Analysis and Interpretation: Descriptive Statistics

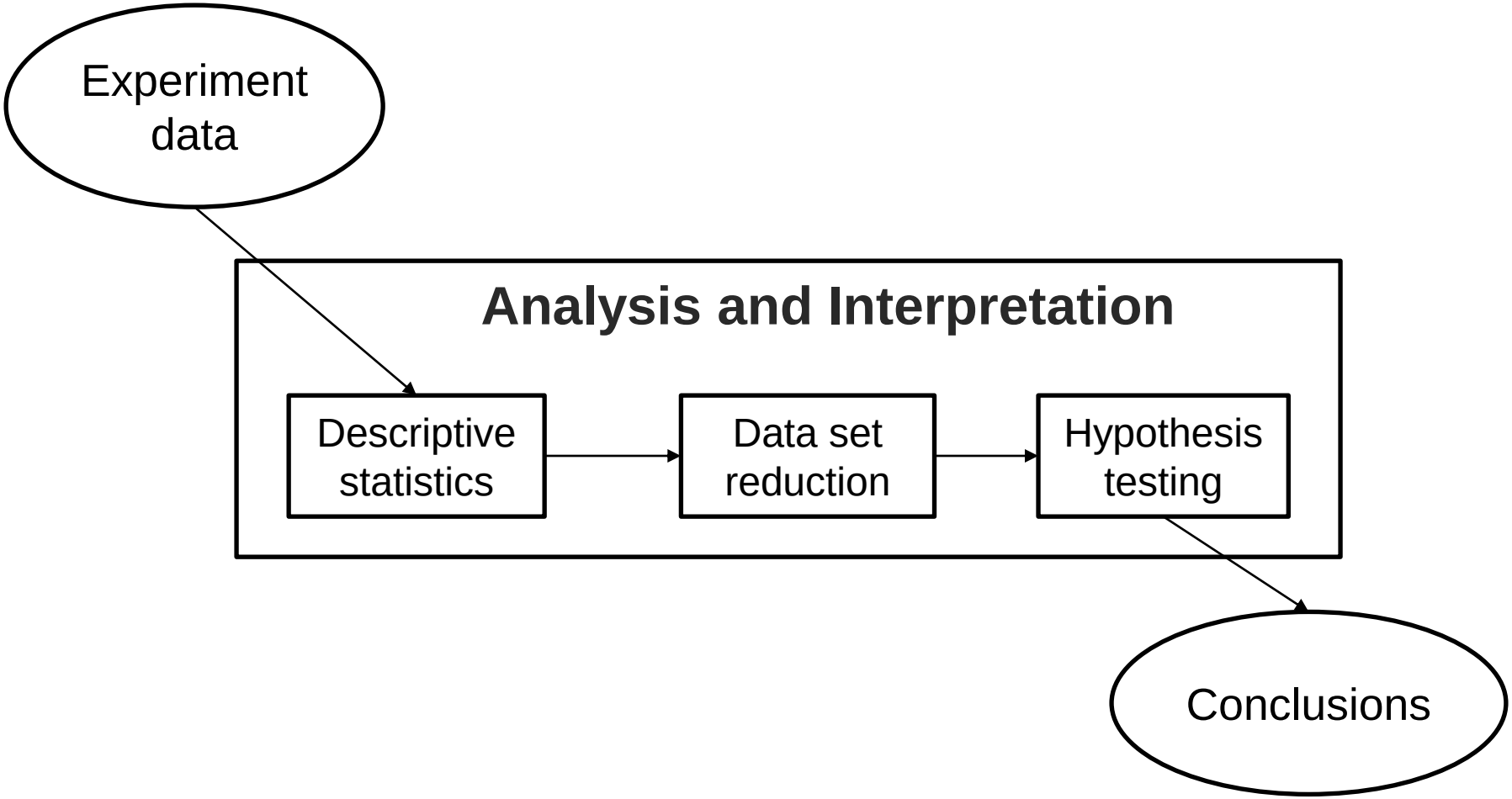
Eduardo Figueiredo

http://www.dcc.ufmg.br/~figueiredo

Analysis and Operation

- After collecting data in the operation phase, we want to draw conclusions
  - The analysis and operation phase aims to interpret the collected experimental data
- This phase has three main steps
  - **Descriptive statistics**
  - Data set reduction
  - Hypothesis testing

Analysis and Operation Overview



Descriptive Statistics

- Descriptive statistics are used to describe and graphically present interesting aspects of the data set
  - They allow identifying abnormal or false data points (called outliers)
- The scale of measurements restricts the type of statistics

Statistics for each Scale

	Central Tendency	**Dispersion**	**Dependency**
Nominal	Mode	Frequency	
Ordinal	Median, Percentile	Interval of Variation	Spearman, Kendall
Interval	Mean, Variance, Range	Standard Deviation	Pearson
Ratio	Geometric mean	Coefficient of variation	

Measures of Central Tendency

- These measures indicate the "middle" of the data set

- Common measures
  - Mean
  - Median
  - Percentile
  - Mode

Mean and Median

- Mean ($\bar{x}$)
  - It is meaningful for interval and ratio scales
  - Mean(1, 1, 2, 4) = 2.0
- Median ($\tilde{x}$)
  - The same number of samples are higher and lower than the median (middle value)
  - Well defined when n is odd (ordinal scale)
  - Median(1, 1, 2, 4) = 1.5

valid only for interval and ratio scales

Percentile and Mode

- Percentile
  - Median is a special case of percentile (50%)
  - Other common percentiles: 25% and 75%

- Mode
  - It represents the most common value
  - Valid for all scale types
  - Mode(1, 1, 2, 4) = 1

Measures of Dispersion

- They measure the variation from the central tendency

- Common measures
  - Variance
  - Standard Deviation
  - Range

Variance and Standard Deviation

- Variance (s^2)
  - It is the mean of the square distance from the sample mean

$$\text{Variance} = \frac{1}{n-1}\sum (Xi - \overline{X})^2$$

- Standard Deviation (s)
  - It is the square root of the variance
  - It has the same unit of the sample data

$$\text{Standard deviation} = \sqrt{\frac{1}{n-1}\sum (Xi - \overline{X})^2}$$

Range

- Range
  - It is the distance between the maximum and minimum data values
  - Meaningful for interval and ratio scales

$$\text{range} = x_{max} - x_{min}$$

  - Range(1, 1, 2, 4) = 4 – 1 = 3

Measures of Dependency

- When the data set consists of related samples in pairs (x_i, y_i), it is often interesting to analyze the dependency

- Common measures
  - Linear regression
  - Covariance
  - Correlation (Pearson, Spearman, Kendall)

Graphical Visualization

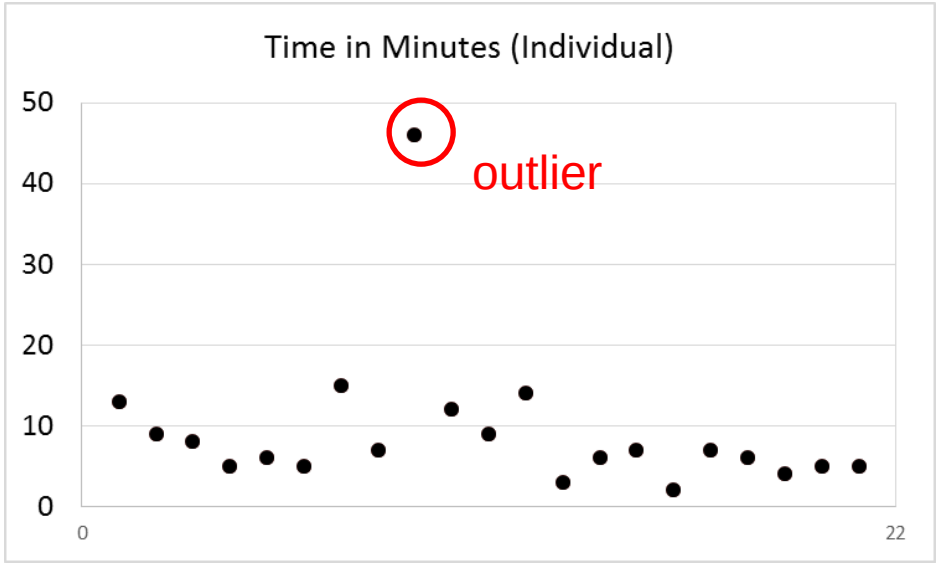
Graphical Visualization

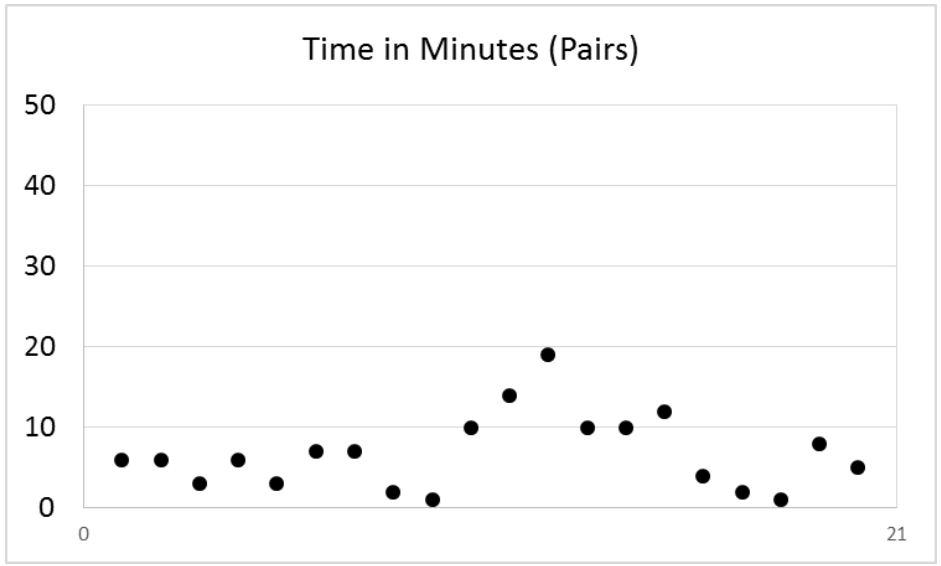
- Graphical visualization is useful to analyze measures of central tendency, dispersion, and dependency

- Common charts
  - Scatter plot
  - Box plot
  - Histogram
  - Pie chart

Scatter Plot

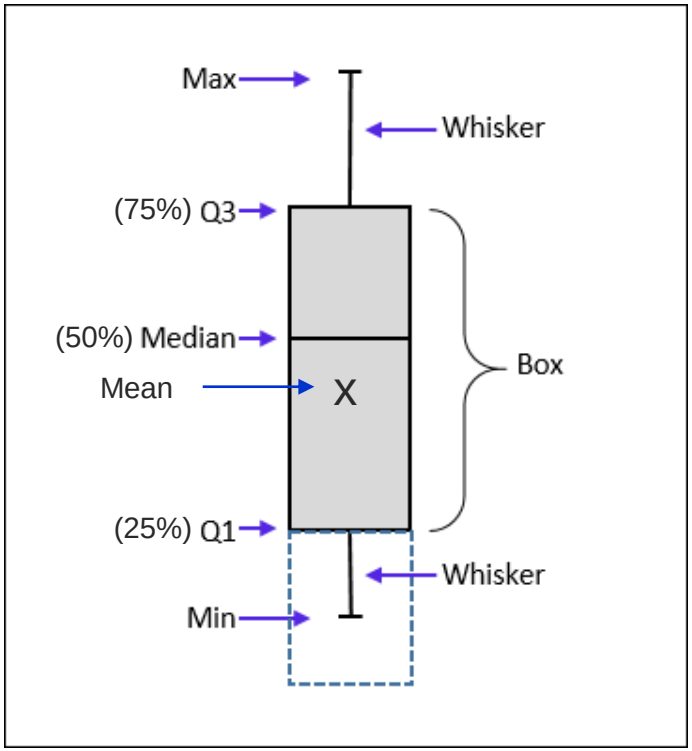
- Scatter plot is useful for assessing dependencies between two variables
  - It also shows how spread or concentrated

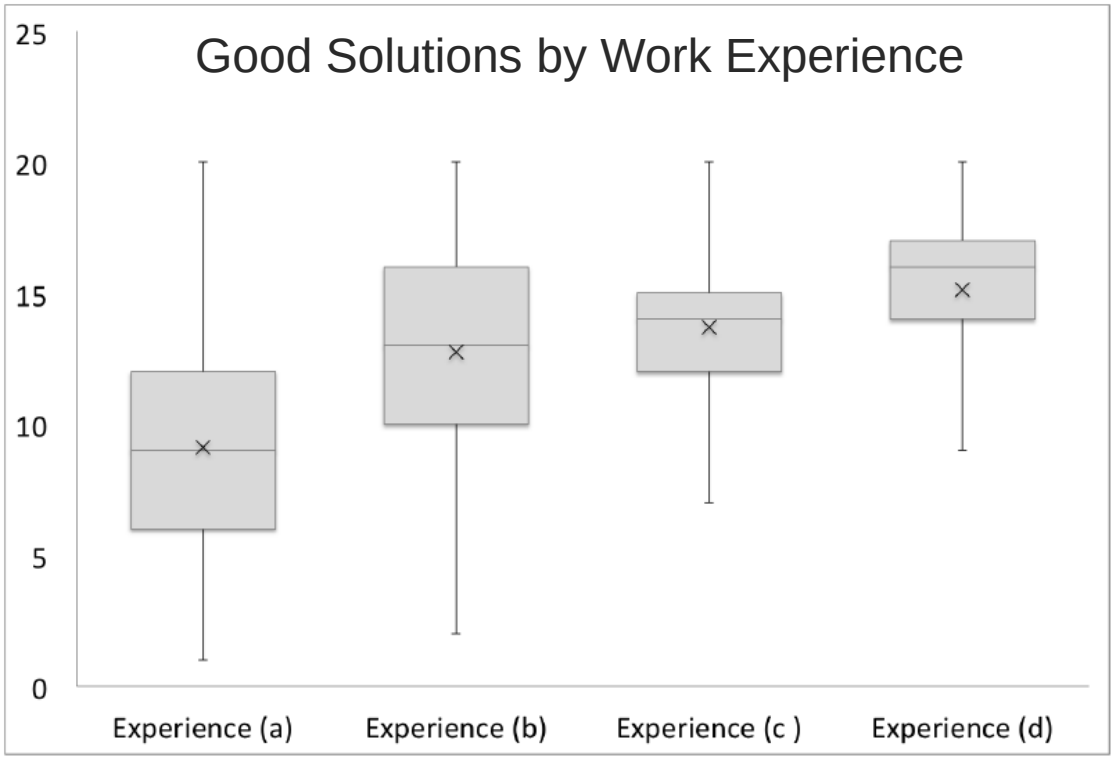




Box Plot

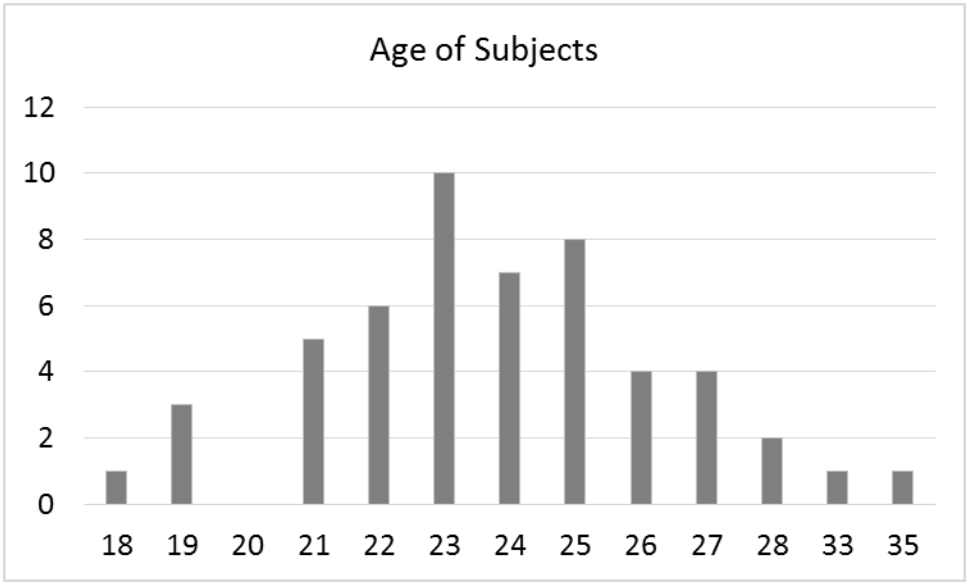
- Box plot is good for visualizing dispersion of sample data





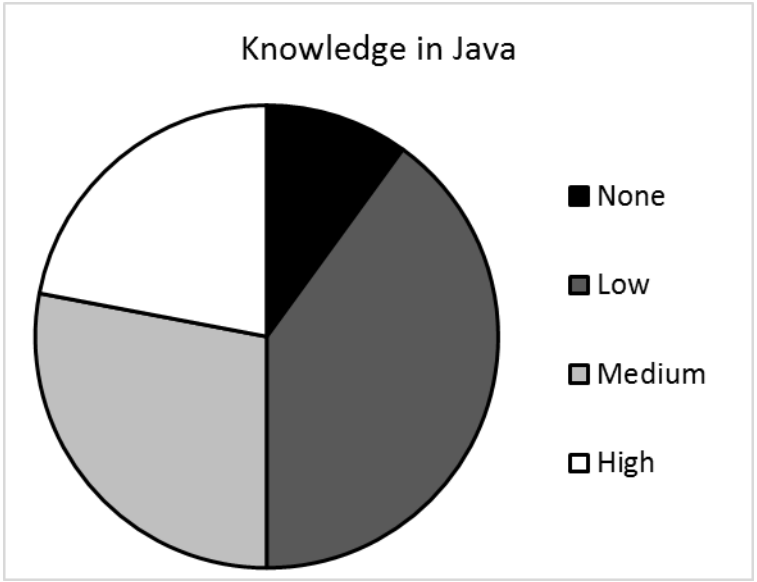
Histogram

- Histogram is used to give you an overview of the distribution density
  - It often presents the frequency of each value
  - It is useful to test normal distribution



Pie Chart

- Pie chart shows the relative frequency of the data values (in percentage)
  - It is useful to represent predominant values in the nominal scale



Bibliography

- C. Wohlin et al. **Experimentation in Software Engineering**, Springer. 2012.
  - Chapter 10 – Analysis and Interpretation (Section 10.1 Descriptive Statistics)