Seminar (graduate)

Spring 2024

Miran Kim (mirankim316@gmail.com)

Seminar Topics:

Part 1. Introduction

Part 2. Descriptive Statistics

Part 3. Chi-squared test with Frequency Data

Part 1. Introduction

1. Why statistics?
2. Steps of statistical approach
3. Types of data
4. Software

[1] Why statistics?

A. Data Analysis and Interpretation:

B. Evidence-Based Decision Making:

C. Quantitative Research and Validity:

A. Data Analysis and Interpretation:

- Statistics provides **a systematic framework for collecting, organizing, and analyzing data** in humanities and education research. It allows researchers to make sense of **complex datasets** and **draw meaningful conclusions** from them.

e.g., in education studies, statistics can be used to analyze test scores, student performance data, and educational outcomes to identify trends, patterns, and factors influencing student achievement.

B. Evidence-Based Decision Making:

- By analyzing data, we can **identify** effective teaching strategies, **assess** the impact of educational interventions, and **make informed choices** about curriculum development, resource allocation, and educational policy changes.

- This leads to more effective and efficient educational practices and policies.

C. Quantitative Research and Validity:

- In humanities and education studies, quantitative research often requires the use of statistical methods to ensure the **validity and reliability of findings**.

- Statistics allows researchers to **test hypotheses**, **measure the strength of relationships between variables**, and **determine the significance of results**.

[2] Steps of statistical approach

- **Gathering** data:

  e.g., survey, measurements, text essays, Corpus, etc.

- **Describing** and **visualizing** data:

  e.g., statistic (t-test, F-test, Chi-Square, Correlation, Regression,), representative values (mean, median, mode, etc.), various plots (bar, box, line, dot, distribution, etc.)

- **Statistical** analysis: parametric vs. non-parametric

- **Interpreting data and drawing conclusions**:

  e.g., hypothesis & interpretation

[3] Important concepts:

- **Parameter**:

  - A parameter refers to **a numerical value** or **characteristic** that summarizes a population or probability distribution.

  - For instance, in parametric statistics, the goal is often to **estimate these population parameters from a sample** and **make inferences** about the population based on the sample data.

e.g., Parametric statistical methods assume that **the data follows a specific probability distribution (e.g., normal distribution**) with known or estimated parameters, which allows for hypothesis testing and making predictions about the population.

A. Parametric vs. non-parametric statistics

- **How to decide?**

  (1) **Data Distribution:**

  Parametric tests often assume that the data follows a specific distribution, typically the **normal distribution**. If your data approximately follows a normal distribution (or can be transformed to approximate normality), parametric tests are more appropriate. (e.g., t-test, ANOVA)

  Non-parametric tests, on the other hand, do not make distributional assumptions and are robust to deviations from normality. (e.g., Chi-square test with categorical data, Mann-Whitney U Test, Friedman Test, Cramer's V, etc.)

A. Parametric vs. non-parametric statistics

- **How to decide?**

  (2) **Measurement Scale:**

  The measurement scale of your data matters.

  Parametric tests are usually designed for interval or ratio data, which have a clear order and meaningful distances between values.

  Non-parametric tests can be used with nominal or ordinal data, which have categories or rankings but lack meaningful numerical distances.

A. Parametric vs. non-parametric statistics

- **How to decide?**

(3) **Sample Size:**

For small sample sizes, **parametric tests** may not perform well, especially if the assumptions are violated.

Non-parametric tests are often more robust in such cases and can provide valid results with smaller sample sizes.

A. Parametric vs. non-parametric statistics

- Summary: **How to decide?**

  If the **mean** of the data more accurately represents the centre of the distribution, and the sample size is large enough, we can use the **parametric** test.

  - Whereas, if the median of the data more accurately represents the centre of the distribution, and the sample size is large, we can use **non-parametric** distribution. (e.g., Kruskal Wallis Test, Sign Test, Mann Whitney U test, Wilcoxon signed-rank test)

  e.g., The **Chi-square test** is also a non-parametric test in statistics, and it is often considered as a distribution-free test.

B. Data types

- Understanding of the different **data types**, also called **measurement scales**, is a crucial prerequisite for doing Exploratory Data Analysis (EDA), since **you can use certain statistical measurements only for specific data types**.

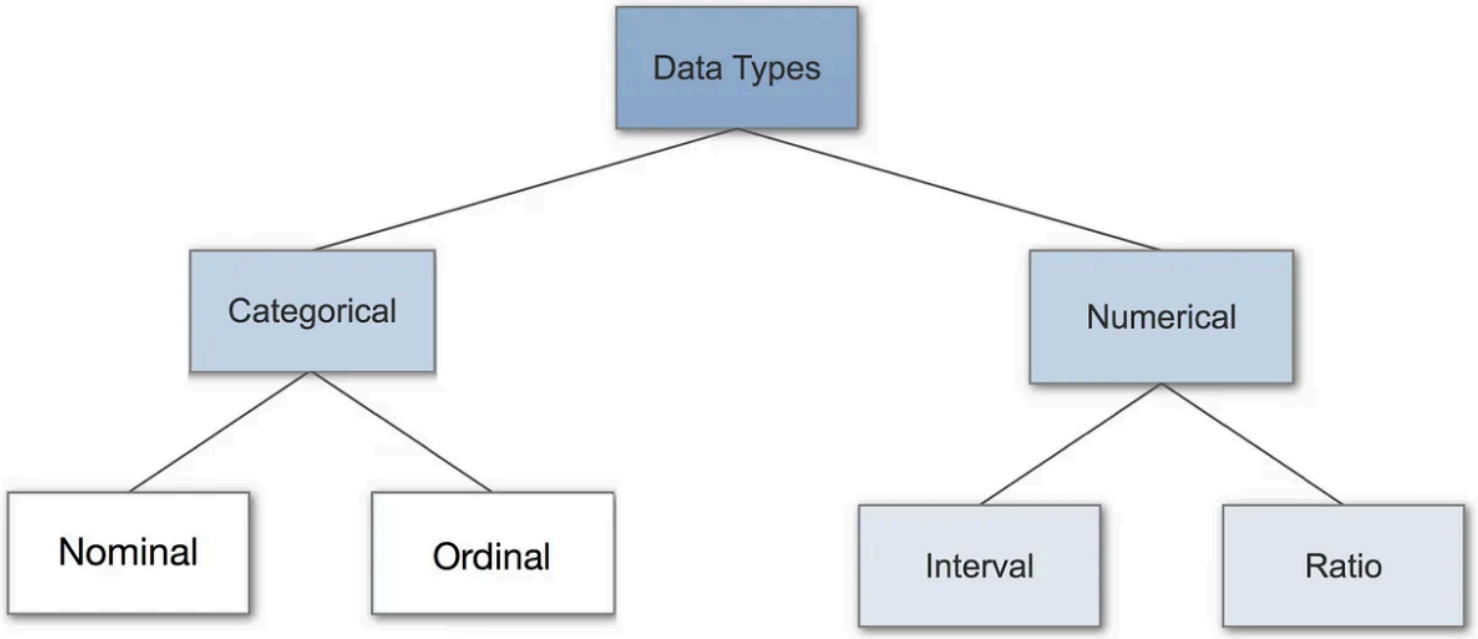


Image Source (https://towardsdatascience.com/data-types-in-statistics-347e152e8bee)

B. Data types

Categorical

- Categorical data represents **characteristics**. Therefore it can represent things like a person's gender, language etc. Categorical data can also take on numerical values.

1) Nominal

- Nominal values represent **discrete units** and are used to **label** variables, that have no quantitative value. Note that nominal data that has no order.

e.g., YES/NO, Gender, Language types, etc.
e.g., 1 for female and 0 for male

2) Ordinal

- Ordinal values represent **discrete and ordered units**. It is therefore nearly the same as nominal data, except that it's **ordering matters**.

e.g., Grade levels,

B. Data types

Numerical

- This is a type of data that is quantifiable and can be measured. Numerical data is often used in mathematical calculations and statistical analysis.

Interval

- This is a type of numerical data where the intervals between values are meaningful. However, it does not have a true zero point. A common example is temperature. The difference between 10°C and 20°C is the same as between 20°C and 30°C, but 0°C does not mean 'no temperature'.

e.g., time duration, distance, income, etc.

Ratio

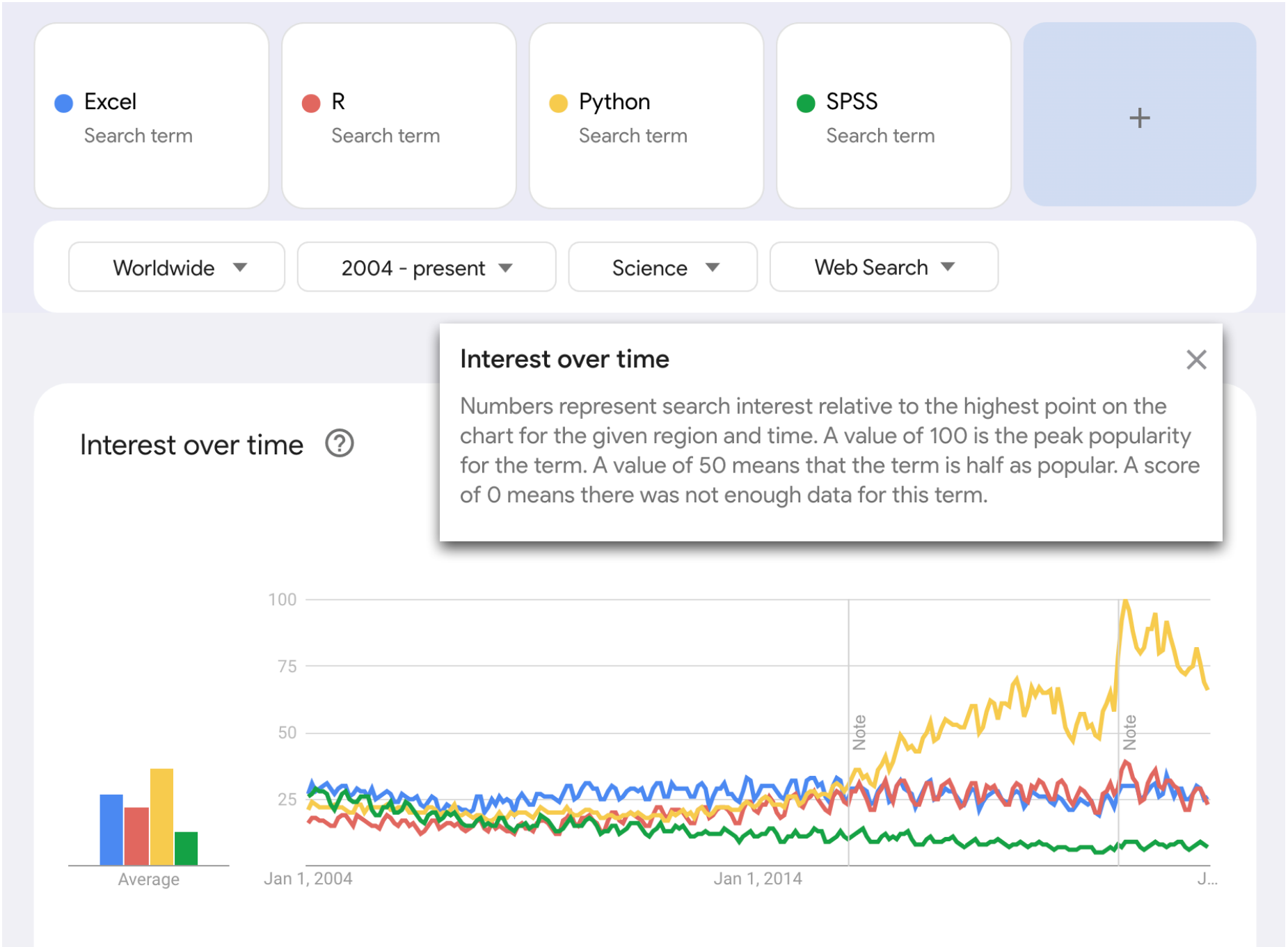
- This is also numerical data, but unlike interval data, it has a true zero point. This zero point means the absence of the quantity being measured. 0 kg means there is no weight, and this allows for comparisons like 'twice as heavy' to be meaningful.

e.g., weight, height, and duration.

[4] Software

- Familiarize yourself with statistical software or programming languages like **Excel**, **R, Python**, **SPSS**, etc., which can efficiently perform descriptive statistical analysis on large datasets.

Google trend (as of Jan.20, 2024)

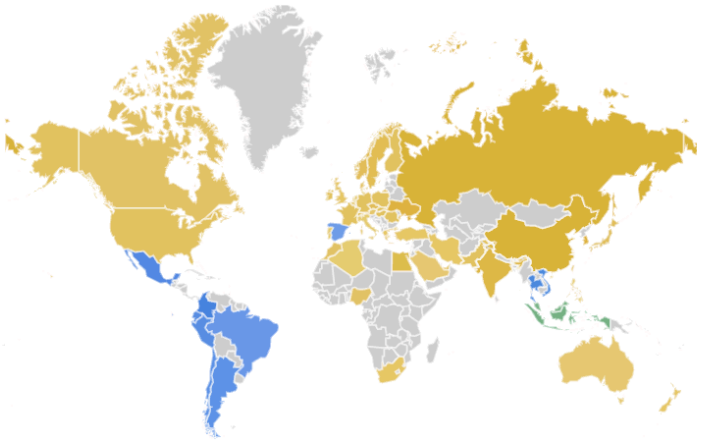


Google trend (as of Jan.20, 2024)

Compared breakdown by region

Region

Excel R Python SPSS



Sort: Interest for Python

1	Israel	
2	China	
3	Russia	
4	Ukraine	
5	South Korea	