Course 02402 Introduction to Statistics Lecture 1:

Introduction to Statistics

Per Bruun Brockhoff

DTU Compute
Danish Technical University
2800 Lyngby – Denmark
e-mail: perbb@dtu.dk

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, quantiles
   - Covariance and correlation
5. Software: R

Practical course information

- Teaching module: Tuesdays 13-17
- Generic weekly agenda:
  - BEFORE teaching module: Read announced stuff
  - 2 hours long lectures (curriculum of the week)
  - 2 hours of exercises (Mix of: Enote and online quiz-questions)
  - AFTER teaching module: Test yourself by online exam quiz.
- Exam: 4 hour multiple choice, Sunday 28/05
- MANDATORY projects: 2 must be approved to be able to go to the exam.
  - Each project will have 4 optional versions!

Practical Information

- Homepage: 02402.compute.dtu.dk
  - Online book (website or via blue lix)
  - Syllabus, Lecture plan
  - Exercises & solutions
  - Slides
  - Podcasts of lectures (In English AND Danish)
  - Quizzes
- Campusnet: www.campusnet.dtu.dk
  - Messages and (certain) file sharings
  - Links to interesting stories
  - Projects - description AND submission
- The blue lix eBook portal
  - eBook portal with marking and annotation features.

Introduction to Statistics - a primer

New England Journal of medicine:

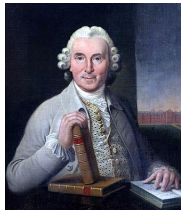
EDITORIAL: Looking Back on the Millennium in Medicine, *N Engl J Med*, 342:42-49, January 6, 2000.
`http://www.nejm.org/doi/full/10.1056/NEJM200001063420108`

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

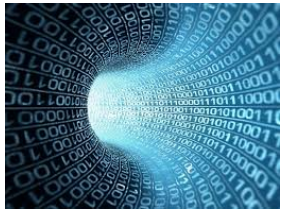
The origin of modern epidemiology is often traced to 1854, when John Snow demonstrated the transmission of cholera from contaminated water by analyzing disease rates among citizens served by the Broad Street Pump in London's Golden Square. He arrested the further spread of the disease by removing the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

Google - *Big Data*

A quote from New York Times, 5. August 2009, from the article titled For Today's Graduate, Just One Word: Statistics is:

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.

(And Politiken, 12/2 2014 - see links in CampusNet)



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.

Statistics and Engineers

- How to treat (or analyse) data?
- What is random variation?
- Statistics is an important tool in problem solving
- Data analysis
- Quality improvement
- Design of experiments
- Predictions of future values
- .. and much more!

Statistics at DTU (mostly Compute)

- Energy Systems:
  - Prognoses of sun and wind power
  - Optimization of storage of energy e.g. in buildings
  - Modelling of human behaviour waste water treatment plant
- Control:
  - Robot navigation
  - Mechanical systems (e.g. cars, ships, wind turbines, etc)
- Medicine, Food and Pharma (Compute):
  - Statistics of clinical trials
  - Artificial pancreas
  - Human perceptual data in industrial product development
  - Pharmakokinetic and dynamic modelling
- Image Analysis
  - Image data is used more and more
  - X-rays, scannings, satelite photos, etc – videos

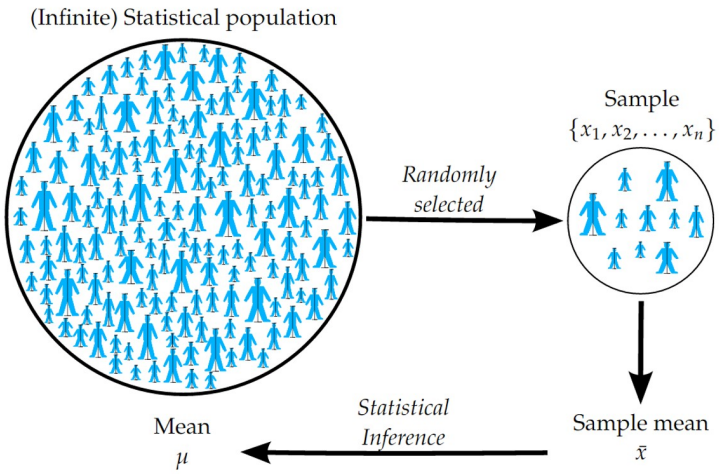
Statistics at DTU (mostly Compute)

- Signal processing:
  - Electrical systemes (filters, amplifiers, ...)
- Computer science:
  - Internet data (trafic, Google, Facebook, etc.)
  - Text recognition and mining
  - Security: Server attacks etc.
  - Software testing
- Civil Engineering
  - Tests of material properties and constructions
  - Production methods, e.g. casting of concrete
  - Energy systemes and indoor climate testing
- Management:
  - Financal Engineering, questionnaire surveys, ...
- Chemistry, Physics, Environment, Food, Vet, Aqua, etc

Statistics

- Statistics is often about analyzing a *sample*, that is taken from a *population*
- Based on the sample, we try to generalize (or comment on) the population
- Therefore it is important that the sample is *representative* of the population

Statistics



Summary statistics

We use a number of summary statistics to summarize and describe data (stochastic variables)

- Mean $\bar{x}$ and Median
- Variance s^2 and Standard deviation s
- Percentiles/quantiles
- Covariance and Correlation

Mean, Definition 1.4

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

- The mean:

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

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

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

- Median:

  The observation in the middle (in sorted order)

Example: Student heights:

- Sample:

```
x <- c(185, 184, 194, 180, 182)
```

  n=5

- **mean**:

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

- **Median**, first order data: 180 182 184 185 194

  And then chosse the middle (since n is uneven)(3'th) number: 184

- What if a person on 235cm is added to the data:

  Median = 184

  Mean = 193

Variance and standard deviation, Definition 1.10

The variance (or the standard deviation) indicates the 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:

- **Data** n=5:
  185 184 194 180 182
- **Varians**, $s^2 =$

$$\frac{1}{4}((185-185)^2 + (184-185)^2 + (194-185)^2 + (180-185)^2 + (182-185)^2) = 29$$

- **Standard deviation**, $s = \sqrt{s^2} =$

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

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, the coefficient of variation:

$$V = \frac{s}{\bar{x}} \cdot 100$$

Percentiles, quantiles

The median it the point that divides the data into two halves. It is of course possible to find other points that divide the data in other parts, they are called percentiles.

Often calculated percentiles are

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

Note: the 50% percentile is the median

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:

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

4. If pn is a non-integer, take the "next one" in the ordered list:

$$\text{The } p\text{'th quantile } = x_{(\lceil np \rceil)} \tag{2}$$

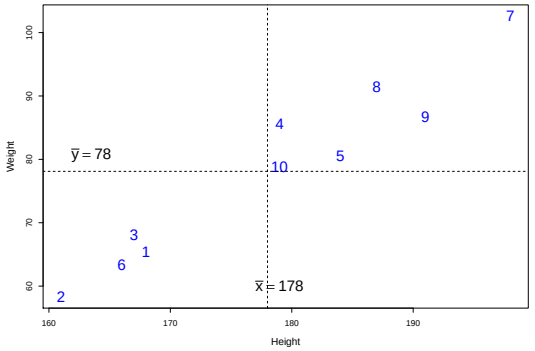
where $\lceil np \rceil$ is the *ceiling* of np, that is, the smallest integer larger than np.

Example: Student heights:

- **Data,** n=5:
  185 184 194 180 182
- **Lower quartile**, Q_1, first order the data: 180 182 184 185 194
  Then choose the right based on $np = 1.25$:
  $Q_1 = 182$
- **Upper quartile**, Q_3, first order the data: 180 182 184 185 194
  Then choose the right based on $np = 3.75$:
  $Q_3 = 185$

Covariance and correlation - measuring relation

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



Covariance and correlation, Definitions 1.18 and 1.19

The sample covariance is given by

$$s_{xy} = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y}) \tag{3}$$

The sample correlation coefficient is given by

$$r = \frac{1}{n-1}\sum_{i=1}^{n}\left(\frac{x_i - \bar{x}}{s_x}\right)\left(\frac{y_i - \bar{y}}{s_y}\right) = \frac{s_{xy}}{s_x \cdot s_y} \tag{4}$$

where s_x and s_y is the sample standard deviation for x and y respectively.

Covariance and correlation - measuring relation

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

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

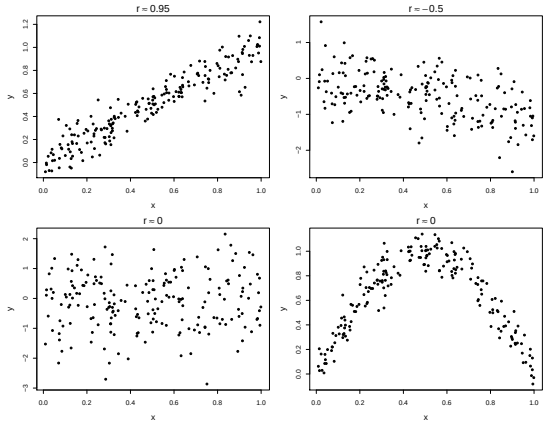
$$s_x = 12.21, \quad \text{and} \quad s_y = 14.07$$

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

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



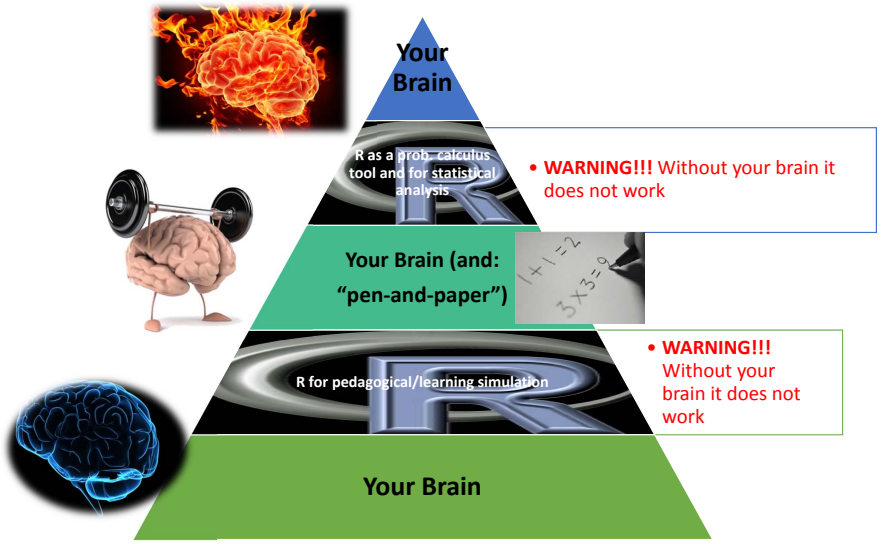
Figures/Tables

- Quantitative data:
  - Scatter plot (xy plot)
  - Histogram
  - Cumulative distribution
  - Boxplots
- Count data:
  - Bar charts
  - Pie charts

Software: R

- Install R and Rstudio
- Intro to basic computing
- Introduced in the eNote
- We use in an integrated way throughout the course and material
- Globalt rapidly growing open source computing environment
- WAARRRNIING: R CANNOT substitute our brains!!!! (Note Section 1.5)

Use your Brain!!!



Software: R

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

[1] 5

```
> y <- 3
```

```
> x <- c(1, 4, 6, 2)
> x
```

[1] 1 4 6 2

```
> x <- 1:10
> x
```

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

Software: R

```
## Sample Mean and Median (data from eNote)
x <- c(168,161,167,179,184,166,198,187,191,179)
mean(x)
```

[1] 178

```
median(x)
```

[1] 179

```
## Sample variance and standard deviation
var(x)
```

[1] 149

```
sd(x)
```

[1] 12

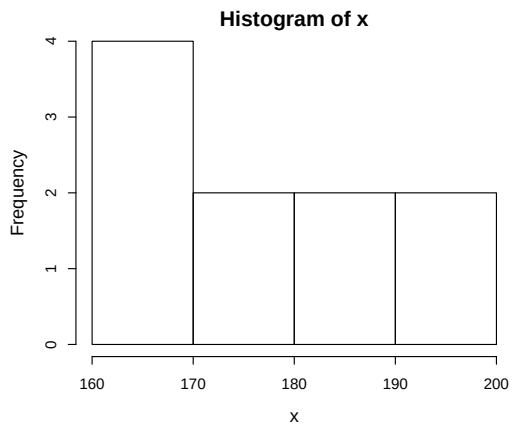
Software: R

```
## 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
```

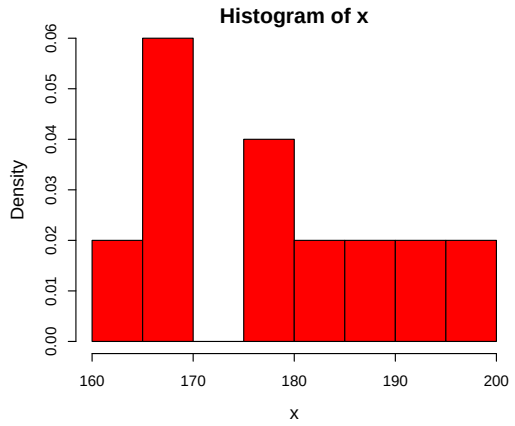

Software: R

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



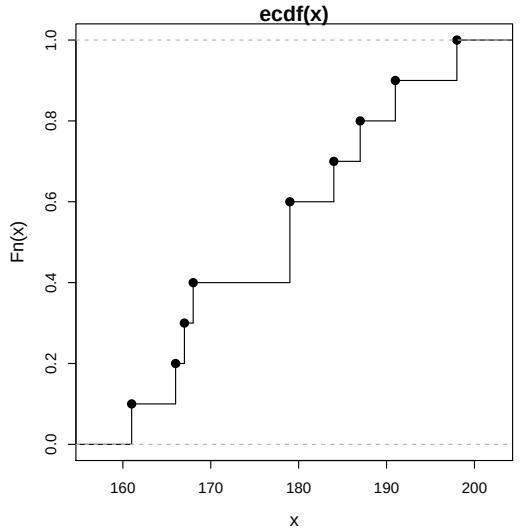
Software: R

```
## A density histogram of the heights:
hist(x,freq=FALSE,col="red",nclass=8)
```



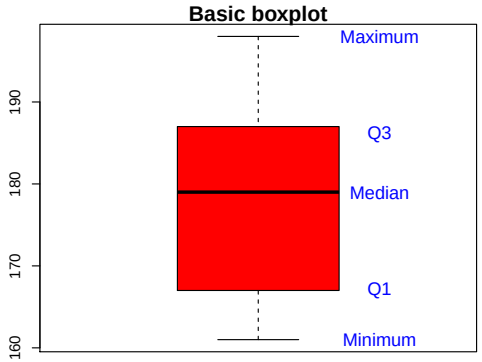
Software: R

```
plot(ecdf(x),verticals=TRUE)
```



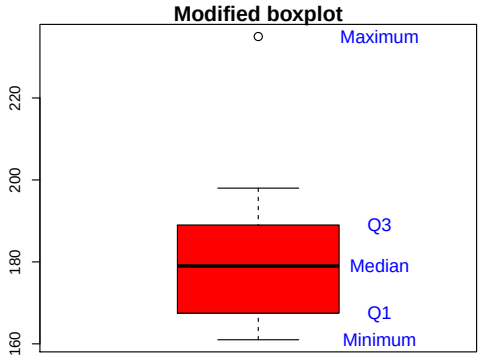
Software: R

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

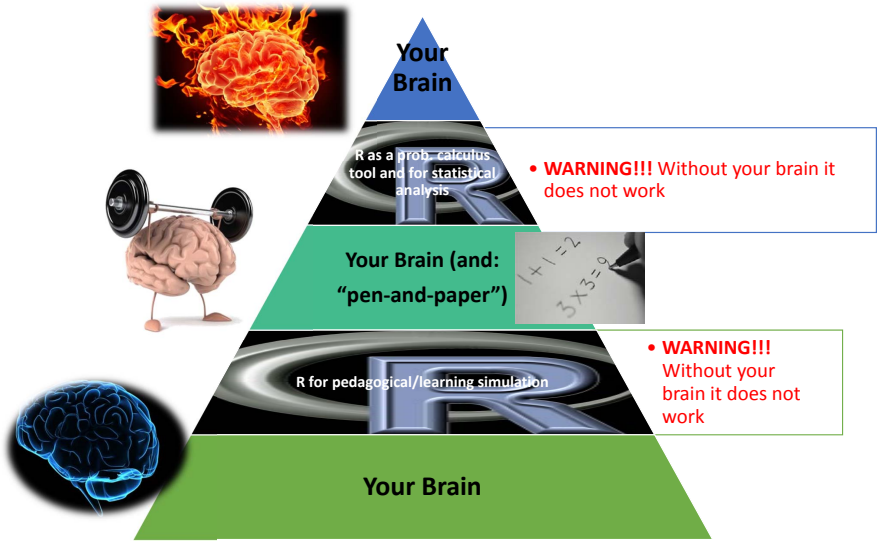


Software: R

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



Use your Brain!!!



Next week:

- Probability, part 1 - eNote 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, quantiles
   - Covariance and correlation
5. Software: R