Probability and Statistics

Chapter 3: Descriptive Statistics

Joris Vankerschaver

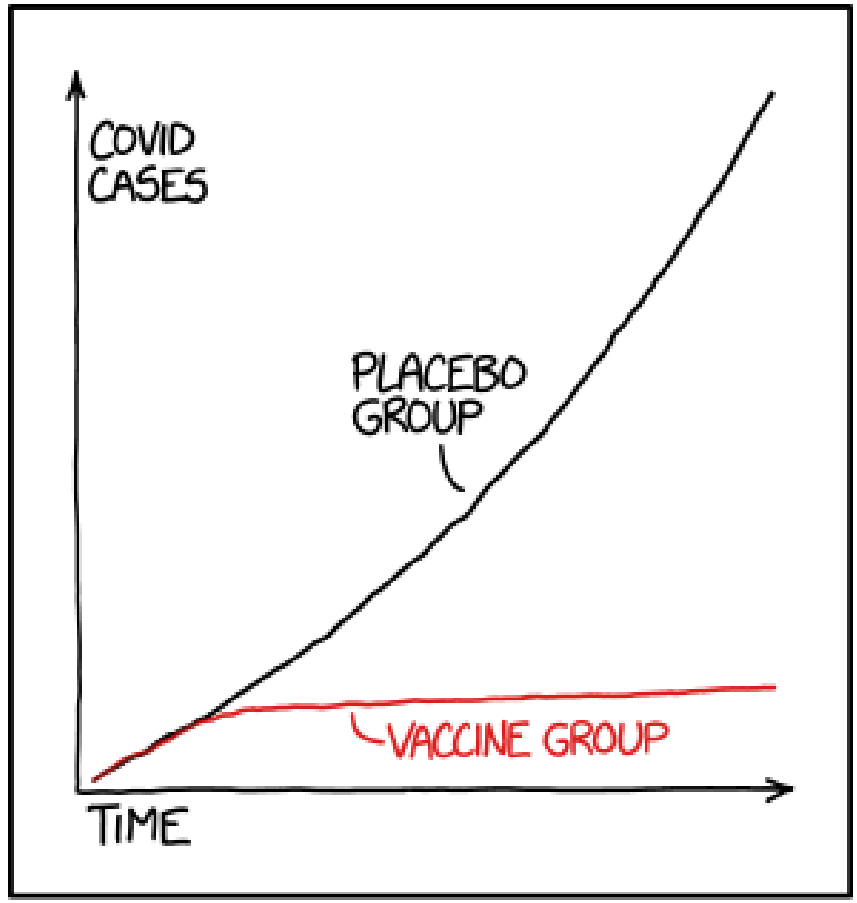
Goal of this chapter

Obtain insight and information from datasets:

- Summarize location, spread, etc
- Recognize associations between variables
- Build compelling visualizations

with a goal of:

- Confirming existing research hypotheses
- Generating new hypotheses



STATISTICS TIP: ALWAYS TRY TO GET DATA THAT'S GOOD ENOUGH THAT YOU DON'T NEED TO DO STATISTICS ON IT

Section 1

Introduction

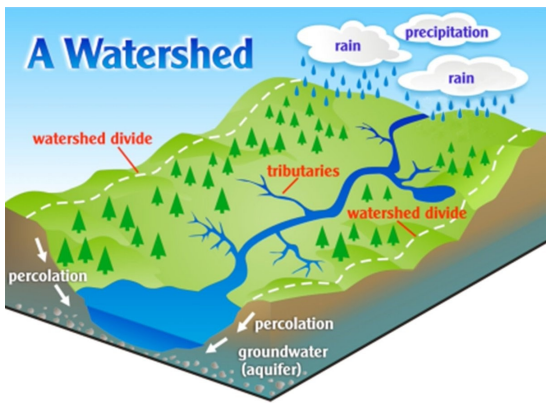
Overview

1. **Introduction**
    1. Catskill Mountains Data Set
    2. Types of Data
2. Graphical representation
    1. Categorical Variables
    2. Numerical Variables
3. Descriptive Statistics
    1. Measures of Location: Mean, Median, Geometric Mean, Mode
    2. Measures of Spread: Variance, Standard Deviation, IQR
    3. Measures of Association: Pearson Correlation
4. Case study: Captopril
5. Spurious correlations (optional)

Composition of forested watersheds

- Chemical composition of streams in Catskill Mountains (NY State, USA)
- 38 sites:
  - Concentration of 10 chemical variables (e.g., SO_4^{2-} and Cl^-), average over 3 years
  - 4 characteristics: max. height, height where sample was drawn, length, area of watershed

Goal: Describe "normal'' concentrations SO_4^{2-} and Cl^- in these mountains.



Subset of the data

STREAM	MAX	SAMP	SO4	CL	HEIGHT
Santa Cruz	1006	680	50.6	15.5	1
Colgate	1216	628	55.4	16.4	1
Halsey	1204	625	56.5	17.1	1
Batavia Hill	1213	663	57.5	16.8	1
Windham Ridg	1074	616	58.3	18.3	1
Silver Sprin	1113	451	63.0	15.7	0
Little Timbe	1027	463	66.5	26.9	0
Hunter	1234	634	64.5	22.0	1
West Kill	1234	658	63.4	21.3	1
Mill	1137	674	58.4	29.8	1
Kelly Hollow	1061	533	70.6	18.4	1
Pigeon	1173	619	56.9	16.6	1

Different types of variables

1. **Numerical** (quantitative): lots of different values possible
   - Discrete: e.g., number of salamanders of species P. jordani on 1 km^2
   - Continuous: e.g., age, weight
2. **Categorical** (qualitative): limited number of values possible
   - Ordinal: e.g., height below 500 m ($X = 0$), height between 500 m and 750 m ($X = 1$), height above 750 m ($X = 2$)
   - Nominal: e.g., salamander of species P. jordani ($X = 0$), P. metcalfi ($X = 1$), or P. shermani ($X = 2$)

Hint

Can you do math with it: add, subtract, multiply, take the mean, ...? Then it is probably a numerical variable.

Variables in data set

- STREAM:
- MAX:
- SAMP:
- SO4:
- CL:
- HEIGHT:

Variables in data set

- STREAM: Categorical, nominal
- MAX: Numerical, continuous
- SAMP: Numerical, continuous
- SO4: Numerical, continuous
- CL: Numerical, continuous
- HEIGHT: Categorical, ordinal

Section 2

Representing statistical variables

Overview

1. Introduction
    1. Catskill Mountains Data Set
    2. Types of Data
2. **Graphical representation**
    1. Categorical Variables
    2. Numerical Variables
3. Descriptive Statistics
    1. Measures of Location: Mean, Median, Geometric Mean, Mode
    2. Measures of Spread: Variance, Standard Deviation, IQR
    3. Measures of Association: Pearson Correlation
4. Case study: Captopril
5. Spurious correlations (optional)

Textual representation of categorical variables

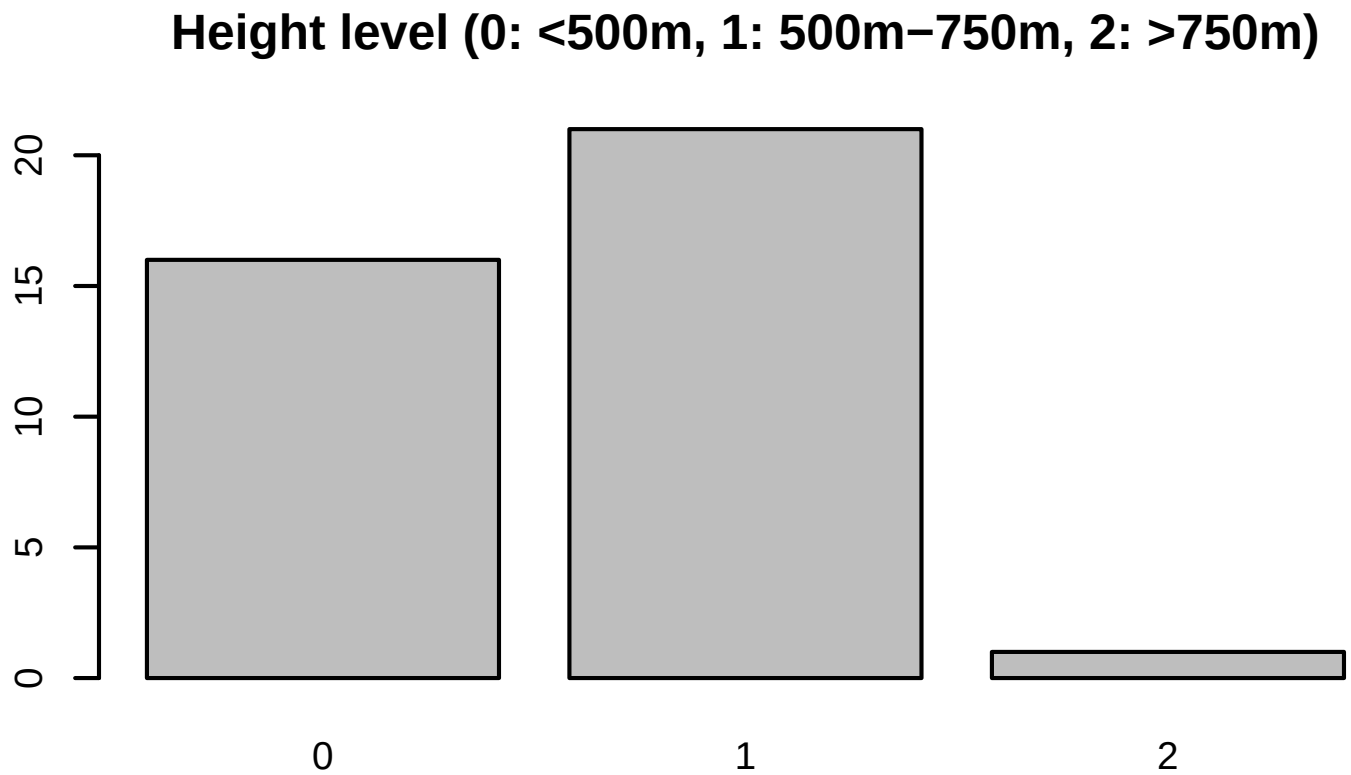
```
summary(lovett$HEIGHT2)
##  0  1  2
## 16 21  1
```

```
knitr::kable(summary(lovett$HEIGHT2),
             col.names = c("HEIGHT2"))
```

	HEIGHT2
0	16
1	21
2	1

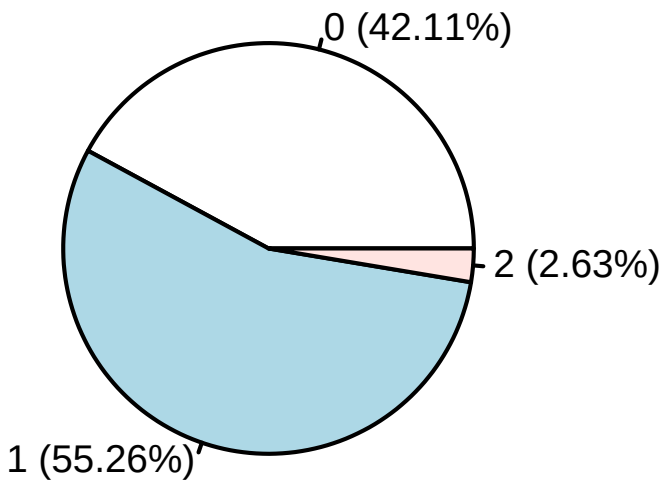
- Tables are good when not too many rows (5-10)
- Consider if a plot would be more informative

Graphical representation of categorical variables: barplot




```
barplot(c(16,21,1),
  main = "Height level (0: <500m, 1: 500m-750m, 2: >750m)",
  names.arg=c(0, 1, 2))
```

Graphical representation of categorical variables: pie chart



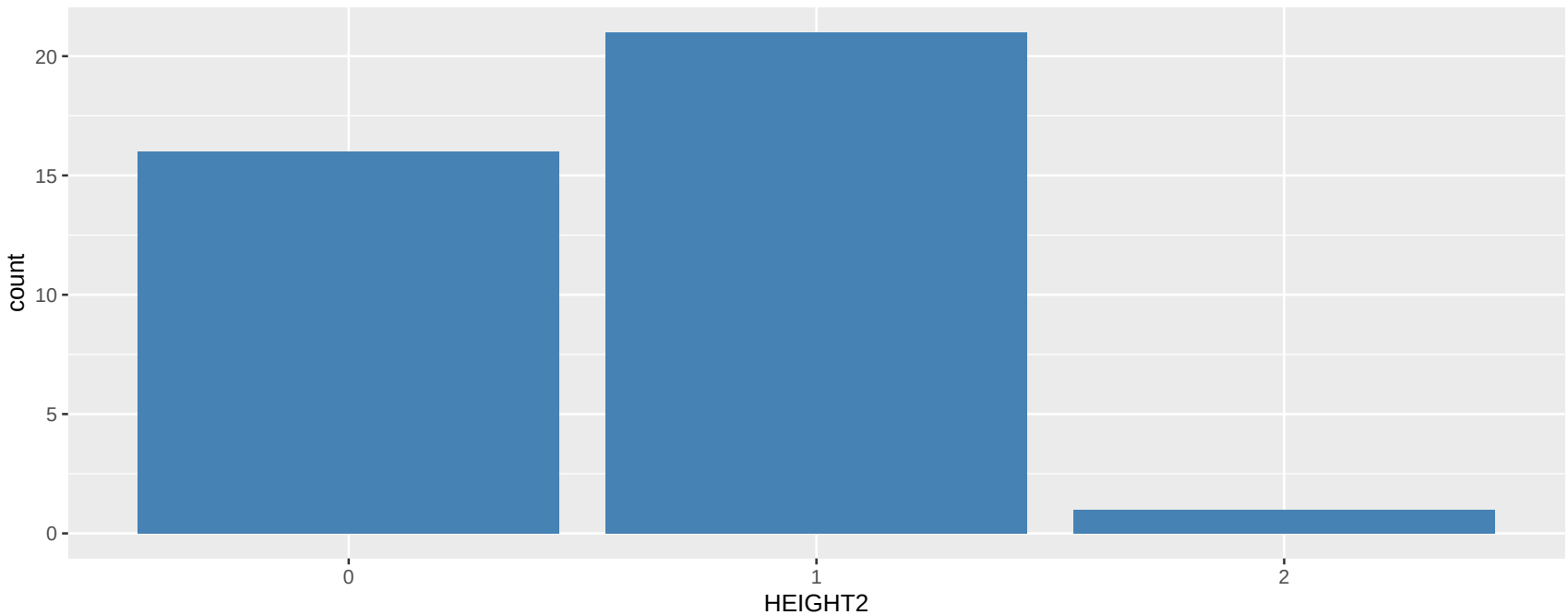

```
pie(c(16,21,1),
  labels = c('0 (42.11%)', '1 (55.26%)', '2 (2.63%)'),
  main = 'Height level (0: <500m, 1: 500m-750m, 2: >750m)')
```

Warning

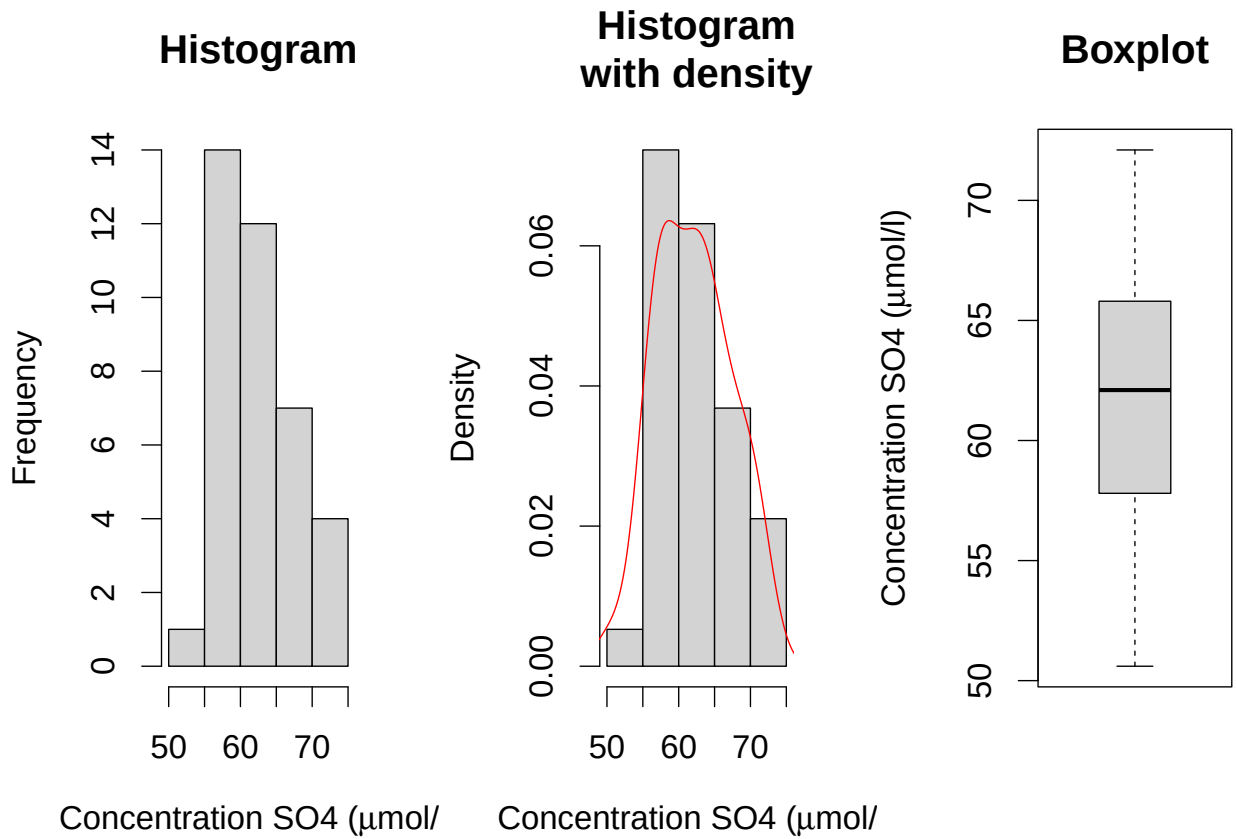
Hard to compare size of pie wedges!

Aside: Prettier graphics with ggplot2

```
library(ggplot2)
ggplot(lovett, aes(x=HEIGHT2)) +
  geom_bar(stat="count", fill="steelblue")
```



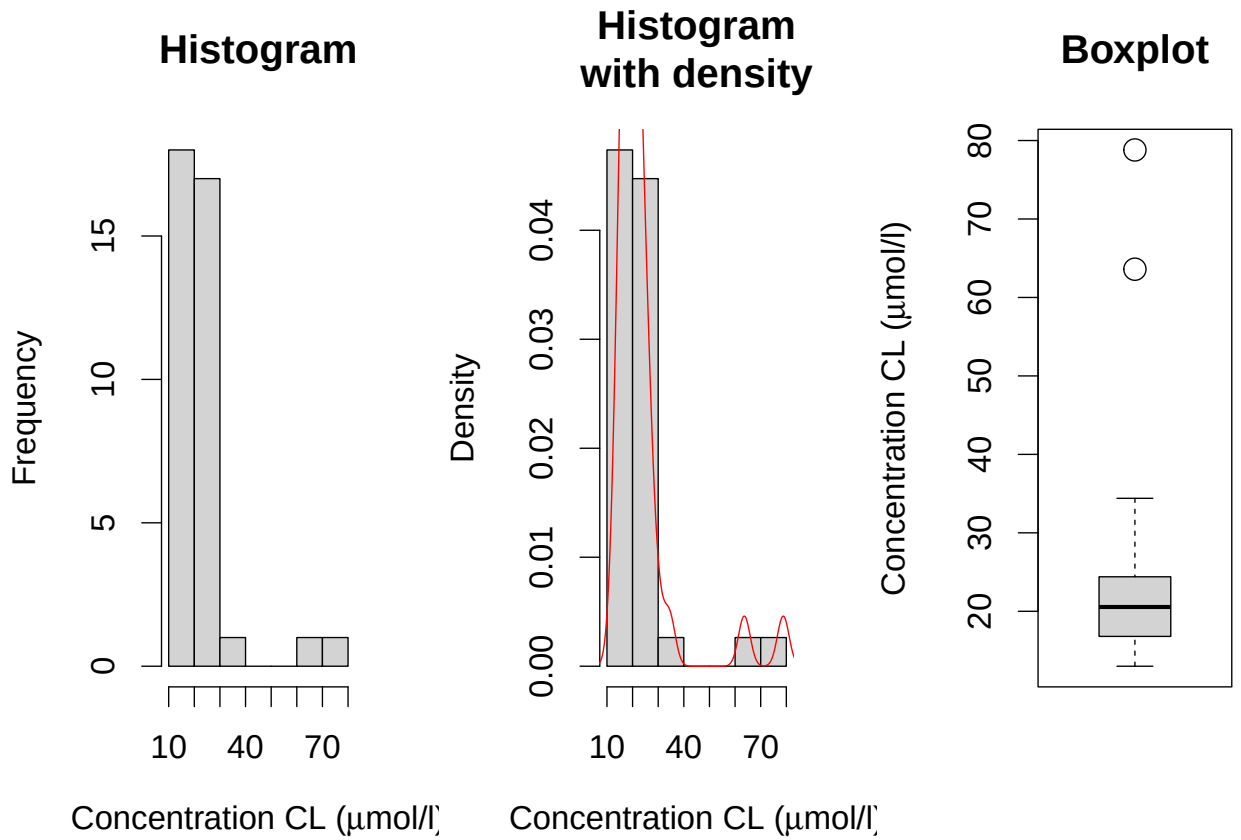
Graphical representation of numerical variables: SO4




```
summary(lovett$SO4)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   50.60   57.92   62.10   62.08   65.72   72.10
```

Graphical representation of numerical variables: CL




```
summary(lovett$CL)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   13.00   16.88   20.55   23.07   23.90   78.80
```

Section 3

Descriptive statistics

Overview

1. Introduction
    1. Catskill Mountains Data Set
    2. Types of Data
2. Graphical representation
    1. Categorical Variables
    2. Numerical Variables
3. **Descriptive Statistics**
    1. Measures of Location: Mean, Median, Geometric Mean, Mode
    2. Measures of Spread: Variance, Standard Deviation, IQR
    3. Measures of Association: Pearson Correlation
4. Case study: Captopril
5. Spurious correlations (optional)

Descriptive statistics

Characterize the distribution of the data in a concise way:

- What is center?
- How do data vary around center?
- Symmetrical or not?
- Is there an association between variables?

Measures of Location: Center

- Suppose we have n observations $x_1, \ldots, x_n$
- How can we summarize these observations in **one single value**?
- Many different options:
  - Mean
  - Median
  - Geometric mean
  - Mode

SO_4^{2-} concentration of $n = 38$ watersheds

Mean: 62.08, median: 62.10, geometric mean: 61.87, mode: 63.40

Mean (= mathematical average)

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

Examples:

- Mean of 1, 2, 3, 4, 5, 6 equals ...
- Mean concentration SO_4^{2-} is 62.08 μmol/l
- Mean desired number of partners over 30 years (Miller and Fishkin, 1997):
  - Men: 64.3
  - Women: 2.8

Median or 50% percentile

The number $\mathbf{x_{50}}$ such that

- at least half of the observations is larger or equal to $\mathbf{x}_{50}$ **and**
- at least half of the observations is smaller or equal to $\mathbf{x}_{50}$

Calculation:

1. Sort the data
2. Choose correct value
   - if n odd: middle observation
   - if n even: average of the middle 2 observations

Median or 50% percentile

Examples:

- Median of 1, 2, 3, 4, 5, 6 equals . . .
- Median of 1, 2, 3, 4, 60 equals . . . (mean = . . .)
- Median concentration SO_4^{2-} is 62.10 μmol/l
- In a time span of 30 year, the median of the desired number of partners is 1 for both men and women (Miller and Fishkin, 1997)

Also makes sense for censored data:

- Lab animals that had chemotherapy are observed until death
- Study runs for 12 weeks
- Observations for the 6 animals:
  1, 3, 4, 7, $>$12, $>$12 weeks

Mean or median?

- The mean is very **sensitive** to outliers
- The median is **robust** to outliers

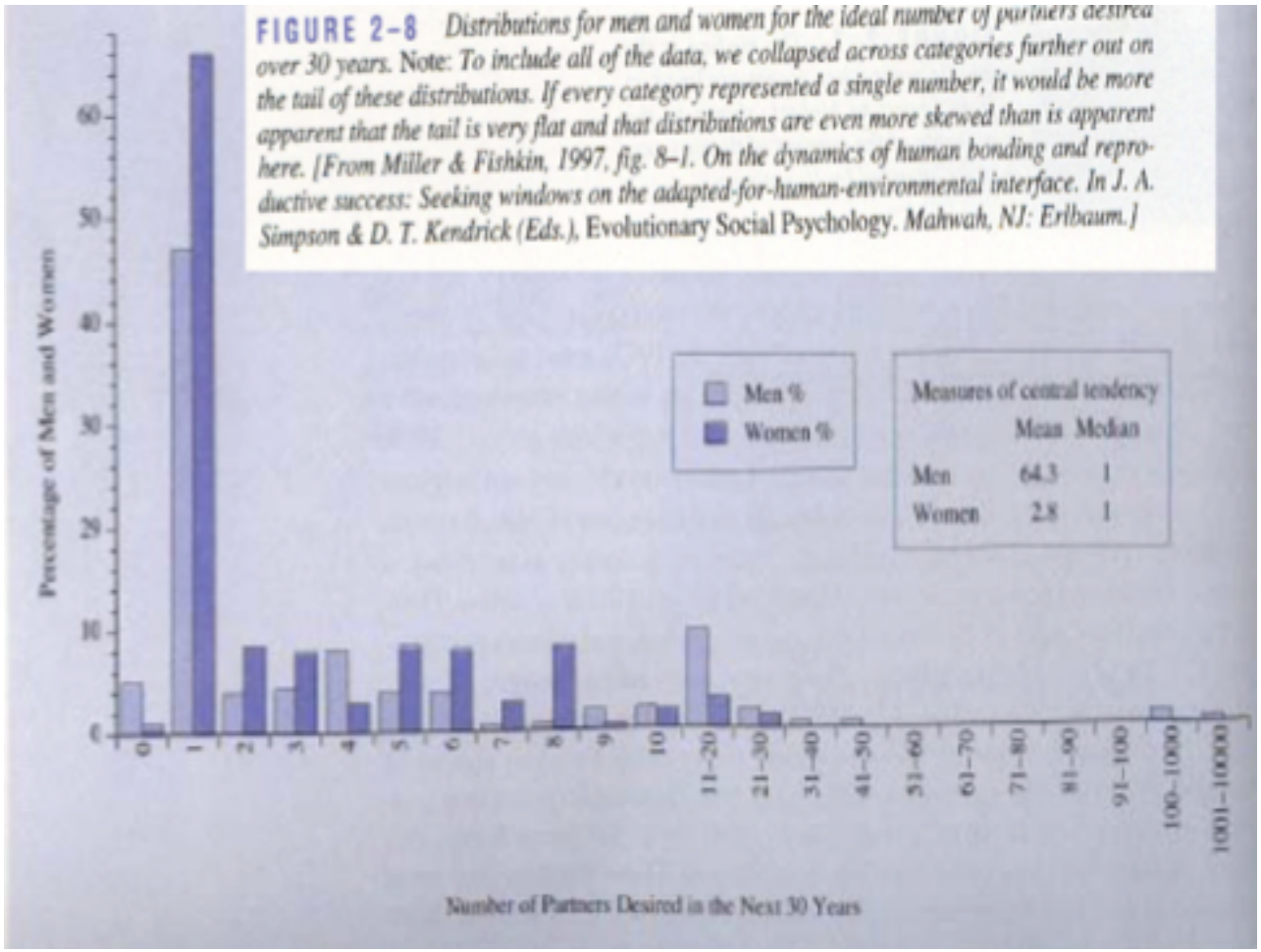
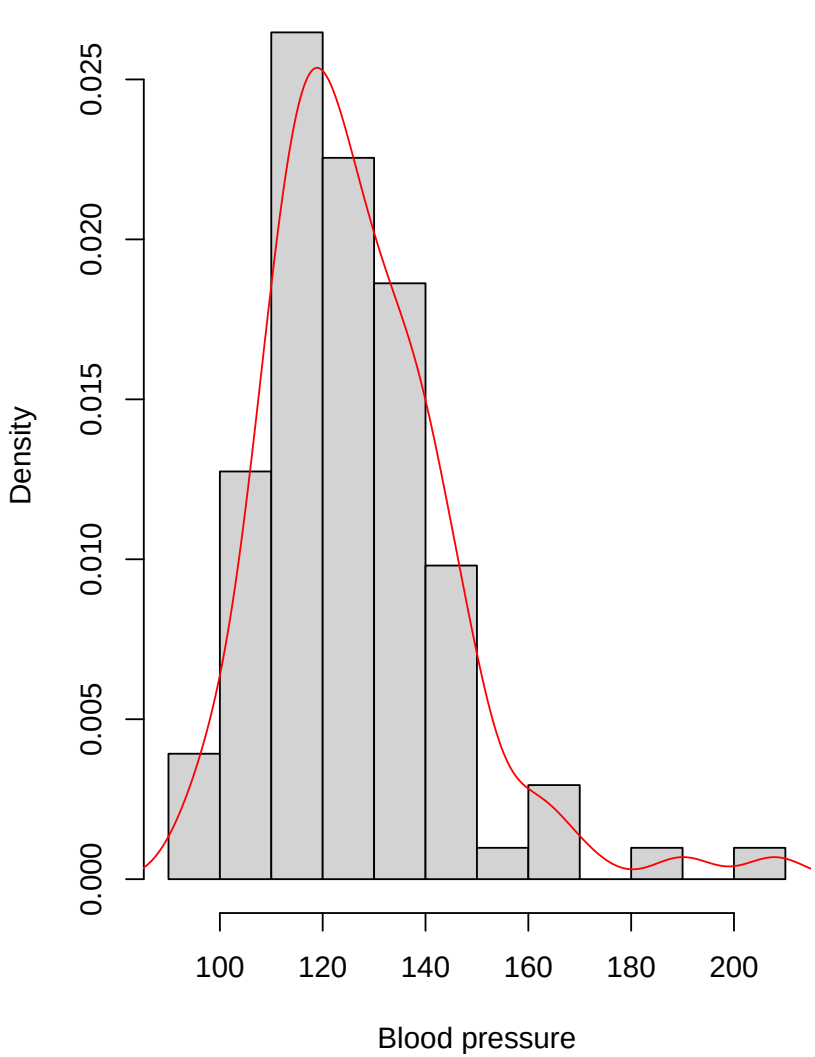


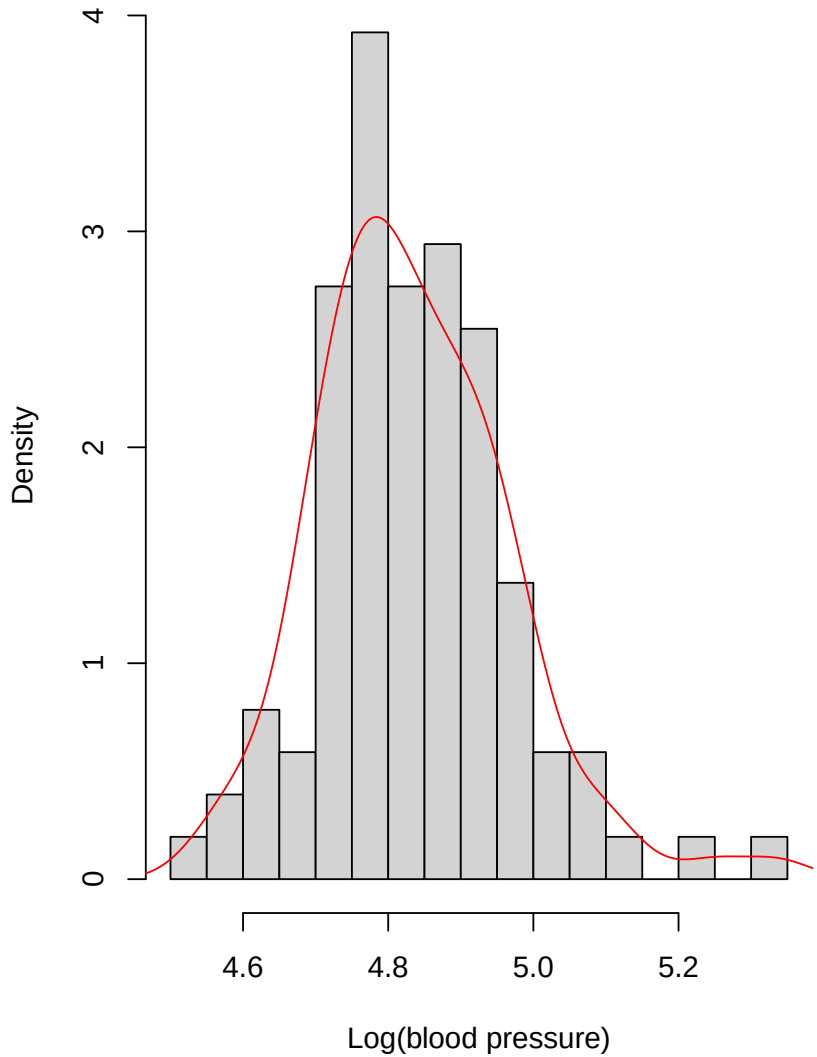
FIGURE 2–8 *Distributions for men and women for the ideal number of partners desired over 30 years. Note: To include all of the data, we collapsed across categories further out on the tail of these distributions. If every category represented a single number, it would be more apparent that the tail is very flat and that distributions are even more skewed than is apparent here. [From Miller & Fishkin, 1997, fig. 8–1. On the dynamics of human bonding and reproductive success: Seeking windows on the adapted-for-human-environmental interface. In J. A. Simpson & D. T. Kendrick (Eds.),* Evolutionary Social Psychology. *Mahwah, NJ: Erlbaum.]*

Mean or median?

1. Symmetrical distributions: mean and median **equal** (**in theory**)
   - Report **mean**:
     - More precise: extracts more information from data
     - More flexible
2. Skewed distributions or outliers: mean attracted by extreme values
   - Report **median** or transform data to more symmetric distribution (where mean works)

Blood pressure in obese patients





Geometric mean

- Blood pressure: **right skewed**:
  - Mean: 127 mm Hg, median: 124 mm Hg
- Log-transform: more symmetric
  - Mean: 4.84, median: 4.82
- Transform back using exponential function

$$\exp(4.84) = 126.5\,\text{mg/cl}$$

- Result is called geometric mean
- Preferable over median because
  - Usually more precise
  - More flexible since based on mathematical average

Measures of spread

- Not only center, but also how data are spread
- Sometimes primary goal of analysis

Examples:

- Determine 'normal' concentrations SO_4^{2-} and cholesterol
- Determine precision of measurements (e.g., alcohol control)

- For every analysis: also secondary goal
- Information on reliability of data in sample

Variance

- Spread: deviations $x_i - \bar{x}$
- Average of these is 0
- Use square!

Variance: average of the squared deviations

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

Standard deviation

Disadvantage: different scale

Example:

- Variance of concentration SO_4^{2-} is 27.46 $(\mu mol/l)^2$

Standard deviation (SD)

$$s_x = \sqrt{s_x^2}$$

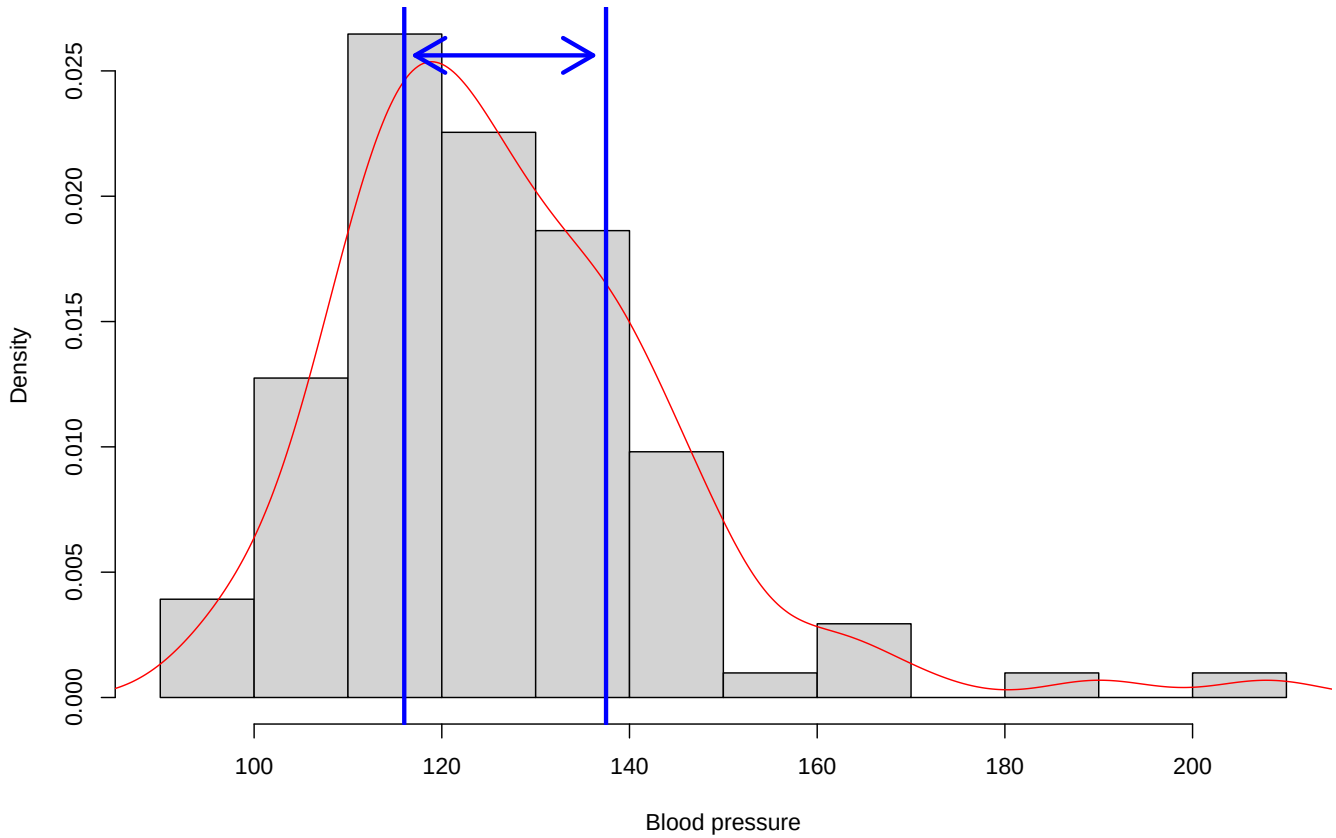
Example:

- SD of concentration SO_4^{2-} is 5.24 $\mu mol/l$

Interquartile range (IQR)

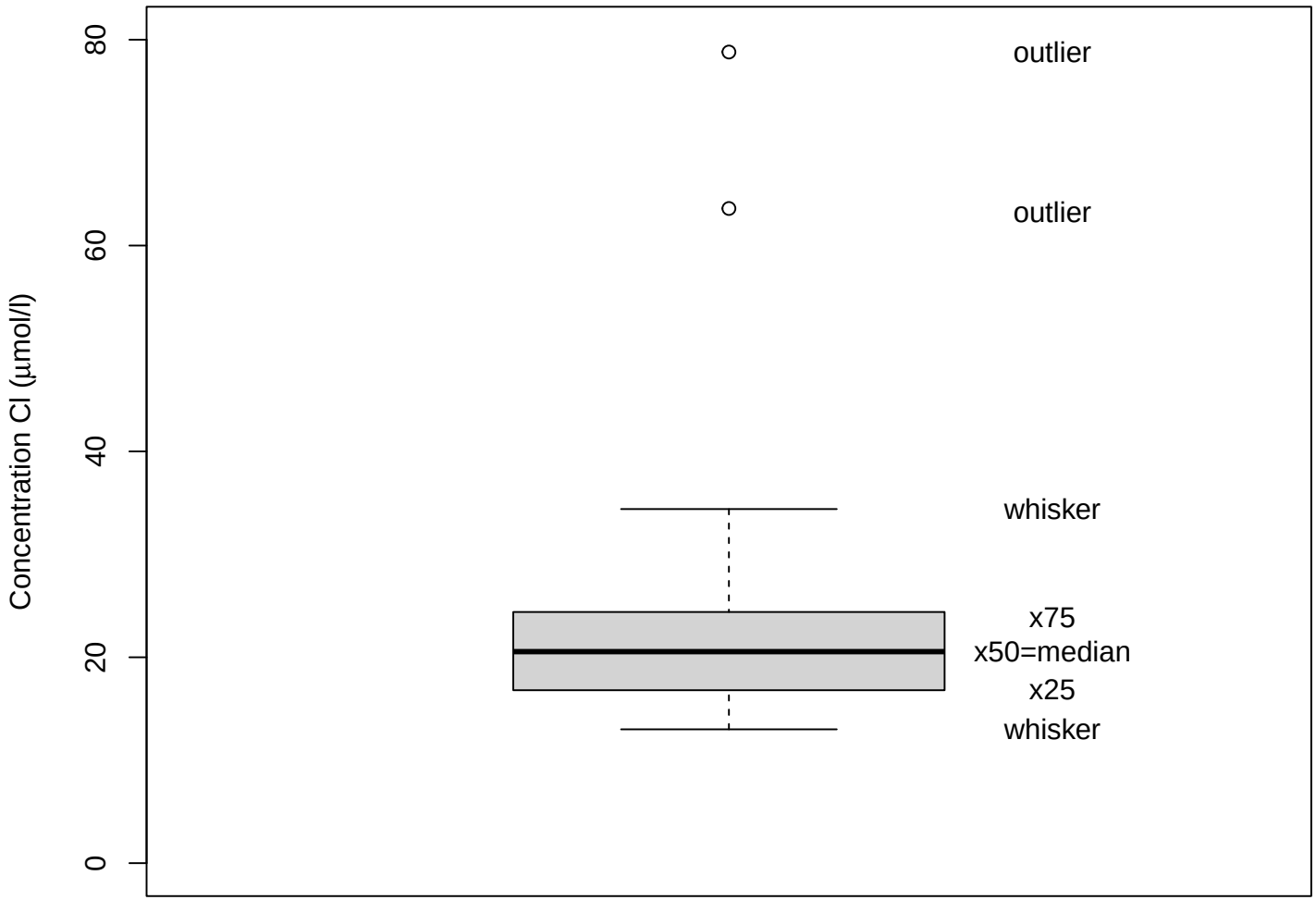
IQR

- Middle 50% of the data
- Difference of 75% and 25% percentile



Mean $\leftrightarrow$ SD, median $\leftrightarrow$ IQR

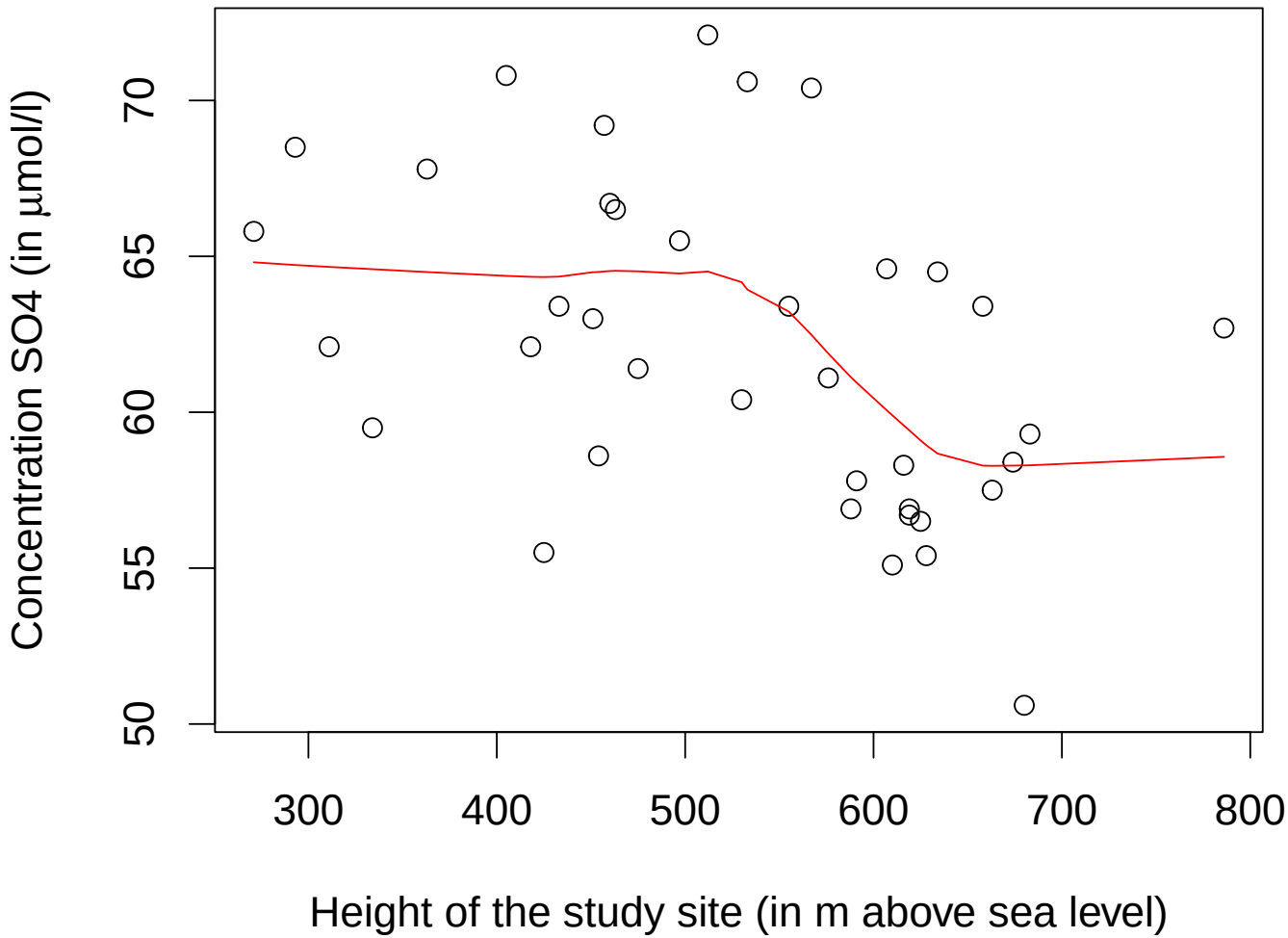
Boxplot Cl^-


Concentration Cl (µmol/l)
0
20
40
60
80
outlier
outlier
whisker
x75
x50=median
x25
whisker

Association concentration SO_4^{2-} and height

- What is association between concentration SO_4^{2-} and height of the stream?
- Regression curve, such as scatterplot smoother

Scatterplot



The red is optional and represents the trend in the data, if any.

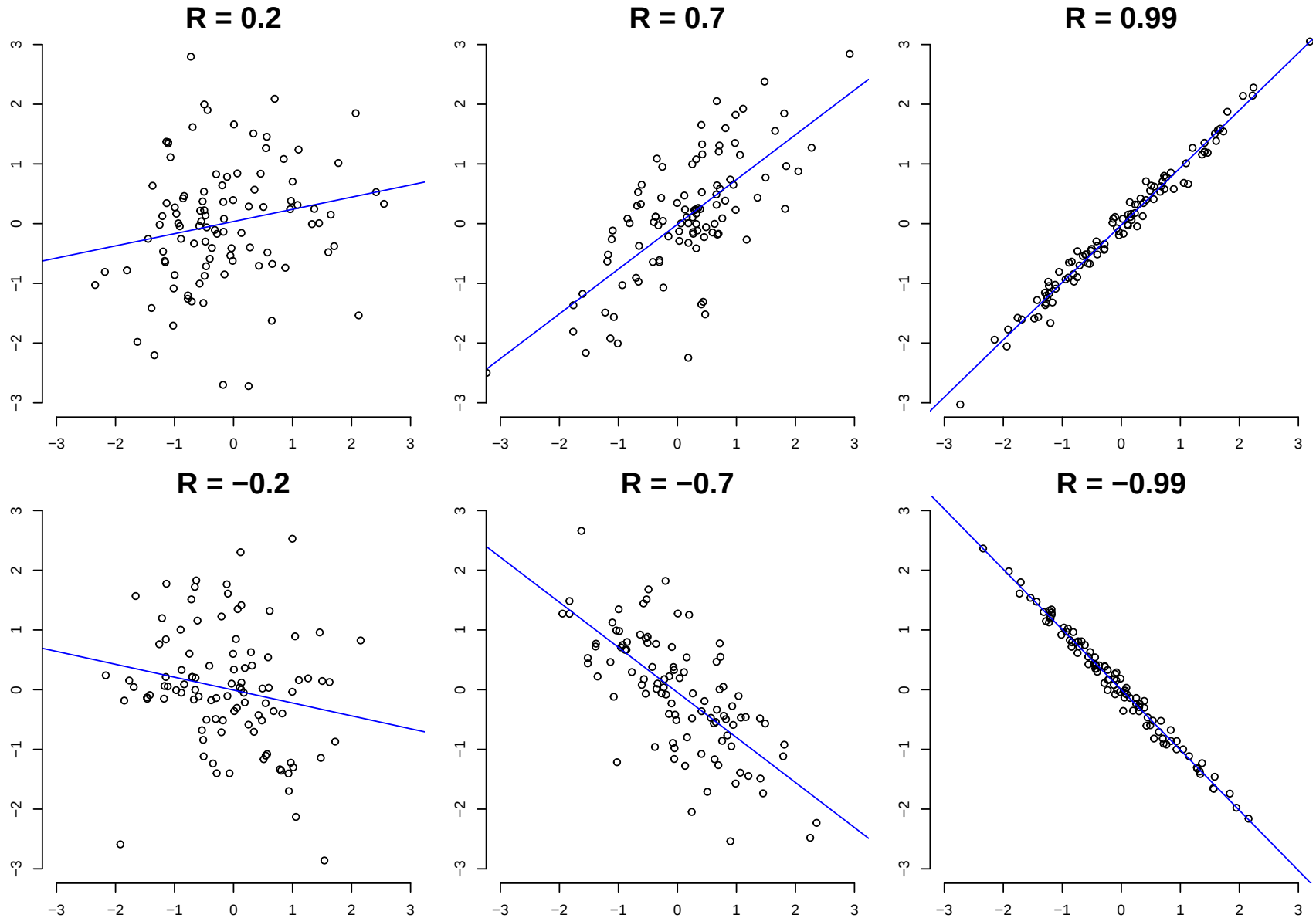
Pearson correlation

Expresses the **linear** association between 2 variables

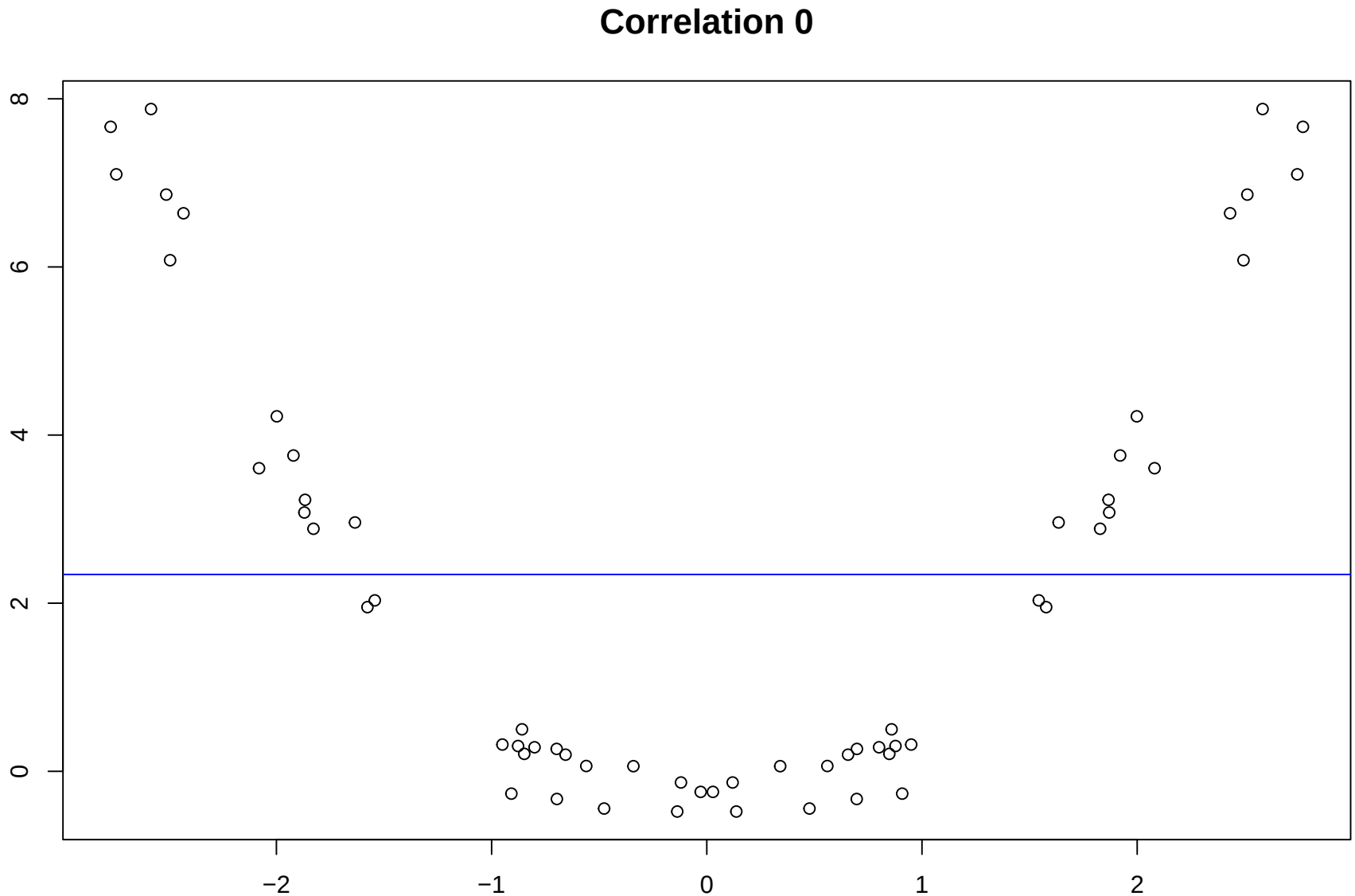
$$\text{Cor}(x, y) = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{(n-1)s_x s_y}$$

- Positive correlation: $x \nearrow \Rightarrow y \nearrow$
- Negative correlation: $x \nearrow \Rightarrow y \searrow$
- Always between -1 and 1

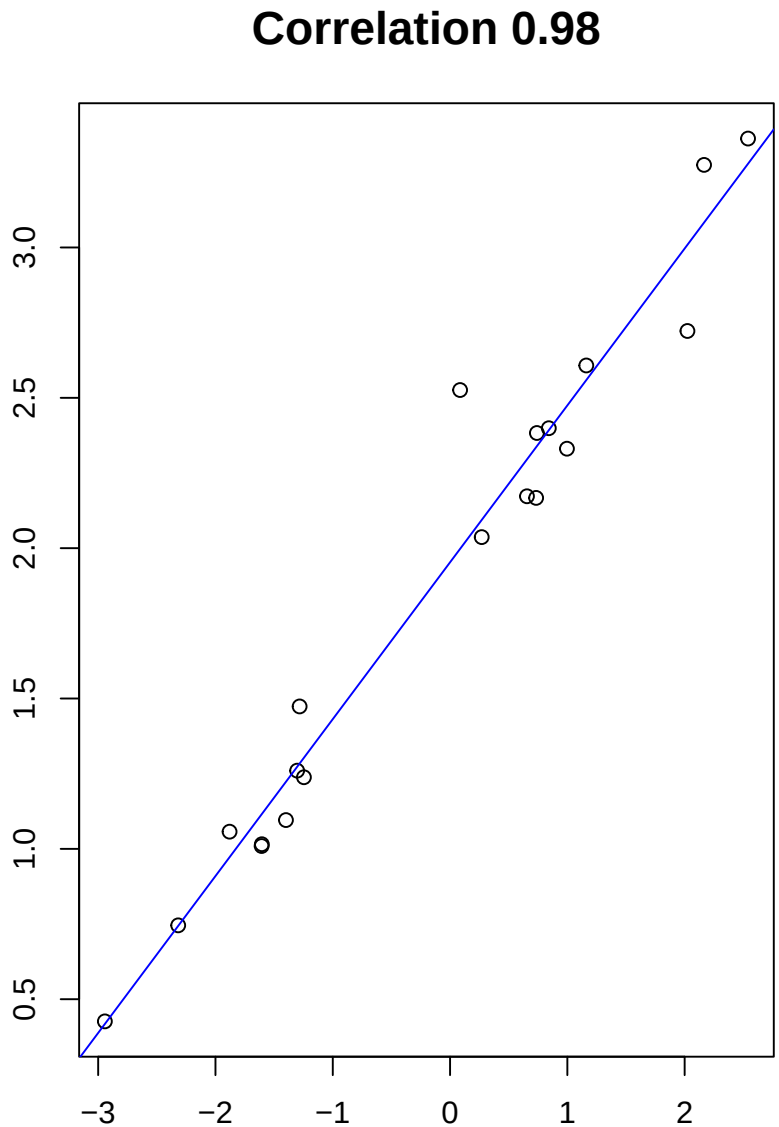
Pearson correlation

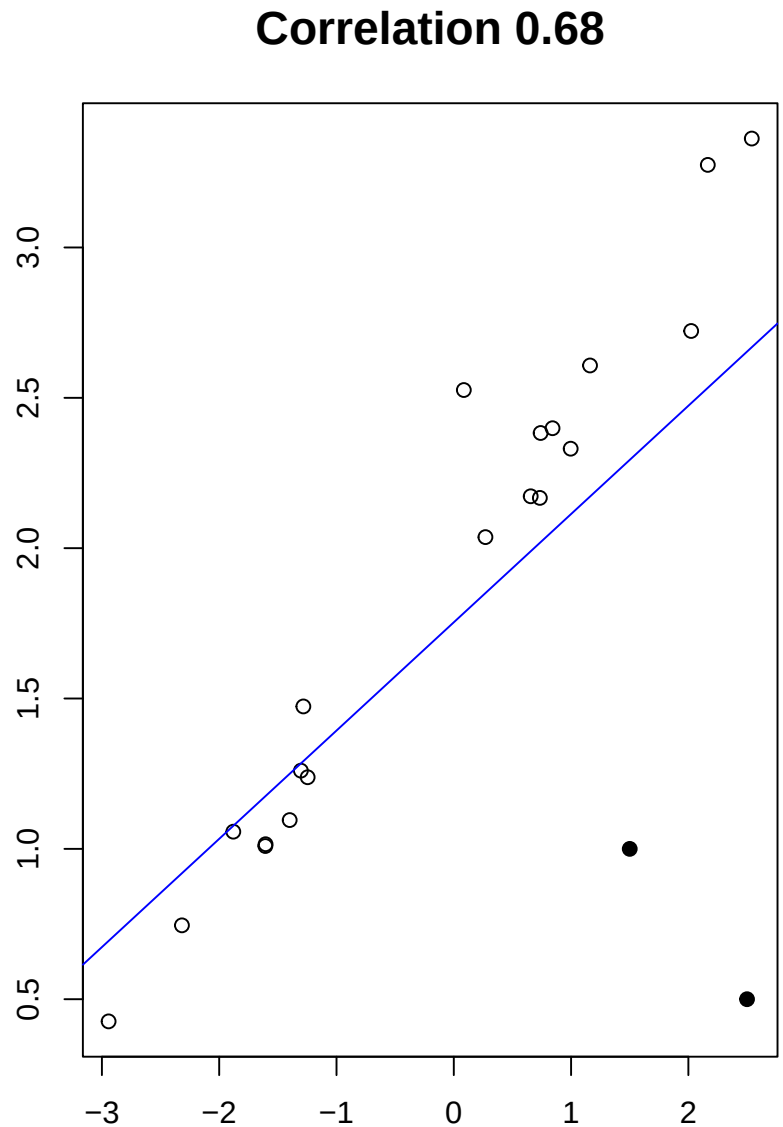


Correlation 0 means 'no **linear** association'



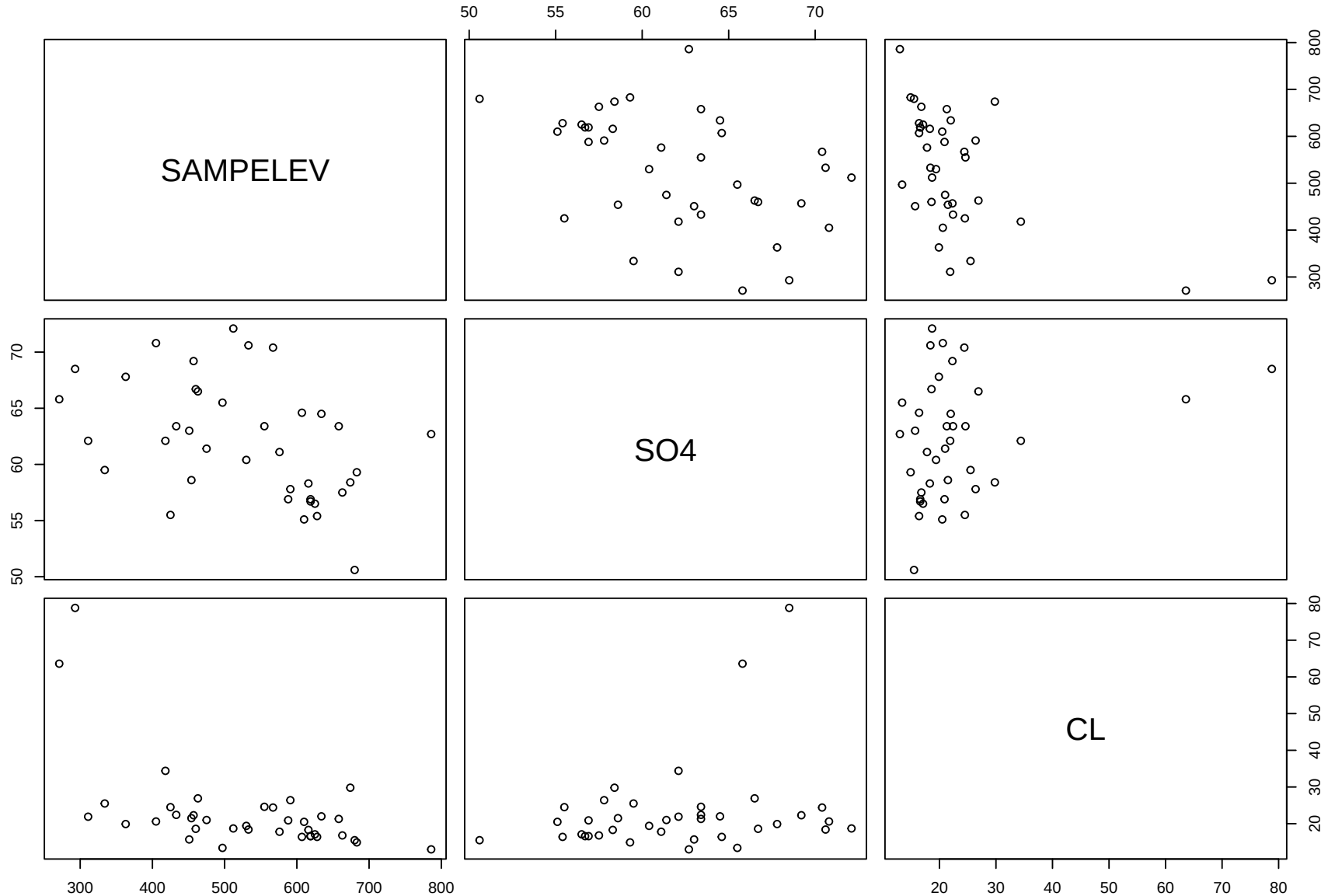
Correlation and outliers





Graphical representation of multiple numerical variables

```
pairs(lovett[,c("SAMPELEV", "SO4", "CL")])
```



Section 4

Case study: Captopril

Overview

1. Introduction
    1. Catskill Mountains Data Set
    2. Types of Data
2. Graphical representation
    1. Categorical Variables
    2. Numerical Variables
3. Descriptive Statistics
    1. Measures of Location: Mean, Median, Geometric Mean, Mode
    2. Measures of Spread: Variance, Standard Deviation, IQR
    3. Measures of Association: Pearson Correlation
4. **Case study: Captopril**
5. Spurious correlations (optional)

Summary so far

Research question

Is there a significant change in blood pressure before/after administration of Captopril?

Study

- Sample of 15 patients
- Blood pressure before/after intervention (systolic and diastolic)

The data

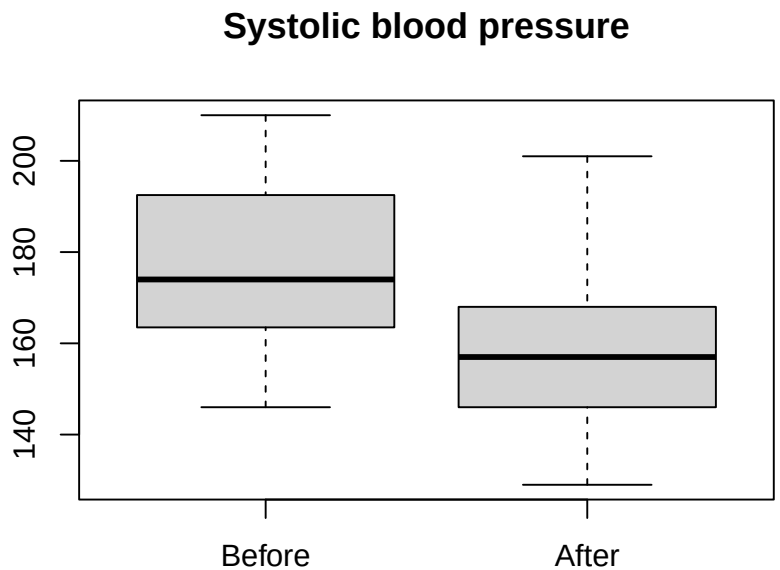
	Before		After	
Patient	SBP	DBP	SBP	DBP
1	210	130	201	125
2	169	122	165	121
3	187	124	166	121
4	160	104	157	106
5	167	112	147	101
6	176	101	145	85
7	185	121	168	98
8	206	124	180	105
9	173	115	147	103
10	146	102	136	98
11	174	98	151	90
12	201	119	168	98
13	198	106	179	110
14	148	107	129	103
15	154	100	131	82

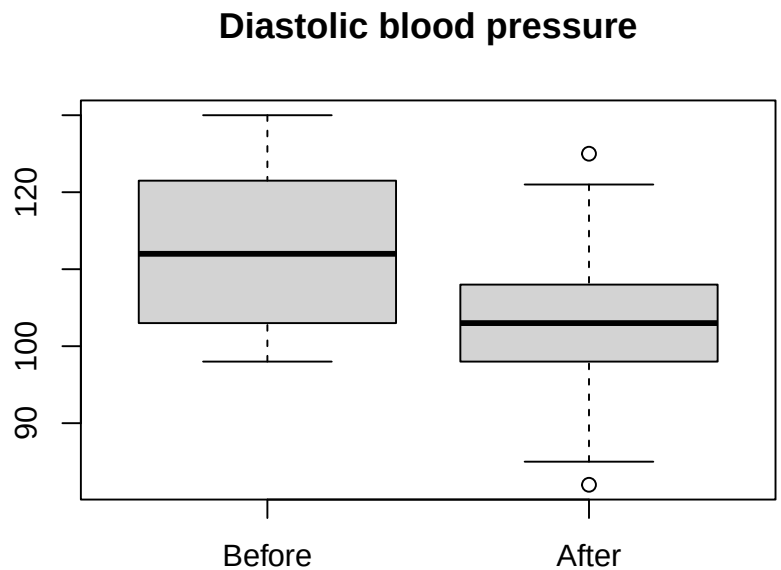
Goal of descriptive statistics

- Answer specific questions:
  - What is the average blood pressure before/after?
  - What is the spread in each group?
- Gain support for our research hypothesis
- Gather additional research questions
  - "That's weird..."
- Verify that data has been loaded correctly

SBP/DBP before/after administration

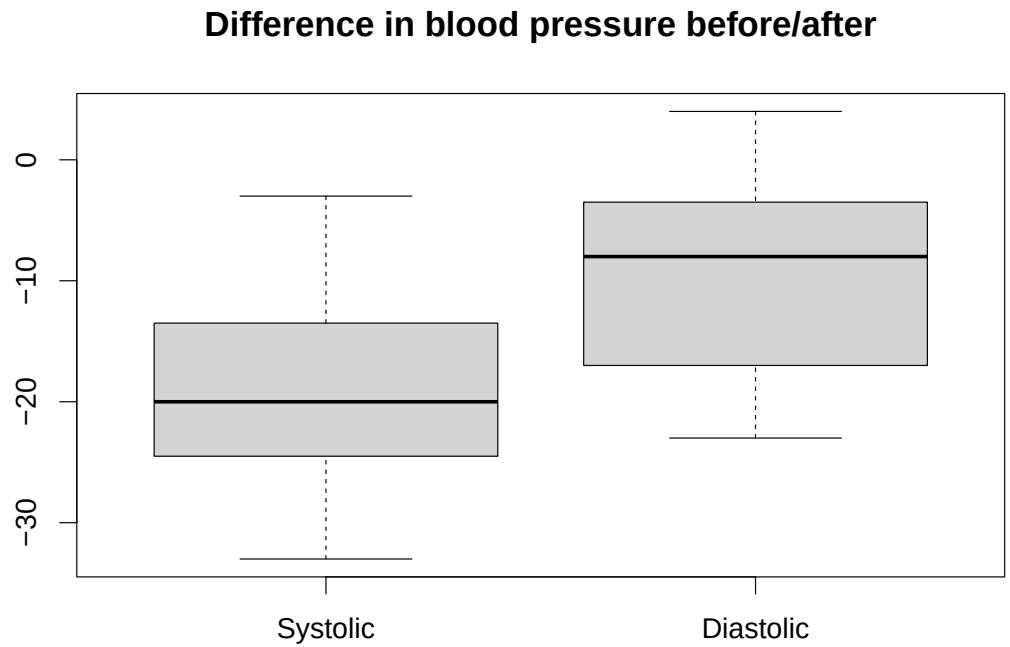
```
par(mfrow=c(1, 2))
boxplot(captopril$SBPb, captopril$SBPa,
        names=c("Before", "After"), main="Systolic blood pressure")
boxplot(captopril$DBPb, captopril$DBPa,
        names=c("Before", "After"), main="Diastolic blood pressure")
```





Difference before/after

```
par(cex=1.5)  # For slides only
boxplot(captopril$SBPa - captopril$SBPb,
        captopril$DBPa - captopril$DBPb,
        names=c("Systolic", "Diastolic"),
        main="Difference in blood pressure before/after")
```



Section 5

Spurious correlations (optional)

Overview

1. Introduction
    1. Catskill Mountains Data Set
    2. Types of Data
2. Graphical representation
    1. Categorical Variables
    2. Numerical Variables
3. Descriptive Statistics
    1. Measures of Location: Mean, Median, Geometric Mean, Mode
    2. Measures of Spread: Variance, Standard Deviation, IQR
    3. Measures of Association: Pearson Correlation
4. Case study: Captopril
5. **Spurious correlations (optional)**

Are attractive people mean?

The claim

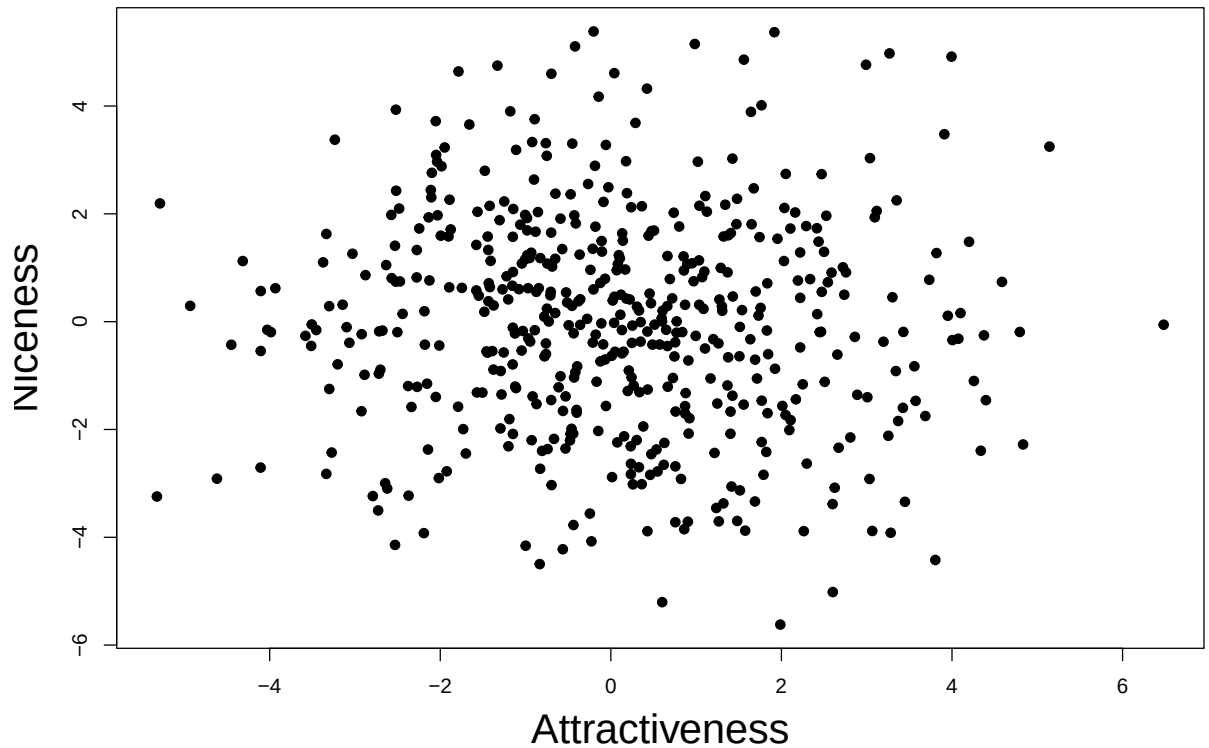
Among the people that we interact with, attractive people are meaner, on average. Conversely, less attractive people are nicer.

Is there any truth to this?

- No biological basis for such a claim
- Spurious correlations can appear due to how data is generated

Let's generate some data

Let's look at a simulated population of 500 individuals for which attractiveness and niceness are independent.



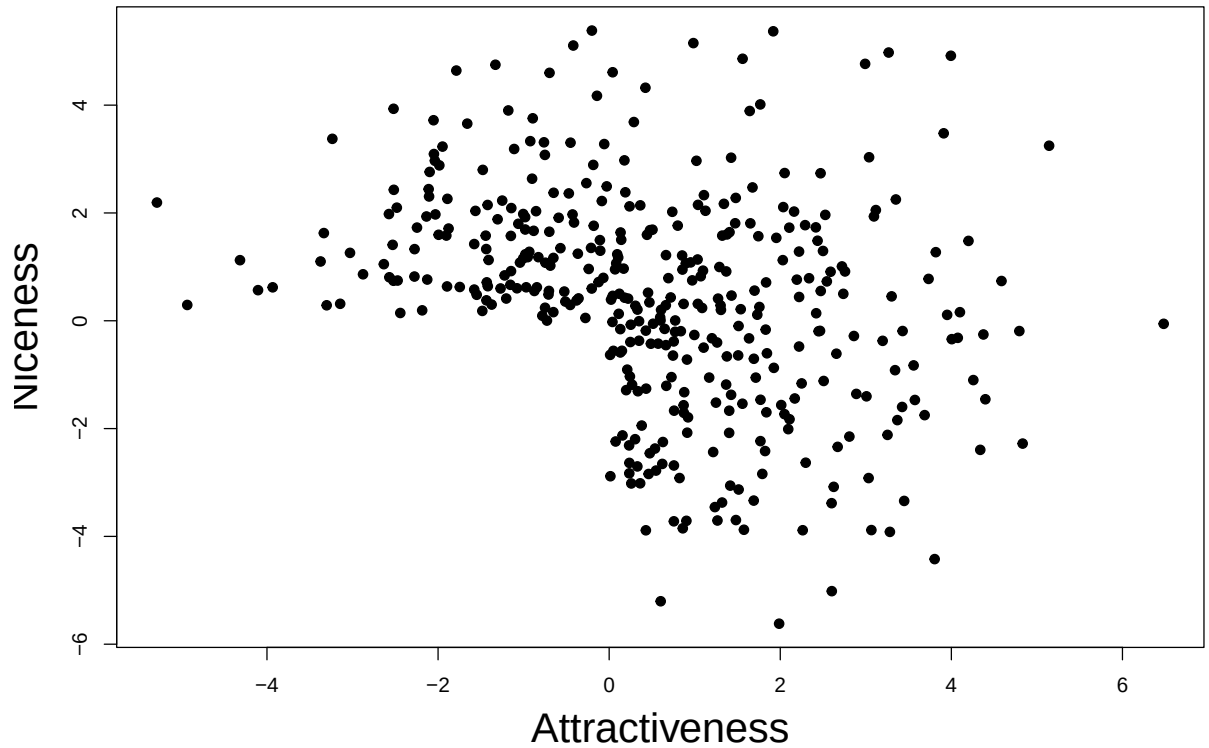
- Sample correlation: Cor = -0.05.
- No association between attractiveness and niceness.

Selection bias

We don't want to hang out with people who are both mean (nice $<$ 0) and non-attractive (attractive $<$ 0). Let's exclude those people.

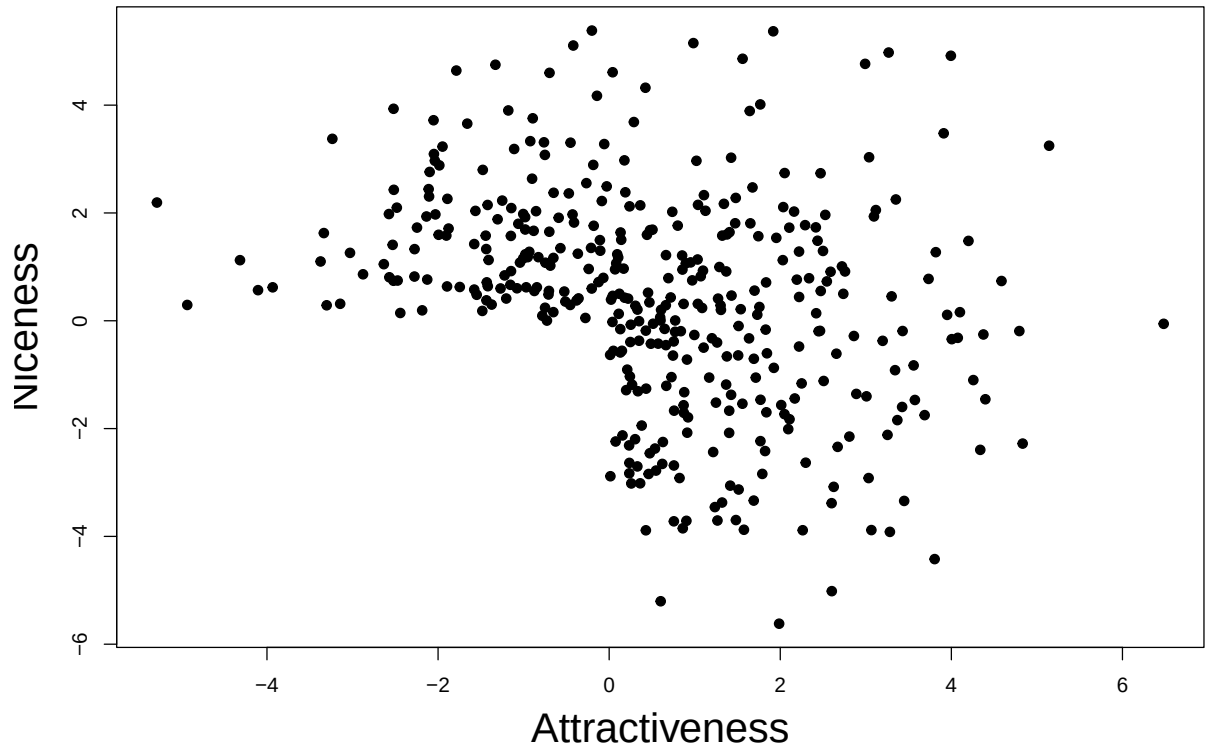
Selection bias

We don't want to hang out with people who are both mean (nice $<$ 0) and non-attractive (attractive $<$ 0). Let's exclude those people.



Selection bias

We don't want to hang out with people who are both mean (nice $<$ 0) and non-attractive (attractive $<$ 0). Let's exclude those people.



- Sample correlation: Cor = -0.32.
- An association appears, *just because of how we look at the data.*