Inferens 1, F1

Rolf Larsson

Uppsala Universitet

1 november 2022

Course overview

- Alm och Britton (AB):
  Stokastik - Sannolikhetsteori och statistikteori med tillämpningar. Liber 2008.
- Included in this course:
  - Dataanalys, kap. 6
  - Statistisk inferens, kap. 7
  - Icke-parametriska metoder, kap. 8
- Wackerly, Mendenhall, Shaeffer (WMS):
  Mathematical Statistics with applications, 7th ed. Duxbury 2008.
- Included in this course:
  - What is statistics?, chap. 1
  - Sampling distributions and the central limit theorem, chap. 7
  - Estimation, chap. 8
  - Properties of point estimators and methods of estimation, chap. 9
  - Hypothesis testing, chap. 10
  - Analysis of categorical data, chap. 14
  - Nonparametric statistics, chap. 15

Course overview

- Written exam.
- Two hand-in assignments giving bonus points on the ordinary exam.
- 12 lectures
- 4 problem solving sessions, including a guest lecture
- Two computer labs.
- Quizzes for practice, one for each lecture.
- Project (*obligatory!*)
- Om Studium: Schedule, slides, hand-ins, quizzes...

Today

- Introduction
- chap. 6 (AB): Data analysis (chap. 1 in WMS)
    - 6.1: Introduction
    - 6.2: Location and Dispersion measures
    - 6.3: Graphical illustration
    - 6.4: Data materials in several dimensions
- Gapminder

Some examples
Opinion polls

https://www.riksdagen.se/sv/ledamoter-partier/ledamot/annie-loofe234524d-e04b-448a-b609-05926a0a8b36

- In the opinion poll by Statistics Sweden in May 2022, 4274 voters were asked about (and replied on) their political sympathies.
- C got 6.6%.
- In the previous poll in November 2021, (4319 voters asked and replied), C got 8.5%.
- Is this a statistically significant change?

Some examples

Drug testing

http://livetskemi.se/?p=92

- A drug is supposed to lower the blood pressure. One group of patients gets the drug, and another group gets placebo.

Group	Pressure decrease									
Drug	8	4	6	-3	10	5	-1	2	9	7
Placebo	2	3	-2	0	1	1	-1	3	0	

- Does the drug have any effect?

Some examples

Quality control

http://www.100innovationer.com/svensk/innovationerna/innovationer/glodlampan.218.html

- A certain type of light bulb is supposed to work one year (365 days) "on average".
- A batch of 60 light bulbs is tested. The mean life length is 300 days.
- Does this disprove that the lamps work for one year on average?

Some examples

Salary statistics

http://hok.se/ny-sedlar-och-mynt-i-sverige-2015-158/

Monthly salaries in kkr for randomly selected mathematical statisticians in the public and private sector are given in the table below. (Fictive data.)

public	40	42	33	55	34
private	43	44	70	56	

Are the salaries for mathematical statisticians in the public and the private sector on average the same or different?

Some examples

Opinion polls, revisited

http://www.biblioteksforeningen.org/2013/10/24/riksdagen-beslutade-ny-bibliotekslag/

The opinion polls of Statistics Sweden in November-21 and May-22, respectively, gave the following results in per cent (4319 replies in November, 4274 in May) .

	SD	M	KD	L	C	MP	S	V	others
Nov.	17.0	22.4	4.4	3.0	8.5	4.0	30.1	9.6	1.0
May	16.5	21.4	5.1	3.6	6.6	3.4	33.0	8.4	1.9

Did the opinion change? (Or are the changes in these numbers just random?)

Some examples

Labor market for statisticians

- Official statistics
- Drug companies
- Insurance
- Industry
- Medicine
- Finance
- ...

Data analysis

- Location measures
- Dispersion measures
- Graphical illustration
- Data materials in several dimensions

Data analysis

Location measures

Data $x_1, x_2, ..., x_n$

Definition (6.1)

The sample mean is given by $\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i = \frac{1}{n}(x_1 + x_2 + ... + x_n)$

Definition (6.2)

The median is the "middle value" of the sorted data.
If n is even, the median is defined as the mean of the two middle values.

Definition (6.3)

The mode ("typvärdet") is the most common data value.

Data analysis

Location measures

Age for the Swedish parliament members from the Uppsala county (after election).

32 34 41 44 45 50 50 54 55 57 58 60 63

Mean 49.5

Median 50

Mode 50

Data analysis

Location measures

Age for all Swedish parliament members.

23 24 26 26 26 27 27 27 27 27 28 28 28 28 28 29 29 29 29 29 29 30 30 30 30 30 30 31 31 31
31 31 31 31 31 32 32 32 32 32 32 32 32 32 32 33 33 33 33 33 33 33 33 33 33 33 34 34 34 34
34 34 34 34 34 35 35 35 35 35 35 35 36 36 36 36 36 36 36 37 37 37 37 37 37 37 37 37 38 38
38 38 38 38 39 39 39 39 39 39 39 39 39 39 39 39 39 39 39 40 40 40 40 40 40 41 41 41 41 41
41 41 41 41 42 42 42 42 42 42 42 42 42 42 42 42 42 42 42 42 43 43 43 43 43 43 43 43 43 44
44 44 45 45 45 45 45 45 45 45 45 46 46 46 46 46 46 46 46 46 46 46 46 46 46 46 46 47 47 47
47 47 47 47 47 47 47 48 48 48 48 48 48 48 48 48 49 49 49 49 49 49 49 49 50 50 50 50 50 50
50 50 50 50 50 50 50 50 50 50 50 50 50 50 51 51 51 51 51 51 51 51 52 52 52 52 52 52 52 52
52 52 52 53 53 53 53 53 53 53 53 53 53 53 54 54 54 54 54 54 54 54 54 54 55 55 55 55 55 55
55 55 55 55 56 56 56 56 56 56 56 56 57 57 57 57 57 57 57 57 57 58 58 58 58 58 58 58 58 59
59 59 59 59 59 59 59 59 59 60 60 60 60 60 61 61 61 61 61 61 61 61 62 62 63 63 63 63 63 64
64 64 64 64 65 65 65 65 65 66 67 67 68 71 72 72 75 76 78

Mean 46.2, Median 46, Mode 50 (20 people)

Data analysis

Dispersion measures

Definition (6.4)

The sample variance is given by

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

Definition (6.4)

The sample standard deviation is given by

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

Data analysis

Dispersion measures

Data 0 0 1 2 2

Is the sample variance 0.5, 1 or 2?

Data 5 5 6 7 7

What is the sample variance?

Data analysis

Dispersion measures

Definition (6.5)

The range ("variationsbredden") is the difference between the largest and the smallest values of the data.

Definition (6.6)

The inter quartile range ("kvartilavståndet") is the difference between the upper and lower quartiles.

Definition

The lower quartile is the median of the lower half of the data material *including the median if n is odd*.
The upper quartile is the median of the upper half of the data material *including the median if n is odd*.

Data analysis

Dispersion measures

Data 0 0 1 2 2

What is the inter quartile range?

Data 0 0 1 1 2 2

What is the inter quartile range?

Data 0 0 1 1 1 1 1 2 2

What is the inter quartile range?

Data analysis

Dispersion measures

Age for the Swedish parliament members from the Uppsala county.

32 34 41 44 45 50 50 54 55 57 58 60 63

Standard deviation 9.8

Range $63 - 32 = 31$

Inter quartile range $57 - 44 = 13$

Data analysis

Dispersion measures

Data 1:

32 34 41 44 45 50 50 54 55 57 58 60 63

Mean 49.5, Median 50

Standard deviation 9.8

Range $63 - 32 = 31$

Inter quartile range $57 - 44 = 13$

Data 2:

32 34 41 44 45 50 50 54 55 57 58 60 83

Mean 51.0, Median 50

Standard deviation 13.1

Range $83 - 32 = 51$

Inter quartile range $57 - 44 = 13$

Age for the Swedish parliament members from the Uppsala county.

32 34 41 44 45 50 50 54 55 57 58 60 63

Stem and leaf plot (Stam-bladdiagram)

```
> u=c(32, 34, 41, 44, 45, 50, 50, 54, 55, 57, 58, 60, 63)
> stem(u)

  The decimal point is 1 digit(s) to the right of the |

  3 | 24
  4 | 145
  5 | 004578
  6 | 03
```

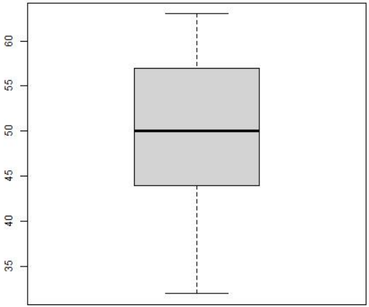
Stem and leaf plot for all Swedish parliament members.

```
> stem(x)

  The decimal point is 1 digit(s) to the right of the |

  2 | 34
  2 | 6667777788888999999
  3 | 00000011111111222222222233333333333444444444
  3 | 55555556666666777777777888888999999999999999
  4 | 0000001111111112222222222222222333333333444
  4 | 5555555556666666666666666777777777788888888899999999
  5 | 000000000000000000001111111222222222222333333333334444444444
  5 | 555555555566666666777777777888888888999999999
  6 | 0000011111111223333344444
  6 | 555556778
  7 | 122
  7 | 568
```

Data analysis

Graphical illustration

Box plot (Lådagram), Uppsala county

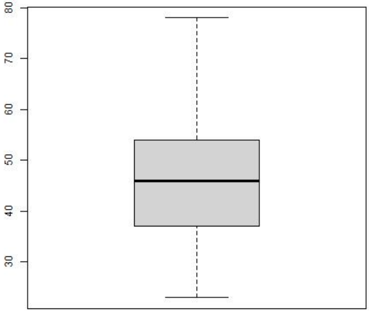


Max 63, upper quartile 57, median 50, lower quartile 44, min 32

Data analysis

Graphical illustration

Box plot, Sweden

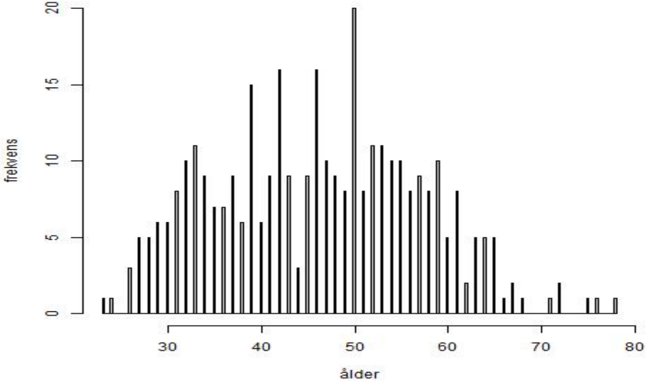


Max 78, upper quartile 54, median 46, lower quartile 37, min 23

Data analysis

Graphical illustration

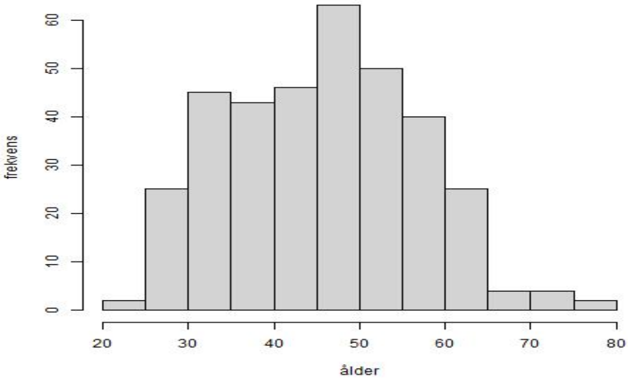
Bar chart (stapeldiagram)



Data analysis

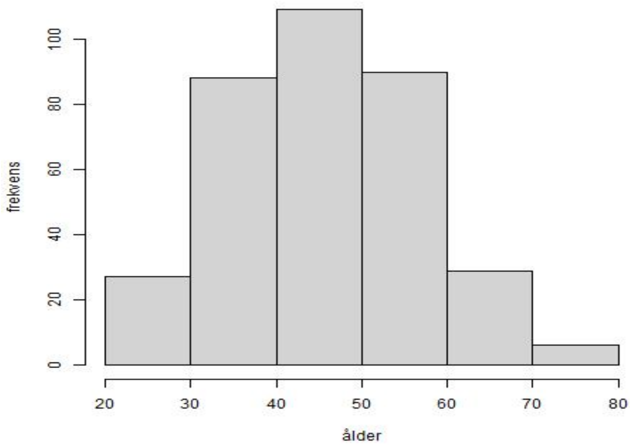
Graphical illustration

Histogram



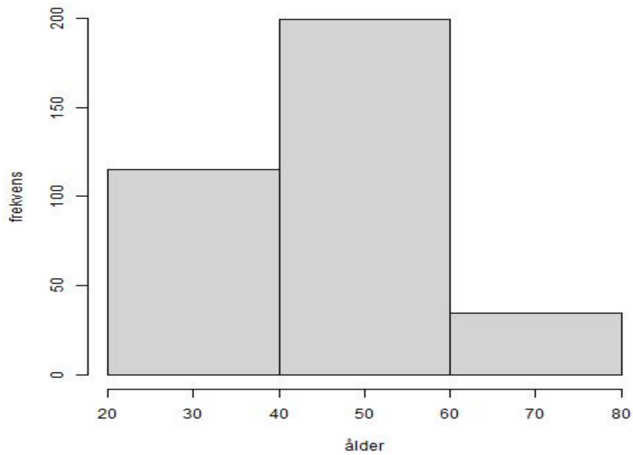
Data analysis

Graphical illustration



Data analysis

Graphical illustration



Data analysis

Graphical illustration

Is the data normally distributed?

Construction of *QQ-plot*: (example: Uppsala)

Start with the ordered sample $x_{(1)}, x_{(2)}, ..., x_{(13)}$.

i	1	2	3	4	5	6	7	8	9	10	11	12	13
$x_{(i)}$	32	34	41	44	45	50	50	54	55	57	58	60	63
$\Phi(z)$	.038	.115	.192	.269	.346	.423	.500	.577	.654	.731	.808	.885	.962
z	-1.77	-1.20	-0.87	-0.62	-0.40	-0.19	0.00	0.19	0.40	0.62	0.87	1.20	1.77

$$\Phi(z) = \frac{i - 0.5}{13}, \quad i = 1, 2, ..., 13.$$

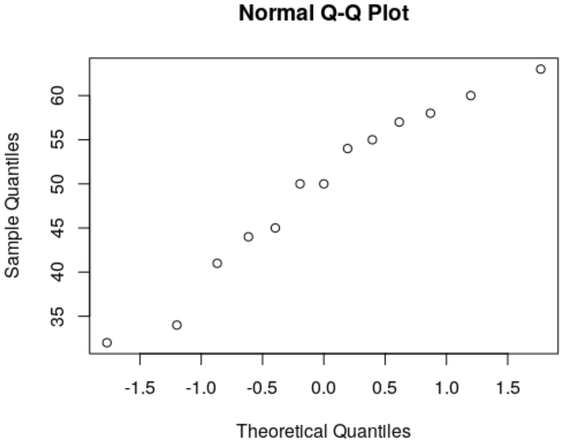
If data was perfectly normal, $x_{(i)}$ would be a linear function of z.

Plot z on the x axis and $x_{(i)}$ on the y axis.

Data analysis

Graphical illustration

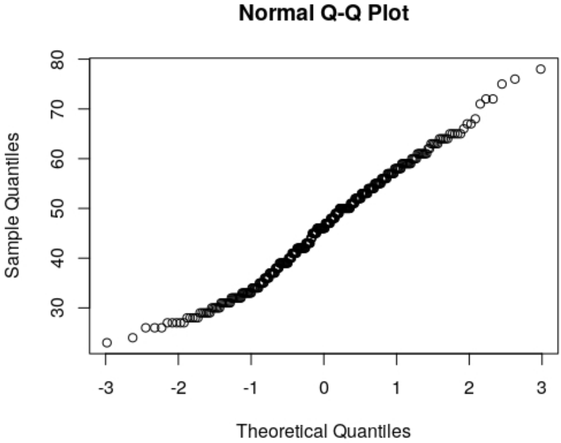
QQ-plot for Uppsala:



Data analysis

Graphical illustration

QQ-plot for Sweden:



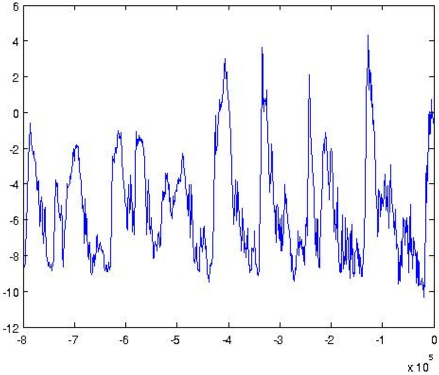
In R:

```
> x=read.table("riksdag.dat")$V1
> stem(x)
> boxplot(x)
> hist(x,main='',xlab='ålder',ylab='frekvens',breaks=349)
> hist(x,main='',xlab='ålder',ylab='frekvens',breaks=10)
> hist(x,main='',xlab='ålder',ylab='frekvens',breaks=5)
> hist(x,main='',xlab='ålder',ylab='frekvens',breaks=3)
> qqnorm(x)
```

Data analysis

Data materials in several dimensions

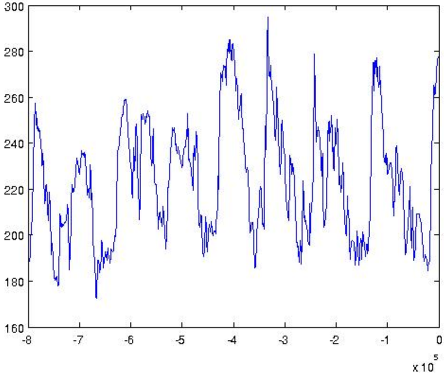
Proxies of temperatures from ice core data from Antarctica.
(Time: 800 000 years back up to now.)



Data analysis

Data materials in several dimensions

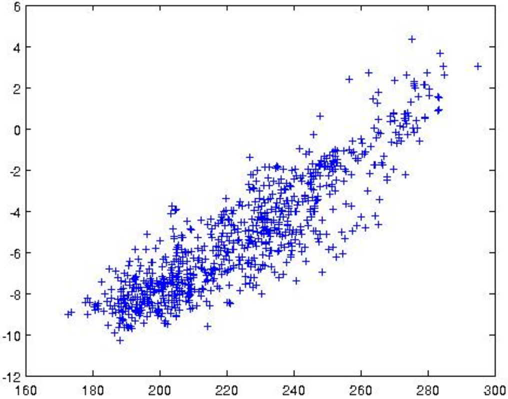
Proxies of Carbon Dioxide concentrations from ice core data from Antarctica.
(Time: 800 000 years back up to now.)



Data analysis

Data materials in several dimensions

Ice core data: Carbon dioxide concentration (x) and temperature (y)



Data analysis

Data materials in several dimensions

Let $(x_1, y_1), (x_2, y_2), ..., (x_n, y_n)$ be a two dimensional data material. How do we measure the covariation?

Definition (6.7)

The sample covariance is defined as

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

Not scale invariant!

Data analysis

Data materials in several dimensions

Definition (6.8)

The sample correlation coefficient is defined as

$$r_{xy} = \frac{c_{xy}}{s_x s_y}$$

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

Theorem (Sats 6.1)

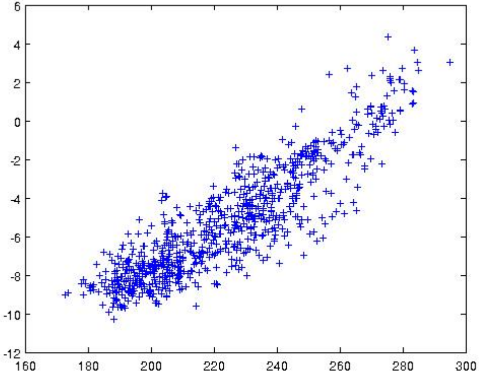
The sample correlation coefficient satisfies

$$-1 \leq r_{xy} \leq 1.$$

Data analysis

Data materials in several dimensions

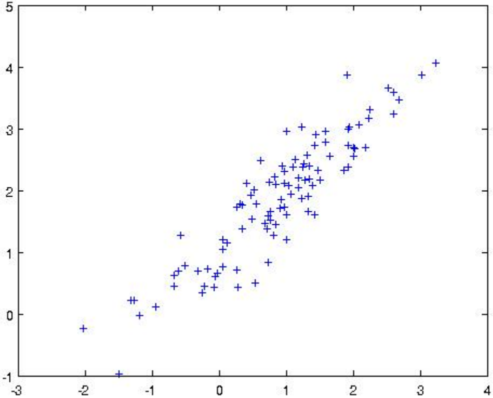
Ice core data: Carbon dioxide concentration (x) and temperature (y)



$r_{xy} = 0.89$

Data analysis

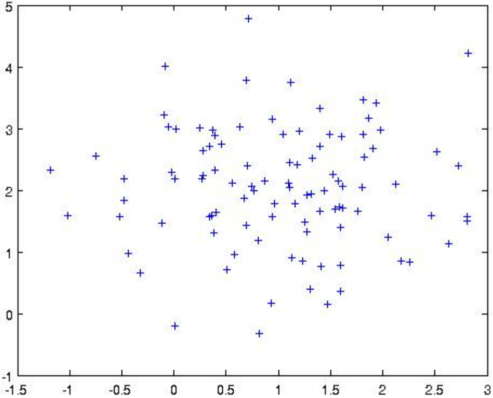
Data materials in several dimensions



Is the sample correlation -0.9, 0 or 0.9?

Data analysis

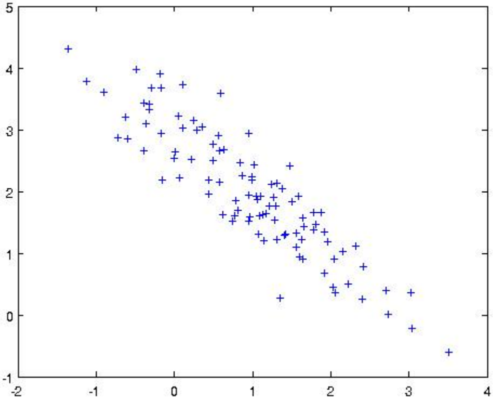
Data materials in several dimensions



Is the sample correlation -0.9, 0 or 0.9?

Data analysis

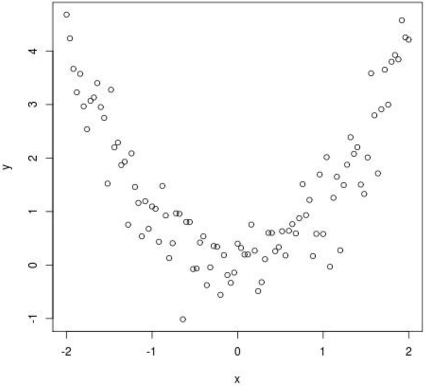
Data materials in several dimensions



Is the sample correlation -0.9, 0 or 0.9?

Data analysis

Data materials in several dimensions



Is the sample correlation -0.9, 0 or 0.9?

News of today

- Measures of location:
  - Sample mean
  - Median
  - Mode (typvärde)
- Measures of dispersion:
  - Sample variance
  - Sample standard deviation
  - Range (variationsbredd)
  - Inter quartile range (kvartilavstånd)
- Graphics:
  - Stem and leaf plot (Stam-bladdiagram)
  - Box plot (Lådagram)
  - Bar chart (Stapeldiagram)
  - Histogram
  - QQ plot
- Two dimensional:
  - Sample covariance, correlation coefficient

Problems: 6.4.3, 601, 602, 605