Continuous Mathematical Foundations: Week 1 - Introduction to Statistics

Dr. Georgios Stagakis

City College

October 2021

Statistics

Statistics is the discipline that concerns the collection, organization, analysis, interpretation, and presentation of data. [1]

[1] A Dictionary of Statistics (2 rev. ed.)

Goals of the Week

- You will learn to extract and present, simplistically, information from loads of data.

- You will learn to make comparisons for cases where there is uncertainty.

Basic Tools in Statistics

- Plots (Pie Charts - Histograms - Scatter Plots)
- Measures (of Centrality and Dispersion)

Plots

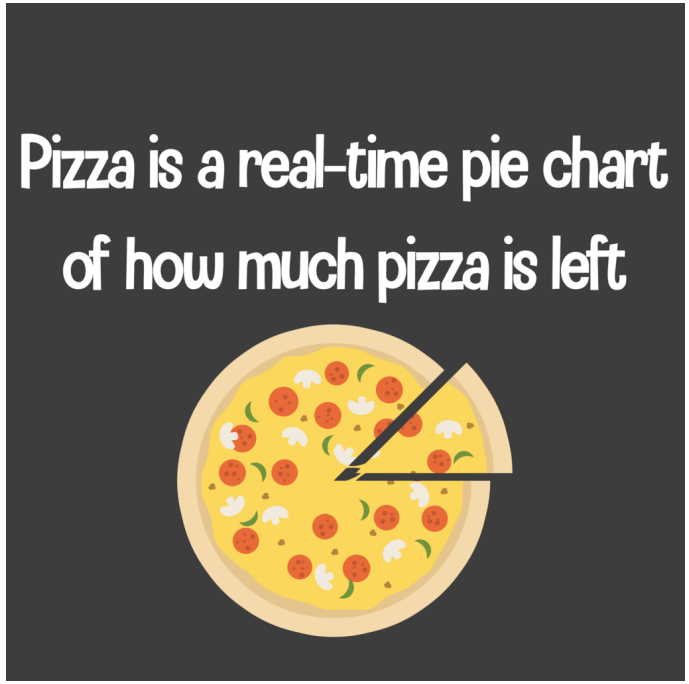
- Pie Chart: For discrete observations with their respective frequency, i.e. the percentage of votes of each political party.

- Histogram: For continuous observations with their respective frequency,
  i.e. the weight of the population of a country.

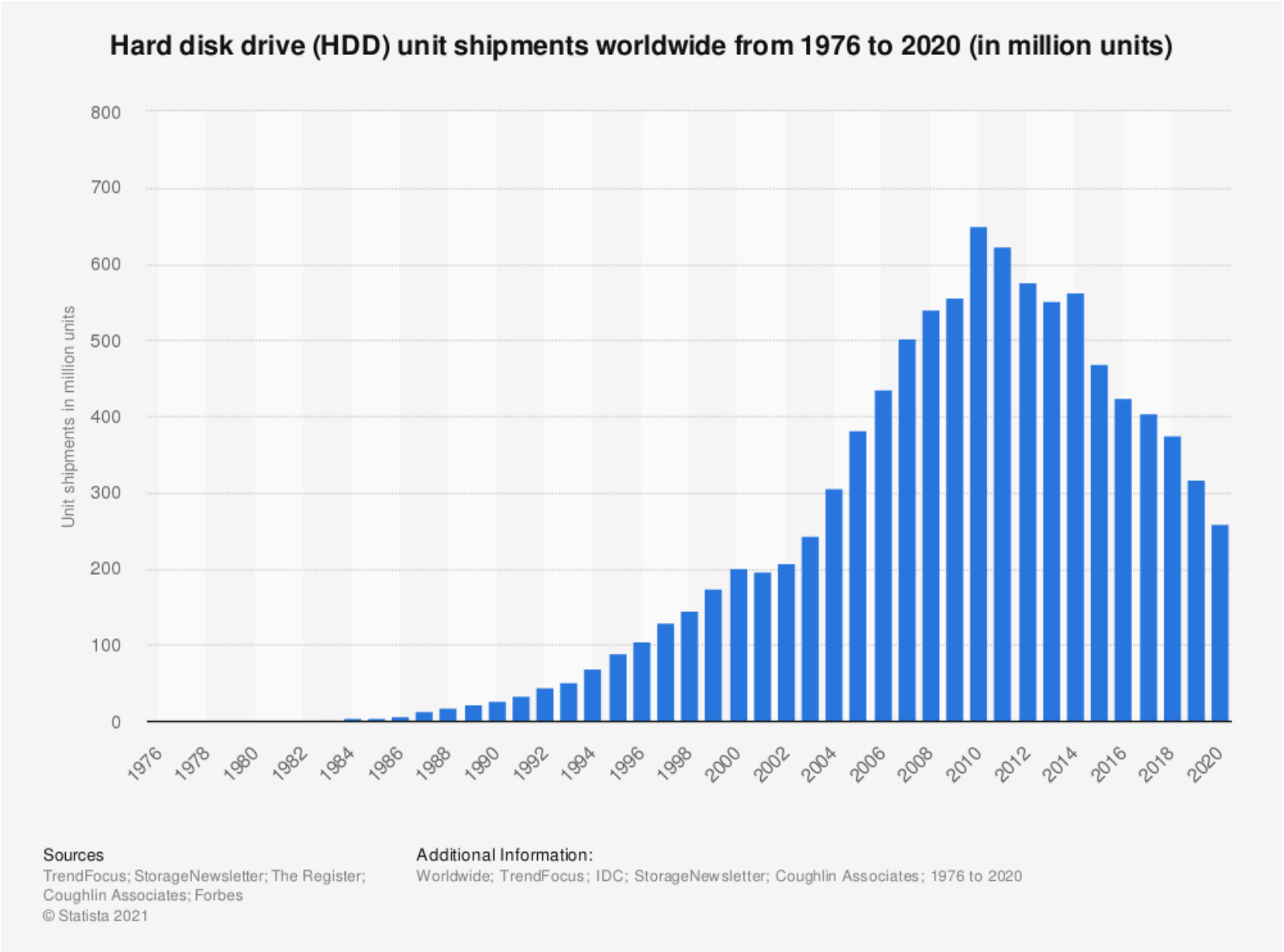
- Scatter Plot: For bivariate observations,
  i.e. the weight and height of the population of a country.

Pie Chart


13th Cyprus Summit
C-suite executives
Editors/analysts
Diplomats
Academics
Politicians
Managers/Advisers
various

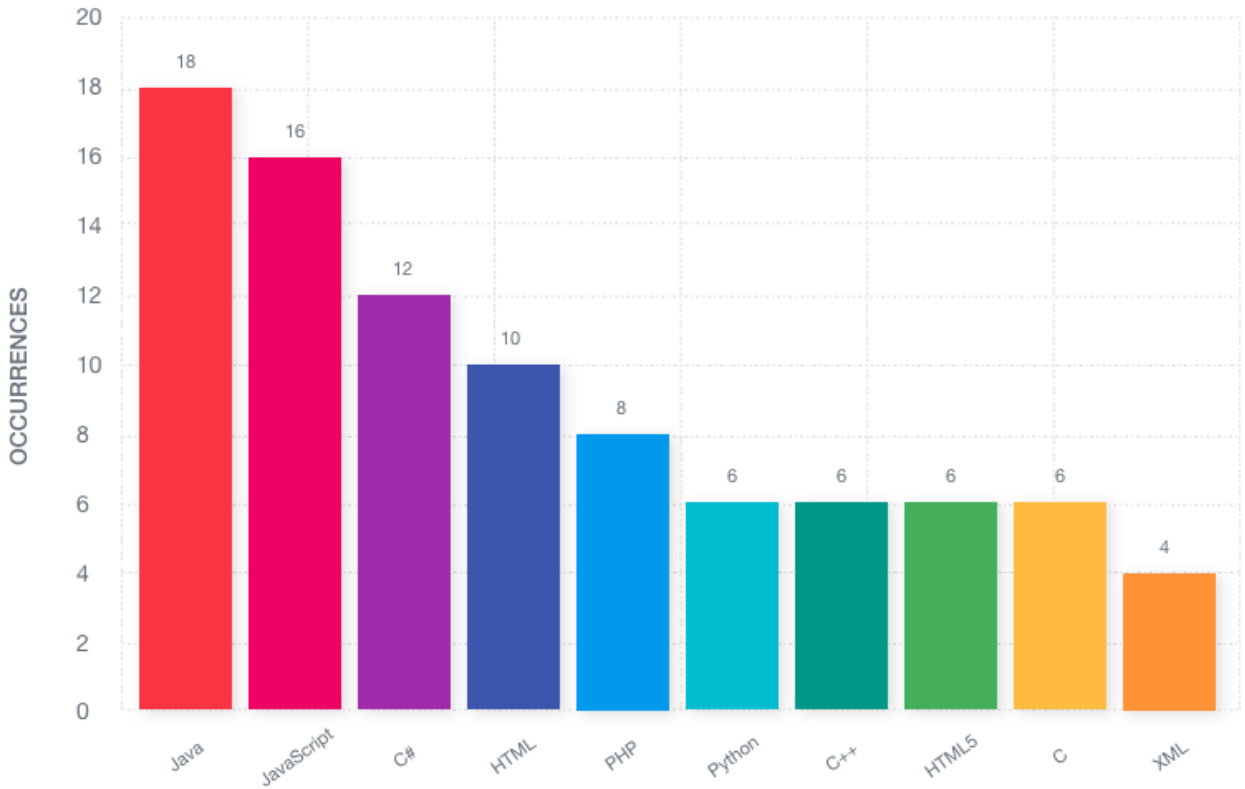

Pizza is a real-time pie chart
of how much pizza is left

Histogram

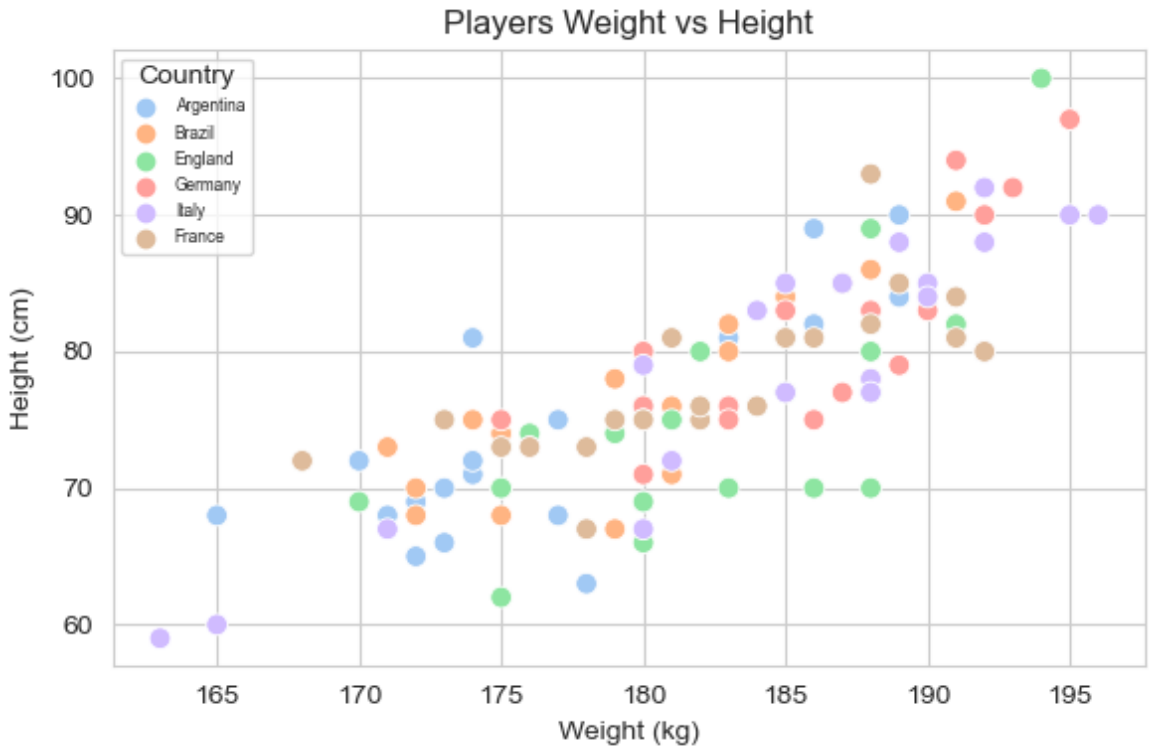

Hard disk drive (HDD) unit shipments worldwide from 1976 to 2020 (in million units)
Unit shipments in million units
800
700
600
500
400
300
200
100
0
1976
1978
1980
1982
1984
1986
1988
1990
1992
1994
1996
1998
2000
2002
2004
2006
2008
2010
2012
2014
2016
2018
2020
Sources
TrendFocus; StorageNewsletter; The Register; Coughlin Associates; Forbes
© Statista 2021
Additional Information:
Worldwide; TrendFocus; IDC; StorageNewsletter; Coughlin Associates; 1976 to 2020

Attention!!!

Not a histogram, Bargraph!!!!



Scatter plot



Measures of Centrality

- Measures of Centrality are used in order to calculate values that represent the main values of a sample. They are also used for predictive purposes.

- In the next slides we will see three Measures of Centrality, Mean ($\bar{x}$), Median ($\hat{\mu}_{1/2}$) and Mode (m).

- For the next slides we will use "n" as notation for the number of observations in a sample and $x_1, x_2, ..., x_n$ for notation of the sample values.

Mean $\bar{x}$

Mean is calculated as the sum of the observations of a sample $(x_1, x_2, ..., x_n)$, divided by the number of observations (n),

$$\bar{x} = \frac{x_1 + x_2 + ... + x_n}{n}.$$

Mean example

Example

In our class, we have a sample from 5 students about how many hard disks they have included in their systems. The observations are,

$$1, 3, 1, 1, 2.$$

The mean of the disks in our sample is

$$\bar{x} = \frac{x_1 + x_2 + ... + x_n}{n} = \frac{1 + 3 + 1 + 1 + 2}{5} = 1.6 \approx 2.$$

Through $\bar{x}$ we can approximate the number of disks per person in the class.

Median $\hat{\mu}_{1/2}$

Median is the value that is placed in the middle of the sample, if it is ordered. If there is no middle value (n is even) it is the mean of the two middle values.

Median example

In our previous example (page 13) in order to find median first we have to order the sample from lowest to highest or vice versa, i.e. $1, 1, 1, 2, 3$. The value in the middle is the third value and it is equal to 1, or else $\hat{\mu}_{1/2}$=1.

The previous sample had 5 values (odd). If it had 6 (even), i.e. the sample was $1, 1, 1, 2, 2, 3$, then the median would be the mean of the third and forth value, $\hat{\mu}_{1/2} = \frac{1+2}{2} = 1.5$.

Mode m

- Mode is the value that appears most often in a sample.
- In our previous example (page 13), 1 appears more times than any other value in the sample, as a results m=1.
- Mode can be more than one value, i.e. for the sample $1, 1, 1, 2, 2, 2, 3$, m=1 and 2.

Extreme Value example

We have the following sample

$$1, 1, 1, 1, 1, 1, 2, 2, 2, 2, 2, 3, 50,$$

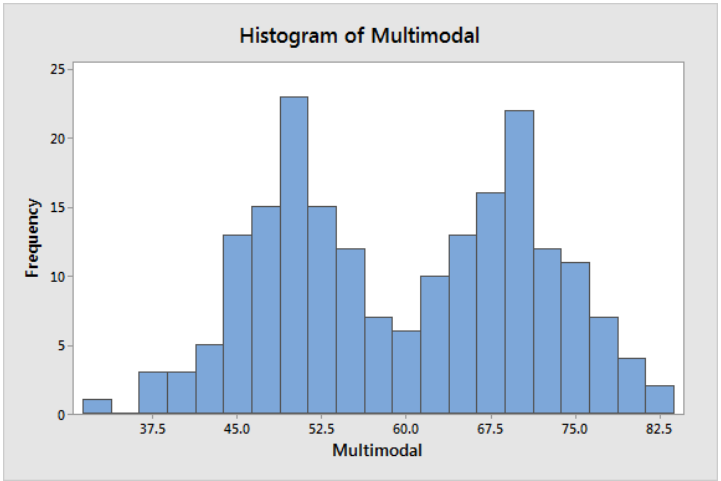
that represents the number of hard disks in computer systems per household. The last value probably belongs to a person working with an HPC (high performance computer) for professional purposes.

Please check that for this sample $\bar{x} \approx 5.31$, $\hat{\mu}_{1/2} = 2$ and $m = 1$.

Which do you believe is the most appropriate measure of centrality to describe the number of hard disks in systems per household?

Comments on Centrality Measures

- Mean is highly affected by extreme values.
- Median and Mode are not affected at all by extreme values.
- Mean is easy to calculate and bears many beneficial mathematical properties.
- Median is good replacement for mean in case of extreme values.
- Mode should be used when we have many different high likelihood values and low likelihood in between, such as in the plot below.



Measures of Dispersion

- Measures of Dispersion are used in order to calculate a scalar value representing the deviation of the sample from the Measures of Centrality. They are usually used for comparison of dispersion between samples.
- The only Measures of Dispersion we will focus on are Variance (s^2) and range (r). The root of variance is called standard deviation ($s = \sqrt{s^2}$).
- Variance is calculated as

$$s^2 = \frac{(x_1 - \bar{x})^2 + (x_2 - \bar{x})^2 + ... + (x_n - \bar{x})^2}{n - 1}.$$

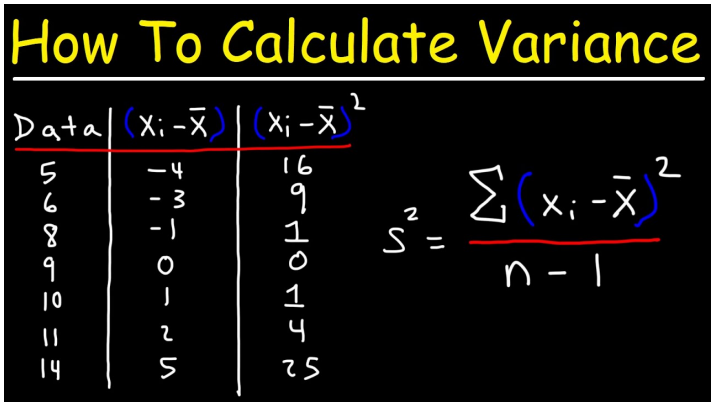
- Range is calculated as the maximum minus the minimum value of the sample,

$$r = max\{x_1, x_2, ..., x_n\} - min\{x_1, x_2, ..., x_n\}.$$

Variance Calculation

Below algorithmic steps are given for the calculation of the variance,

- Calculate the mean value $\bar{x}$ of the sample.
- Transform the sample from $x_1, x_2, \ldots, x_n$, to $y_i = x_i - \bar{x}$.
- Transform the sample from $y_1, y_2, \ldots, y_n$, to $z_i = y_i^2$.
- Sum $z_1, z_2, \ldots, z_n$ and divide by n-1.



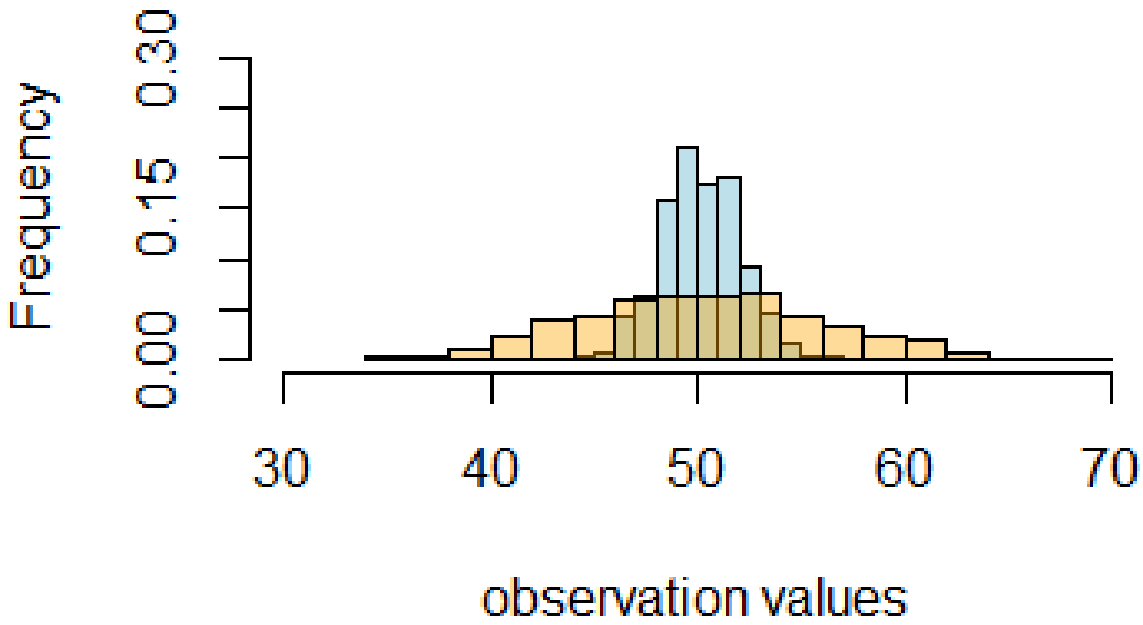
Example Description

We want to buy a software service that checks for malware in our databases. We received demos from two companies, Enohart and Pear, in order to make tests for a trial period and check if what was provided can cope with our needs.

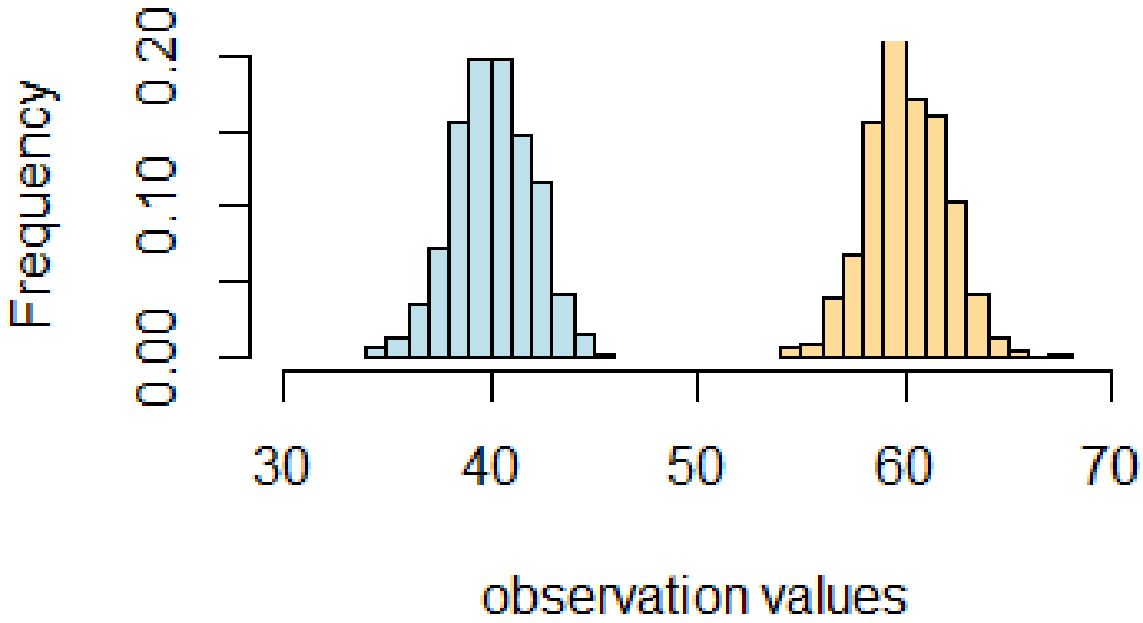
In the next pages we see how both software corresponded in different days, the amount of time they needed to run a full scan.

Enoh@art

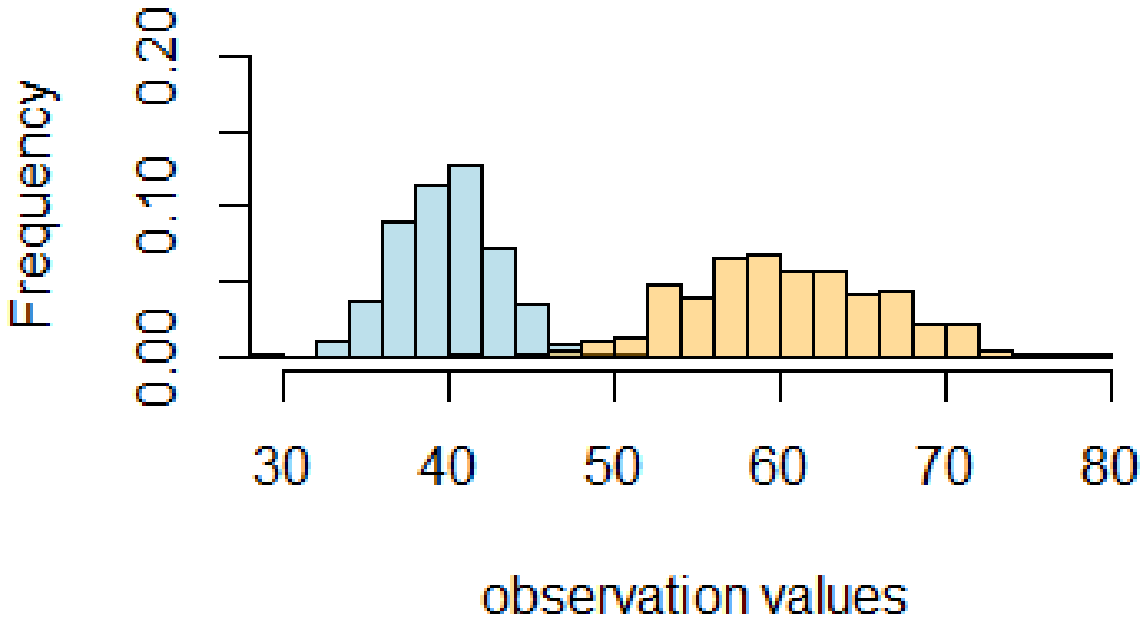
Example 1:
Wednesday noon- Equal means, different variances



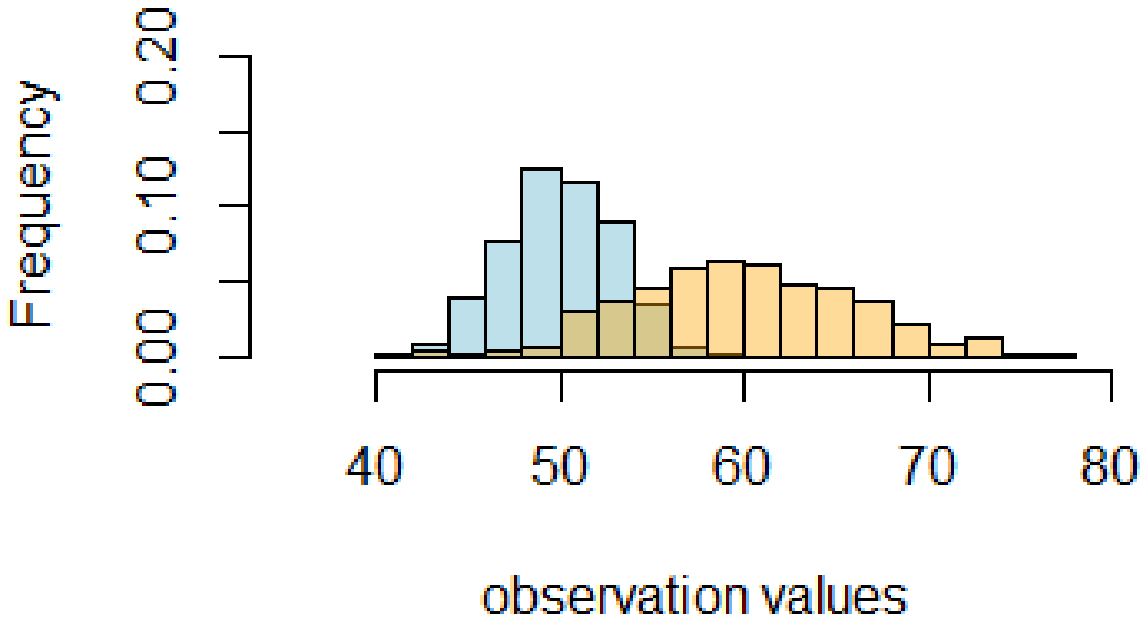
Example 2:
Thursday night- Different means, equal variances



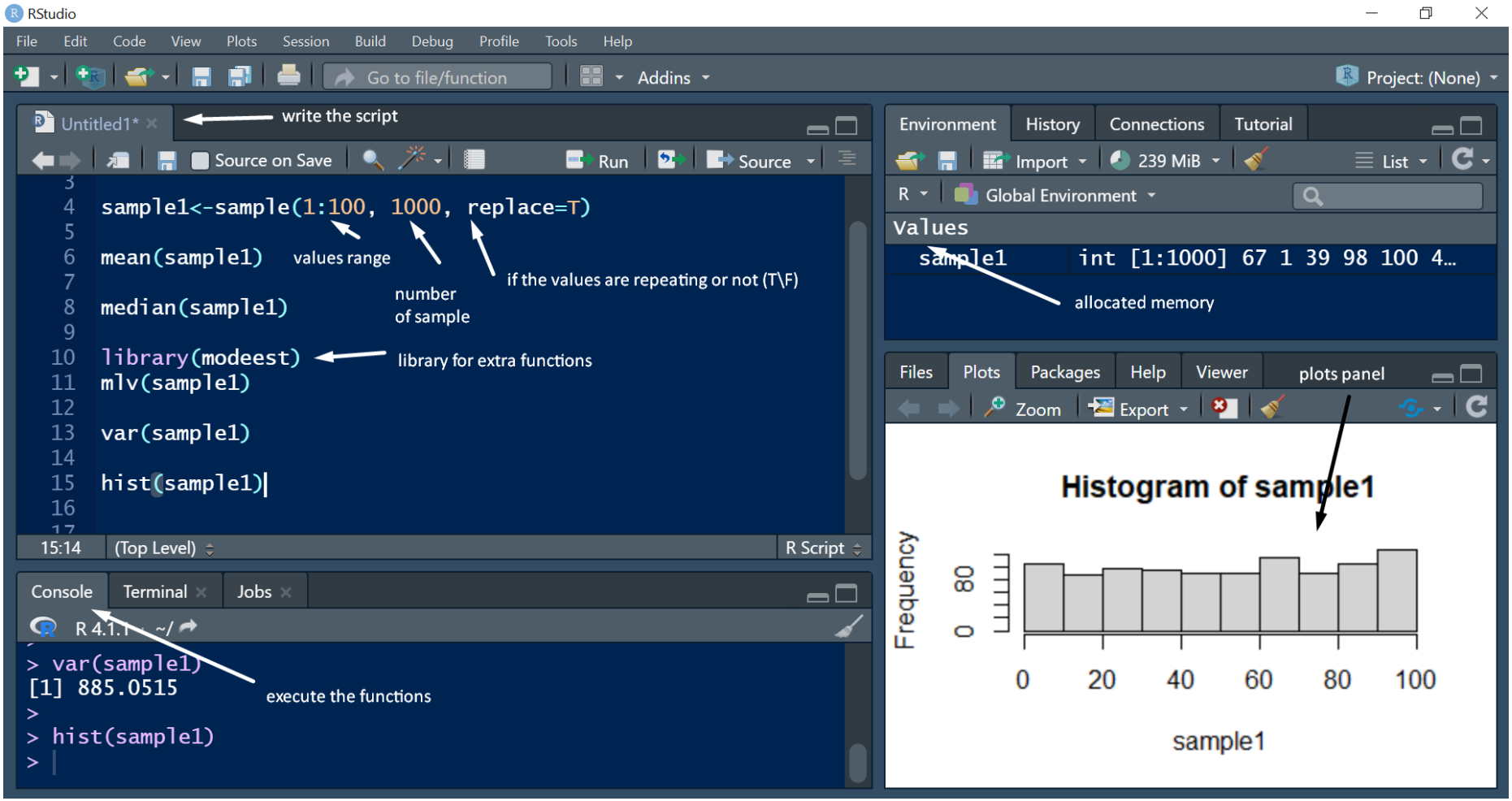
Example 3:
Saturday morning-Different means, different variances



Example 4:
Saturday night-Different means, different variances



Statistics in R



Summing up

After Week 1 you have to know, how to

- identify and understand pie charts, histograms and scatter plots.
- calculate and interpret mean, median, mode and variance.
- make comparisons between samples based on mean and variance.