Workshop 8 - Statistics 1

Dr. Markus Hofmann

Overview

- Generate Summaries and make general statements about the data, and its relationships within the data is the heart of Exploratory Data Analysis
- We generally make assumptions on the entire population but mostly just work with small samples. Why are we allowed to do this???
- Two important definitions:
  - Population: A precise definition of all possible outcomes, measurements or values for which inference will be made about.
  - Sample: A portion of the population which is representative of the population (at least ideally)

Overview

- Parameters: numbers that characterise a population
- Statistics: summaries that characterise a sample which was taken from the population
- A sample should be an unbiased, random sample from the population
- Two statistical methods play an important role throughout the Exploratory Data Analysis stage:
  - **Summarising the data:** summarising sample data sets and making confident statements about the entire population
  - **Characterising the data:** Characterise the variables and the relationships between them
  - Making statements about "hidden" facts: Once a group of observations has been identified statistics give us the ability to make confident statements about these groups

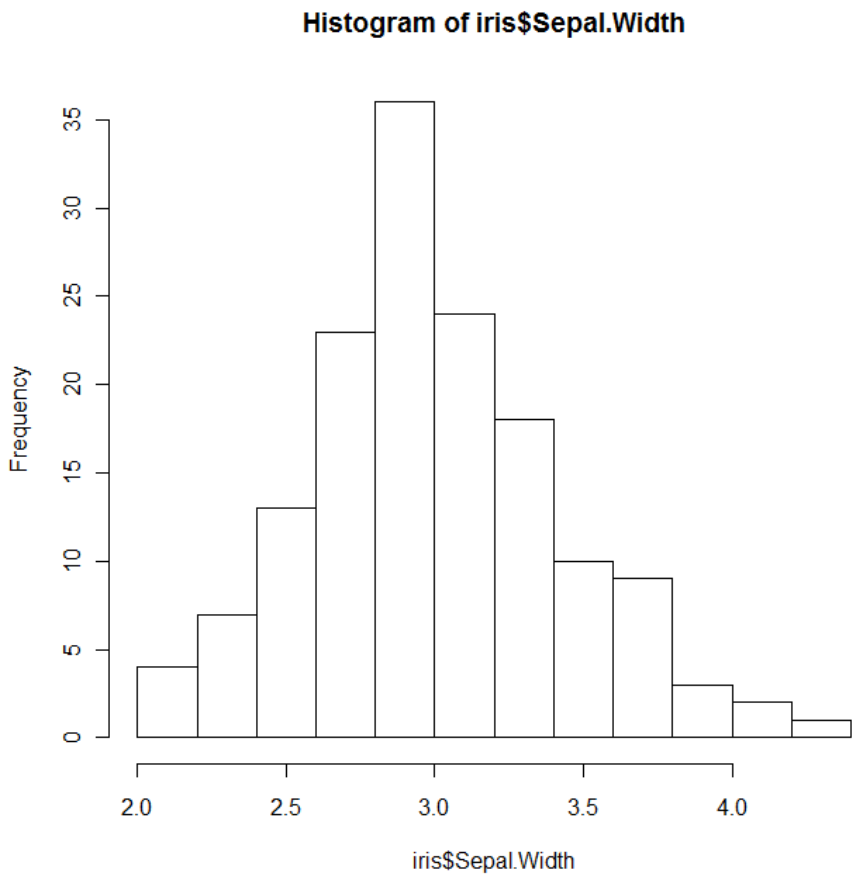
Overview

- Numerical scores are called: measurement or scale data.
- Rankings are called: ordinal data
- Categories such as gender are called: nominal data.

- Download and install the R project statistical package
  - http://www.r-project.org/
  - Also on my student share
    - Just copy the folder onto your machine

Descriptive Statistics

- Descriptive statistics describes the variables in a number of ways.
- E.g. The histogram of the variable Sepal Width shows a number of approx. descriptive statistics such as mean, median, skewness, distribution, etc…



Descriptive Statistics

- It allows us to calculate specific and precise measures of a variable.
- The following main areas are part of descriptive statistics:
  - Central Tendency Measures
    - Arithmetic Mean, Median, Mode
  - Measures of Variation
    - Range, Quartiles, Variance, Standard Deviation, z-score
  - Measures of Shape
    - Skewness, Kurtosis

Central Tendency and Variability

- **Measures of central tendency**
  - tell us about the most **typical scores**.
- **Measures of variability**
  - tell us about how the **scores** are **spread out**.

- Let's have a look at some measures….

Central Tendency - Mean

- Arithmetic Mean is more commonly known as Average
- It is defined by the sum of all values divided by the number of all values:

$$\bar{x} = \frac{1}{n} \cdot \sum_{i=1}^{n} x_i$$

- The mean is commonly defined as x bar ($\bar{x}$). n is the number of observations (examples). x_i is the individual value.
- In R: > mean(iris$Sepal.Width)
  - This results in a single return value representing the mean of the variable SepalWidth within the Iris dataset
- Other Means also exist such as Geometric Mean, Weighted Mean, Harmonic Mean, and Truncated Mean

Central Tendency - Median

- Median is the middle value of a variable once it has been sorted in ascending order (low to high).
- For even number of observations the two middle numbers need to be averaged.
- E.g. 2,2,4,6,7,**8**,**9**,9,11,14,16, 20
  8 and 9 are the two middle numbers therefore the median is 8.5 ((8+9)/2)
- Median can be calculated on ordinal, interval and ratio scales.
- In particular well suited for variables measured on the ordinal scale
- Unlike the average, the median will not be influenced by extreme values
- In R: > median(iris$Sepal.Length)

Central Tendency - Mode

- The mode is the most frequent value of a variable
- E.g. 2,2,4,6,7,8,9,9,11,14,16, 20
  The mode here is {2 , 9} as 2 and 9 both occur twice
- Mode provides the only measure for nominal type attributes
- Can also be calculated for ordinal, interval and ratio scales
- Provided by RapidMiner in the Meta Data View (for nominal attributes)
- Note: mode() gives you the storage mode and not the Mode
- In R:
  - my_mode <- table(iris$Sepal.Width)
  - my_mode
  - my_mode[which(my_mode ==max(my_mode))]

Measures of Variation - Range

- The Range is simply the difference between min and max value of a variable
- In R: > min(iris$Sepal.Width)
  > max(iris$Sepal.Width)
  > range(iris$Sepal.Width)

- Range can be used on Ordinal, Ratio and Interval scales

Measure of Variation - Quartile

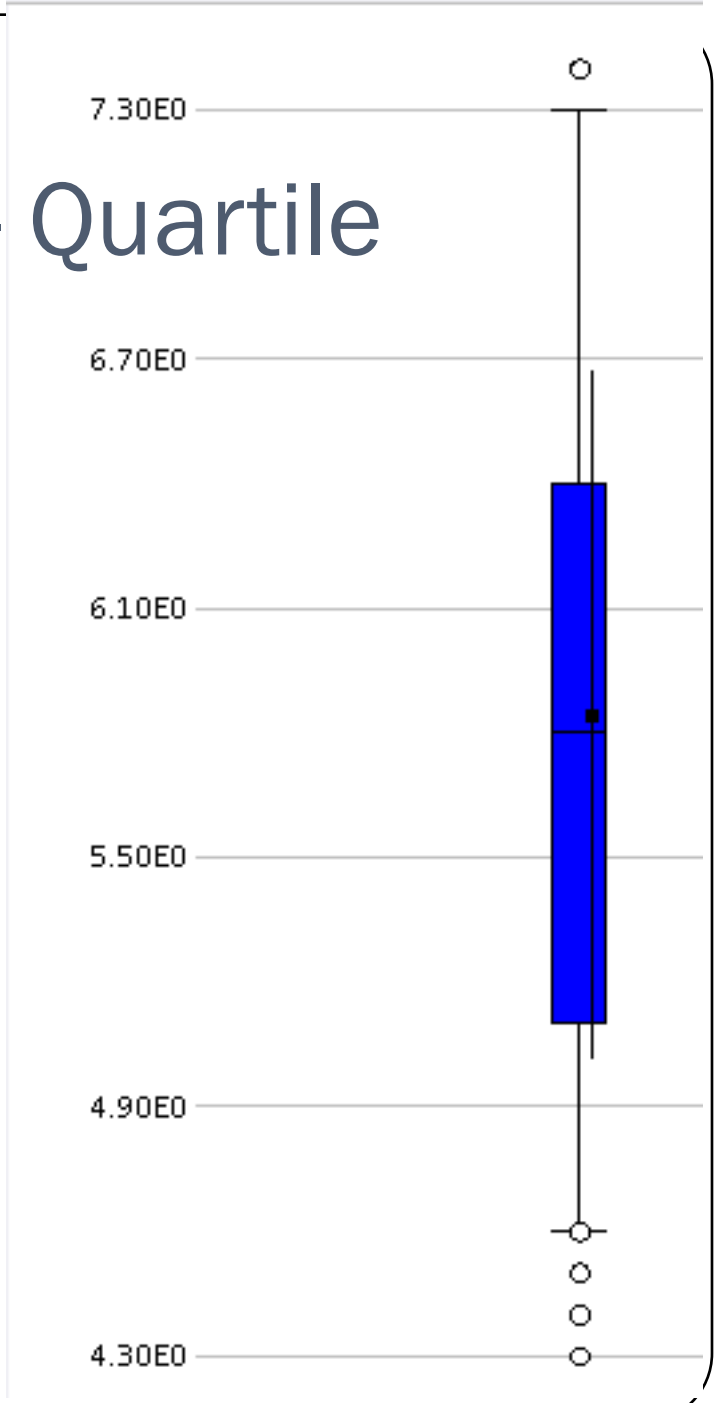
- Quartiles divide a variable into four even segments based on the number of observations.
- E.g. 100 observations the there will be 4 groups of 25 numbers
- First quartile (Q1) –25% mark
- Second quartile (Q2) – 50% mark (equal to median)
- Third quartile (Q3) – 75% mark
- Calculated similarly to Median. All values are ordered from low to high. The middle number(s) of the 0% to 50% numbers will become Q1 and the middle number(s) of 50%-100% numbers will become Q3
- Also known as quantiles or percentiles

Measure of Variation - Quartile

- In R:

```
> quantile(iris$Sepal.Length)
  0%  25%  50%  75% 100%
 4.3  5.1  5.8  6.4  7.9
```

- 0% and 100% are equivalent to min max values.
- BoxPlot shows this also
- In R:
- boxplot(iris$Sepal.Length)



Measure of Variation - Variance

- Variance describes the spread of the data
- It is a measure of deviation of a variable from the arithmetic mean
- For sample data the formula to calculate the variance is

$$s^2 = \frac{\sum_{i=1}^{n} \left(x_i - \bar{x}\right)^2}{n-1}$$

- S^2 is the sample variance. x_i is the actual value of example i, x bar is the arithmetic mean and n is the number of examples in the sample set.
- In R: > var(iris$Sepal.Length)

Measure of Variation – St. Deviation

- The Standard Deviation is the square root of the variance
- This measure is the most widely used to express deviation from the mean in a variable
- The higher the value the more widely distributed are the variable data values around the mean
- Assuming the frequency distributions approximately normal, about 68% of all observations are within +/- 1 standard deviation
- Approximately 95% of all observations fall within two standard deviations of the mean (if data is normally distributed).
- In R: > sd(iris$Sepal.Length)

Measure of Variation - z- score

- A way to compare two measurements originating from different distributions
  - Is 60% in Web Development a better or worse grade than 60% in Operating Systems?
- Z-Scores can be turned into percentages.
- Distribution needs to be 'normal'

Measure of Variation - z- score

- z-score represents how far from the mean a particular value is based on the number of standard deviations.
- The formula to calculate the z-score is
  - $$z = \frac{x - \bar{x}}{s}$$
  - where x is the data value of x_i, x bar is the arithmetic mean and s is the standard deviation
- z-scores are also known as standardized residuals
- Note: mean and standard deviation are sensitive to outliers
- In R:

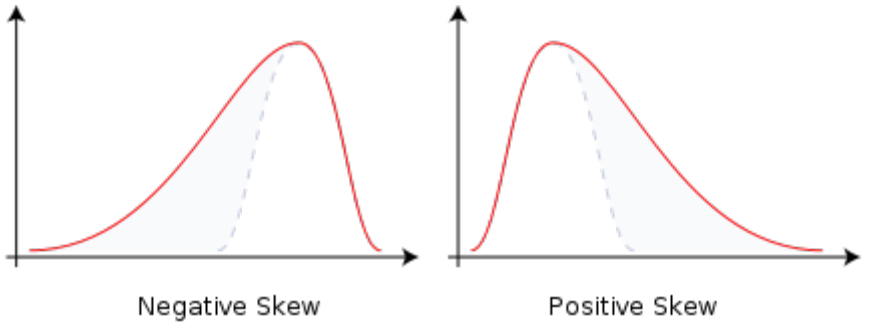
```
> x<-((iris$Sepal.Width)-mean(iris$Sepal.Width))/sd(iris$Sepal.Width)
> x
```

Shape of Distribution - Skewness

- Skewness is a method for quantifying the lack of symmetry in the distribution of a variable
- Skewness can be calculated using this formula:

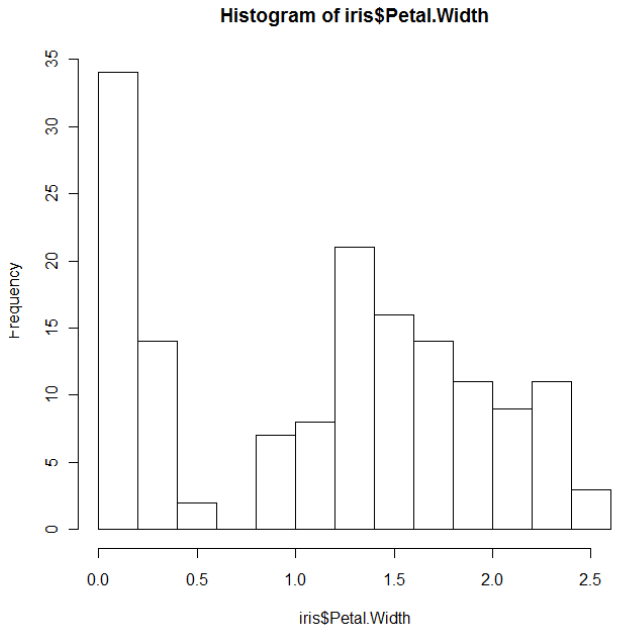
$$skewness = \frac{\sum_{i=1}^{n}(x_i - \bar{x})^3}{(n-1)s^3}$$

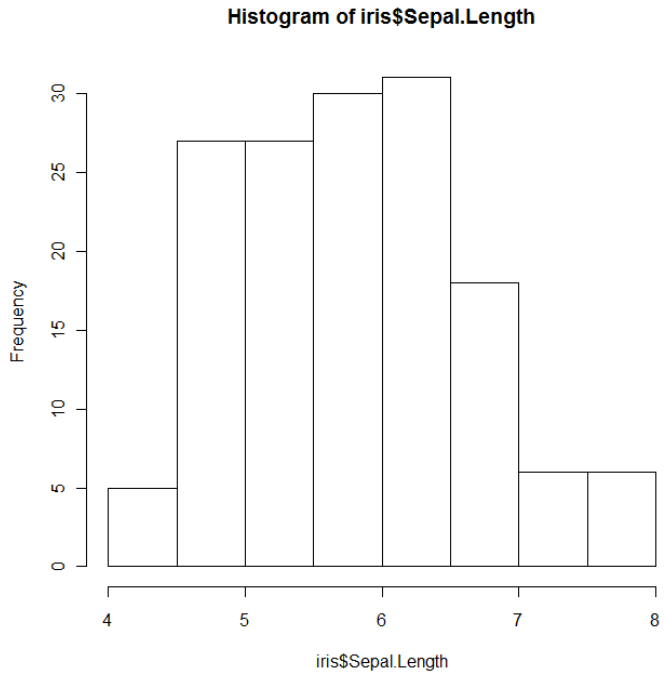
- Skewness value of zero indicates that the variable is distributed symmetrically. Positive number indicate asymmetry to the left, negative number indicates asymmetry to the right



Shape of Distribution - Skewness

- Shape functions are not part of the Basic package in R. The 'moments' package needs to be loaded
- In R: > skewness(iris$Sepal.Width)
- Skewness: -0.1019342 0.3117531





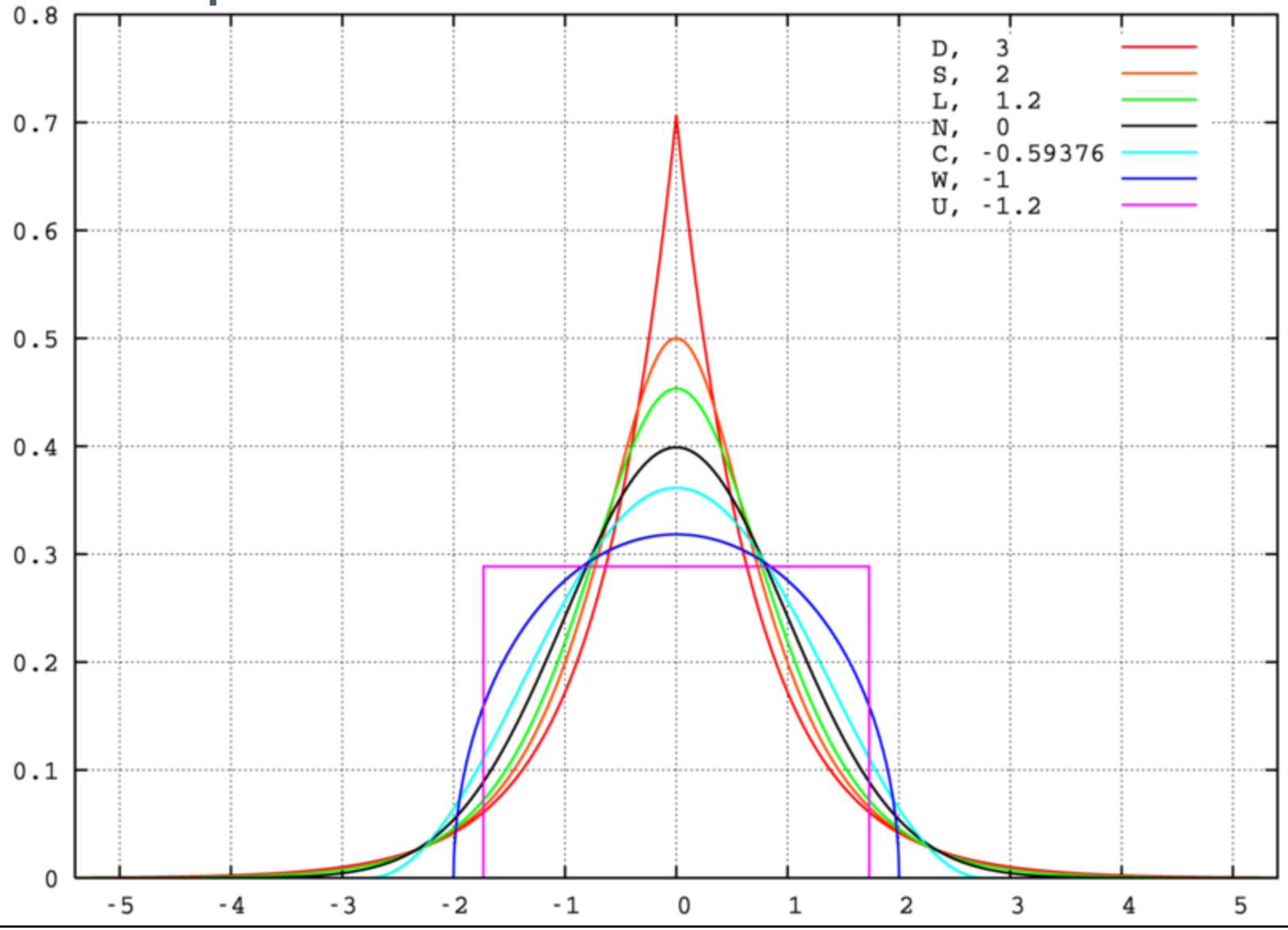
Shape of Distribution - Kurtosis

- Kurtosis is a measure that gives indication in terms of the peak of the distribution
- The following formula can be used to calculate the kurtosis:

$$kurtosis = \frac{\sum_{i=1}^{n}(x_i - \bar{x})^4}{(n-1)s^4}$$

- Variables with a pronounced peak toward the mean have a high Kurtosis score and variables with a flat peak have a low Kurtosis score.

Shape of Distribution - Kurtosis



Things to know

- **The mean is the centre** of normal distributions.
- **For skewed distributions** - report both the mean and median.
- **For normal distributions** - the mean, median and mode are equal.
- **The mean is unbiased** – on the average, sample means aren't always too high or too low.
- For non-normal distributions use the **range** and the **median** to summarize the distribution.