ΤΜΗΜΑ ΙΑΤΡΙΚΗΣ
Σχολή Επιστημών Υγείας
Πανεπιστήμιο Θεσσαλίας

Medical and Computer Statistics
Department of Biomathematics

Mprotsis Theodoros, MSc, PhD
Teacher & Research Fellow

Statistical Notes

Medical and Computer Statistics

Mprotsis Theodoros, MSc, PhD
Teacher & Research Fellow

A. Types of Variables:

- **Qualitative:** The data belong to categories or groups (2 or more). The order of categorization may matter (ordinal data, e.g., disease severity: mild < moderate < severe) or may not matter (nominal data, e.g., smoker vs. non-smoker). In practice, SPSS uses "nominal" in the variable view for these types of data.
- **Quantitative:** The data take numerical values, either discrete (discrete data, e.g., the absolute number of births in Larissa in 2023) or continuous (continuous data, e.g., blood cholesterol level). In SPSS, the corresponding option for both types is “Scale.”

B. Analysis of the Relationship Between Two Variables:

1. One Qualitative (with 2 Categories or Groups) and One Quantitative:

Example: "Drug" (A or B) and "blood cholesterol level."

- **Measurement on Different People:** Some participants received Drug A, while others received Drug B, and the blood cholesterol level was measured in each group.
    - **Question:** Is there a difference between Drug A and Drug B in their effectiveness at regulating blood cholesterol levels in the population?
    - **Answer:** Perform a **t-test for unpaired observations (two-sample t-test)** or the corresponding non-parametric test (Mann-Whitney U test), depending on whether the quantitative variable is normally distributed in the population. Note that some variables, such as CRP or blood glucose, are known for their skewed distributions. Additionally, consider the sample size: if each category of the qualitative variable has a sample size greater than 30, a parametric test is generally more appropriate.
- **Measurement on the Same People:** The same subjects first received Drug A, and their cholesterol levels were measured. After sufficient time for the effect of Drug A to wear off, the subjects received Drug B, and their cholesterol levels were measured again.
    - **Question:** Is there a difference between Drug A and Drug B in their effectiveness at regulating blood cholesterol levels in the population?
    - **Answer:** Perform a **t-test for paired observations (paired t-test)** or the corresponding **non-parametric test (Wilcoxon signed-rank test)**. The same considerations for choosing between parametric and non-parametric tests apply here, based on the distribution of the quantitative variable and the sample size.

2. One Qualitative (with 3 or More Categories) and One Quantitative:

Example: "Drug" (A, B, or C) and "blood cholesterol level."

- **Measurement on Different People:** Some participants received Drug A, some received Drug B, and others received Drug C, and their blood cholesterol levels were measured.
    - **Question:** Is there a difference in the effectiveness of Drugs A, B, and C in regulating blood cholesterol levels in the population?
    - **Answer:** Perform a one-way ANOVA test or the corresponding non-parametric test (Kruskal-Wallis test). Note that one-way ANOVA or Kruskal-Wallis will indicate whether there is a difference among the groups, but it does not specify which groups differ (e.g., whether A differs from B, A from C, or B from C). To determine this, perform post-hoc tests or Mann-Whitney U tests.

3. One Qualitative Variable with Another Qualitative Variable:

Example: "Drug" (A or B) and "Treatment Achievement" (YES or NO) (or with more categories if applicable).

- **Question:** Is there a relationship between the choice of drug and the achievement of treatment?
  Or equivalently, is there a difference in cure rate between Drug A and Drug B?
- **Answer**: Perform a **Chi-square test (χ^2 test)**. To quantitatively assess which drug is superior, calculate the odds ratio.

4. One Quantitative Variable with Another Quantitative Variable:

Example: "Age" and "blood cholesterol level."

- **Question:** Is there a linear relationship between age and cholesterol level?
- **Answer:** Calculate the **correlation coefficient (r)** and perform a **simple linear regression**. The slope coefficient obtained from the linear regression (even if statistically significant) indicates how much, on average, the cholesterol level increases (or decreases) with each one-year increase in age.

Mprotsis Theodoros, MSc, PhD
Teacher & Research Fellow

C. Analysis of the Relationship Between More Than Two Variables

When analyzing the relationship between a quantitative dependent variable and multiple independent variables (which can be either quantitative or qualitative), we use various statistical methods depending on the nature of the data.

- **Example with Two Quantitative Independent Variables:**
  Continuing with the example above, we use **multivariable regression** to understand how the dependent variable (cholesterol level) changes with the change in age, while controlling for the effect of another variable, such as weight. The coefficient (slope) obtained from the regression tells us how much the average cholesterol level increases (or decreases) with each additional year of age, assuming weight is included in the regression model. Essentially, this is like comparing individuals of the same weight to see the "net" effect of age on cholesterol levels.
- **Example with Two Qualitative Independent Variables:**
  Consider age and weight as categorical variables (e.g., over or under 40 years old, and obese or normal weight) and their relationship to cholesterol levels. In this case, we could use either **multivariable linear regression** or **two-way ANOVA** without interaction. The coefficient (slope) from the regression model indicates how much cholesterol levels increase (or decrease) on average for individuals over 40 compared to those under 40, while controlling for weight. This is akin to comparing people within the same weight category to observe the "net" effect of the age category on cholesterol levels.
- **Examining Interactions Between Two Independent Variables:**
  If we want to investigate whether there is an interaction between the two independent variables, such as whether the relationship between age (quantitative or categorical) and cholesterol level varies by different levels of weight (quantitative or categorical), we use **multivariable linear regression with interaction terms** or **two-way ANOVA with interaction**. In the statistical output (e.g., from SPSS), we examine whether the p-value associated with the interaction term (e.g., AGE*WEIGHT) is less than 0.05 (indicating interaction) or greater than 0.05 (indicating no interaction). Essentially, if no interaction is present, it suggests that the relationship between age and cholesterol is consistent across all weight levels. If an interaction is detected, we must consider the weight level to accurately describe the relationship between age and cholesterol levels.