Introduction to Statistics

ES 2016/2017

Adam Belloum

a.s.z.belloum@uva.nl

Content

- Definitions
- Sample and Population
- Type of Variables: categorical, Ordinal, ratio
- Data Collection: sampling
- Summary measure
- Data Representation: frequency table, histograms, Bar Chart and Frequency polygons, Box-plot, …
- Data Collection: Observational studies and Experiments

Objective of this lecture

- the objective of this introduction is to cover basic statistical knowledge about the collection and interpretation of data :
  - display and summarise large amounts of quantitative information, before undertaking a more sophisticated analysis.

What is statistical analysis?

- Statistics:

  "… determination of the probable from the possible"

  Davis, Statistics and data analysis in geology, p. 6

- Inferential statistics:
  - from samples to populations → what could have been or will be observed if we have analysed the entire population
- Descriptive statistics:
  - numerical summaries of samples → what was observed

Why use statistical analysis?

- Descriptive statistics: summarize some data in a **shorter** form
- Inferential statistics: understand some process and possible predict based the outcome
  - We need a conceptual representation (model) from which we infer the process.
  - We need to know if the model is "**correct**"?
    - Are we imagining relations where there are none?
    - Are there true relations we haven't found?

Statistical analysis gives a way to quantify the confidence we can have in our inferences.

What is "statistical analysis"?

This term refers to a wide range of techniques to...

- Describe
- Explore
- Understand
- Prove
- Predict

...based on sample datasets collected from populations, using some sampling strategy.

Populations and samples

- Population: a **set** of all elements (individuals)
  - Finite vs. "infinite"
- Sample: a **subset** of elements taken from a population

We make inferences about a population from a sample taken from it.

- If it is feasible to examine the entire population; then there is no inference from a sample.
  - Example: all pixels in an image.
- Questions: **is the sample** Representative or biased?

Populations and samples

Representative vs. biased

- For example, if we are interested in conducting a survey of the amount of *physical exercise* undertaken by the general public.

- surveying persons entering and leaving a **gym** would provide:
  - a biased **sample** of the **population**,
  - and the results obtained cannot generalise to the population

Populations and samples

Representative vs. biased

An Example of a non-representative sample: The *Literary Digest* Poll on the 1936 U.S. Presidential elections

Literary digest magazine sent survey to 10 million people who where subscribers or owned car or telephones

- ✓ 2.3 million people responded
- ✓ The results incorrectly predicted a landside for the republican candidate

Populations and samples

An Example of a could be non-representative sample: The change in Canada in 2010 of the National Household Survey (NHS).

- *long-form census* was given to a random sample of 1/5 of household and was mandatory.
- NHS is given to 1/3 of household and is voluntary

- How might voluntary response affect the quality of data received

Alison Gibbs and Jeffrey Rosenthal, Statistics Making sense of data, university of Toronto,[coursera]

Types of Bias of a sample

- **Selection Bias**: Occurs when the sample is selected in such away that it systematically excludes on under-represent part of the population
- **Measurement or Response Bias**: Occurs when the data are collected in such a way that it tends to result in observed values that are different from the actual value in some systematic way
- **Nonresponse Bias**: Occurs when responses are not obtained from all individuals selected for inclusion in a sample

Alison Gibbs and Jeffrey Rosenthal, Statistics Making sense of data, university of Toronto,[coursera]

How to select a representative Sample

Some method of non-random sample selection

- Systematic sampling:
  - select every k^{th} individual from a list of the population,
  - where the position of the first person chosen is randomly selected form the first k individuals.

Note: This will give a non-representative sample if there is a structure in the list

- Convenience or Volunteer Sampling:
  - Use the first n individuals that are available or the individuals who volunteer to participate.

Note: This is almost sure to give a non-representative sample which cannot be generalised to the population

Alison Gibbs and Jeffrey Rosenthal, Statistics Making sense of data, university of Toronto,[courera

How to select a representative Sample

Use Randomisation

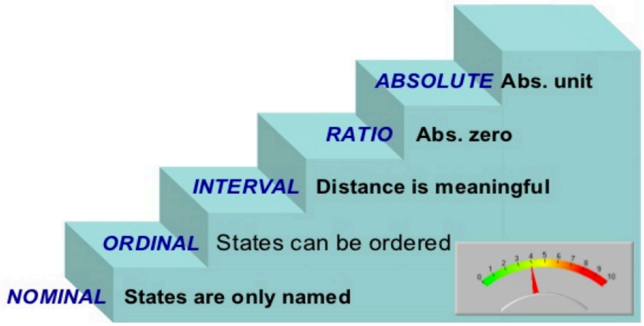
Some methods of random sample selection:

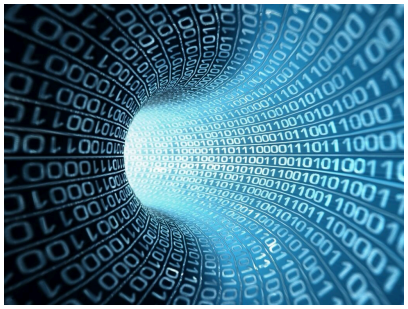
- **Sample Random Sampling** (SRS): Each possible sample of size n from the populations is equally likely to be the sample that is chosen.
- **Stratified sampling**: Divide the population into non-overlapping subgroups called starts and choose a SRS within each subgroup.
- **Cluster Sampling**: dived the population into non-overlapping subgroups called clusters, select clusters at random, and include all individuals in the chosen cluster in the sample

Alison Gibbs and Jeffrey Rosenthal, Statistics Making sense of data, university of Toronto,[coursera]

Variables, Observations, and data

- Variables are quantities measured on a sample
- Observation a particular outcome.
- Data is a collection of Several observations.





Type of Variables

- Qualitative variables (Nominal or Categorical) have:
  - non-numeric outcomes,
  - with no **natural** ordering.

  For example, gender, disease status, and type of car

- Quantitative variables (ordinal, Interval, Ratio) have
  - numeric outcomes
  - can be **discrete** or **continuous**.

  For example, survival time, height, age, number of children, and number of faults

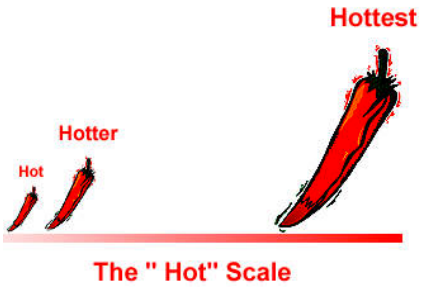
Qualitative variables: Nominal (Categorical)


Sorting into categories...

- Values are from a set of classes with **no natural** ordering
  - Example: Land uses (agriculture, forestry, residential . . .)
- Can determine **equality**, but **not rank**
- Meaningful sample statistics:
  - mode (class with most observations);
  - frequency distribution (how many observations in each class)

- **Note**:
  - Numbers may be used to label the classes but these are arbitrary and have no numeric meaning (the "first" class could just as well be the "third");
  - ordering is by convenience (e.g. alphabetic)

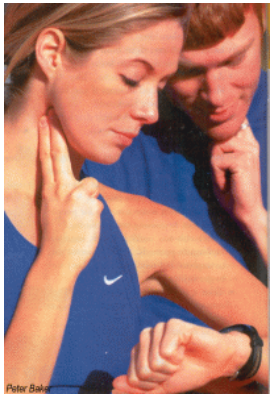
Quantitative variables: Ordinal



- Values are from a set of naturally ordered classes with no meaningful units of measurement
  - Can determine rank (greater, less than)
  - This ordering is an intrinsic part of the class definition
  - Example: Soil structural grade (0 = structureless, 1 = very weak, 2 = weak, 3 =medium, 4 = strong, 5= very strong)
- Meaningful sample statistics:
  - mode;
  - frequency distribution

Note: Numbers may be used to label the classes; their order is meaningful, but not the intervals between adjacent classes are not defined (e.g. the interval from 1 to 2 vs. that from 2 to 3)

Quantitative variables: Interval



- Values are measured on a **continuous** scale with well-defined units of measurement
  - no **natural origin** of the scale,
  - i.e. the zero is arbitrary,
- Example: Temperature in C.
  - "**It is twice as warm yesterday as today**" is meaningless, even though "Today it is 20C and yesterday it was 10C" may be true.
- Meaningful statistics: quantiles, mean, variance

Quantitative variables: Ratio

- Values are measured on a continuous scale with well-defined units of measurement
  - a **natural origin** of the scale, i.e. the zero is meaningful
- Examples: Temperature in K; concentration of a chemical in solution
  - "There is twice a much heat in this system as that" is meaningful, if one system is at 300K and the other at 150K
- Meaningful statistics:
  - Quantile, mean, variance;
  - the coefficient of variation. (CV = SD / Mean; is a ratio).

Summary measures

- A set of data on its own is very hard to interpret it is often useful to obtain **quantitative summaries** of certain aspects of the data.

- Summary measurements are divided in two types:
  - quantities which are "**typical**" of the data, known as **measures of location**
  - quantities which **summarise** the **variability** of the data, Known as **measures of spread**.

Summary measures

- Measures of location
  - Sample mean
  - Sample median
  - Sample mode
- Measures of spread
  - Range
  - Mean absolute deviation (M.A.D.)
  - Sample variance and standard deviation
  - Quartiles and the interquartile range
  - Coefficient of variation

Summary measures: Measures of location

- Sample **mean**:

$$\bar{x} = \frac{x_1 + x_2 + \cdots + x_n}{n} = \frac{1}{n}\sum_{i=1}^{n} x_i.$$

- Sample **median**: is the **middle** observation
  when the data are **ranked** in increasing order.

$$\text{SampleMedian} = \begin{cases} x_{\left(\frac{n+1}{2}\right)}, & n \text{ odd}, \\ \frac{1}{2}x_{\left(\frac{n}{2}\right)} + \frac{1}{2}x_{\left(\frac{n}{2}+1\right)}, & n \text{ even}. \end{cases}$$

- Sample **mode**: is the value which occurs with the **greatest frequency**

$$\text{SampleMode} = \{y_k | f_k = \max_i \{f_i\}\}.$$

Summary measures: Measures of spread

- **Range**: **difference** between the **largest** and **smallest** observation

$$\text{Range} = x_{(n)} - x_{(1)}$$

- **Mean absolute deviation** (M.A.D.): average **absolute deviation** from the **sample mean**

$$\text{M.A.D.} = \frac{|x_1 - \bar{x}| + \cdots + |x_n - \bar{x}|}{n} = \frac{1}{n}\sum_{i=1}^{n} |x_i - \bar{x}|$$

- Sample **variance** & **standard deviation**: **average squared distance** of the observations from their mean value

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

Summary measures: Measures of spread

- **Quartiles** and the **interquartile range**
  - **lower quartile** (LQ) has a quarter of the data less than it→ (n+1)/4th
  - **upper quartile** (UQ) has a quarter of the data above it.→ 3(n+1)/4th
  - **inter-quartile** (IRQ) range is the difference between the upper and lower quartiles

Calculating lower quartiles

$n = 15$	LQ at $(15+1)/4 = 4$	LQ is $x_{(4)}$
$n = 16$	LQ at $(16+1)/4 = 4\frac{1}{4}$	LQ is $\frac{3}{4}x_{(4)} + \frac{1}{4}x_{(5)}$
$n = 17$	LQ at $(17+1)/4 = 4\frac{1}{2}$	LQ is $\frac{1}{2}x_{(4)} + \frac{1}{2}x_{(5)}$
$n = 18$	LQ at $(18+1)/4 = 4\frac{3}{4}$	LQ is $\frac{1}{4}x_{(4)} + \frac{3}{4}x_{(5)}$
$n = 19$	LQ at $(19+1)/4 = 5$	LQ is $x_{(5)}$

$$\text{IQR} = \text{UQ} - \text{LQ}.$$

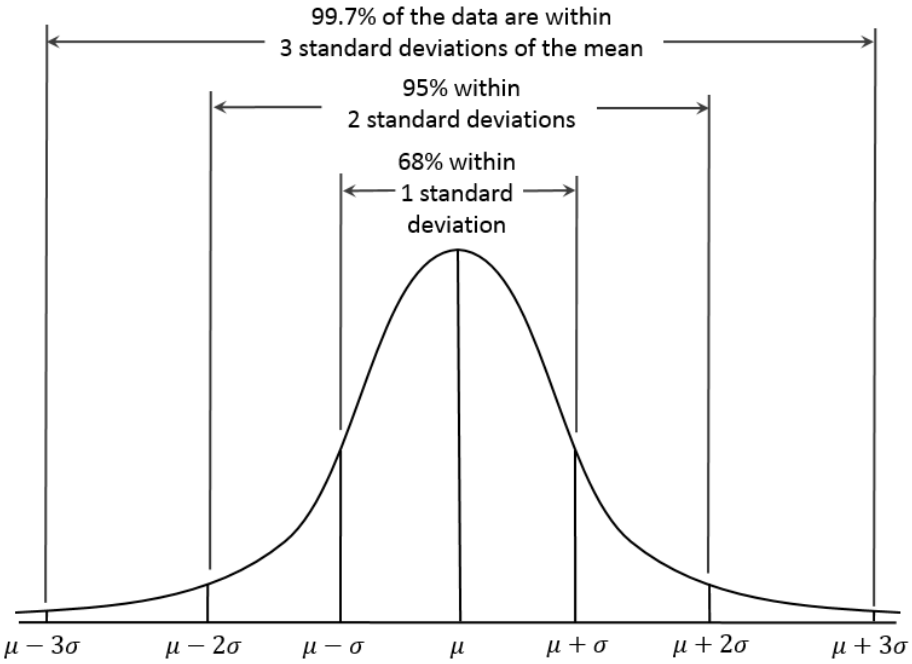
- **Coefficient of variation**: is the ratio of the standard deviation to the mean

$$\text{Coefficient of variation} = \frac{s}{\bar{x}},$$

The empirical rule

- If the data distribution is unimodal and symmetric → There is empirical rule

The empirical rule states that for a normal distribution, nearly all of the data will fall within three standard deviations of the mean.

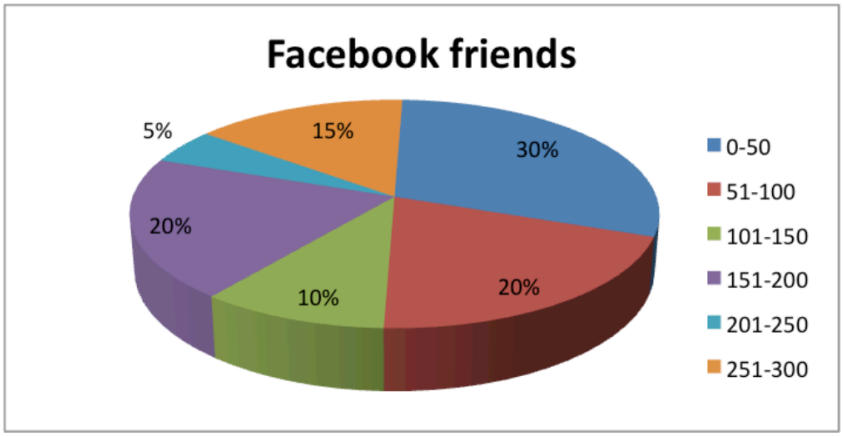


Data presentation

- Frequency tables
- Histogram
- Bar Chart and Frequency polygons
- Box-and-whisker plots
- …

Pie chart

- Suitable to represent categorical data;
- Used to show percentages;
- Areas are proportional to value of category



Data presentation: Frequency tables

Frequency table shows a tally of the number of data observations in different categories

- For **discrete** **qualitative** & **quantitative** data, we use all of the observed values as categories.
  - if the number of observations is large, consecutive observations may be grouped to form **combined categories**
- For **continuous** data, the choice of **categories** is more arbitrary.
  - Usually 8 to 12 non-overlapping consecutive intervals of equal width are used

Data presentation: Frequency tables

The number of calls from motorists per day for roadside service was recorded for the month of December

Class interval	Tally	Frequency
0 - 39	I	1
40 - 79	𝍸	5
80 - 119	𝍸 𝍸 II	12
120 - 159	𝍸 III	8
160 - 199	IIII	4
200 - 239	I	1
	Sum =	31

(germinating seeds), we can construct the following frequency table.

No. germinating	85	86	87	88	89	90	91	92	93	94
Frequency	3	1	5	2	3	6	11	4	4	1

$n = 40$

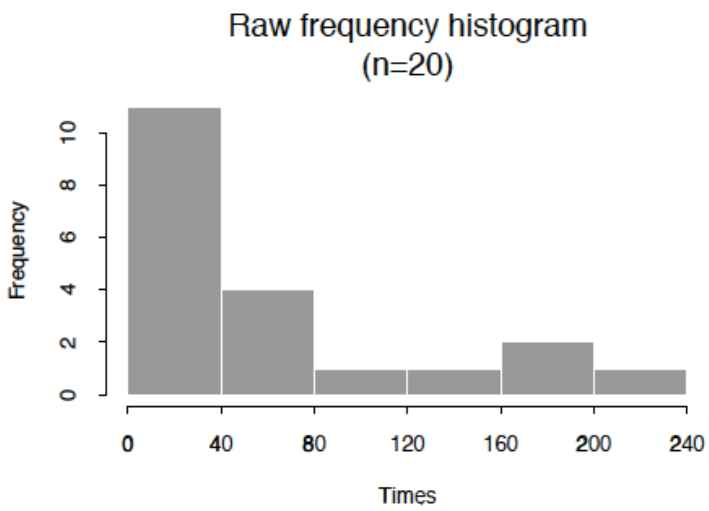
Data presentation: Histogram

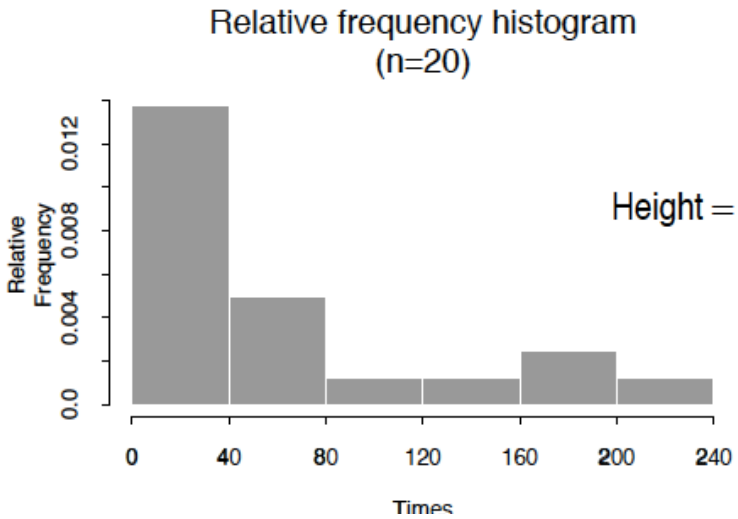
Once the frequency table has been constructed, pictorial representation can be considered.

For most **continuous** data sets, the best diagram to use is a **histogram**

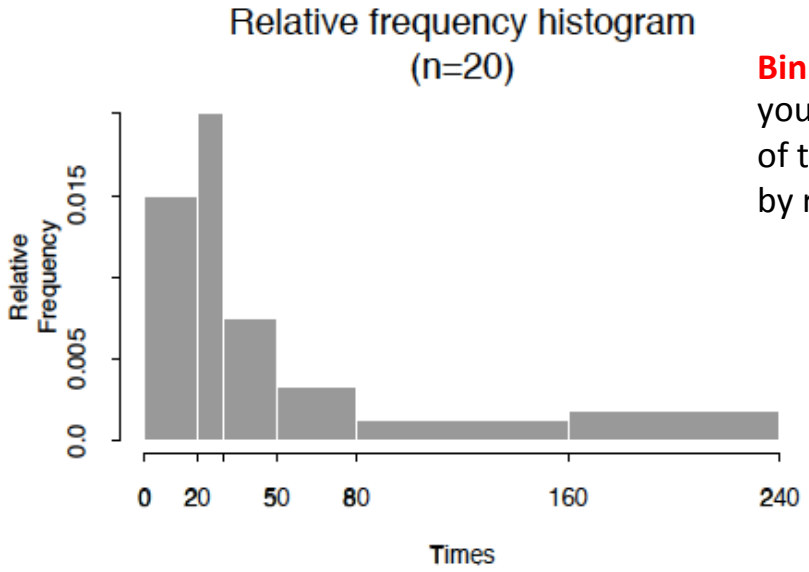
- Rectangles are drawn on this base with their areas proportional to the frequencies
  - classification intervals are represented to scale on the abscissa (x-axis)
  - heights of the rectangles are proportional to the frequencies iff. class intervals are of equal width

Data presentation: Histogram





$$\text{Height} = \frac{\text{Frequency}}{n \times \text{BinWidth}}.$$



Bin widths should be chosen so that you get a good idea of the distribution of the data, without being swamped by random variation.

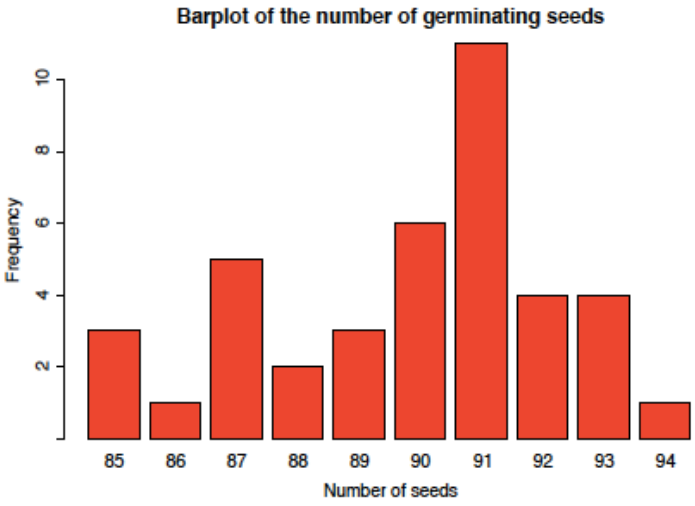
Data presentation: Bar Chart and Frequency polygons

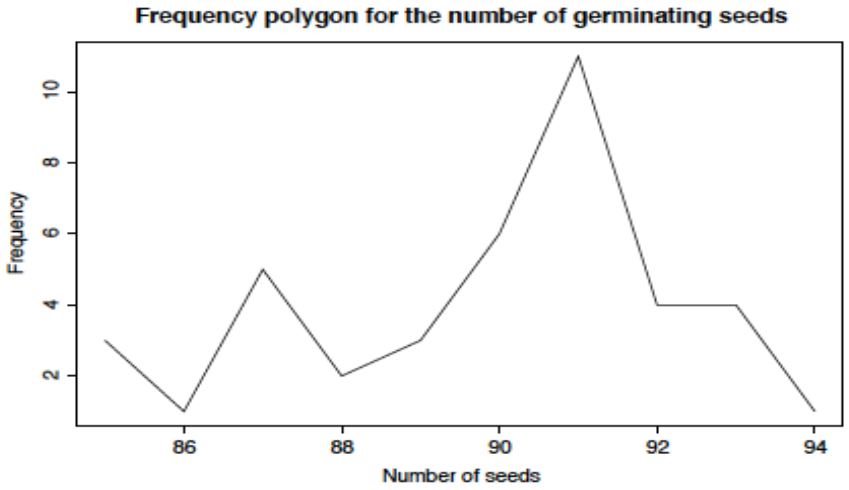
When the data are **discrete** and the frequencies refer to **individual** values, we can display them using:

- **Bar chart** with **heights** of bars representing frequencies
- **Frequency polygon** in which only the **tops** of the bars are marked, and then these points are joined by straight lines.

Data presentation: Histogram

Using the frequency table constructed earlier, we can construct a Bar Chart and Frequency Polygon as follows.

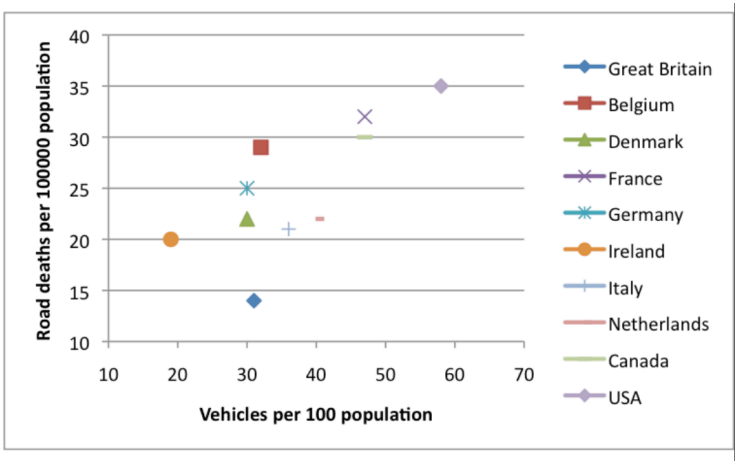




$$n = 40$$

Scatter Plot

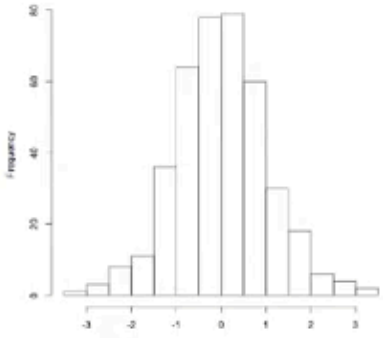
- Displays values for two variables for a set of data;
- The independent variable is plotted on the horizontal axis, the dependent variable on the vertical axis;
- It allows to determine correlation
  - Positive (bottom left -> top right)
  - Negative (top left -> bottom right)
  -



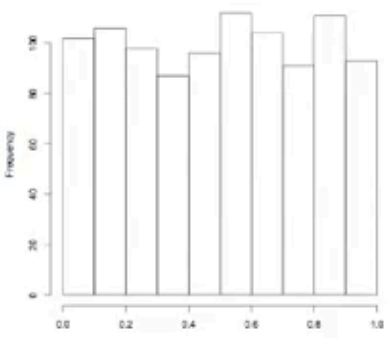
Data presentation: Box-and-whisker plots

- description of the main **features** of a the observations.
- There are many variations on the box plot. The simplest form is constructed a fellows:

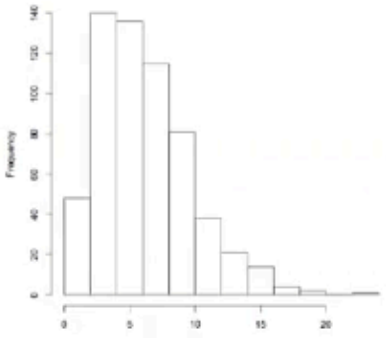
1. draw a rectangular box which stretches from the **lower quartile** to the **upper quartile**,
2. divide in two at the **median**.
3. From each end of the box, a line is drawn to the **maximum** and **minimum** observations (these lines are sometimes called whiskers, hence the name)



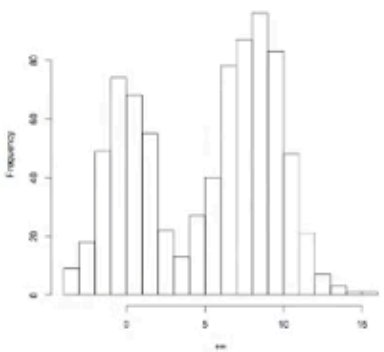
Symmetric
Unimodal



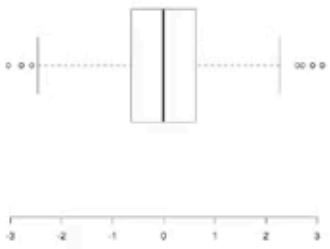
Symmetric
Uniform

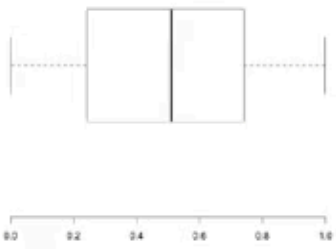


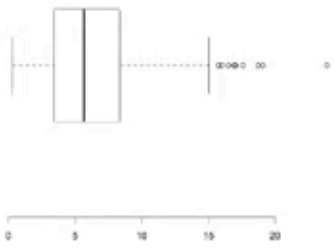
Right-skewed
Unimodal

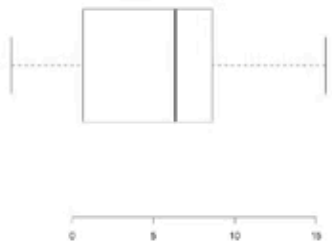


Bimodal









Data Collection: Observational Studies & Experiments

- Anecdotes
- Observations Studies
- Experiments

Alison Gibbs and Jeffrey Rosenthal, Statistics Making sense of data, university of Toronto, [coursera]

Observation vs. Experiments

- **Observation** of the characteristics of an existing sample
  - The goal is either to draw conclusion:
    - about the population
    - differences between two or more groups
    - The relationships between variables
  - The investigator has **no control** over which individuals are in which group any other of their characteristics.

- **Experiments**:
  - **Interventions** are imposed by the investigator in the subjects.

Alison Gibbs and Jeffrey Rroenthal, Statistics Making sense of data,university of Toronto,[coursera]

Observational studies

observational study is the mechanisms that can result in an **observed association** between **an explanatory variable** and the out come:

- Changes in the explanatory variable cause the outcome to change

- Reverse causation causes the explanatory variable to change

1. The association is coincidence
2. There is a common cause that causes both the explanatory and the outcome to change
3. A **confounding** variable is associated with the explanatory variable and causes the outcome to change

Alison Gibbs and Jeffrey Rroenthal, Statistics Making sense of data, university of Toronto,[coursera]

Confounding variables: lurking variable

- Lurking variables are variables that are not considered in the study but may affect the nature of the relationship between the **explanatory variable** and the **outcome**

- A lurking variable can be a confounding variable or the source of a common response or another variable that, when considered, changes the nature of the association

Alison Gibbs and Jeffrey Rroenthal, Statistics Making sense of data, university of Toronto,[coursera]

Experimental studies

- Response variable (or dependent variables) the outcome of interest measured each subject, on entity participating in the study
- Explanatory variables (predictor or Independent variable): A variable that we think that might explain the value of the response variable
- experiments: the experimenter manipulate the explanatory variable to see the effect on the response