Data analysis

Purpose: Analyze collected data to answer research questions

Steps:

1) Descriptive statistics
   - * is data sensible?
   - * unusual values
   - * missing values, outliers & inconsistent values
2) Inferential Statistics
3) Interpret results

Step 1: Descriptive Statistics

* Frequencies
  - * Absolute: Number of times a value appears in a dataset
  - * Relative: Proportion or % a value appears respect dataset
* Measures of central tendency
  - * Mode: Most common value
  - * Median: Middle value
  - * Mean: Arithmetic average
* Measures of distribution
  - * Maximum/minimum: Self-descriptive
  - * Interquartile range (IQR): Range between Q_1 (25%) & Q_3 (75%), spread of the middle 50% of the data.
  - * Variance: Spread of a set of values around their mean. High value = data very spread respect mean
  - * Std. Deviation: $\sqrt{\text{variance}}$. Average distance of data points respect the mean. Large value = variability among data points

Data Validation

* Check consistency & credibility
* Ensuring correctness & completeness of collected data
* Identifying missing variables
* Identifying outliers → How to deal? →
  - Exclude
  - Report two analysis (w & w/o)
  - Transform variable
  - Gather more data
  - Create box-plots
  - Analyse if there are systematic explanations for these values
  - Perform sensitive analysis (e.g. scatterplot)

Dispersion example

We will compute the Standar deviation
- From a population (σ)
- From a sample (s)

Dataset = $\{1, 2, 2, 2, 3, 14\}$

1) Computation Std dev σ

1) Compute mean of the dataset

$$\bar{x} = \frac{1+2+2+2+3+14}{6} = 4$$

2) Compute difference between each point and the mean, and square it

$(x_i - \bar{x})^2$:

$(1-4)^2 = 9$
$(2-4)^2 = 4$
$(2-4)^2 = 4$
$(2-4)^2 = 4$
$(3-4)^2 = 1$
$(14-4)^2 = 100$

3) Compute the variance

$\frac{\sum (x_i - \bar{x})^2}{N}$ → If we compute over population we use N ⚠

$$\text{variance} = \frac{(9+4+4+4+1+100)}{6} = \frac{122}{6} = 20'33$$

4) Compute σ

$$\sigma = \sqrt{\text{variance}} \approx 4'51$$

2) Computation Std dev s

1) Compute mean of the dataset

$\bar{x} = 4$

2) Compute difference between each point and the mean, and square it

$(x_i - \bar{x})^2$: $\{9, 4, 4, 4, 1, 10\}$

3) Compute the variance

$\frac{\sum (x_i - \bar{x})^2}{N-1}$ → If we compute over sample we use $N-1$ ⚠

$$\text{variance} = \frac{(9+4+4+4+1+100)}{5} = 24'4$$

4) Compute s

$$s = \sqrt{\text{variance}} \approx 4'94$$

3) Computing IQR

Methods:

- Tukey:
- Moore & McCabe:
- Mendenhall & Sincich:
- Freund & Perles:
- Minitab:

Handling dependency

- 2 variables

  Relationship between X & Y ⟶

  - linear regression: $y = a + bx$
  - Correlation coefficient (Pearson)

$$r = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2}\sqrt{\sum_{i=1}^{n}(y_i - \bar{y})^2}}$$ = degree of linear dependence between 2 variables

- > 2 variables
  - Multivariate analysis
  - Principal component

  ...

Step 2: Statistical analysis

Hypothesis testing

* **Definition**

  Procedure for determining wether experimental results of a sample provide support for the validity of a thesis/hypothesi

* **Basic questions**

  1) Is the relationship between 2 o+ variables due to random chance?
  2) If chance is excluded, what does it mean?
     * There might be a meaningful association between variables

* **Approach (Falsification principle)**

  Opposite-of-what-you-predict reasoning

  Draw conclusions by

  - rejecting the inverse hypothesis although is true (< 0'05)

* **Procedure**

  1) Formulate alternative & null hypothesis
  2) Select a statistical test

     ⚠ considering data distribution → * Normal distribution → Parametric test
     * non-normal/ordinal/nominal distribution → Non-parametric test
  3) Apply test
  4) Use a significance level (α)

     It represents the probability of rejecting H_0 when is actually true → 0'01 or 0'05 → Type I error
  5) Perform power analysis (β)

     Probability of failing to reject H_0 when is actually false → 0'2 → Type II error
  6) Interpretation

     If H_0 rejected:

     There is an effect
     ↳ calculate real effect size
     ↳ Is effect in concordance with the theory?
     ↳ Modify theory?

     If !H_0 rejected:

     - It is not possible to conclude there is no effect
     - There is no sufficient evidence to accept that there is an effect
     - Discuss descriptive statistics
     - Calculate real effect size → Determine sample size to reject H_0
     - Formulate new Hypothesis

Example on experiments: Simple test where 2 techiques are applied to students. Results are very similar. B seems better (but slightly), How do we make a decision?

1) Compute difference between means
   diff = $(\bar{x}_b - \bar{x}_a)$

2) Locate diff in the histogram

3) Calculate area that falls in the right side of the histogram

   That area is the probability that we could obtain a result $\geq (\bar{x}_b - \bar{x}_a)$ → p-value

4) Analyze
   If p-value $< \alpha$
   Techniques are not alike → significant result

Parametric VS non-parametric

* T-Test → parametric
* Main diff → Assumption of the distribution of the sample
  Non parametric: does not make any assumption

Parametric testing

1) T-Test (independent sample)

1.1) One sample t-test

Compare $\bar{x}$ response of a group against a specific value

$$t = \frac{\bar{X} - \mu_0}{s/\sqrt{n}}$$

$\bar{X}$ = mean of group
μ_0 = specified value
n = # subjects group
s = std dev group

→ look in t-distribution table to obtain p-value

⚠ Requirements

If any does not hold → non-parametric test

1) Samples must be independent & identically distributed
2) Mean estimator should be normally distributed
3) Response variables measured on ratio scales
   ⚠ ordinal metrics can't be used

1.2) Two sample t-test

- Checks → Statistical significance of the difference between the mean responses of two levels of a factor
  → H_0 of 2 subpopulations where the mean is the same

$$H_0: \mu_1 = \mu_2$$
$$\Theta_1 = \Theta_2$$

- Pre-requisites:
  * The two sample sizes are equal
  * The two distributions have the same variance

$$t = \frac{\bar{X}_1 - \bar{X}_2}{s_{X_1 X_2} \cdot \sqrt{2/n}} \qquad s_{X_1 X_2} = \sqrt{\frac{1}{2}\left(s^2_{X_1} + s^2_{X_2}\right)}$$

- $\bar{X}$ = mean
- s = std dev
- n = # subjects (equal in both groups)
- s^2 = unbiased estimators of the variance

1.3) Special cases

1.3.1) Unequal sample size

⚠ Equal variance assumed

$$t = \frac{\bar{X}_1 - \bar{X}_2}{s_{X_1 X_2} \cdot \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}$$

$$s_{X_1 X_2} = \sqrt{\frac{(n_1 - 1)s^2_{X_1} + (n_2 - 1)s^2_{X_2}}{n_1 + n_2 - 2}}$$

1.3.2) Equal/Unequal sample sizes & unequal variances

$$t = \frac{\bar{X}_1 - \bar{X}_2}{s_{\bar{X}_1 - \bar{X}_2}}$$

$$s_{\bar{X}_1 - \bar{X}_2} = \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}$$

2) Paired T-tests (Dependant sample)

Repeated measures or matched pairs

⚠ Diff between all pairs must be calculated

$$t = \frac{\bar{X}_D - \mu_0}{s_D/\sqrt{n}} \qquad \bar{X}_D = \text{mean of diff between pairs}$$

3) One factor ANOVA

↗ "Analysis Of Variance"

Checks the statistical significance of the difference between the mean responses of one factor with several levels

* Factor: independant variable selected to determine its effect on the response variable
* Level of factor: Different categories / distinct values that the factor can take on

Steps

1) Identify mathematical model

$Y_{ij} = \mu + \alpha_j + e_{ij}$ (j: level of factor, i: observation)

- μ: overall mean response
- α_j: effect of the j level of factor
- e_{ij}: error

Where $\alpha_j = \bar{Y}_{\cdot j} - \mu = \bar{Y}_{\cdot j} - \bar{Y}_{\cdot\cdot}$

2) Validation of the basic model that relates the experiment variables

Confirm assumptions of ANOVA are met

3) Calculate the factor-induced variation in the response variable

ANOVA decomposes total variation in the response variable into different components:

1) Variation between groups (induced by the factor)
2) Variation within groups (random error)

} How much of the variability in the response variable is attributable to differences between the levels of factor

4) Calculate the statistical significance of the factor-induced variable

- Asses whether the differences in mean responses among the levels of factor are significant
- F-test → if f-ratio large & f-ratio > f-distribution → diff between means among levels of the factor

5) Establish recommendations on the alternative that provides the best response variable values

After determining the statistical significance of the factor-induced variation we draw conclusions.

If H_0 rejected:

≥1 level of factor differs from the others

Non-parametric testing

1) MANN-WHITNEY-U Test (Independant sample)

- For independent groups
- It can be used instead of t-test if data is not normally distributed
- Robust against outliers

Pre requisites

- Responses are on an ordinal scale
- Distributions of both groups are equal under the H_0

1.1) Method 1

Small samples.

Choose smaller sample (for easy computation)

For each observation in sample 1
    count smaller rank obs in sample 2
} sum result = U

1.2) Method 2

Large samples

$$U_1 = n_1 \cdot n_2 \frac{n_1(n_1+1)}{2} - R_1 \quad \leftarrow \text{sum of ranks of group 1}$$

$$U_2 = n_1 \cdot n_2 \cdot \frac{n_2(n_2-1)}{2} - R_2$$

Reject H_0 if:

$\min(U_1, U_2) \leq$ critical value α

2) Wilcoxon signed rank test

Alternative to paired t-test

Pre-requisite

It must be possible to determine which value is larger and to rank the differences

3) Sign-test

- Alternative to paired t-test ⚠ Less powerful than Wilcoxon
- Use if:
  - Not possible to rank the differences
  - ⚠ Must be on ordinal scale at least
- Formula: $p = \frac{1}{2^N} \sum_{i=0}^{n} \binom{N}{i}$