Course 02402 Introduction to Statistics

Lecture 1: Introduction and R

DTU Compute
Technical University of Denmark
2800 Lyngby – Denmark

Agenda

1. Practical course information
2. Introduction to Statistics - a primer
3. Statistics and Engineers
4. Descriptive Statistics
   - Mean and median
   - Variance and standard deviation
   - Percentiles and quantiles
   - Covariance and correlation
5. Software: R & RStudio

Overview

1. Practical course information
2. Introduction to Statistics - a primer
3. Statistics and Engineers
4. Descriptive Statistics
   - Mean and median
   - Variance and standard deviation
   - Percentiles and quantiles
   - Covariance and correlation
5. Software: R & RStudio

Practical course information

- Teaching module
  - Lectures: Tuesday 13-15 at Zoom Channel introstat_F21_02402
  - Exercises: Tuesday 15-17 at
    - Zoom Channels. See Email or Course Website.
    - Quizzes and Area9 Rhapsode
- Exam
  - Sunday 30 May 2021 9am-1pm
  - 4 hour multiple choice
- Mandatory projects
  - 2 projects must be approved in order to participate in the exam.
  - For each project, choose between one of four topics.

Practical course information

- Generic weekly agenda
  - Before teaching: Read relevant chapters/sections in eNote/book
  - Lectures: 2 hours, curriculum of the week
  - Exercises: 2 hours, exercises and online quizzes
  - After teaching: Online "exam quiz" (test yourself)
- Teaching material
  - Available under *Material* on course website
  - Optional: Use the available R script to download all the material in one go (but beware that it overwrites changes made to previously downloaded files)
  - Lecture slides and R code will be updated shortly before each lecture (remember to refresh browser)

Practical Information

- Homepage: 02402.compute.dtu.dk
  - Online book (website or via lix)
  - Syllabus
  - Lecture plan / agenda
  - Exercises & solutions
  - Slides
  - Discussion forum
  - Podcasts of previous years' lectures (English and Danish)
  - Quizzes
- Learn: `learn.inside.dtu.dk/d2l/home/60261`
  - Announcements (Must subscribe to get email)
  - Projects - description and submission

Overview

Introduction to Statistics - a primer

New England Journal of Medicine:

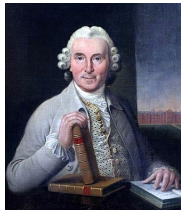
EDITORIAL: Looking Back on the Millennium in Medicine,
N Engl J Med, 342:42-49, January 6, 2000.
`http://www.nejm.org/doi/full/10.1056/NEJM200001063420108`

Millennium list

- Elucidation of human anatomy and physiology
- Discovery of cells and their substructures
- Elucidation of the chemistry of life
- **Application of statistics to medicine**
- Development of anesthesia
- Discovery of the relation of microbes to disease
- Elucidation of inheritance and genetics
- Knowledge of the immune system
- Development of body imaging
- Discovery of antimicrobial agents
- Development of molecular pharmacotherapy

James Lind

One of the earliest clinical trials took place in 1747, when James Lind treated 12 scorbutic ship passengers with cider, an elixir of vitriol, vinegar, sea water, oranges and lemons, or an electuary recommended by the ship's surgeon. The success of the citrus-containing treatment eventually led the British Admiralty to mandate the provision of lime juice to all sailors, thereby eliminating scurvy from the navy.
(See also `http://en.wikipedia.org/wiki/James_Lind`).



John Snow

The origin of modern epidemiology is often traced to 1854, when John Snow demonstrated the transmission of cholera from contaminated water by analyzing disease rates among citizens served by the Broad Street Pump in London's Golden Square. He arrested the further spread of the disease by removing the pump handle from the polluted well.
(See also `http://en.wikipedia.org/wiki/John_Snow_(physician)`).

Google - *Big Data*

A quote from the New York Times article titled *For Today's Graduate, Just One Word: Statistics* (5 August 2009)
`http://www.nytimes.com/2009/08/06/technology/06stats.html`

I keep saying that the sexy job in the next 10 years will be statisticians, said Hal Varian, chief economist at Google. And I'm not kidding.

IBM - *Big Data*

The key is to let computers do what they are good at, which is trawling these massive data sets for something that is mathematically odd, said Daniel Gruhl, an I.B.M. researcher whose recent work includes mining medical data to improve treatment. And that makes it easier for humans to do what they are good at - explain those anomalies.

Intro Case stories: IBM big data, Novo Nordisk small data, Skive fjord

- Presentation by Senior Scientist Hanne Refsgaard, Novo Nordisk A/S
- IBM Social Media podcast by Henrik H. Eliassen, IBM.
- Skive Fjord podcasts, by Jan K. Møller, DTU.

Overview

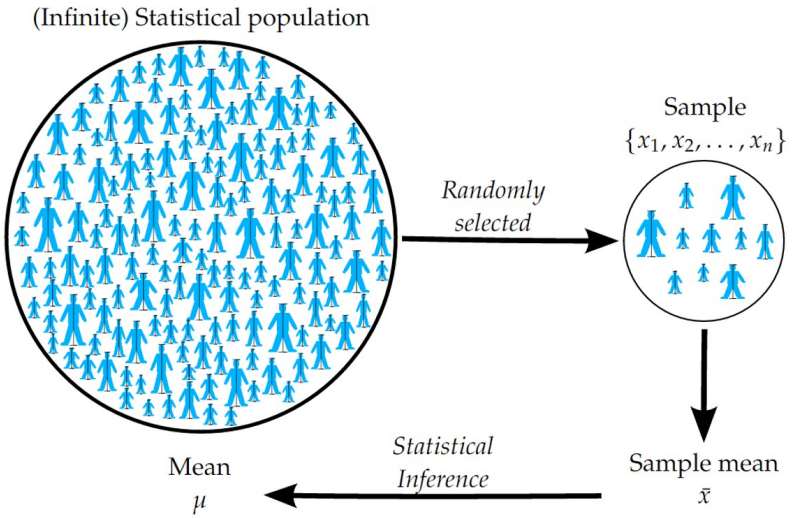
Statistics and Engineers

- Analysis of data ("both small & big")
- Understanding random variation
- Understanding the advantages (and limitations) of statistics for problem solving
- Quality improvement
- Design of experiments
- Prediction of future values
- ... and much more!

Statistics

- Descriptive statistics vs. *statistical inference*
- Statistics is often about analyzing a *sample*, taken from a *population*.
- Based on the sample, we try to generalize to the population.
- Therefore, it is important that the sample is *representative* of the population.

Statistics



Overview

Summary statistics

We use *summary statistics* to summarize and describe data (stochastic variables)

- Measures of centrality
  - e.g.: mean ($\bar{x}$) and median
- Measures of dispersion
  - e.g.: variance (s^2) and standard deviation (s)
- Measures of relation
  - e.g.: covariance and correlation

Note the difference between, e.g., the *(sample)* mean $\bar{x}$ and the *(population)* mean μ.

Mean, Definition 1.4

The mean value is a key number which indicates the centre of gravity or centering of the data.

The sample mean (average):

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

We say that $\bar{x}$ is an *estimate* of the population mean.

Median, Definition 1.5

The median is also a key number indicating the center of the data.

In some cases, for example in the case of extreme values, the median is preferable to the mean.

Sample median:
The observation in the middle (in sorted order).

Example: Student heights

- **Sample:** Student heights in cm, $n = 5$.

$$(x_1, x_2, x_3, x_4, x_5) = (185, 184, 194, 180, 182)$$

- **Mean:**

$$\bar{x} = \frac{1}{5}(185 + 184 + 194 + 180 + 182) = 185$$

- **Median:**
  - First order the data: $180, 182, 184, 185, 194$.
  - Then choose the third/middle number (n uneven): 184
- If a person with height 235 cm is added to the data:
  - *Mean:* 193
  - *Median:* 184.5

Variance and standard deviation, Definition 1.10

The variance and the standard deviation indicate the dispersion ("spread") of the data:

- Variance

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

- Standard deviation

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

Example: Student heights

- **Sample:** Student heights in cm, $n = 5$.

$$(x_1, x_2, x_3, x_4, x_5) = (185, 184, 194, 180, 182)$$

- **Variance:**

$$s^2 = \frac{1}{4}((185-185)^2 + (184-185)^2 + \cdots + (182-185)^2) = 29$$

- **Standard deviation:**

$$s = \sqrt{29} = 5.385$$

The coefficient of variation, Definition 1.12

The standard deviation and the variance are key numbers for absolute variation.

If it is of interest to compare variation between different data sets, it might be a good idea to use a *relative* key number.

Coefficient of variation:

$$V = \frac{s}{\bar{x}} \tag{1}$$

Percentiles and quantiles

The median it the value that divides the data into two halves.

More generally, we may compute *percentiles*, e.g.:

- $0, 25, 50, 75, 100$ % percentiles and/or
- $0, 10, 20, 30, 40, 50, 60, 70, 80, 90, 100$ % percentiles

Note:

- The median is the 50% percentile.
- The 25, 50, 75 % percentiles are often referred to as the first, second and third quartiles, and denoted $Q1$, $Q2$, and $Q3$, respectively.
- Inter Quartile Range (IQR): $Q3 - Q1$

Quantiles, Definition 1.7

The p'th *quantile*, also named the $100p$'th *p*ercentile, can be defined by the following procedure:

1. Order the n observations from smallest to largest: $x_{(1)}, \ldots, x_{(n)}$.
2. Compute pn.
3. If pn is an integer: Average the pn'th and $(pn+1)$'th ordered observations:

$$\text{The } p\text{'th quantile } = \left(x_{(np)} + x_{(np+1)}\right)/2$$

4. If pn is a non-integer, take the next ordered observation:

$$\text{The } p\text{'th quantile } = x_{(\lceil np \rceil)}$$

where $\lceil np \rceil$ is the *ceiling* of np, that is, the smallest integer larger than np.

Example: Student heights

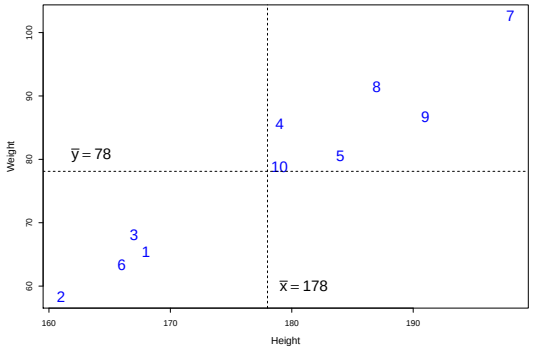
- **Sample:** *Ordered* student heights in cm.

$$(x_{(1)}, x_{(2)}, x_{(3)}, x_{(4)}, x_{(5)}) = (180, 182, 184, 185, 194)$$

- **Lower quartile, Q1:**
  - Establish that $np = 1.25$, as $p = 0.25$ and $n = 5$.
  - The smallest integer larger than np is 2.
  - $Q1 = x_{(2)} = 182$.
- **Upper quartile, Q3:**
  - Establish that $np = 3.75$, as $p = 0.75$ and $n = 5$.
  - The smallest integer larger than np is 4.
  - $Q3 = x_{(4)} = 185$.
- **IQR:**
  - $Q3 - Q1 = 3$

Covariance and correlation

Height (x_i)	168	161	167	179	184	166	198	187	191	179
Weight (y_i)	65.5	58.3	68.1	85.7	80.5	63.4	102.6	91.4	86.7	78.9



Covariance and correlation, Definitions 1.18 and 1.19

The sample covariance is given by

$$s_{xy} = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})$$

The sample correlation coefficient is given by

$$r = \frac{1}{n-1}\sum_{i=1}^{n}\left(\frac{x_i - \bar{x}}{s_x}\right)\left(\frac{y_i - \bar{y}}{s_y}\right) = \frac{s_{xy}}{s_x \cdot s_y}$$

where s_x and s_y are the sample standard deviations for x and y respectively.

Covariance and correlation

Student	1	2	3	4	5	6	7	8	9	10
Heights (x_i)	168	161	167	179	184	166	198	187	191	179
Weights (y_i)	65.5	58.3	68.1	85.7	80.5	63.4	102.6	91.4	86.7	78.9
$(x_i - \bar{x})$	-10	-17	-11	1	6	-12	20	9	13	1
$(y_i - \bar{y})$	-12.6	-19.8	-10	7.6	2.4	-14.7	24.5	13.3	8.6	0.8
$(x_i - \bar{x})(y_i - \bar{y})$	126.1	336.8	110.1	7.6	14.3	176.5	489.8	119.6	111.7	0.8

$$\begin{aligned} s_{xy} &= \frac{1}{9}(126.1 + 336.8 + 110.1 + 7.6 + 14.3 + 176.5 + 489.8 \\ &\quad + 119.6 + 111.7 + 0.8) \\ &= \frac{1}{9} \cdot 1493.3 \\ &= 165.9 \end{aligned}$$

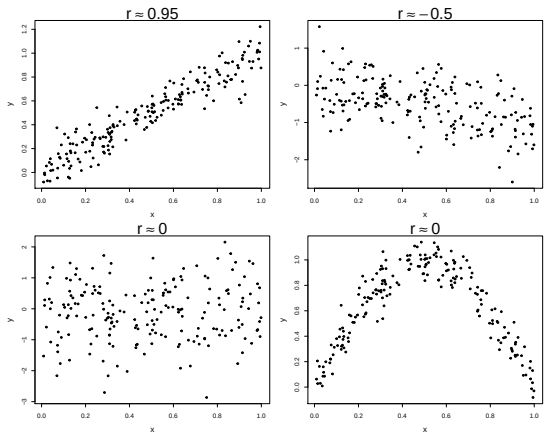
$$s_x = 12.21, \quad \text{and} \quad s_y = 14.07$$

$$r = \frac{165.9}{12.21 \cdot 14.07} = 0.97$$

Correlation - properties

- r is always between -1 and 1: $-1 \leq r \leq 1$.
- r measures the degree of linear relation between x and y.
- $r = \pm 1$ if and only if all points in the scatterplot are exactly on a line.
- $r > 0$ if and only if the general trend in the scatterplot is positive.
- $r < 0$ if and only if the general trend in the scatterplot is negative.

Correlation



Figures/Tables

- Quantitative data
  - Scatter plot (xy plot)
  - Histogram
  - Cumulative distribution
  - Box plot
- Count data
  - Bar chart
  - Pie chart

Overview

Software: R & RStudio

- R: Software/language for statistical analysis and visualization of data.
- R & RStudio: Free to download, can be installed on Linux, Mac, Windows.
- R, RStudio, extra packages under continuous development.
- Introduction in the book.
- Intregrated in the course material and teaching.
- Learning by doing. Also: use Google!

Software: R

```
> # Adding numbers in the console
> 2 + 3

## [1] 5
```

```
> # Assigning a number to a variable
> x <- 3
> x

## [1] 3
```

```
> # Assigning a vector to a variable
> x <- c(1, 4, 6, 2); x

## [1] 1 4 6 2
```

```
> # A vector of integers from 1 to 10
> ( x <- 1:10 )

##  [1]  1  2  3  4  5  6  7  8  9 10
```


Software: R

```
# Height data from before
x <- c(168, 161, 167, 179, 184, 166, 198, 187, 191, 179)
```

```
# Sample mean
mean(x)

## [1] 178
```

```
# Sample median
median(x)

## [1] 179
```

```
# Sample variance
var(x)

## [1] 149
```


Software: R

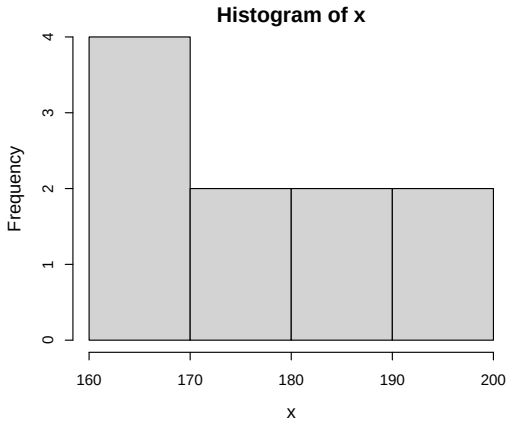
```
# Sample standard deviation
sd(x)

## [1] 12
```

```
# Sample quartiles
quantile(x, type = 2)

##   0%  25%  50%  75% 100%
##  161  167  179  187  198
```

```
# Sample quantiles 0%, 10%,..,90%, 100%
quantile(x, probs = seq(0, 1, by = 0.10), type = 2)

##   0%  10%  20%  30%  40%  50%  60%  70%  80%  90% 100%
##  161  164  166  168  174  179  184  187  189  194  198
```

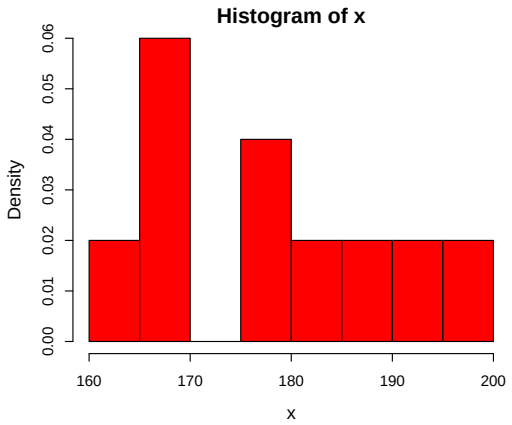
Software: R

```
# A histogram of the heights
hist(x)
```



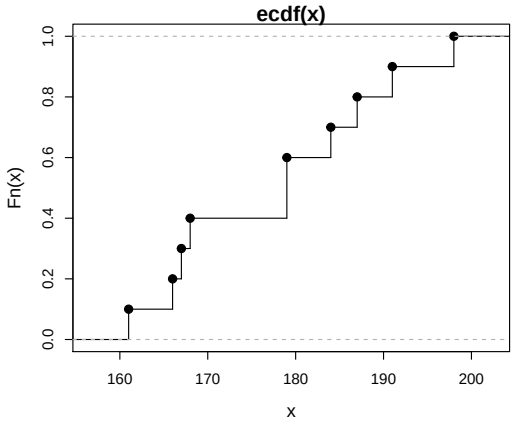
Software: R

```
# A density histogram of the heights
hist(x, prob = TRUE, col = "red", nclass = 8)
```



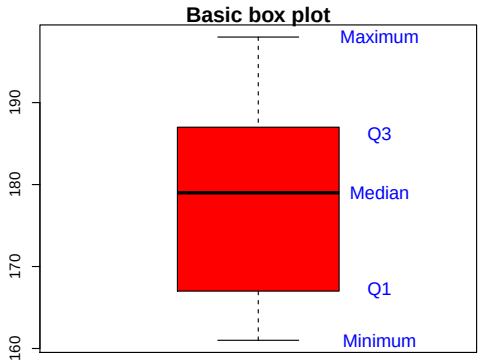
Software: R

```
# Empirical cumulative distribution function of the heights
plot(ecdf(x), verticals = TRUE)
```



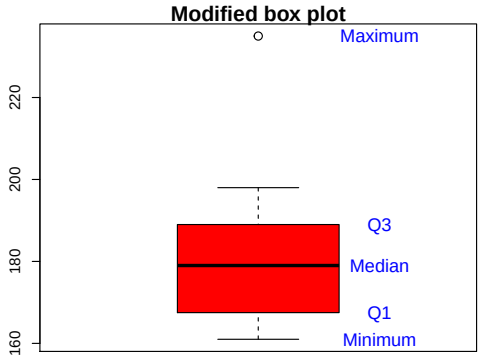
Software: R

```
# Basic box plot of the heights ('range = 0' makes it "basic")
boxplot(x, range = 0, col = "red", main = "Basic box plot")
text(1.3, quantile(x), c("Minimum", "Q1", "Median", "Q3", "Maximum"), col = "blue")
```



Software: R

```
# Modified box plot of heights with an additional extreme observation (235 cm).
# The modified version is the default.
boxplot(c(x, 235), col = "red", main = "Modified box plot")
text(1.3, quantile(c(x, 235)), c("Minimum", "Q1", "Median", "Q3", "Maximum"), col = "blue")
```



Next week:

- Probability, part 1 - eNote/book chapter 2.

Agenda

1. Practical course information
2. Introduction to Statistics - a primer
3. Statistics and Engineers
4. Descriptive Statistics
   - Mean and median
   - Variance and standard deviation
   - Percentiles and quantiles
   - Covariance and correlation
5. Software: R & RStudio