and its standard error ($s_{\bar{Y}}$).
- It is a **random interval** because it varies with different samples.
- Different samples yield different $\bar{Y}$ and $s_{\bar{Y}}$, and thus different intervals.

Confidence interval for μ (cont'd)

- The probability (e.g., 95%) refers to the likelihood that the **sample** produces an interval that contains the true μ.
- It's about the process of repeatedly taking samples and constructing intervals.
- In the long run, 95% of the intervals will contain μ.

Confidence interval for μ (cont'd)

The confidence coefficient $(1 - \alpha)$...

- represents the probability a sample will produce an interval that covers the true population mean μ.
- measures our "confidence" in the sampling procedure and interval construction.

Confidence interval for μ (cont'd)

- The width of the confidence interval depends on the confidence coefficient and the size of the sample.
    - For a 90% interval would narrower than a 95% interval. For the 90% interval, we may be a little more stringent with the width as we aren't requiring such a high a level confidence.
    - If we wish greater confidence of 95%, we must widen the interval in order to be "more confident" that it will cover the fixed value μ.

Exercise: Dissolved Oxygen Concentration

- Data: 6 measurements of dissolved oxygen (mg/L).
- Values: 2.62, 2.65, 2.79, 2.83, 2.91, 3.57
- Assume data follows a normal distribution $N(\mu, \sigma^2)$.
- Goal: Find a confidence interval for the population mean μ.

Exercise: Dissolved Oxygen Concentration (cont'd)

- Sample size: $n = 6$
- Sample mean:

$$\bar{Y} = \frac{2.62 + 2.65 + 2.79 + 2.83 + 2.91 + 3.57}{6} = 2.895 \text{ mg/L}$$

- Sample variance:

$$s^2 = \frac{0.075625 + 0.059025 + 0.011025 + 0.004225 + 0.000225 + 0.455625}{5} = 0.12115$$

- Sample standard deviation: $s = 0.348$ mg/L
- Standard error of the mean: $s_{\bar{Y}} = \frac{0.348}{\sqrt{6}} = 0.142$ mg/L

Exercise: Dissolved Oxygen Concentration

95% Confidence Interval

- Using $t_{5,0.025} = 2.571$ (t-distribution with 5 degrees of freedom).
- Confidence interval:

$$(2.895 \pm 2.571 \times 0.142) = (2.529, 3.261) \text{ mg/L}$$

Exercise: Dissolved Oxygen Concentration

90% Confidence Interval

- Using $t_{5,0.05} = 2.015$ (t-distribution with 5 degrees of freedom).
- Confidence interval:

$$(2.895 \pm 2.015 \times 0.142) = (2.609, 3.181) \text{ mg/L}$$

Example in R

How can we calculate it using R?

```
# Data
oxygen_values <- c(2.62, 2.65, 2.79, 2.83, 2.91, 3.57)
```

Example in R

How can we calculate it using R?

```
# Sample statistics
n <- length(oxygen_values)  # Sample size
mean_Y <- mean(oxygen_values)  # Sample mean
s <- sd(oxygen_values)  # Sample standard deviation
se_Y <- s / sqrt(n)  # Standard error

df <- n - 1  # Degrees of freedom
```

Example in R

How can we calculate it using R?

```
# 95% Confidence Interval
t_95 <- qt(0.975, df)  # t-value for 95% confidence
i95_lower <- mean_Y - t_95 * se_Y
i95_upper <- mean_Y + t_95 * se_Y
c(i95_lower,i95_upper)
```

Statistical Inference and confidence intervals

Statistical Inference and confidence intervals

- The confidence interval procedure is a form of **statistical inference**.
- We are now defining this term more formally.

Statistical Inference and confidence intervals (cont'd)

- Conclusions about the population are based on **sample evidence**.
- This introduces an element of **uncertainty**.
- We aim to assess and summarize this uncertainty.
- We assess the **quality of our sample evidence**.

Statistical Inference and confidence intervals (cont'd)

- We temper inferences about the population with an indication of our assessment of sample evidence.
- This indication is expressed in terms of **probability**.
- Probability reflects how likely things are in the sample.

Statistical Inference and confidence intervals (cont'd)

We can use confidence intervals as assessment

- A confidence interval is one way to assess sample evidence.
- It is a statement about the **quality of sample evidence** regarding the population mean.

Statistical Inference and Sampling Distributions (cont'd)

- When the population is approximately normal, and we want to make inferences about the population mean (μ) using the sample mean ($\bar{Y}$), we use:

$$\frac{\bar{Y} - \mu}{s_{\bar{Y}}} \sim t_{n-1}$$

- This statistic follows a t-distribution with $n - 1$ degrees of freedom.
- We use this to make probability statements (confidence intervals) about the sample evidence.

Statistical Inference and Sampling Distributions (cont'd)

- The probability distribution of a statistic is central to statistical methodology.
- The distribution of all possible values of a statistic across all possible samples is called a **sampling distribution**.
- Examples:
  - Student's t-distribution
  - Chi-squared (χ^2) distribution

Statistical Inference and Sampling Distributions (cont'd)

- Formal statistical inference is based on comparing statistics to appropriate **sampling distributions**.
- This enables us to make formal probability statements.
- These probability statements characterize the quality of the sample evidence in answering our research question.

Confidence interval for a difference of population means

- We often want to compare **multiple populations**, not just study one.
- Experiments are designed to compare different treatments or conditions.

Confidence interval for a difference of population means (cont'd)

Example: Fertilizer and Yield

- An experiment might compare crop yields with different fertilizer application rates.
- We want to determine if different treatments (fertilizer rates) lead to **truly different responses** (yields).

Experimental Procedure

- Take two random samples of experimental units.
- Apply treatment 1 to the first sample and treatment 2 to the second sample.
- Analyze the difference in responses between the two treatments.

marta.bofill.roig@upc.edu

Example: Toxic Agent on Rats

- Compare two concentrations of a toxic agent on rat weight loss.
- Randomly assign rats to receive either concentration 1 or 2.
- Variable of interest: Y = weight loss.

Two Populations

- Before treatments, rats are from a common population: $Y \sim N(\mu, \sigma^2)$.
- After treatments, consider two populations:
    - Population 1: Y's under treatment 1
    - Population 2: Y's under treatment 2
- Samples are from these respective populations.

Normal Distribution Assumption

- Assume two random variables, Y_1 and Y_2, for each population.
- Assume they are normally distributed:
    - Population 1: $Y_1 \sim N(\mu_1, \sigma_1^2)$
    - Population 2: $Y_2 \sim N(\mu_2, \sigma_2^2)$

Notation for Two Populations

- We introduce notation for observations from two populations.
- Y_{ij} denotes the jth observation from the ith treatment.
- Data Representation:
  - Population 1: $Y_{11}, Y_{12}, \ldots, Y_{1n_1}$ (n_1 observations)
  - Population 2: $Y_{21}, Y_{22}, \ldots, Y_{2n_2}$ (n_2 observations)

Comparing population means

- Question: Is there a difference in response between the two treatments?
- Formally: Is μ_1 different from μ_2?
- Consider the difference: $(\mu_1 - \mu_2)$
  - $(\mu_1 - \mu_2) = 0$: No difference
  - $(\mu_1 - \mu_2) \neq 0$: Real difference

Comparing population means (cont'd)

- Estimate the difference $(\mu_1 - \mu_2)$ using the data from the two samples.

Comparing population means (cont'd)

What's the distribution of the difference?

- If Y_1 and Y_2 are normally distributed, then:

$$(Y_1 - Y_2) \sim N(\mu_1 - \mu_2, \sigma_D^2)$$

  where $\sigma_D^2 = \sigma_1^2 + \sigma_2^2$.

- The difference $(Y_1 - Y_2)$ is also normally distributed.

Comparing population means (cont'd)

- Define $\hat{D} = \bar{Y}_1 - \bar{Y}_2$, where:
  - $\bar{Y}_1 = \frac{1}{n_1} \sum_{j=1}^{n_1} Y_{1j}$
  - $\bar{Y}_2 = \frac{1}{n_2} \sum_{j=1}^{n_2} Y_{2j}$
- $\hat{D}$ is the difference in sample means.

Comparing population means (cont'd)

- $\hat{D}$ is normally distributed:

$$\hat{D} \sim \mathcal{N}(\mu_1 - \mu_2, \sigma_D^2)$$

  where $\sigma_D^2 = \frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}$.

- The mean of the distribution of $\hat{D}$ is $\mu_1 - \mu_2$.

- The statistic $\frac{\hat{D}-(\mu_1-\mu_2)}{\sigma_{\hat{D}}}$ follows a standard normal distribution.

Comparing population means (cont'd)

- Use $\hat{D}$ as an estimator for $\mu_1 - \mu_2$.
- We want to report a confidence interval to assess the quality of the sample evidence.

Comparing population means (cont'd)

Assumptions

- In practice, σ_1^2 and σ_2^2 are usually unknown.
- We consider the simplest case:
  - $n_1 = n_2 = n$ (equal sample sizes).
  - $\sigma_1^2 = \sigma_2^2 = \sigma^2$ (equal variances).

Comparing population means (cont'd)

- We assume equal variances: $\sigma_1^2 = \sigma_2^2 = \sigma^2$.
- Equal sample sizes: $n_1 = n_2 = n$.
- Sample means:
    - $\bar{Y}_1 = \frac{1}{n}\sum_{j=1}^{n} Y_{1j}$
    - $\bar{Y}_2 = \frac{1}{n}\sum_{j=1}^{n} Y_{2j}$
- Variance of the difference: $\sigma_{\bar{D}}^2 = 2(\frac{\sigma^2}{n})$.

Comparing population means (cont'd)

- Since $\sigma_1^2 = \sigma_2^2$, we use a pooled estimate:

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

- s^2 is the average of the two sample variances.

Comparing population means (cont'd)

- Estimator for $\sigma^2_{\bar{D}}$:

$$s^2_{\bar{D}} = 2(\frac{s^2}{n})$$

- Test statistic:

$$\frac{\bar{D} - (\mu_1 - \mu_2)}{s_{\bar{D}}} \sim t_{2(n-1)}$$

where $s_{\bar{D}} = \sqrt{\frac{2}{n}}s$.

Confidence interval for $\mu_1 - \mu_2$

- Confidence interval with confidence coefficient $(1-\alpha)$:

$$\bar{D} \pm t_{2(n-1),\alpha/2} s_{\bar{D}}$$

- Probability statement:

$$P(\bar{D} - t_{2(n-1),\alpha/2} s_{\bar{D}} \leq \mu_1 - \mu_2 \leq \bar{D} + t_{2(n-1),\alpha/2} s_{\bar{D}}) = 1 - \alpha$$

Confidence Interval for $\mu_1 - \mu_2$ (cont'd)

- The confidence interval for the difference of population means $(\mu_1 - \mu_2)$ is:

$$(\bar{D} - t_{2(n-1),\alpha/2} s_{\bar{D}}, \bar{D} + t_{2(n-1),\alpha/2} s_{\bar{D}})$$

- Where:
  - $\bar{D} = \bar{Y}_1 - \bar{Y}_2$ (difference of sample means)
  - $t_{2(n-1),\alpha/2}$ is the critical t-value with $2(n-1)$ degrees of freedom
  - $s_{\bar{D}}$ is the standard error of the difference

Confidence Interval for $\mu_1 - \mu_2$ (cont'd)

- The interpretation is the same as for a single sample and single population.
- The confidence interval is a statement about the quality of the evidence in the samples from the two populations.
- It assesses the evidence about the fixed population parameter $(\mu_1 - \mu_2)$.

Example: Pig Rations

- Two rations, A and B, are tested on pigs for weight gain (lbs).
- 12 pigs per ration ($n = 12$).
- Data:
  - Ration A: 31, 34, 29, 26, 32, 35, 38, 34, 30, 29, 32, 31
  - Ration B: 26, 24, 28, 29, 30, 29, 32, 26, 31, 29, 32, 28

Example: Pig Rations (cont'd)

- $\bar{Y}_1 = 31.75$ (mean for Ration A)
- $\bar{Y}_2 = 28.6667$ (mean for Ration B)
- $(n-1)s_1^2 = 112.25$
- $(n-1)s_2^2 = 66.64$
- Pooled variance: $s^2 = \frac{112.25+66.64}{2\times 11} = 8.1314$
- $\bar{D} = 3.0833$ (difference of means)
- $s_{\bar{D}} = \sqrt{\frac{2(8.1314)}{12}} = 1.1641$

Example: Pig Rations (cont'd)

Based on this, we compute the 95% confidence interval:

- $t_{22,0.025} = 2.074$ (for $\alpha = 0.05$)
- Confidence interval:

$$(3.0833 \pm 2.074 \times 1.1641) = (0.6689, 5.4978)$$

Example: Pig Rations (cont'd)

- The 95% confidence interval $(0.6689, 5.4978)$ does not contain 0.
- This suggests strong evidence that the mean weight gains for the two rations are different.
- The sample evidence supports a real difference in weight gain between the rations.

References and other materials

1. D.R. Cox, E. J. Snell. Applied Statistics - Principles and Examples. Chapman & Hall, 1981.
2. Spiegelhalter, David. The art of statistics: Learning from data. Penguin UK, 2019.
3. James, G., Witten, D., Hastie, T., Tibshirani, R. An introduction to statistical learning. Springer, 2013.
4. Teaching courses from `https://www4.stat.ncsu.edu/~davidian/`
5. Teaching courses from `https://www.drizopoulos.com/`

That's all folks!

UNIVERSITAT POLITÈCNICA
DE CATALUNYA
BARCELONATECH