Basic Methods of Data Analysis
Part 1

Sepp Hochreiter

Institute of Bioinformatics
Johannes Kepler University, Linz, Austria

Course

3 ECTS 2 SWS VO (class)

Basic Course of Master Bioinformatics (mandatory)

Basic Course of Master Computer Science “Intelligent Information Systems” (mandatory)
Basic Course of Master Computer Science “Computational Engineering” (elective)

Class: Thu 15:30-17:00 (MT 226/1)

final exam: 4 times written test (intermediate exams) -> see KUSSS

Other Courses:

- Machine Learning: supervised methods (2VL, Wed 15:30-17:00, HS 5, Ulrich Bodenhofer)
  → Basic Course for Master Bioinformatics

- Sequence Analysis and Phylogenetics (2VL, Mon 15:30-17:00, S2 048)
  → Basic Course for Bachelor Bioinformatics and Complementary in Master Bioinformatics

Course Schedule Bachelor Bioinf 2017 3. Sem.

	MONDAY	TUESDAY	WEDNESDAY	THURSDAY	FRIDAY
8:30-9:15		320.102 Topics in Genetics & Evolution, 2KV		347.310 English for Chemistry 1, 2KV	
9:15-10:00					
10:15-11:00			347.311 English for Chemistry 1, 2KV		**365.062 Sequence Analysis and Phylogenetics, 2UE**
11:00-11:45					
12:00-12:45	326.015 Information systems, 2KV; 344.014 Artificial Intelligence, 2VO				
12:45-13:30					
13:45-14:30	344.021 Artificial Intelligence, 1UE	344.023 Artificial Intelligence, 1UE	347.334 Chemie für Physiker II, 2VO		
14:30-15:15	344.022 Artificial Intelligence, 1UE				
15:30-16:15	**365.060 Sequence Analysis and Phylogenetics, 2VL**				
16:15-17:00					
17:15-18:00	347325 English for Chem. 1, 2KV	320.011 Bioanalytics I, 2VO			
18:00-18:45				347308 English for Chemistry 1, 2KV	
19:00-19:45					

Course Schedule Bachelor Bioinf 2017 3. Sem.

Bioanalytics I (1UE, 470WEBIBA1U14):

The course will be given on the first two days of February 2018

Schedule Master Bioinf 2016 1. Sem.

	MONDAY	TUESDAY	WEDNESDAY	THURSDAY	FRIDAY
8:30-9:15 9:15-10:00		CompIS 342.208 Logic, 2VL	CompIS 365.064 Num. & Symb. Methods 2, 2KV CompIS 353.005 engl Systemnahe Programmierung, 2PR	CompIS 326.011 Algorithmen und Datenstrukturen,, 2KV	
10:15-11:00 11:00-11:45		CompIS 366.554 Statistik 2, 2KV CompIS 342.209 Logic, 1UE	CompIS 376.022 Basics in Chemistry Bioinf., 1KV CompIS 376.022 Basics in Chemistry Bioinf., 1KV CompIS 343.324 Software Engineering, 2VO	365.076 Machine Learning: Supervised Techniques, 1UE	CompIS 365.062 Seq. Analysis & Phylogenetics, 2UE
12:00-12:45 12:45-13:30	CompIS 344.014 Artificial Intell., 2VL CompIS 326.015 InSysteme, 2KV			CompIS 353.068 Comp. Forensics and IT Law, 2VL	
13:45-14:30 14:30-15:15	CompIS 340.023 Algorithmen u. Datens. 2, 2VL CompIS 351.001 InSysteme 1, 2VL		CompIS 347.334 Chemie für Physiker II, 2VL CompIS 364.028 Visual Analytics, 2VL CompIS 343.302 Software Engineering, 1UE	CompIS 351.003 or 351.004 Info-systeme 1, 2UE	
15:30-16:15 16:15-17:00	CompIS 365.060 Sequence Analysis and Phylogenetics, 2VL		365.075 Machine Learning: Supervised Techniques, 2VL CompIS 343.303 Software Engineering, 1UE	CompIS 351.002 & 351.005 Info-ssysteme 1, 2UE 365.074 Basic Methods of Data Analysis, 2KV	
17:15-18:00 18:00-18:45	CompIS 320.007 Molekulare Bio. I, 2VL		CompIS 343.309 Software Eng., 1UE		

Outline

Outline

Outline

3 Summarizing Univariate and Bivariate Data

Outline

Outline

Outline

Literature

- Data Analysis: R. Peck, C. Olsen and J. L. Devore; Introduction to Statistics and Data Analysis, 3rd edition, ISBN: 9780495118732, Brooks/Cole, Belmont, USA, 2009.
- Statistical Data Analysis: B. Shahbaba; Biostatistics with R: An Introduction to Statistics Through Biological Data; Springer, series UseR!, ISBN 9781461413011, New York, 2012.
- Statistical Data Analysis: C. T. Ekstrom and H. Sorensen; Introduction to Statistical Data Analysis for the Life Sciences; CRC Press, Taylor & Francis Group, ISBN: 9781439825556, Boca Raton, USA, 2011.
- Linear Models: A. Dobson; An Introduction to Generalized Linear Models, 2nd edition, ISBN: 1-58488-165-8, Series: Texts in Statistical Science, Chapman & Hall / CRC, Boca Raton, London, New York, Washington D.C., 2002.
- Linear Models: A. C. Rencher and G. B. Schaalje; Linear Models in Statistics, 2nd edition, Wiley, Hoboken, New Jersey, USA, 2008.
- Clustering: L. Kaufman and P. J. Rousseeuw; Finding Groups in Data. An Introduction to Cluster Analysis, Wiley, 1990.

Chapter 1

Introduction

Introduction

Data analysis and visualization are essential to most fields in science and engineering

Goal: basic tool chest of methods for pre-processing, analyzing, and visualizing scientific data

Examples but few theory

Introduction

examples are in **R**

it is not necessary to install R on your computer but might be helpful

R:

- free and open source
- large community
- flexible and extensible
- implementations of major machine learning and statistical methods
- graphics for data visualization
- convenient data handling tools
- matrix and vector calculation tools

See manuscript for instructions to install R or go simply to http://cran.r-project.org/

Introduction

Deductive: human deduces the solution from the problem formulation like during programming

Inductive: knowledge about extracted characteristics, regularities, and structures from data is used to solve the problem

Internet, biology, chemistry, physics, medicine currently produce a huge amount of data

→ statistical methods or a machine that learns: both use data
- Statistics tries to explain variability in the data
- Machine learning tries to find structures in the data

This course: tools and basic techniques for analyzing data with statistical and machine learning methods

Chapter 2

Representing Observations

Representing Observations

Observations and measurements of the real world objects are represented as data on a computer

Subsequently these data are analyzed to explain variation and to find structures in the data

Prediction and classification (supervised):
- predict the outcome of future measurements
- predict future events

Characterize and categorize the objects (unsupervised):
- unknown states of the objects
- relations between the objects and to other objects

Representing Observations

Features or characteristics of objects must be extracted from the original data that are obtained from measurements or recordings of the objects.

Feature extraction: generating features from the raw data

→for example, extraction of features from an image (length or width)

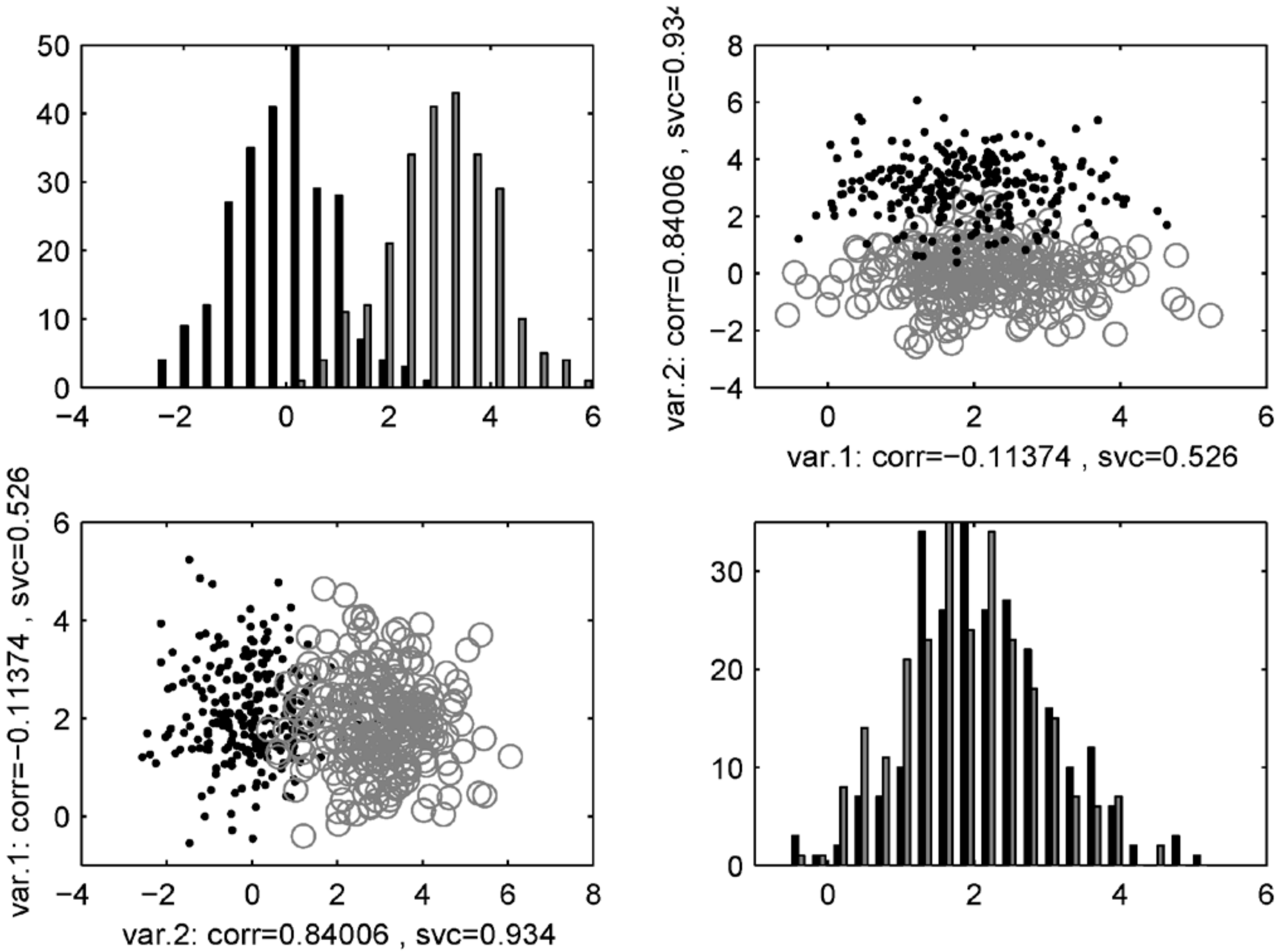
Representing Observations

huge number of features:

- Microarrays: 20,000 genes
- DNA: 1 – 30 million SNPs (sequencing, microarrays)
- Internet: links, web-site users, click-streams

for a specific task many measurements may be irrelevant
e.g. only cancer related genes are of interest for oncolocy

Representing Observations



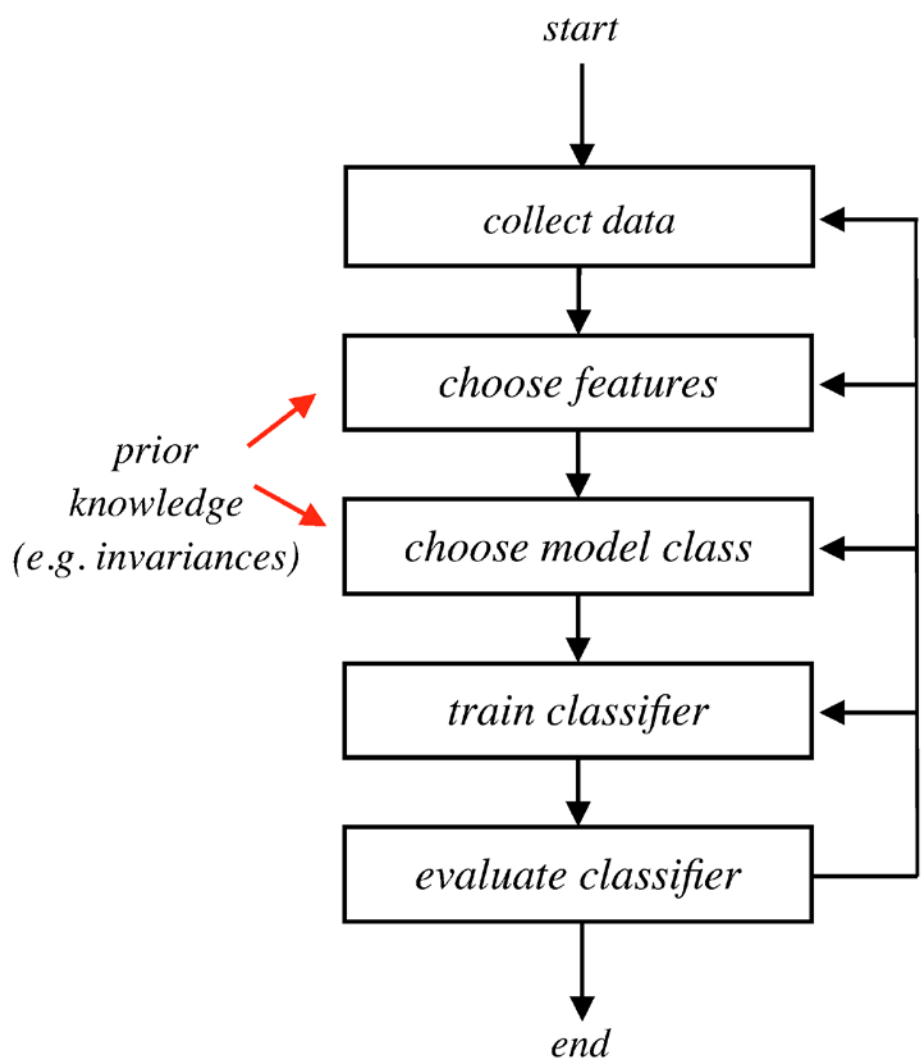
Feature 1 is noise and feature 2 is correlated to the classes.
Between the upper and lower row only the axis are exchanged.

Feature selection: to choose features for a task from a set of features

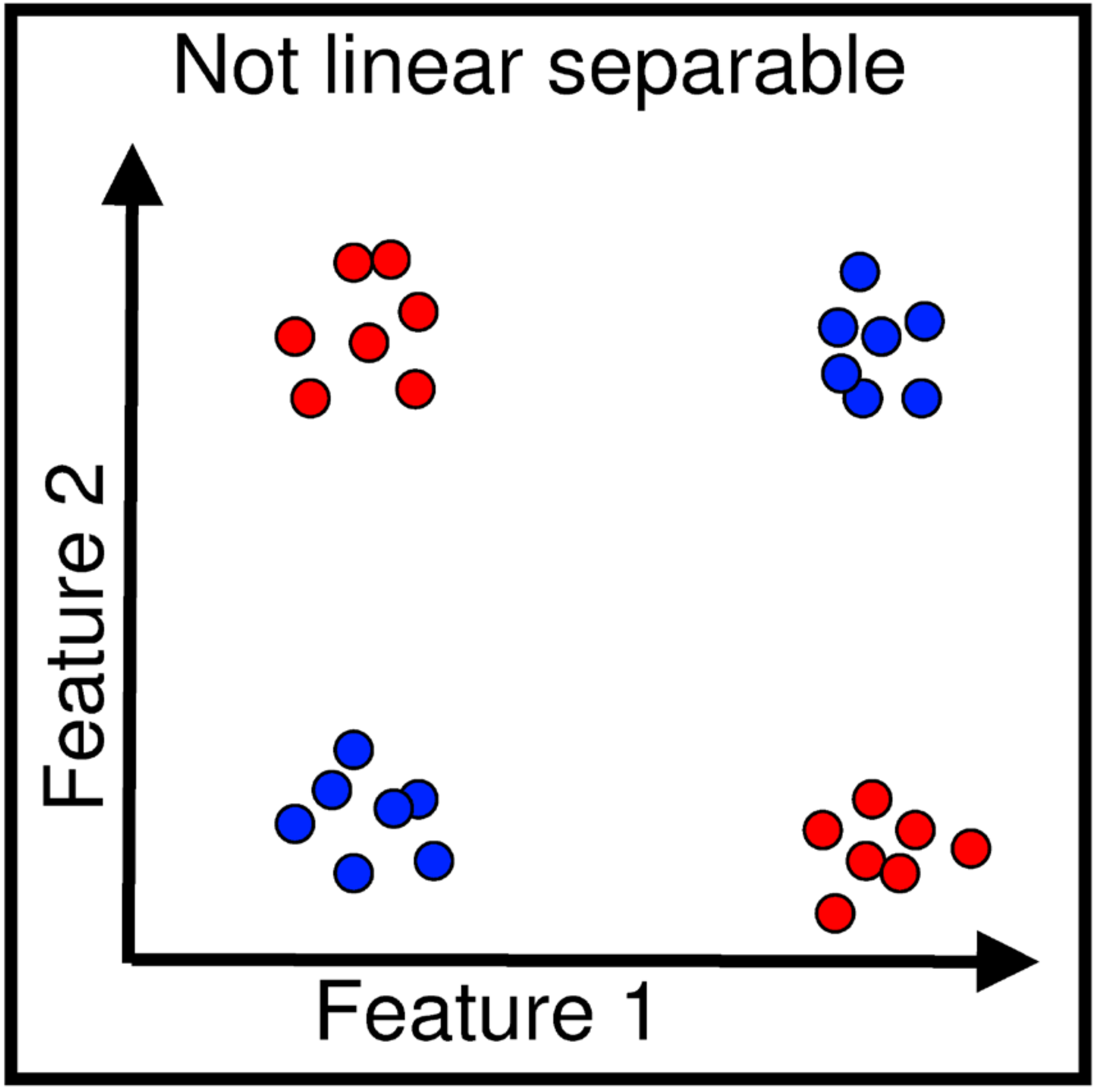
important step to:
- construct appropriate models
- gain insight into real world processes

The first step of data analysis: select the relevant features or chose a model which automatically identifies the relevant features

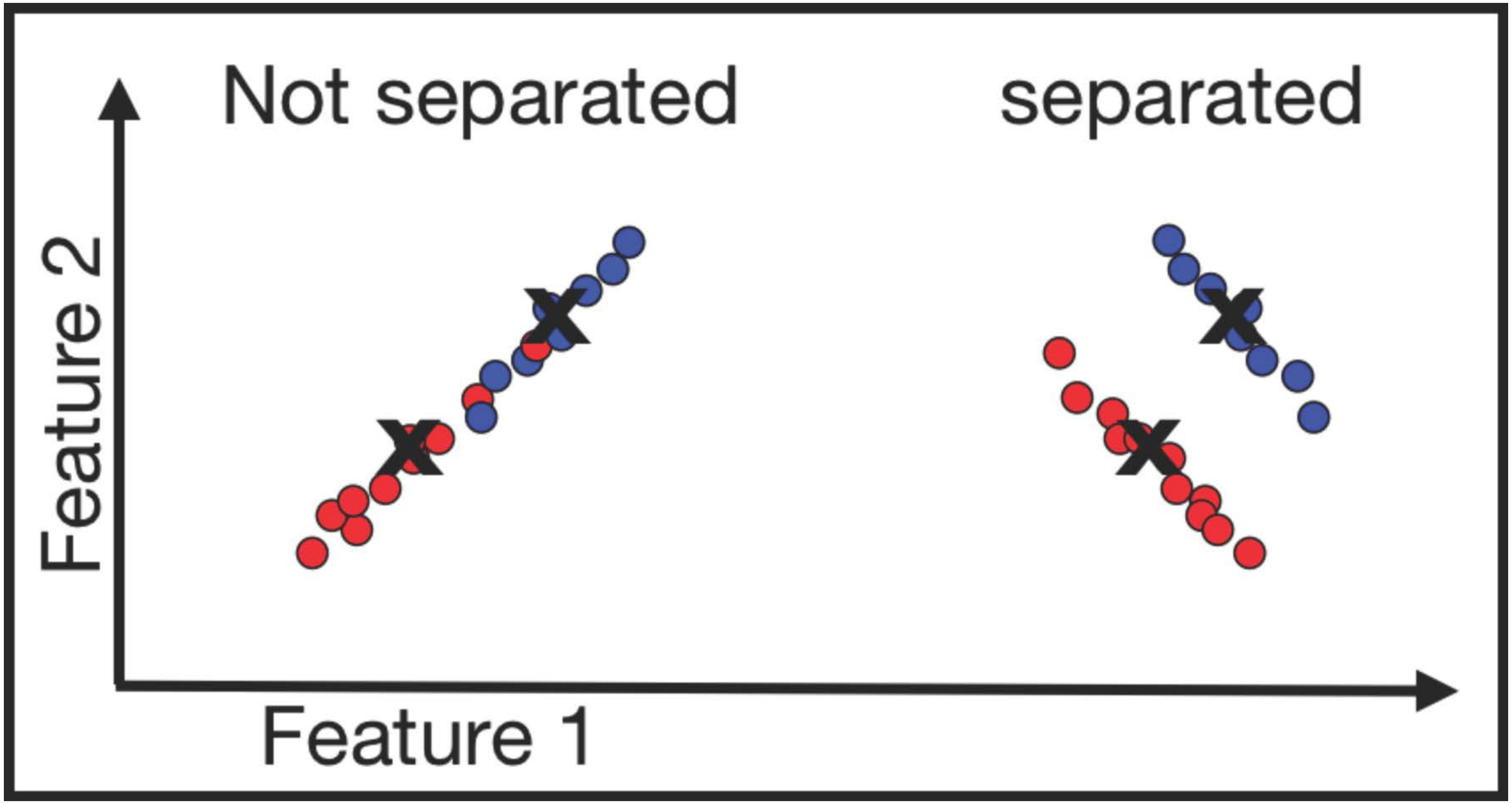
Representing Observations



Representing Observations



Representing Observations



Representing Observations

not correlated with the target: important

large correlation to the target: not important

f_1	f_2	$t = f_1 + f_2$	f_1	f_2	f_3	$t = f_2 + f_3$
-2	3	1	0	-1	0	-1
2	-3	-1	1	1	0	1
-2	1	-1	-1	0	-1	-1
2	-1	1	1	0	1	1

Examples of feature-target correlations.

Left hand side: the target t is $t = f_1 + f_2$, however f_1 is not correlated with t.

Right hand side: $t = f_2 + f_3$, however f_1 has highest correlation coefficient with the target.

Representing Observations

feature construction: create new features from the existing features

- combining correlated features (meta-gene)
- principal component analysis (PCA)
- independent component analysis (ICA)
- kernel methods: feature vector are mapped into new feature space
- non-linear features using prior knowledge: sequence similarity, links between web pages, social networks and user interactions

Representing Observations

We show some typical examples of data sets

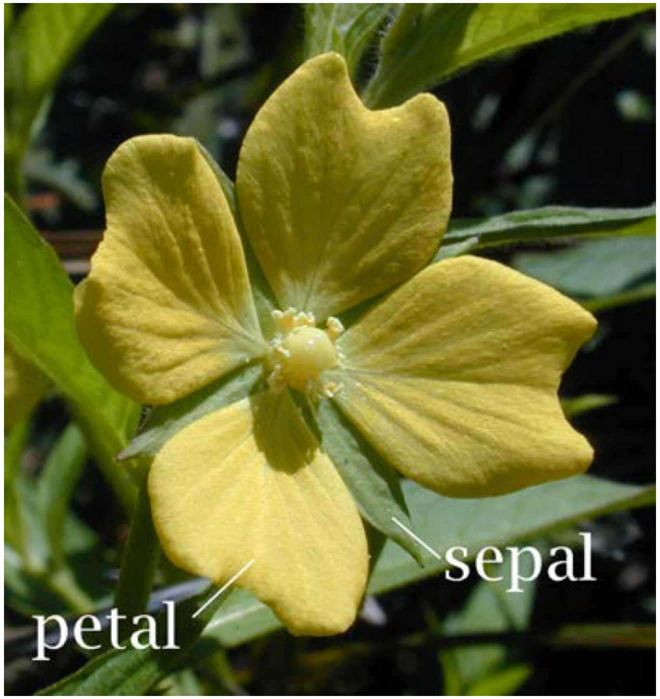
Example 1: Anderson's or Fisher's Iris data set

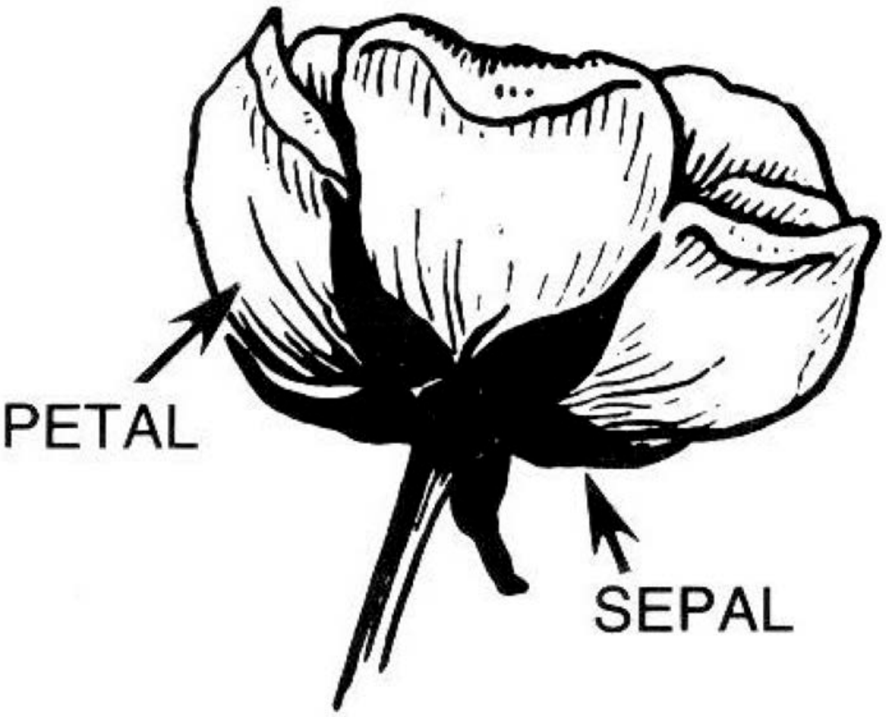
Multivariate data set introduced by Sir Ronald Fisher (1936). Iris is a genus of 260-300 species of flowering plants with showy flowers. The three species of the data set are Iris setosa (Beachhead Iris), Iris versicolor (Larger Blue Flag, Harlequin Blueflag), and Iris virginica (Virginia Iris).

Edgar Anderson collected the data to quantify the morphologic variation of Iris flowers of three related species.

Representing Observations

Representing Observations





Representing Observations

Four features: the length and the width of the sepals and petals (cm)
For each of the three species 50 flowers are measured

No.	Sepal		Petal		Species
	Length	Width	Length	Width	
1	5.1	3.5	1.4	0.2	setosa
2	4.9	3.0	1.4	0.2	setosa
3	4.7	3.2	1.3	0.2	setosa
4	4.6	3.1	1.5	0.2	setosa
5	5.0	3.6	1.4	0.2	setosa
51	7.0	3.2	4.7	1.4	versicolor
52	6.4	3.2	4.5	1.5	versicolor
53	6.9	3.1	4.9	1.5	versicolor
54	5.5	2.3	4.0	1.3	versicolor
55	6.5	2.8	4.6	1.5	versicolor
101	6.3	3.3	6.0	2.5	virginica
102	5.8	2.7	5.1	1.9	virginica
103	7.1	3.0	5.9	2.1	virginica
104	6.3	2.9	5.6	1.8	virginica
105	6.5	3.0	5.8	2.2	virginica

Table 1: Part of the iris data set with features sepal length, sepal width, petal length, and petal width.

Representing Observations

Example 2: Multiple Tissues Microarray Data Set

- Affymetrix microarray data from the Broad Institute
- gene expression profiles from human and mouse samples across a diverse set of tissues, organs, and cell lines
- normal mammalian transcriptome
- insights into gene function, transcriptional regulation, disease
- 102 human and mouse samples
- 5,565 genes selected
- data: 102 x 5,565 matrix of expression values (gene activation)

Four distinct tissue types:

- breast (Br)
- prostate (Pr)
- lung (Lu)
- colon (Co)

Example 3: Breast Cancer Microarray Data Set

microarray data from the Broad Institute:
97 samples for which 1213 gene expression values are

3 subclasses were identified and verified

Example 4: Diffuse Large-B-Cell Lymphoma

Another microarray data set from the Broad Institute:
gene expression profile of diffuse large-B-cell lymphoma (DLBCL)
→ predict the survival after chemotherapy
Data: 180 samples with 661 preselected genes

Three subclasses identified and verified:
- OxPhos: oxidative phosphorylation
- BCR: B-cell response
- HR: host response

Representing Observations

Example 5: US Arrests

arrests per 100,000 residents, for assault, murder, and rape in each of the 50 US states in 1973 plus percent of the population living in urban areas.
Data: 50 observations, 4 features / variables:

- Murder: Murder arrests (per 100,000)
- Assault: Assault arrests (per 100,000)
- UrbanPop: Percent urban population
- Rape: Rape arrests (per 100,000)

Example 6: EU Stock Markets

Time series of the daily closing prices of major European stock indices: Germany DAX (Ibis), Switzerland SMI, France CAC, and UK FTSE. Sampled in business time.

Data: 1860 observations and 4 variables (4 stock indices)

Representing Observations

Example 7: Lung Related Deaths

Time series giving the monthly deaths from lung related diseases bronchitis, emphysema and asthma in the UK during 1974-1979.

Example 8: Sunspots

Monthly mean relative sunspot numbers from 1749 to 1983. During each month the number of sunspots are counted.

Example 9: Revenue Time Series

Freeny's data on quarterly revenue and explanatory variables.
39 observations on quarterly revenue from 1962 to 1971 with explanatory variables:

- price index
- income level
- market potential

Representing Observations

Example 10: Case-Control Study of Infertility

matched case-control study of infertility after spontaneous and induced abortion.

Variables:

- education: 0 = 0-5 years; 1 = 6-11 years; 2 = 12+ years
- age: age in years of case
- parity count
- number of prior induced abortions: 0 = 0; 1 = 1; 2 = 2 or more
- case status: 1 = case; 0 = control
- prior spontaneous abortions: 0 = 0; 1 = 1; 2 = 2 or more
- stratum

Chapter 3

Summarizing Univariate and Bivariate Data

Summarizing Univariate and Bivariate Data

focus on the two most simple cases of data:

- univariate data: set of numbers = scalars = observations
- bivariate data: pairs of numbers; observations have two values

Univariate data are obtained single measurements: weight, height, amplitude, temperature, etc.

Instead of reporting all data points: report summarized data

numerical values:

- data location (the center)
- data variability

Summarizing Univariate and Bivariate Data

univariate data set: $\boldsymbol{x} = \{x_1, x_2, \ldots, x_n\}$

All possible values X with $\Pr(x)$ for the probability of $x \in X$

mean or **expected value**: $$\mu = \sum_{x \in X} x \Pr(x)$$

continuous distributions: $$\mu = \int_X x \Pr(x)\, dx$$

sample mean, **empirical mean**, or **arithmetic mean of samples**

$$\boldsymbol{x} = \{x_1, x_2, \ldots, x_n\} \qquad \bar{x} = \frac{1}{n} \sum_{i=1}^{n} x_i$$

The sample mean approximates the mean.

arithmetic mean ≥ geometric mean ≥ harmonic mean
(average) (log average) (average inverse)

median: separates the higher half of a data from the lower half
continuous case: value, where the probability mass is 0.5

sample median: middle sample / mean of the two middle samples

median m is a robust center as it is not affected by outliers

$$\Pr(X \leq m) \geq \frac{1}{2} \quad \text{and} \quad \Pr(X \geq m) \geq \frac{1}{2}$$

$$\int_{(-\infty,m]} \mathrm{d}F(x) \geq \frac{1}{2} \quad \text{and} \quad \int_{[m,\infty)} \mathrm{d}F(x) \geq \frac{1}{2}$$

unimodal distributions: $\frac{|m - \bar{x}|}{\sigma} \leq (3/5)^{1/2} \approx 0.7746$

distributions with finite variance:

$$\begin{aligned} |\mu - m| = |\mathrm{E}(X - m)| &\leq \mathrm{E}(|X - m|) \\ &\leq \mathrm{E}(|X - \mu|) \\ &\leq \sqrt{\mathrm{E}((X - \mu)^2)} = \sigma \end{aligned}$$

$m = \arg\min_a \mathrm{E}(|X - a|)$

Jensen's inequality

mode: sample that appears most often; most typical sample

Discrete probability distribution $\Pr(x)$ or continuous density $f(x)$:

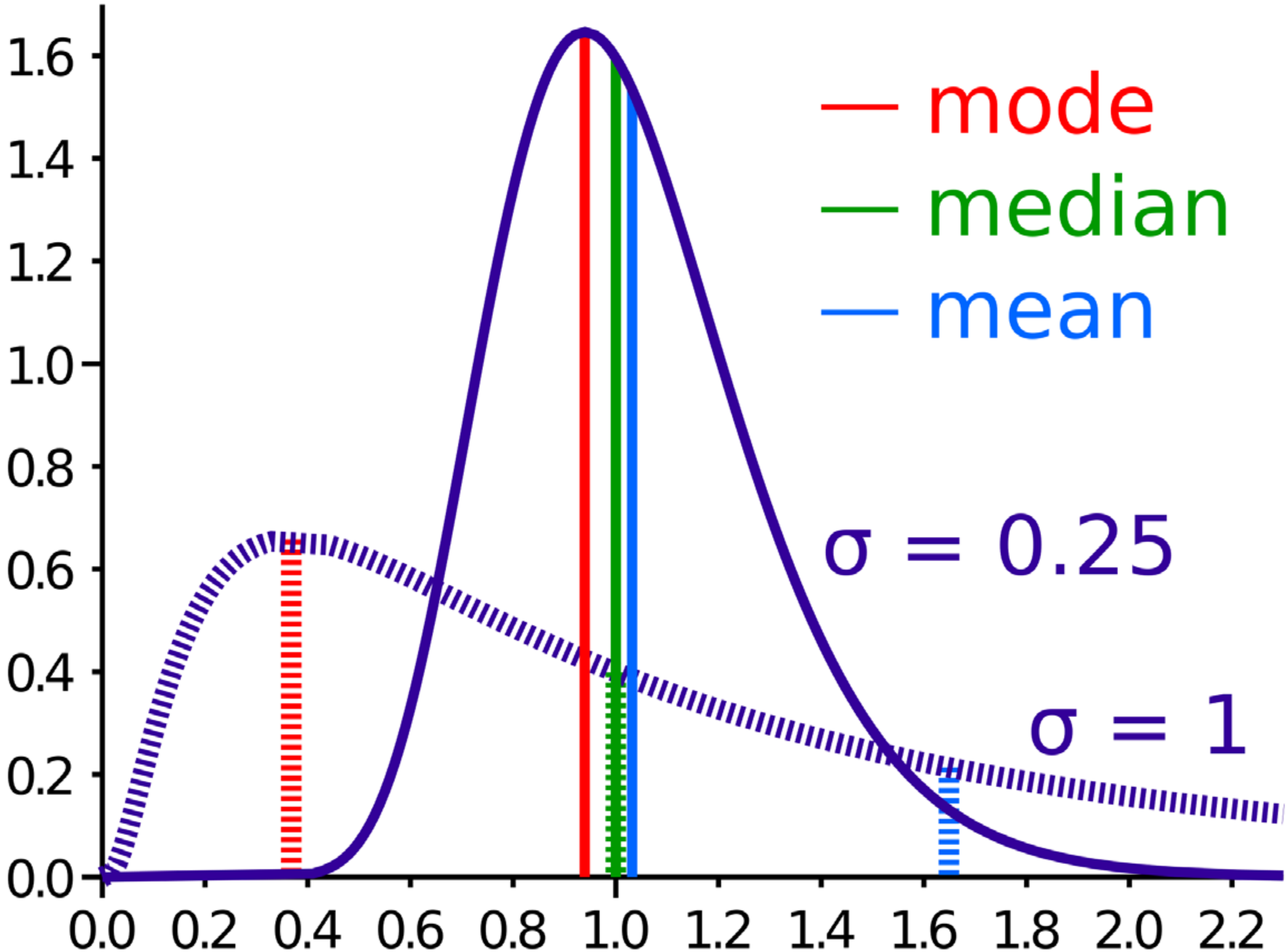
$$\text{mode} \;=\; \arg\max_x \Pr(x) \quad \text{or} \quad \arg\max_x f(x)$$

Inequality: $\frac{|m \;-\; \text{mode}|}{\sigma} \;\leq\; 3^{1/2} \;\approx\; 1.732$

Type	Description	Example	Result
Arithmetic mean	Sum of values of a data set divided by number of values: $\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$	(1+2+2+3+4+7+9) / 7	4
Median	Middle value separating the greater and lesser halves of a data set	1, 2, 2, 3, 4, 7, 9	3
Mode	Most frequent value in a data set	1, 2, 2, 3, 4, 7, 9	2

Table 1: Overview of mean, median, and mode.

Summarizing Univariate and Bivariate Data



Summarizing Univariate and Bivariate Data

- the mean minimizes the average squared deviation: the L^2 norm
- the median minimizes average absolute deviation: the L^1 norm
- the mid-range (0.5 times the range – see later) minimizes the maximum absolute deviation: the L^∞ norm

for symmetric distributions the mean is equal to the median

Gaussian distribution: mean and median should be estimated by the empirical mean

Laplace distribution: mean and median should be estimated by the empirical median

Next feature of the data: spread of the data around the center

range: largest observation minus smallest observation

$$\text{range} = \max \boldsymbol{x} - \min \boldsymbol{x}$$

deviations from the sample mean: $(x_1 - \bar{x}), (x_2 - \bar{x}), \ldots, (x_n - \bar{x})$

The average deviation is zero: $\sum_{i=1}^{n} (x_i - \bar{x}) = \sum_{i=1}^{n} x_i - n\,\bar{x} = n\,\bar{x} - n\,\bar{x} = 0$

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

The data contain (n - 1) pieces of information ((n - 1) degrees of freedom or df) on the deviations. One degree of freedom was used up by the empirical mean.

biased sample variance is $\frac{1}{n} \sum_{i=1}^{n} (x_i - \bar{x})^2$

sample standard deviation (sd): $s = \sqrt{s^2}$

variance and the standard deviation indicate the variability of the data

sd is the size of a typical deviation from the mean

population variance:

discrete $$\sigma^2 = \sum_{x \in X} (x - \mu)^2 \, \Pr(x)$$

continuous $$\sigma^2 = \int_X (x - \mu)^2 \, \Pr(x) \, dx$$

population standard deviation: σ

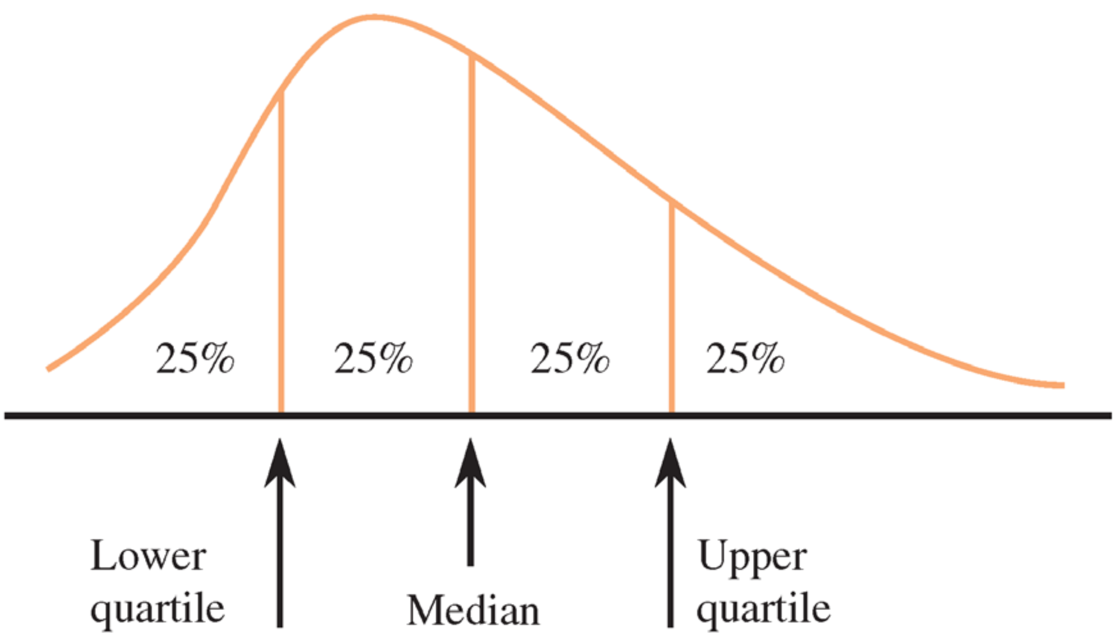
The biased variance has a lower mean squared error than the unbiased variance for Gaussian and Laplace distributions

Summarizing Univariate and Bivariate Data

interquartile range: robust measure of variability

quartiles:

- lower quartile separates the bottom 25% of the data from the upper 75% (the median of the lower half)
- upper quartile separates the top 25% from the bottom 75% (the median of the upper half).
- middle quartile is the median



Summarizing Univariate and Bivariate Data

Iris data set, statistics of sepal length in R:

```
x <- iris[,"Sepal.Length"]
mean(x)
 [1] 5.843333
median(x)
 [1] 5.8
var(x)
 [1] 0.6856935
sd(x)
 [1] 0.8280661
sqrt(var(x))
 [1] 0.8280661
quantile(x)
  0%  25%  50%  75% 100%
 4.3  5.1  5.8  6.4  7.9
summary(x)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  4.300   5.100   5.800   5.843   6.400   7.900
```

The summary for the each iris species shows that the centers of versicolor are larger than those of setosa, and that the centers of virginica are larger than those of versicolor (same for upper quartile):

```
iS <- iris$Species == "setosa"
iV <- iris$Species == "versicolor"
iG <- iris$Species == "virginica"
xS <- x[iS]  ##x <- iris[,"Sepal.Length"]
xV <- x[iV]
xG <- x[iG]
summary(xS)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  4.300   4.800   5.000   5.006   5.200   5.800
summary(xV)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  4.900   5.600   5.900   5.936   6.300   7.000
summary(xG)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  4.900   6.225   6.500   6.588   6.900   7.900
```

Summarizing Univariate and Bivariate Data

The species specific summaries of petal lengths gives a similar figure:

```
x1 <- iris[,"Petal.Length"]
summary(x1)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  1.000   1.600   4.350   3.758   5.100   6.900
x1S <- x1[iS]; x1V <- x1[iV]; x1G <- x1[iG]
summary(x1S)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  1.000   1.400   1.500   1.462   1.575   1.900
summary(x1V)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
   3.00    4.00    4.35    4.26    4.60    5.10
summary(x1G)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  4.500   5.100   5.550   5.552   5.875   6.900
```

- maximum of setosa is below the minimum of virginica and versicolor
- species setosa can be identified by petal length only

z-score or **standardized data**:

$$z = \frac{x - \bar{x}}{s}$$

z-score measures for each observation how many standard deviations it is away from the mean

r-th percentile: value for which r percent of the observations are smaller or equal to this value

- The summary values do not include a reliability value or a variance estimation of the summary itself.
- few observations: high variance → misleading values

Example:
- mean notebook booting time: 10 minutes
- 3 samples: first boot 30 minutes, next two had few seconds
- median: few seconds

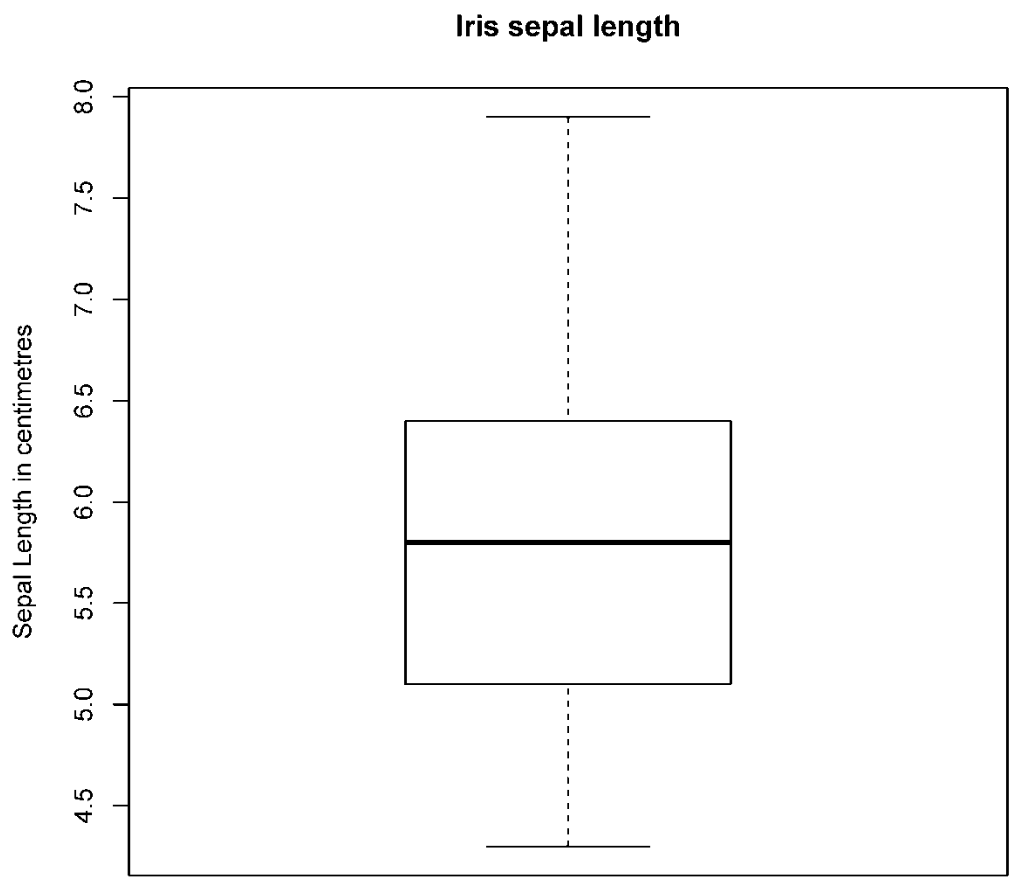
visualizing summary statistics: boxplots

boxplots: box-and-whisker plots of the data with

- median as horizontal bar
- box ranging from the lower to the upper quartile
- whiskers from maximal to minimal value (no outliers!)
- outliers as points; outliers are observations that have larger deviation than `fact` times the interquartile range from the upper or lower quartile. In R default is `fact=1.5`.

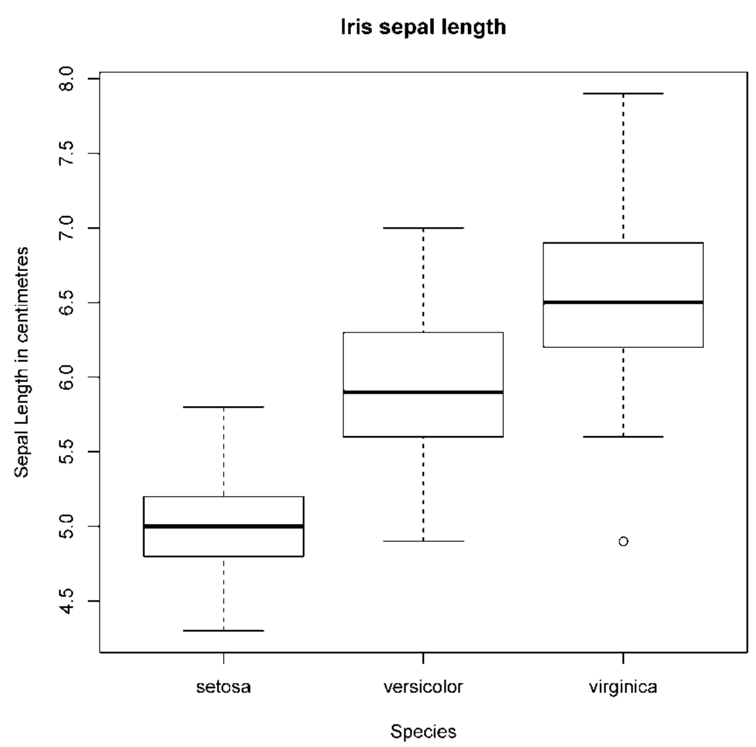
boxplot of the sepal length of the iris data set:

```
boxplot(x,main="Iris sepal length",ylab="Sepal Length in centimetres")
```



boxplots of the sepal length of the iris data set per species

```
boxplot(x ~ unclass(iris$Species),main="Iris sepal length",
+ names=c("setosa","versicolor","virginica"),
+ xlab="Species",ylab="Sepal Length in centimetres")
```

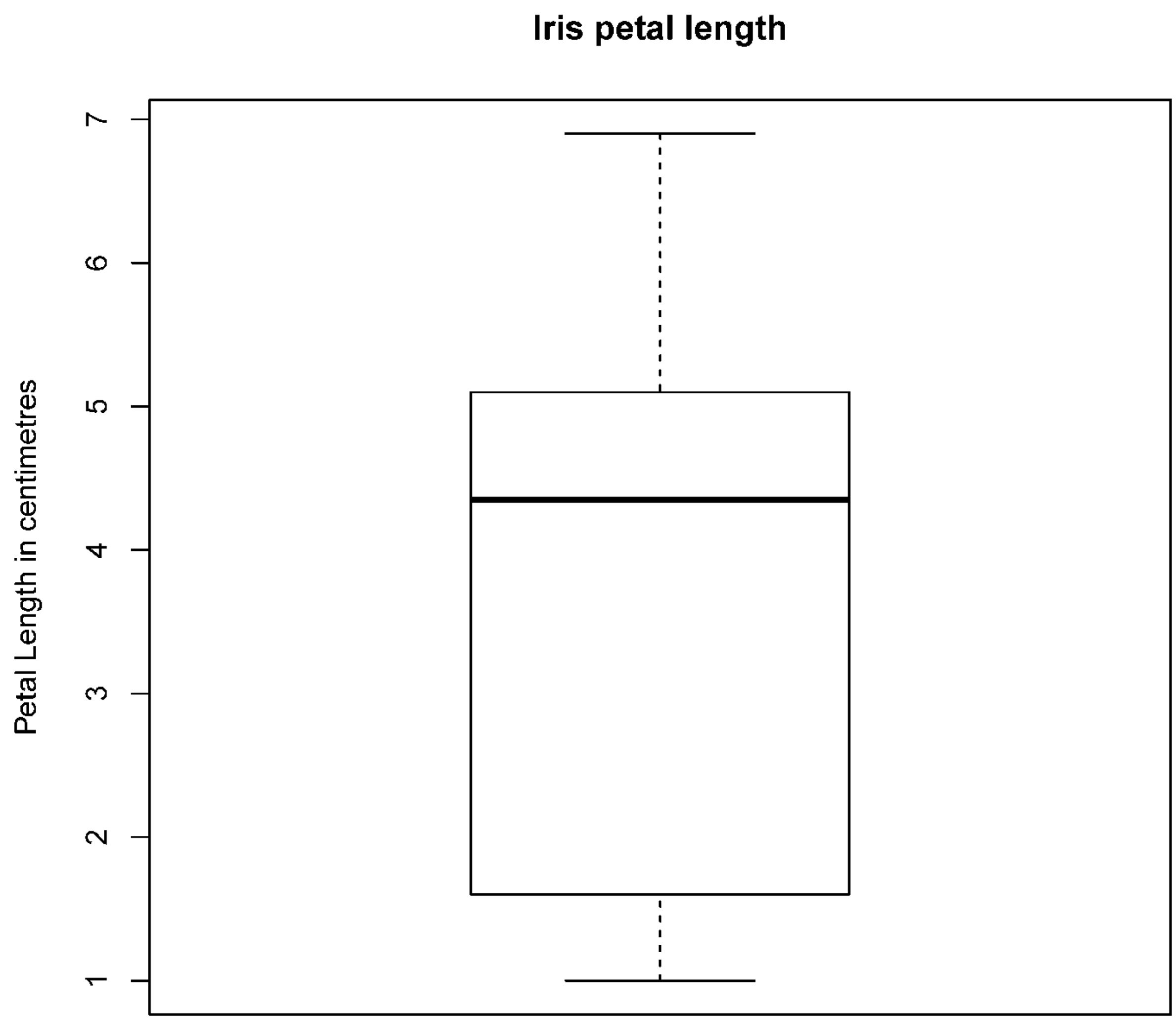


Setosa can be distinguished from the other two species by the sepal length in most cases.

The sepal length of virginica is on average and in most cases larger than the sepal length of versicolor.

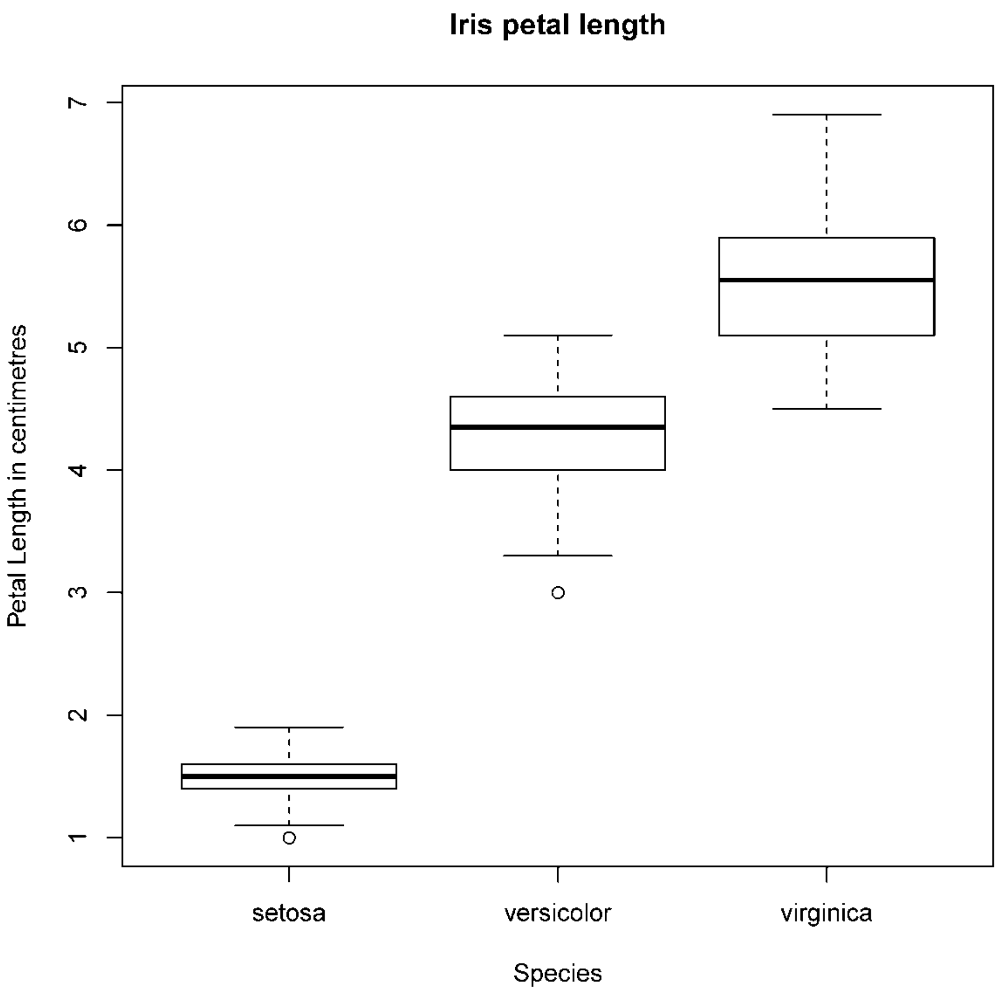
Summarizing Univariate and Bivariate Data

boxplot of the petal length of the iris data set



Summarizing Univariate and Bivariate Data

boxplots of the petal length of the iris data set per species



Setosa can be distinguished from the other two species by the petal length in all cases.

Setosa has clearly shorter sepal lengths than the other two species.

The petal length of virginica allows a better discrimination to versicolor than the sepal length.

Histogram: graphical representation of the data distribution which shows tabulated frequencies as adjacent rectangles which erect over discrete intervals (bins).

- area of the rectangle: equal to the frequency of the observations in the interval
- equidistant bins: heights of the rectangles proportional to frequency of the observations

Histograms help to assess:
- spread or variation
- general shape
- peaks
- low density regions
- outliers

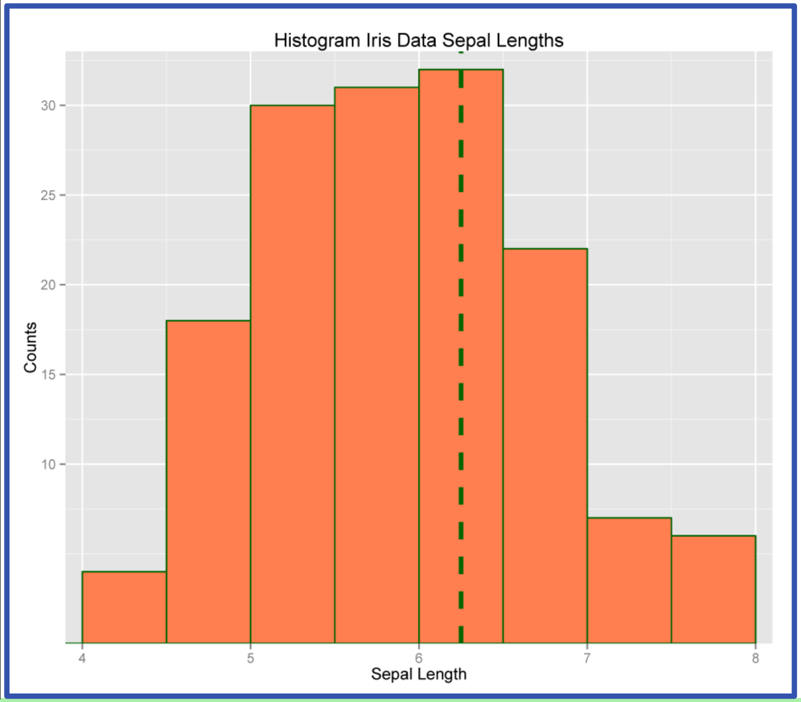
informative overview of the observations
R command `hist()`

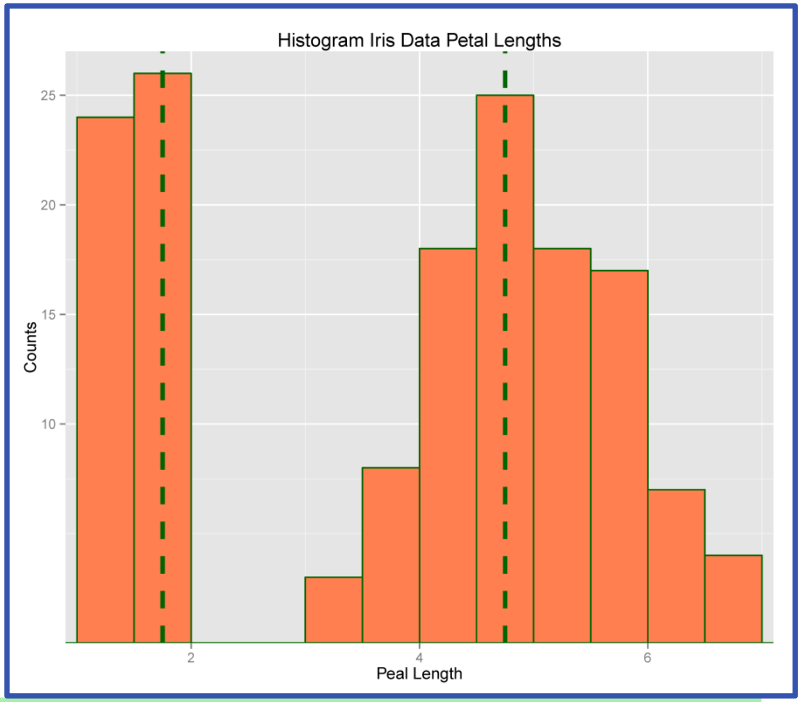
Summarizing Univariate and Bivariate Data

histograms of sepal and petal lengths

- for petal length a gap is visible between short and long petals
- setosa has shorter petals then the other two species

histograms with `ggplot2`





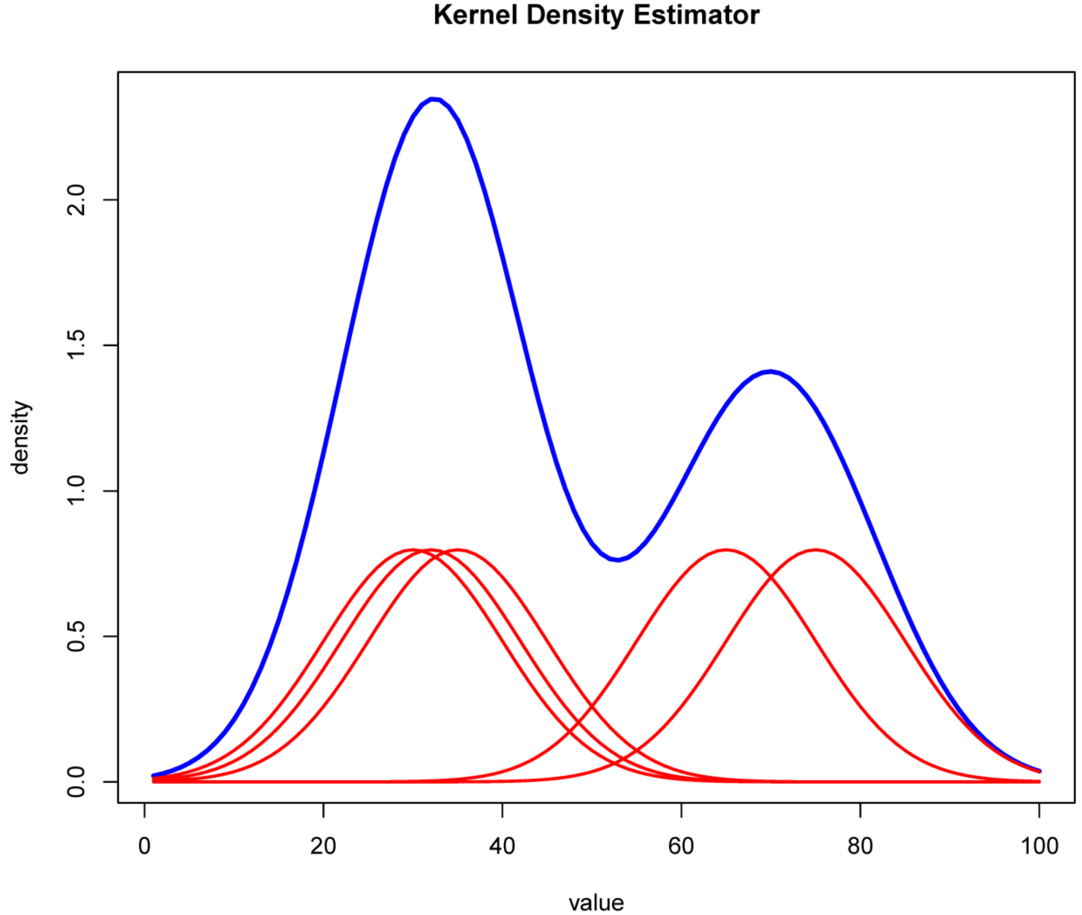
Probability density functions are obtained by kernel density estimation (KDE) which is a non-parametric (except for the bandwidth) method also called Parzen-Rosenblatt window method

kernel density estimator $\hat{f}_h$ has following form:

$$\hat{f}_h(x) = \frac{1}{n} \sum_{i=1}^{n} K_h(x - x_i) = \frac{1}{nh} \sum_{i=1}^{n} K\left(\frac{x - x_i}{h}\right)$$

where $K(.)$ is the kernel (symmetric, positive function that integrates to one) and $h > 0$ is the bandwidth.

kernel density estimator: blue density is approximated by the average of the red kernel densities with locations: 30, 32, 35, 65, 75 and bandwidth is h=10.



The most tricky part of KDE is the bandwidth selection:

- too small: many peaks and wiggly (overfitting)
- too large: peaks vanish and no details (underfitting)

For Gaussian kernels rule-of-thumb (Silverman's rule):

$$h = \left(\frac{4\hat{\sigma}^5}{3n}\right)^{\frac{1}{5}} \approx 1.06\ \hat{\sigma}\ n^{-1/5}$$

where $\hat{\sigma}$ is the standard deviation of the observations.

The closer the true density to a Gaussian,
the better the estimation.

Summarizing Univariate and Bivariate Data

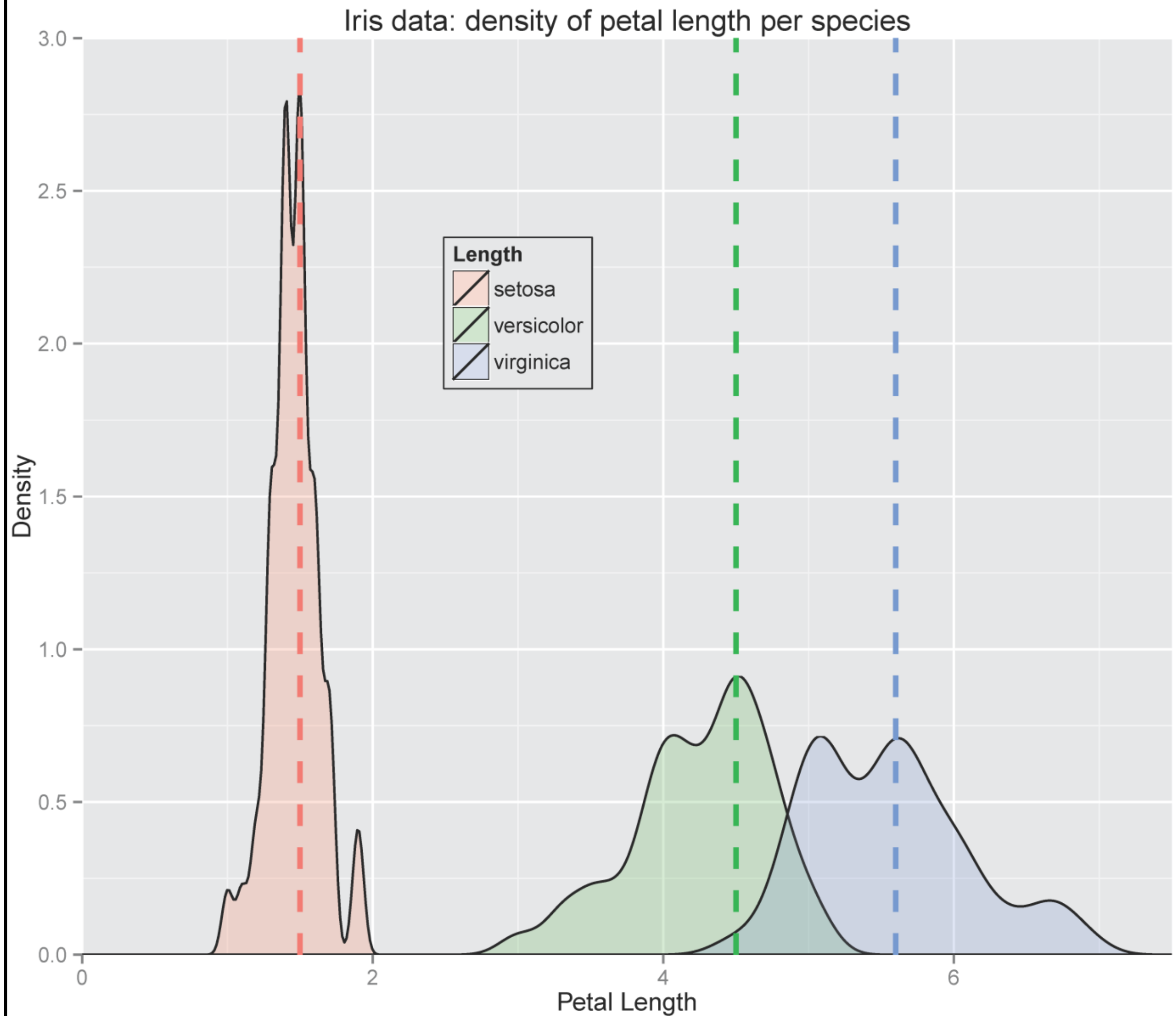
iris data set: densities of sepal lengths per species

Species differ in their sepal length: peaks and location.

Setosa has the least overlap with the other species.

Versicolor and virginica have a considerable overlap of density mass even if their peaks are clearly separated.

iris data set: densities of petal lengths per species

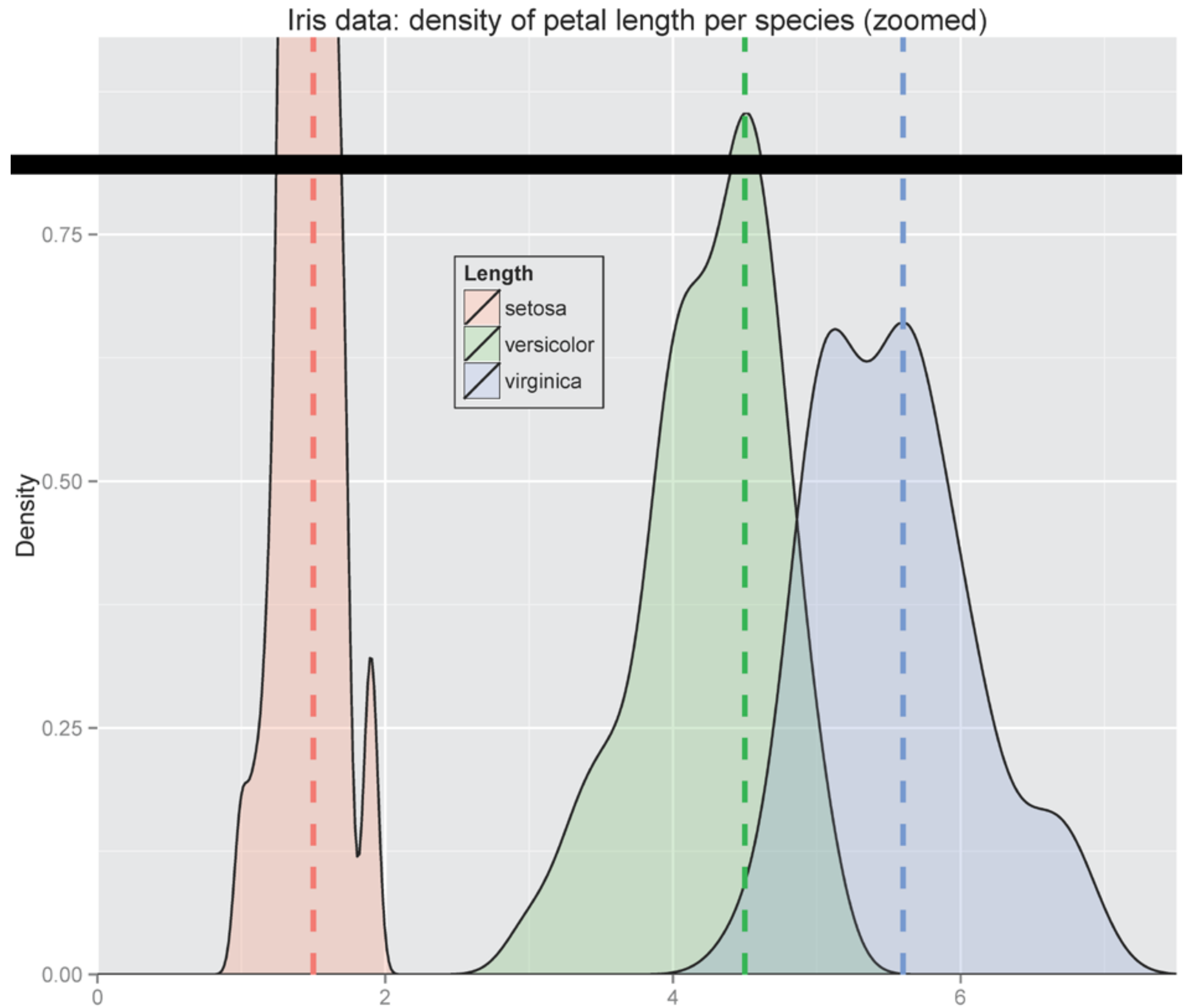


Setosa has no overlap with the other species and the density is very narrow (small variance).

Versicolor and virginica have less overlap than with sepal length and can be separated quite well.

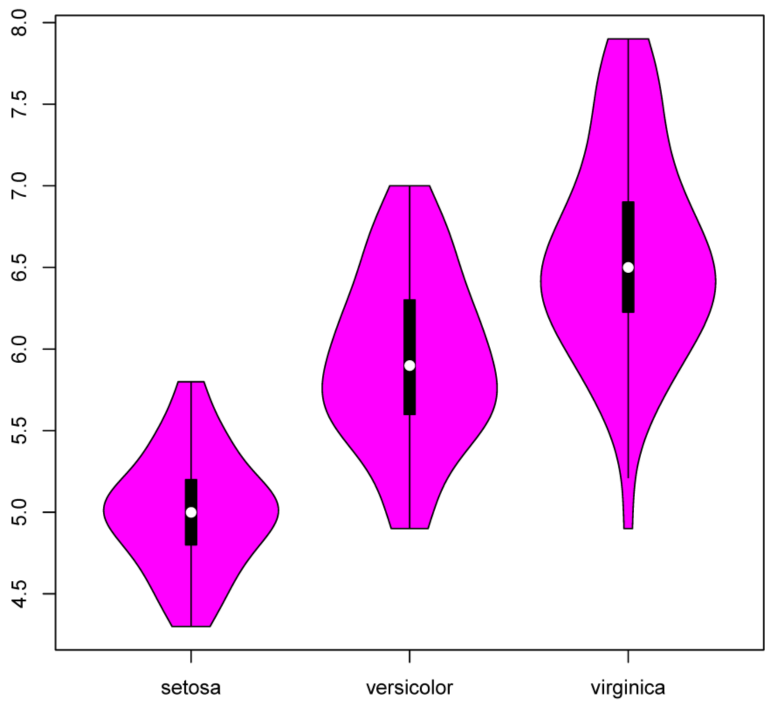
Summarizing Univariate and Bivariate Data

iris data set: zoomed densities of petal lengths per species

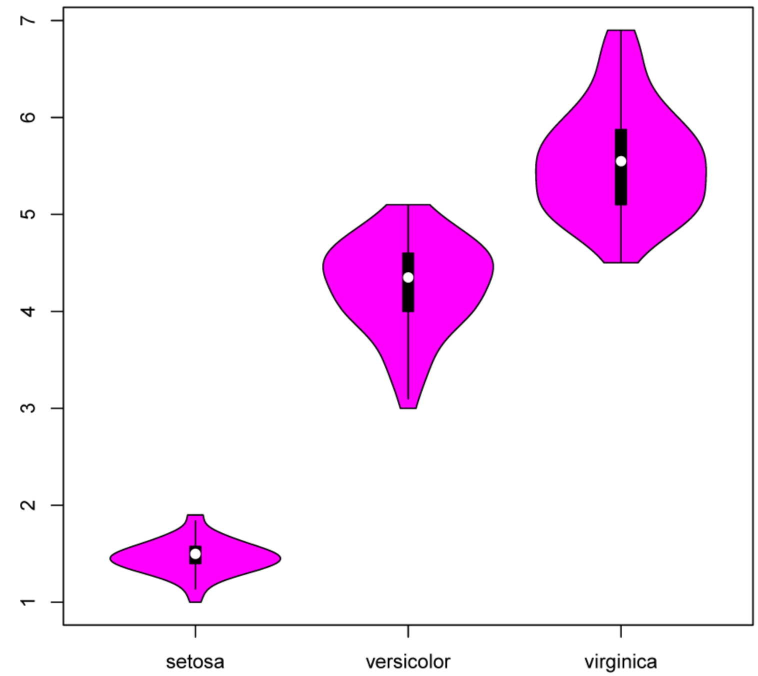


violin plot: combination of boxplot and density estimation
a rotated kernel density at each side of boxplot

iris data sepal length



iris data petal length




```
library(vioplot)
vioplot(x ~ unclass(iris$Species),main="Iris sepal length",
+ names=c("setosa","versicolor","virginica"),
+ xlab="Species",ylab="Sepal Length in centimetres")
```

Summarizing Univariate and Bivariate Data

bivariate data: two scalar variables, pairs of data points

$$\{(y_1, x_1), (y_2, x_2), \ldots, (y_n, x_n)\}$$

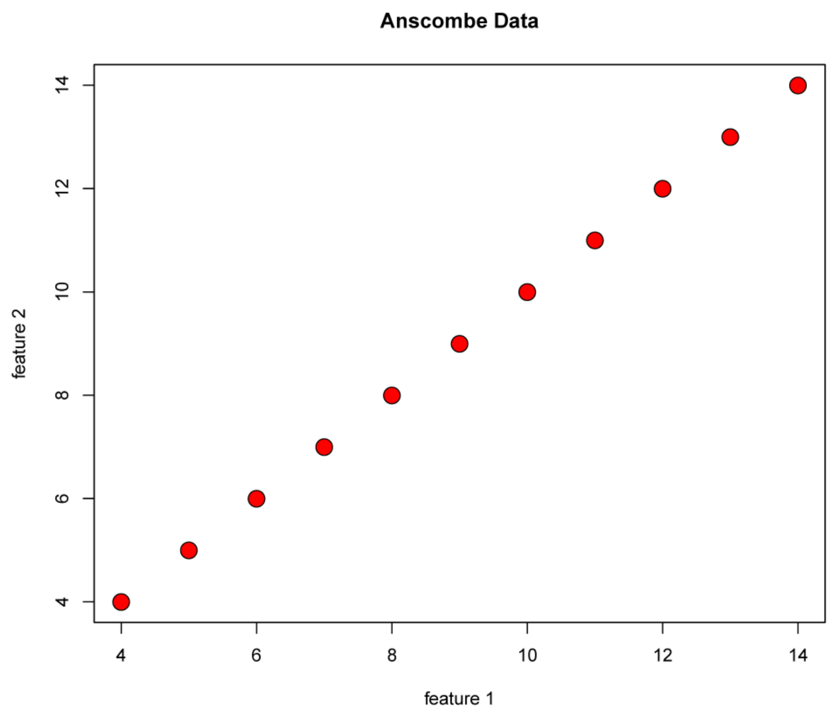
some application: y response or dependent variable
x explanatory variable, independent variable, regressor, feature

response is caused by explanatory variable → causality

statistical or machine learning methods cannot determine causality

scatter plot: shows each observation as a point, where the x-coordinate is the first and the y-coordinate the second variable

```
plot(anscombe[,1:2],main = "Anscombe Data",pch = 21,bg = c("red"),
+ cex=2,xlab="feature 1",ylab="feature 2")
```

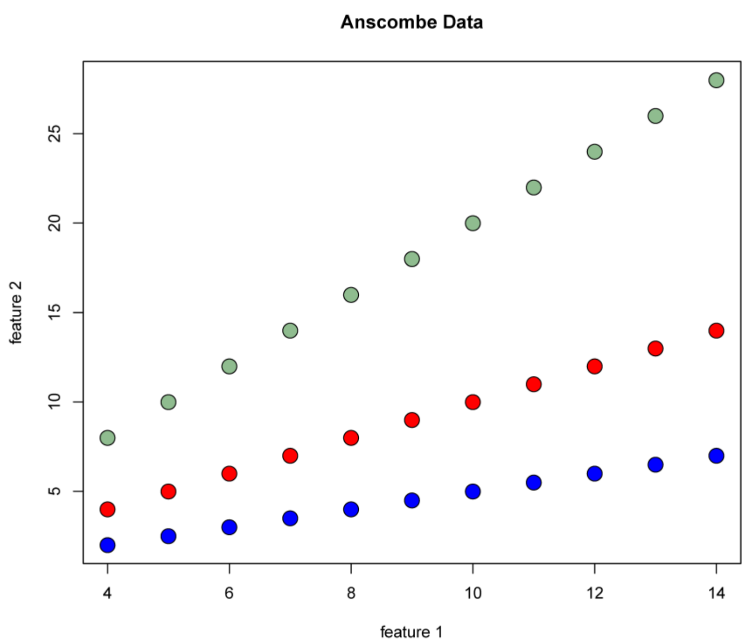


feature 1 and feature 2 are identical: points are on the 45° line

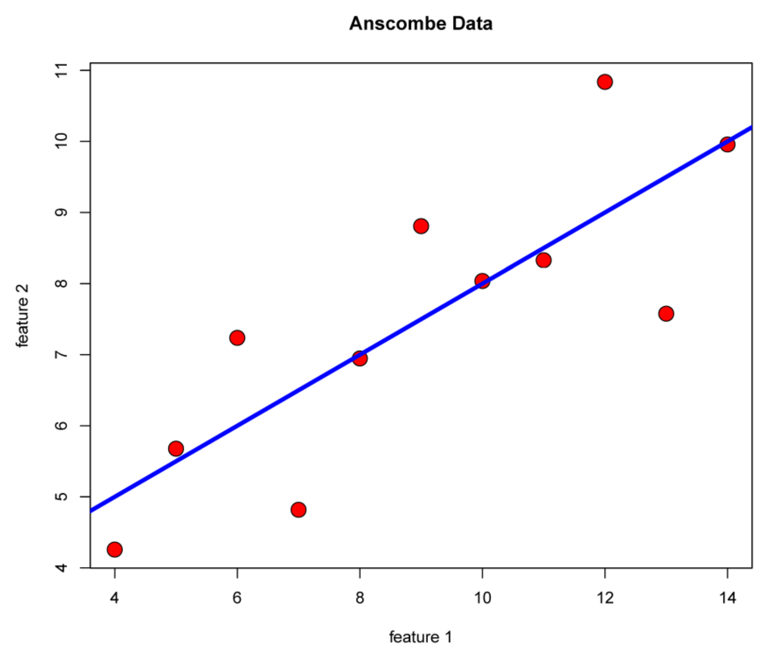
Summarizing Univariate and Bivariate Data

feature 1 and feature 2 are linearly dependent

noise-free



noisy

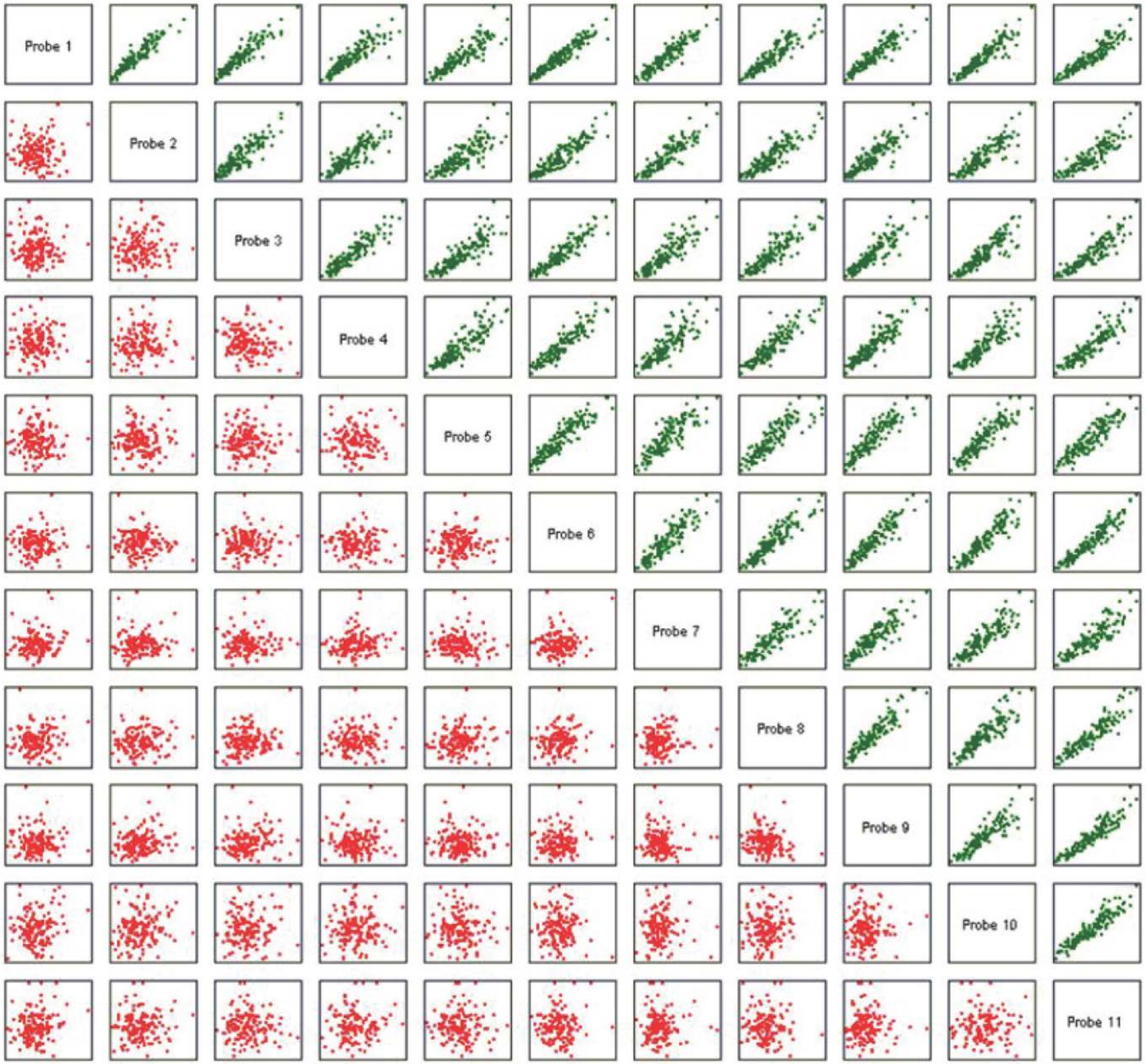


Summarizing Univariate and Bivariate Data

linearly dependent (upper right, green)

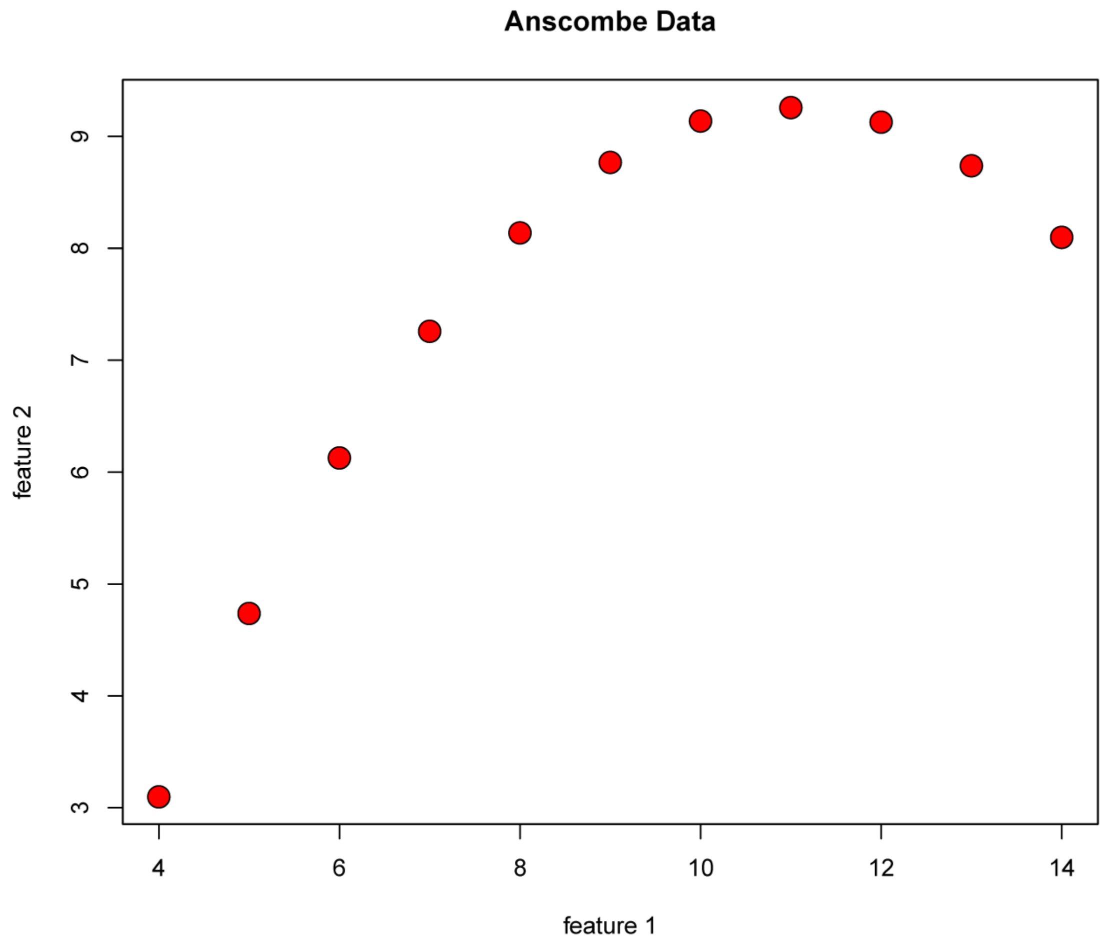
vs.

random (lower left, red)



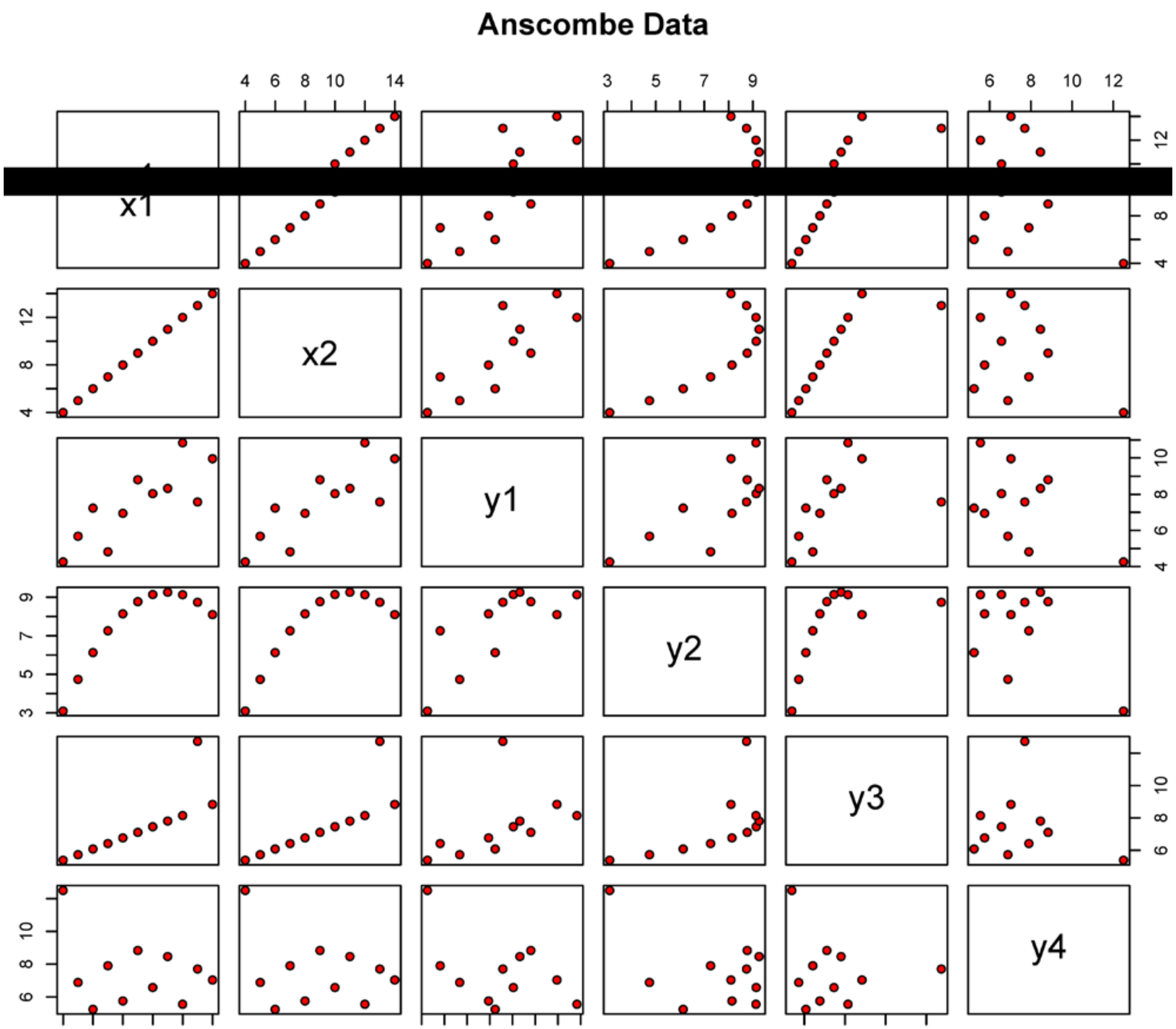
Summarizing Univariate and Bivariate Data

non-linearly dependent features: points are on a one-dimensional curve



matrix of scatter plots:

```
pairs(anscombe[,
c(1,2,5,6,7,8)],
main = "Anscombe
Data", pch = 21,
bg = c("red"))
```



- two variables linearly dependent: points are on a line
- two variables linearly dependent to some degree: points at a line
- the more points are on a line, the higher the linear dependence

Pearson's sample correlation coefficient:
bivariate data $(y_1, x_1), (y_2, x_2), \ldots, (y_n, x_n)$

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

with z-scores

$$r = \frac{1}{n-1} \sum_{i=1}^{n} (z_x)_i (z_y)_i$$

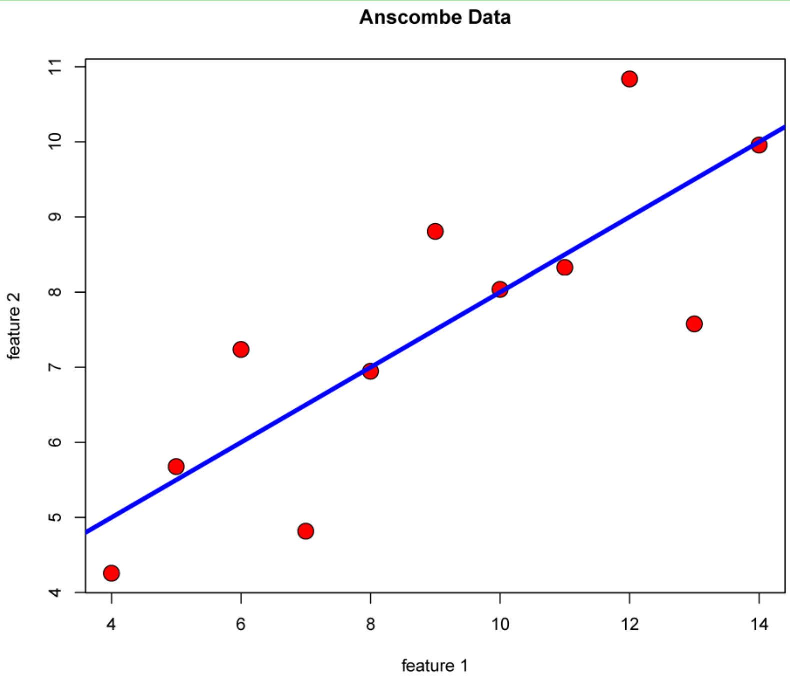
Pearson's population correlation coefficient: ρ
For $x_i = a y_i$ the correlation coefficient is r=1 or r=-1
Since $\bar{x} = a \bar{y}$ and numerator has factor a while denominator $|a|$

r=0.82 obtained by the R code:

```
cor(anscombe[,c(1,5)])
           x1        y1
x1 1.0000000 0.8164205
y1 0.8164205 1.0000000
```

z-scores that is:

```
1/(length(anscombe[, 1])-1)*
crossprod(scale(anscombe[,1]),
scale(anscombe[, 5]))
           [,1]
[1,] 0.8164205
```



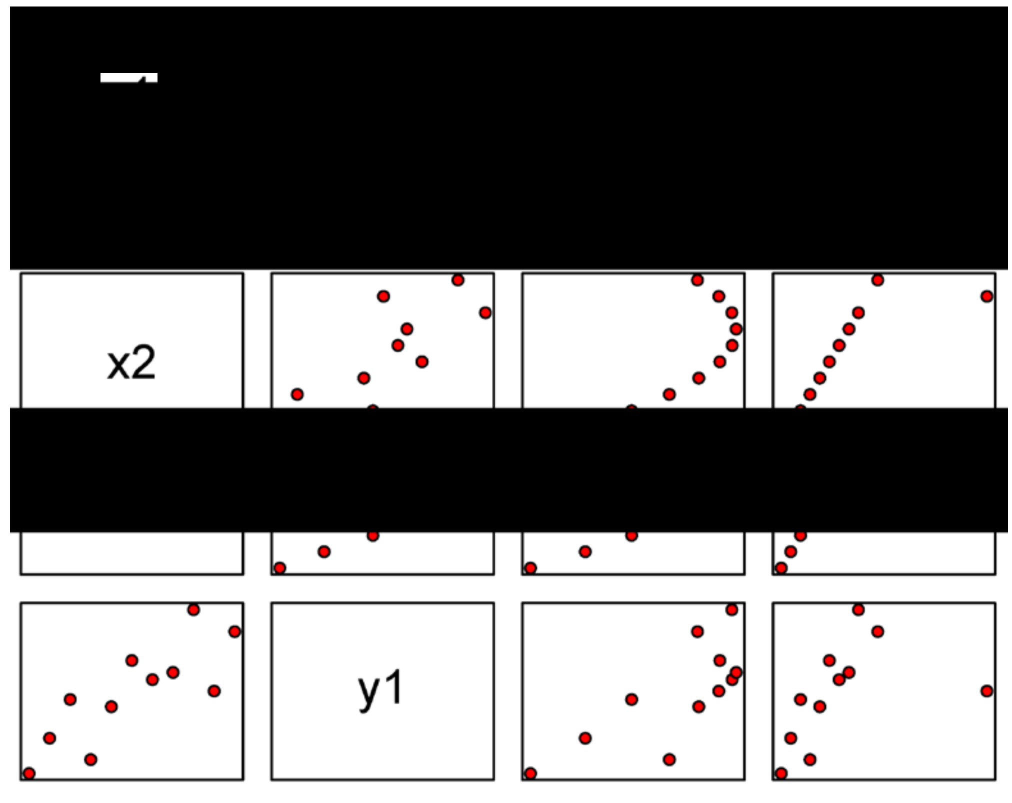
Summarizing Univariate and Bivariate Data


```
cor(anscombe[,c(1,5,6,7)])
          x1        y1        y2        y3
x1 1.0000000 0.8164205 0.8162365 0.8162867
y1 0.8164205 1.0000000 0.7500054 0.4687167
y2 0.8162365 0.7500054 1.0000000 0.5879193
y3 0.8162867 0.4687167 0.5879193 1.0000000
```

Correlation does not imply causality

John Paulos in ABCNews.com:

"Consumption of hot chocolate is correlated with low crime rate, but both are responses to cold weather."



Test for Correlation

Bivariate normal population: test of independence is test for $\rho = 0$

t-test with the test statistic $$t = \frac{r}{\sqrt{\frac{1-r^2}{n-2}}}$$ (r is approx. normal!)

degree of freedom is $\mathrm{df} = n - 2$

Density of Student's t-distribution: $$f(x) = \frac{\Gamma((\mathrm{df}+1)/2)}{\sqrt{\mathrm{df}\pi}\Gamma(\mathrm{df}/2)} \left(1 + \frac{x^2}{\mathrm{df}}\right)^{-(\mathrm{df}+1)/2}$$

In R the p-value can be computed by: `1-pt(t,df=n-2)`

The correlation between x1 and y1 of the Anscombe data set is r=0.8164205 which gives a p-value of:

```
r=0.8164205
t=r/(sqrt((1-r^2)/9))
t
[1] 4.241455
1-pt(t,9)
[1] 0.001084815
```

For y1 and y3 we have r=0.4687167 which gives:

```
r=0.4687167
t=r/(sqrt((1-r^2)/9))
t
[1] 1.591841
1-pt(t,9)
[1] 0.07294216
```

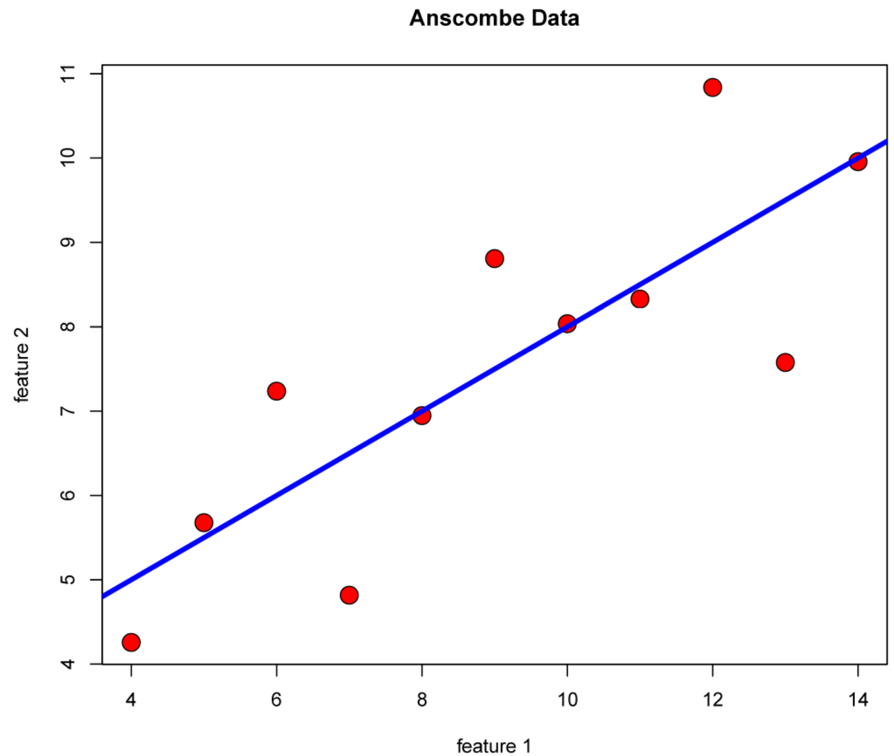
not significant for level 0.05

Linear regression: fit a line to bivariate data

Extract information about the relation of the two variables y and x.

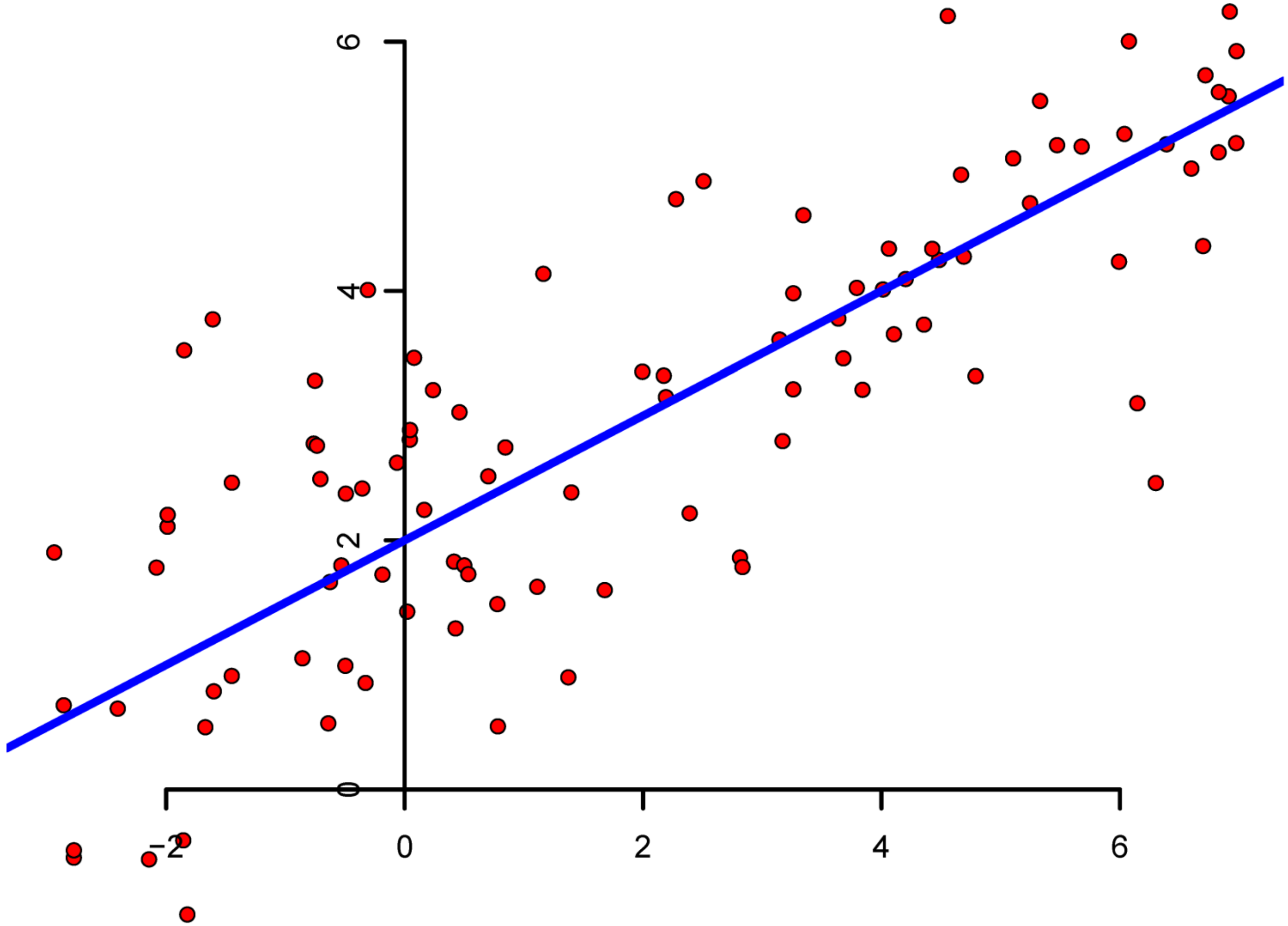
functional relationship: $y \; = \; a \; + \; b \; x$

intercept: a
slope: b



Summarizing Univariate and Bivariate Data

regression curve with a=2 (x=0) and b=0.5 (increase of y relative to x)



BIOINF

goodness of fit criterion or **objective**: quality of fitting
→ find the best fitting line

sum of the squared deviations or **least squares objective**:

$$\sum_{i=1}^{n} \left(y_i \ - \ (\tilde{a} \ + \ \tilde{b} \ x_i) \right)^2$$

$\tilde{a}$ and $\tilde{b}$ are candidate intercept and slope

$\hat{a}$ and $\hat{b}$ that minimize the least squares criterion:

$$\hat{b} \ = \ \frac{\sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n} (x_i - \bar{x})^2} \ = \ \frac{\sum_{i=1}^{n} x_i \ y_i \ - \ \frac{1}{n} \sum_{i=1}^{n} x_i \sum_{j=1}^{n} y_j}{\sum_{i=1}^{n} (x_i^2) \ - \ \frac{1}{n} (\sum_{i=1}^{n} x_i)^2}$$

$$= \ \frac{\overline{xy} \ - \ \bar{x} \ \bar{y}}{\overline{x^2} \ - \ \bar{x}^2} \ = \ \frac{\mathrm{Cov}(x,y)}{\mathrm{Var}(x)} \ = \ r_{xy} \ \frac{s_y}{s_x}$$

$$\hat{a} \ = \ \bar{y} \ - \ \hat{b} \ \bar{x}$$

r_{xy} : correlation coefficient between x and y
s_x : standard deviation of x
s_y : standard deviation of y
$\bar{y}$: mean of y
$\bar{x}$: mean of x

Interchanging x and y: different function

$$y = a + b\,x \;\Rightarrow\; x = \frac{1}{b}(y - a) = -\frac{a}{b} + \frac{1}{b}\,y$$

However this does not hold for the estimates:

$$\hat{b}_y = r_{xy}\; s_y / s_x \qquad \hat{b}_x = r_{xy}\; s_x / s_y$$

$$\hat{b}_y \neq 1/\hat{b}_x \qquad r_{xy} \neq 1/r_{xy}$$

$$y = \hat{a} + \hat{b}\,x \;\Rightarrow\; \frac{y - \bar{y}}{s_y} = r_{xy}\,\frac{x - \bar{x}}{s_x}$$

regression line is reformulated by z-scores:

$z_y = r_{xy}\; z_x$ (no intercept because the data is centered)

BIOINF

error terms normally distributed: $\hat{\varepsilon}_i = y_i - \hat{a} - \hat{b}\, x_i$

$\hat{b}$ is normally distributed with mean b and variance $\sigma^2 / \sum (x_i - \bar{x})^2$, where σ^2 is the variance of the error terms

distribution sum of squared errors: χ^2 with $(n$ - $2)$ degrees of freedom

t-statistic: $t = \dfrac{\hat{b} - b}{s_{\hat{b}}} \sim t_{n-2}$ with $s_{\hat{b}} = \sqrt{\dfrac{\frac{1}{n-2} \sum_{i=1}^{n} \hat{\varepsilon}_i^{\,2}}{\sum_{i=1}^{n} (x_i - \bar{x})^2}}$

has a Student's t-distribution with $(n$ - $2)$ degrees of freedom

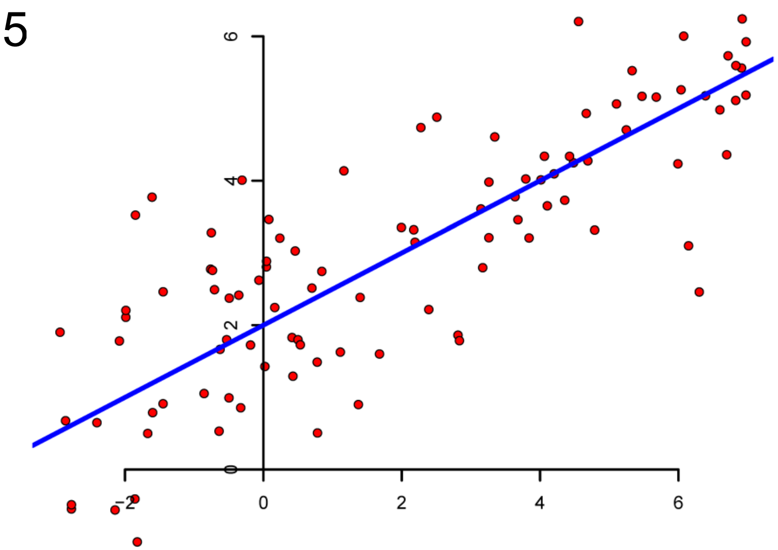
t-statistic allows constructing confidence intervals for a, b, and r_{xy}

R^2 : fraction of variance explained, coefficient of determination

$$R^2 = 1 - \frac{\sum_{i=1}^{n} \hat{\varepsilon}_i^{\,2}}{\sum_{i=1}^{n} (y_i - \bar{y})^2}$$

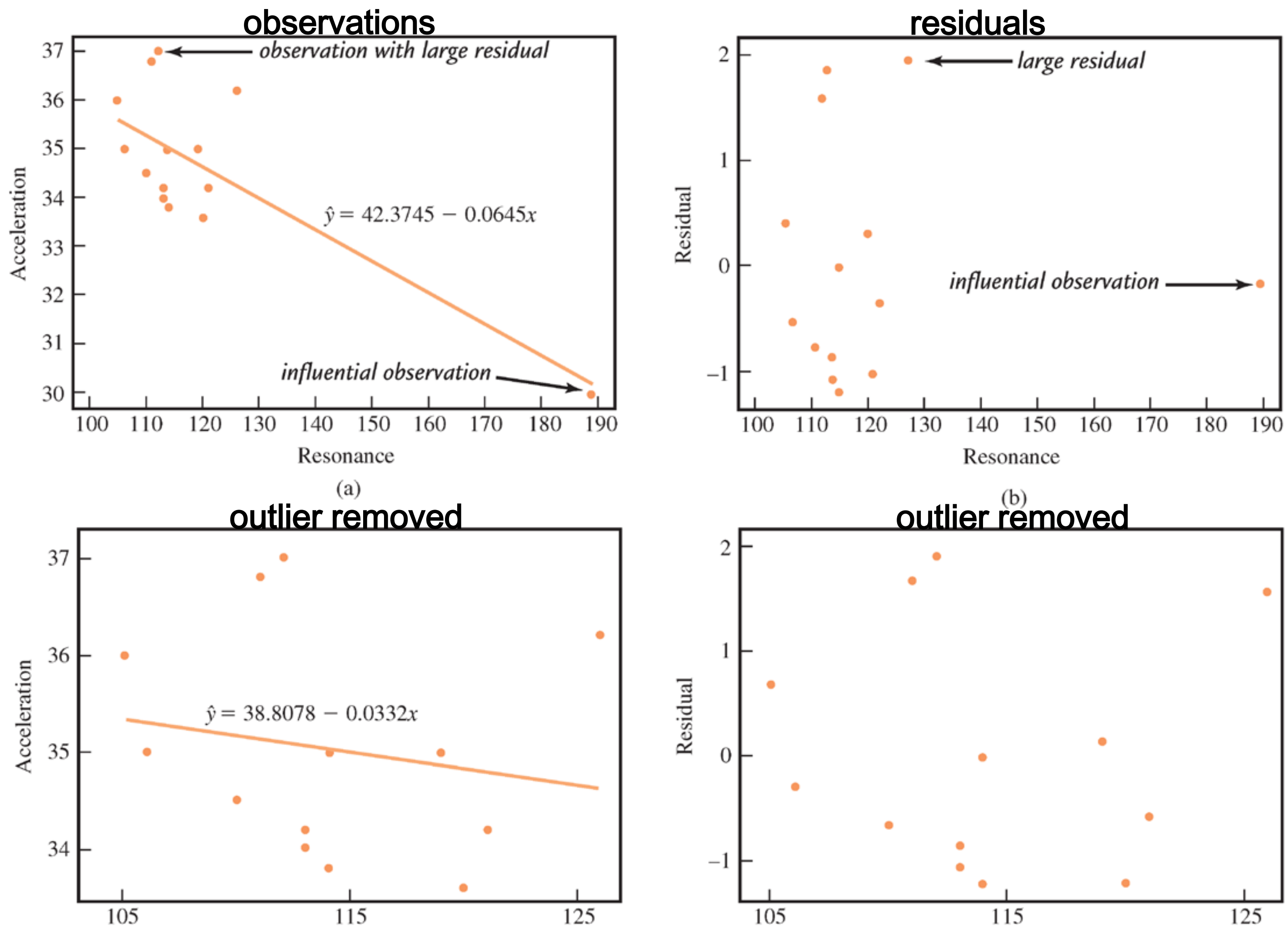
Summarizing Univariate and Bivariate Data

regression curve with a=2 and b=0.5

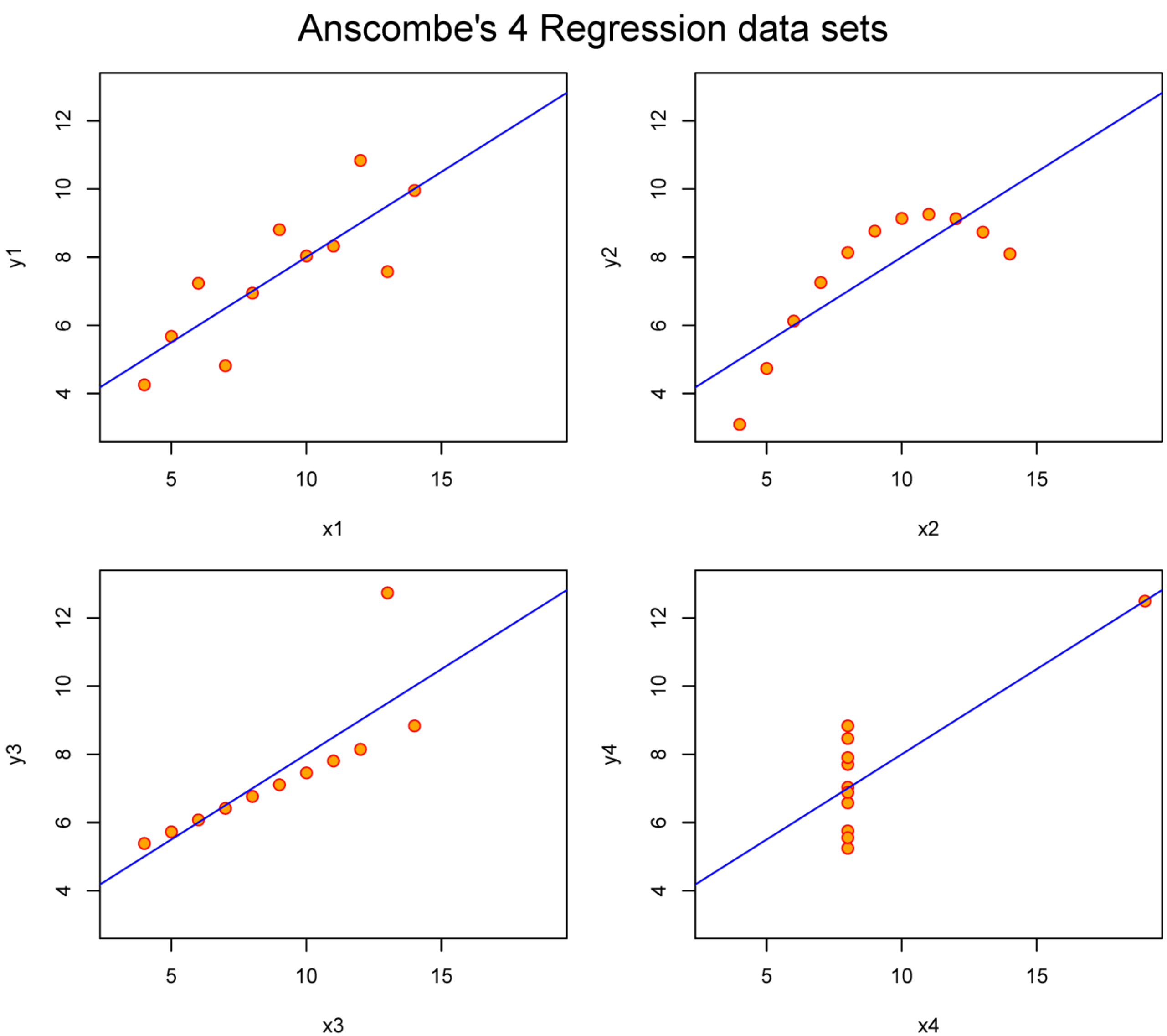



```
res <- lm(y ~ x)
summary(res)
Call:
lm(formula = y ~ x)

Residuals:
     Min       1Q   Median       3Q      Max
-2.55541 -0.64589  0.05834  0.66114  2.42824

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  2.09223    0.12103   17.29   <2e-16 ***
x            0.46427    0.03417   13.59   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1.014 on 98 degrees of freedom
Multiple R-squared:  0.6532,    Adjusted R-squared:  0.6496
F-statistic: 184.6 on 1 and 98 DF,  p-value: < 2.2e-16
```

outliers and influential observations: influential but small error



small error



Anscombe's 4 Regression data sets
y1
x1
y2
x2
y3
x3
y4
x4
4
6
8
10
12
5
10
15

Summarizing Univariate and Bivariate Data


```
$lm1
             Estimate Std. Error  t value    Pr(>|t|)
(Intercept) 3.0000909  1.1247468 2.667348 0.025734051
x1          0.5000909  0.1179055 4.241455 0.002169629

$lm2
            Estimate Std. Error  t value    Pr(>|t|)
(Intercept) 3.000909  1.1253024 2.666758 0.025758941
x2          0.500000  0.1179637 4.238590 0.002178816

$lm3
             Estimate Std. Error  t value    Pr(>|t|)
(Intercept) 3.0024545  1.1244812 2.670080 0.025619109
x3          0.4997273  0.1178777 4.239372 0.002176305

$lm4
             Estimate Std. Error  t value    Pr(>|t|)
(Intercept) 3.0017273  1.1239211 2.670763 0.025590425
x4          0.4999091  0.1178189 4.243028 0.002164602
```

data sets are quite different: same regression line
→ statistical properties do not fully characterize the data