SPSS – Week 1

Tuesday, 9 August 2022 09:49

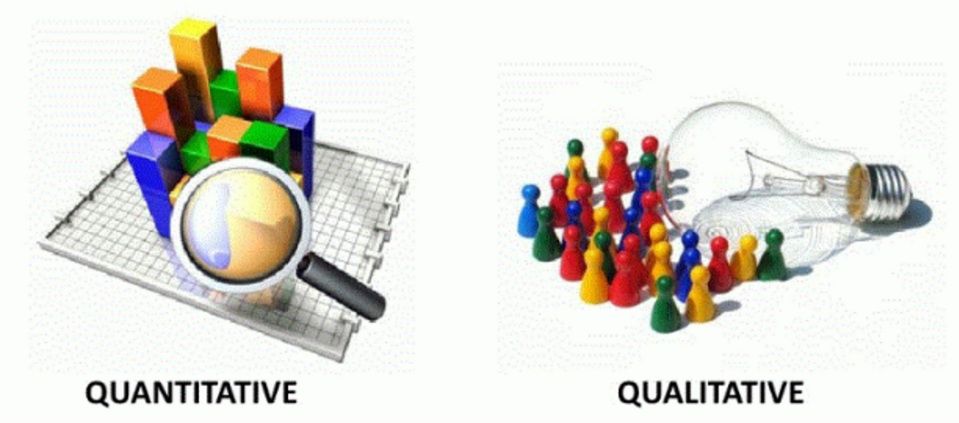
SPSS is a comprehensive statistical analysis software platform designed for ease of use and quick actionable insights to solve business and research problems.

Variables are what defines the column names inside the SPSS table. Can be such as:

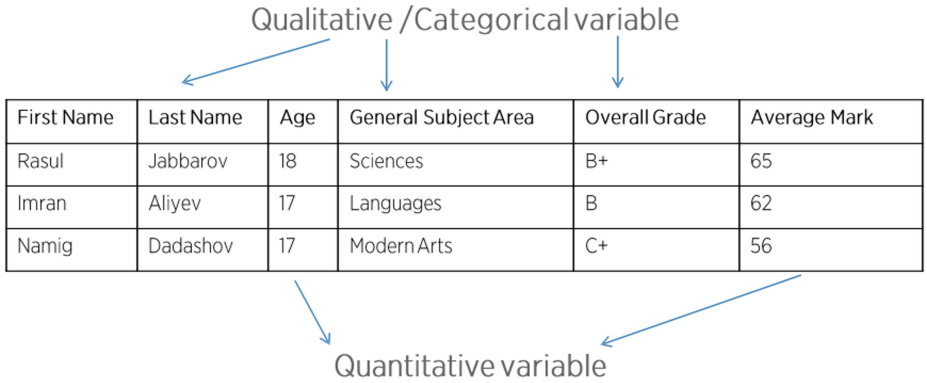
- Age
- Education field
- Name

Data under the variables can come in two main forms:

- Quantitative variables
- Qualitative or categorical variables



Let's understand the variable types by looking at this table:



First Name	Last Name	Age	General Subject Area	Overall Grade	Average Mark
Rasul	Jabbarov	18	Sciences	B+	65
Imran	Aliyev	17	Languages	B	62
Namig	Dadashov	17	Modern Arts	C+	56

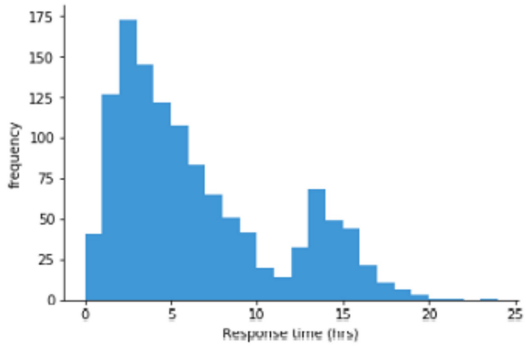
Differences between Nominal and Ordinal data:

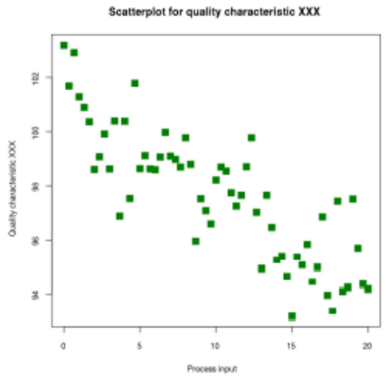
- Nominal scale variables differentiate individual data points e.g. an ID variable or Name. They are always qualitative as they do no describe value, size or

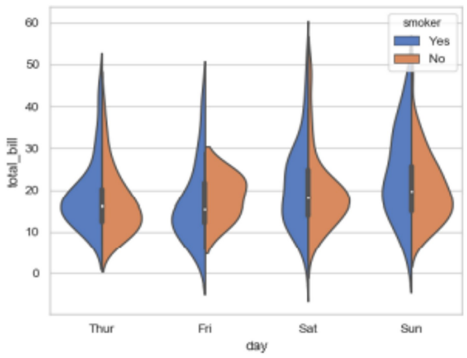
direction
- Ordinal scale variables can measure direction, size and values.

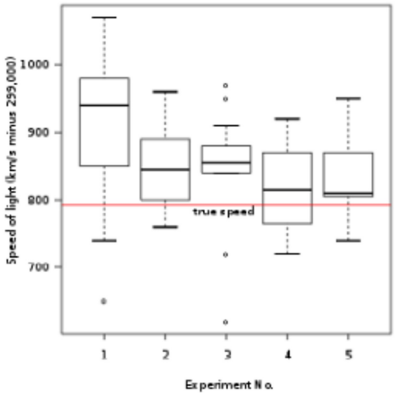
Differences between Continuous and Discrete variables:
- Discrete variables can only be integers, hence they cannot be fraction of a number.
- Continuous variables can be float, such as 100.5, 170.5cm, etc.

Exploratory Data Analysis (EDA) is the process where we visualize, examine, organize and summarize our data. Such methods are:
- Histogram plots
- Scatter plots
- Violin plots
- Boxplots
- Etc









Statistics – everything dealing with forecasting/predicting comes down to statistics. Companies and Researchers leveraging statistical knowledge know:
- What products or movies you like (think Amazon or Netflix)
- Predicting customer demand
- Understanding the cause of certain illnesses

The subfields of Statistics:

- Descriptive statistics – Measures or descriptions used to assess some performance or indicator e.g. Batting Averages, GPA
- Inference – Using knowledge from data to make informed inferences, e.g. answering the questions "How often do people get the common cold" or "How many people can afford to buy a house at the age of 25"
- Risk and Probability – What is likelihood of you rolling a 6 on a dice? Probability is an extensive and important field that is critical for many businesses such as insurance, casinos and finance.
- Correlation and Relationships – How do we know that smoking causes cancer? Extensive statistical studies have to be used for Hypothesis testing.

Sampling summary:

- A subset of population is called sample
- A good sample aims to be a good representative of the entire population
- A sampling error is a statistical error that happens when analyst does not select a sample that does not represent the entire population of data.

Types of sampling that are used to create good representations are:

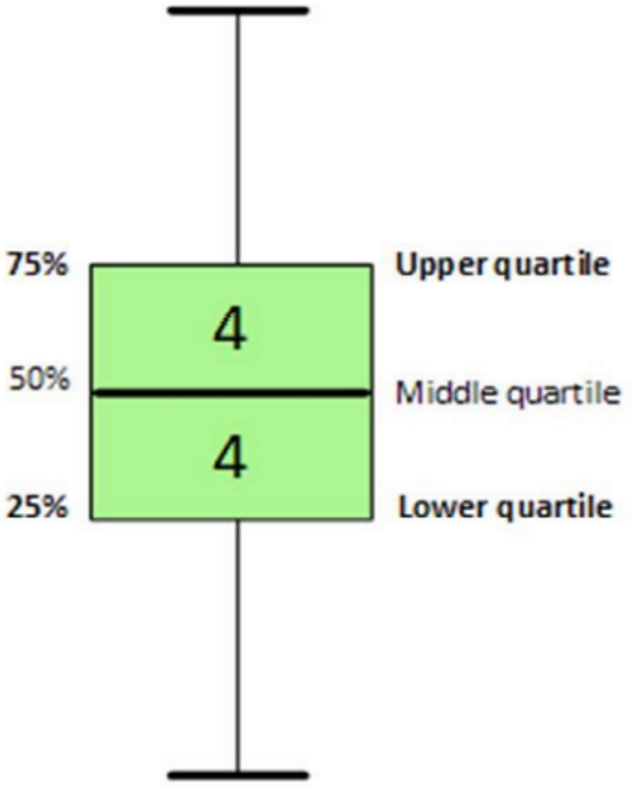
- Random sampling
- Stratified sampling

Descriptive statistics are used to describe or summarize data in ways that are meaningful and useful. Examples:

- Average heights and weights of males
- Average number of items sold per day at a store

Outliers Rule of Thumb

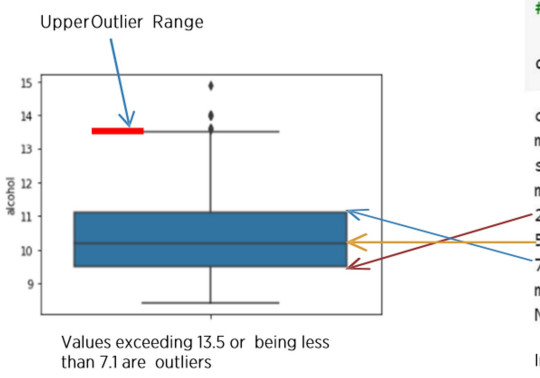
- A value can be considered an outlier if it exceeds 1.5x the difference between the upper quartile and the lower quartile



Inter Quartile Range = Upper quartile - Lower quartile

Measures of location:

- Mean – sum/n
- Median – choose middle number from ascending list
- Mode – most repeating



Red Wines

```
# Showing the quartile ranges of alcohol content

df[df['type'] == 'red']['alcohol'].describe()
```

```
count    1599.000000
mean       10.422983
std         1.065668
min         8.400000
25%         9.500000
50%        10.200000
75%        11.100000
max        14.900000
Name: alcohol, dtype: float64
```

Upper Quartile = 11.1
Lower Quartile = 9.5

Inter Quartile Range = 11.1 – 9.5 = 1.6
Outlier Ranges = 1.6 * 1.5 = 2.4
Upper Outlier Range = 11.1 + 2.4 = 13.5
Lower Outlier Range = 9.5 - 2.4 = 7.1

Detecting Outliers :

- An outlier is an unusually small or unusually large value in a dataset.
- A data with a z-score less than –3 or greater than +3 might be considered an outlier.
- It might be incorrectly recorded data