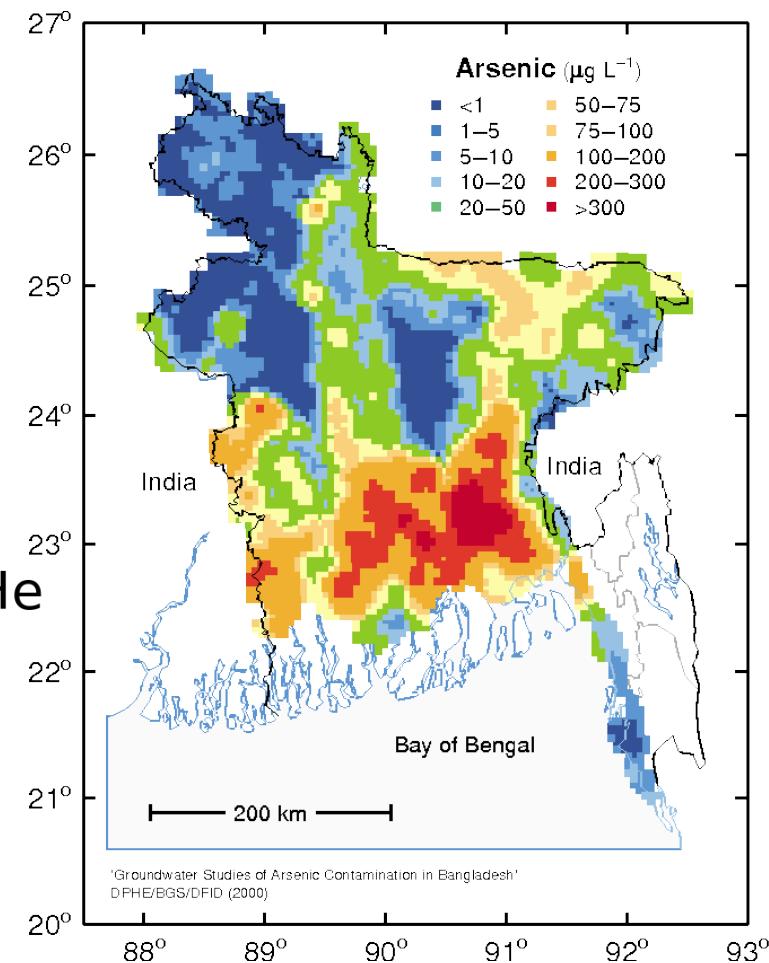
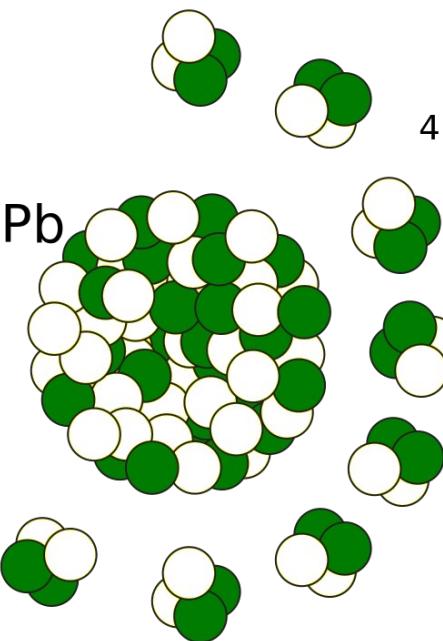
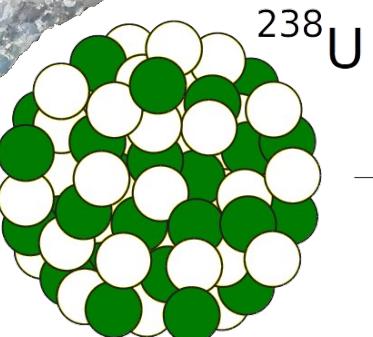


# Statistics for geoscientists

## Introduction

‘statistics’ is:

*The practice or science of collecting and analysing numerical data in large quantities, especially for the purpose of **inferring** proportions in a whole from those in a representative **sample**.*





Mathematical Statistics  
and Data Analysis

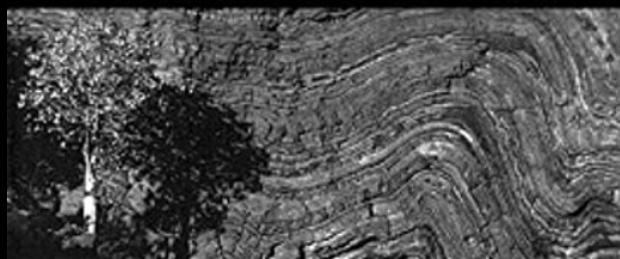
Second Edition

John A. Rice

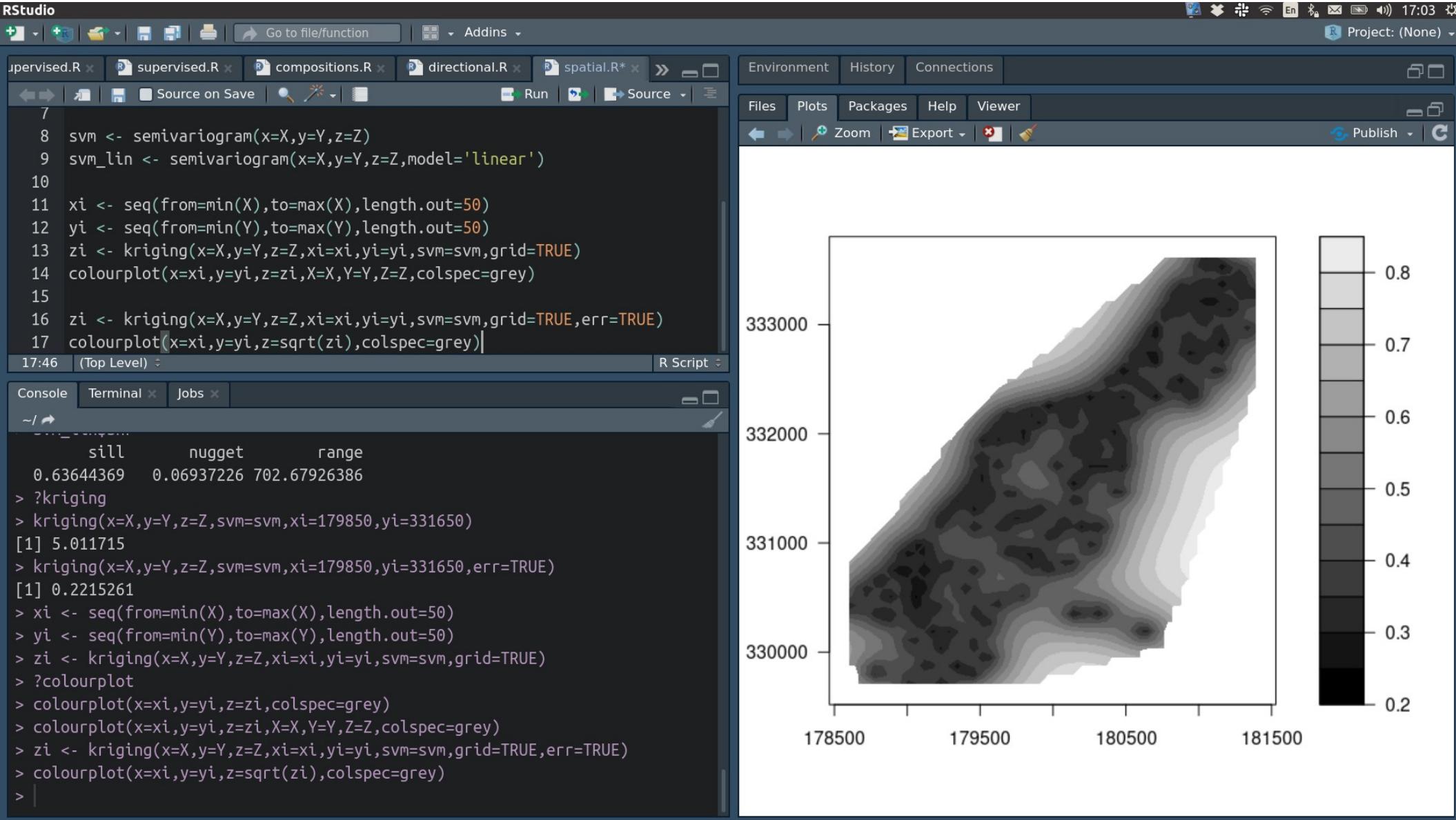
DUXBURY  
PUBLISHERS

STATISTICS  
AND  
DATA ANALYSIS  
IN  
**GEOLOGY**

THIRD EDITION



JOHN C. DAVIS

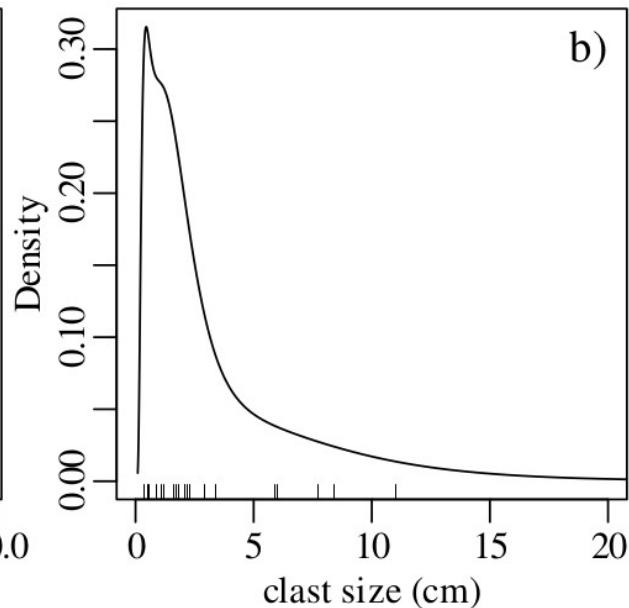
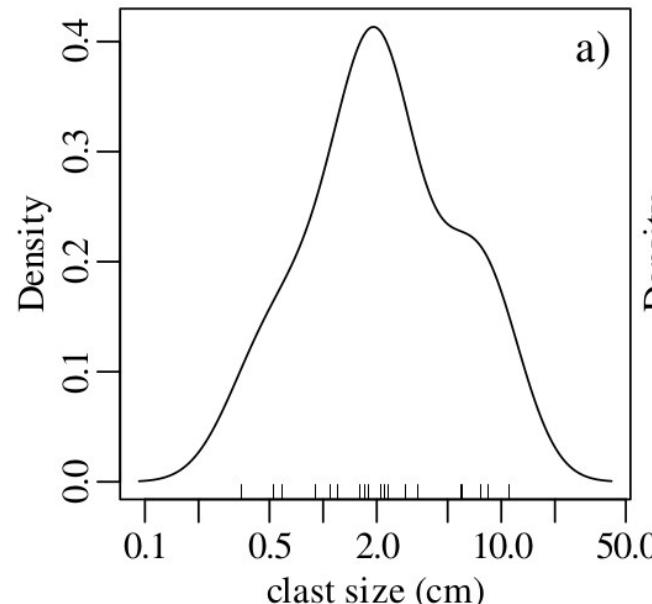
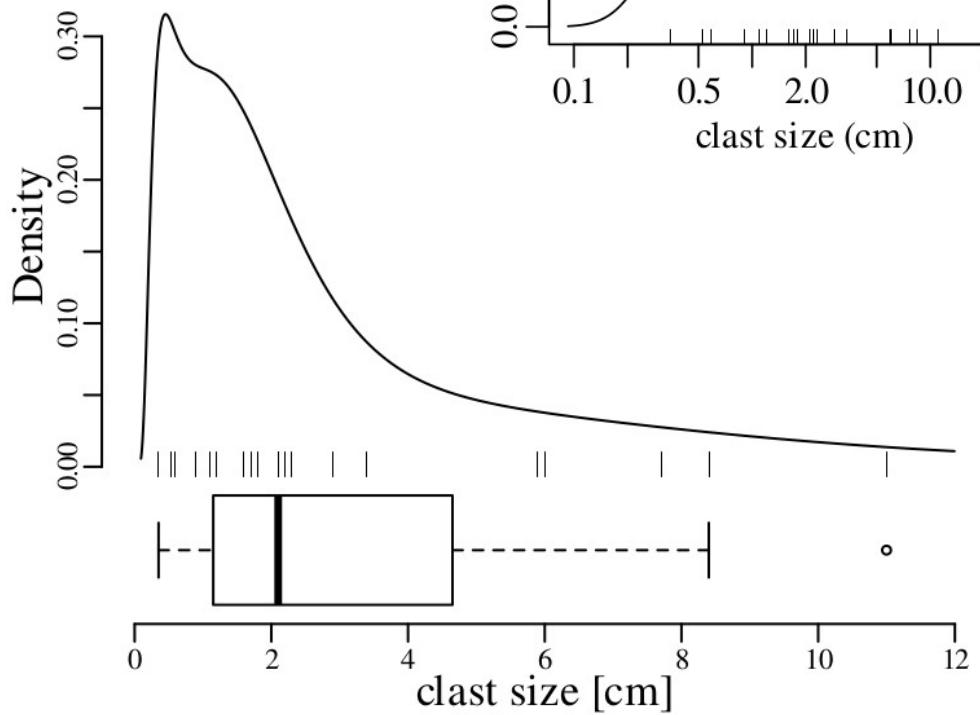


Chapter 2

Plotting data

Chapter 3

Summary statistics



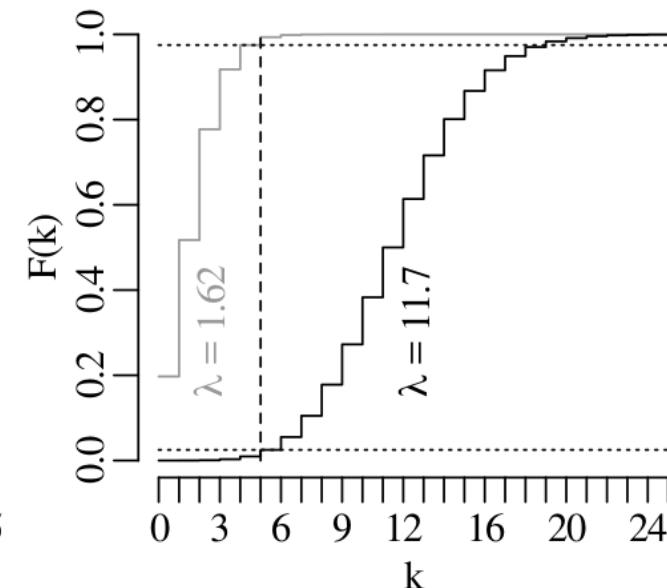
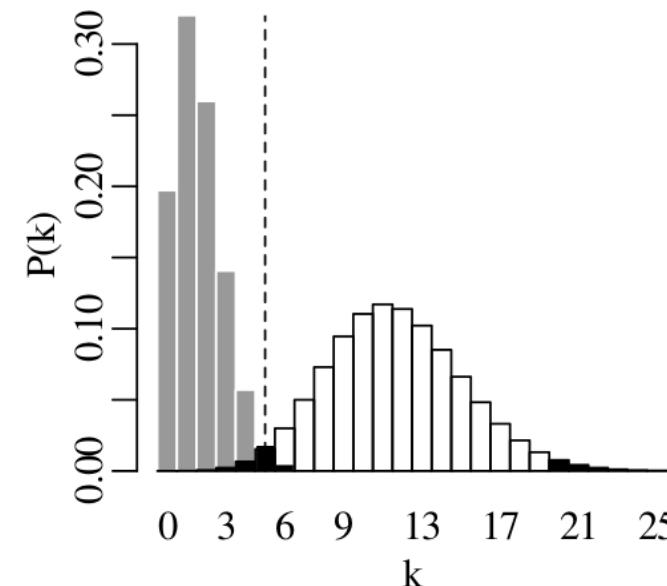
# Probability

# Chapter 5

## The binomial distribution

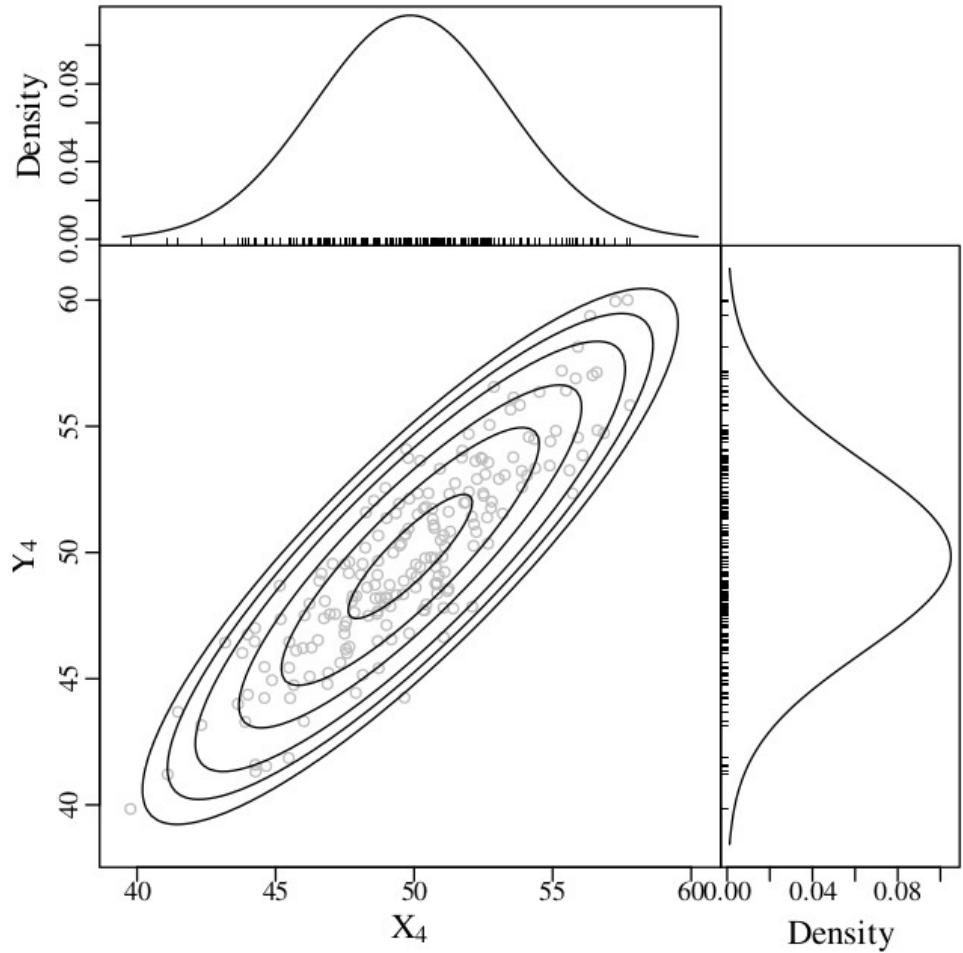
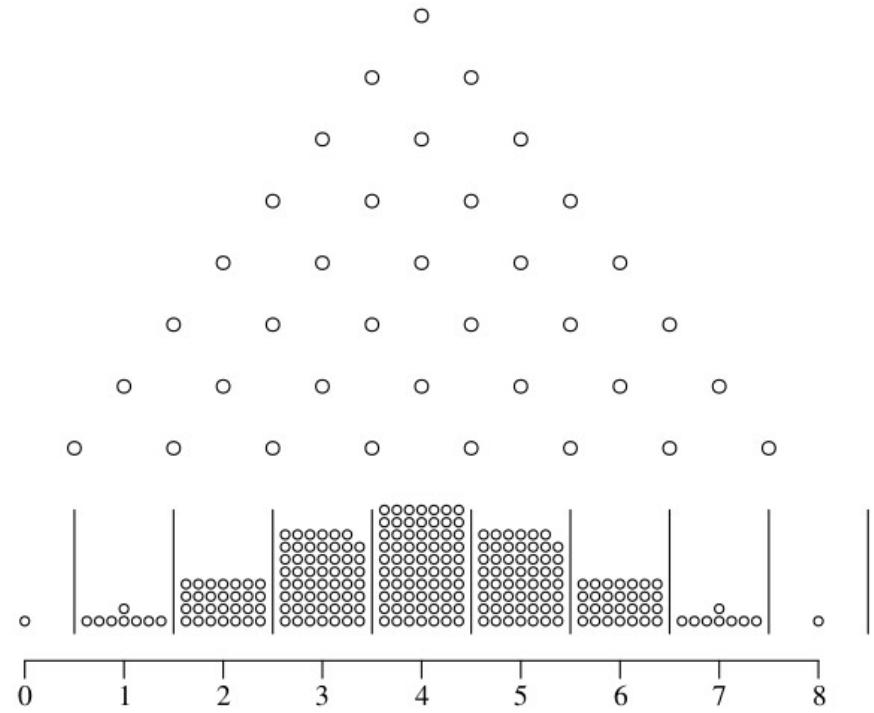
## Chapter 6

## The Poisson distribution



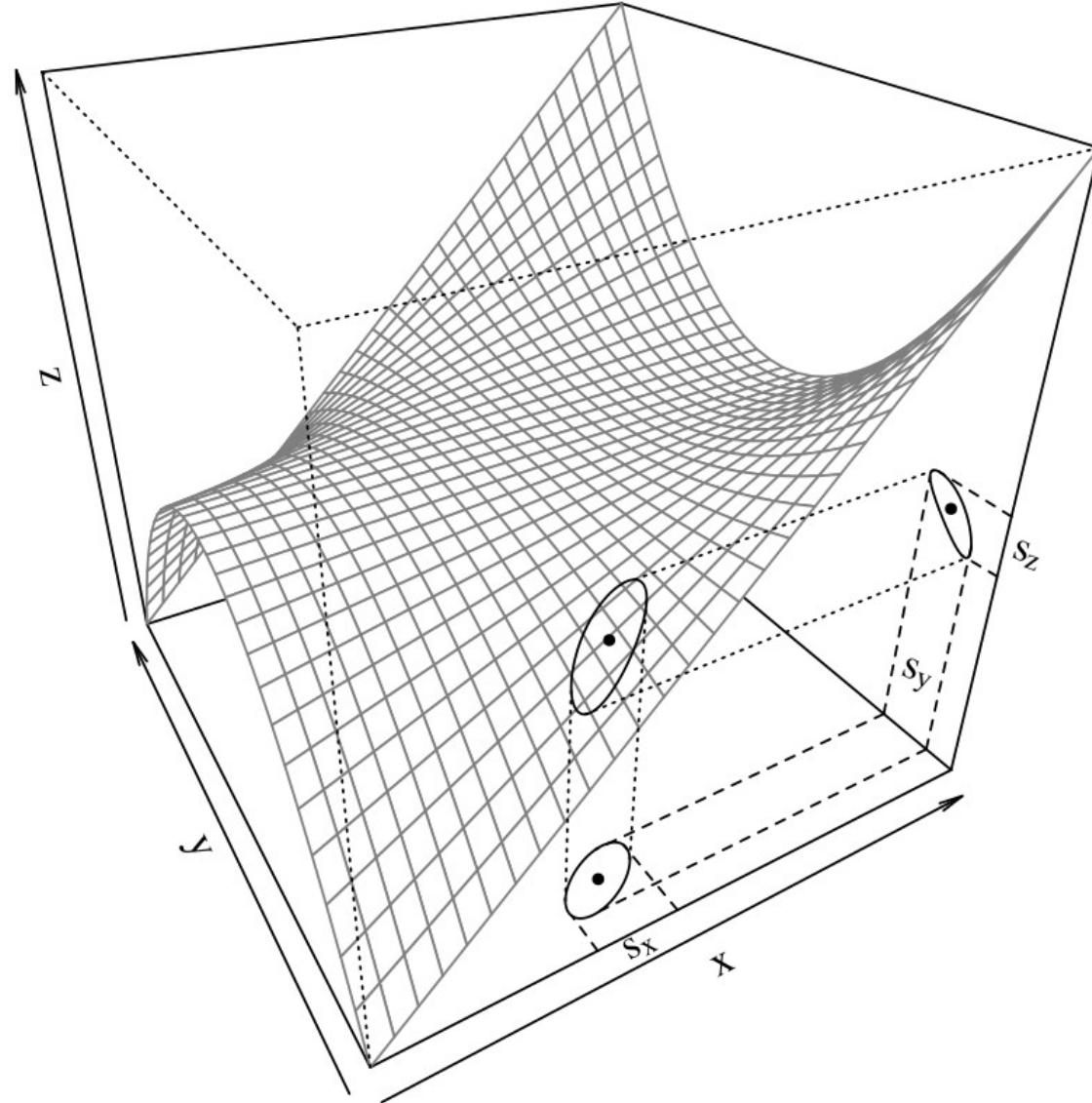
# Chapter 7

## The normal distribution



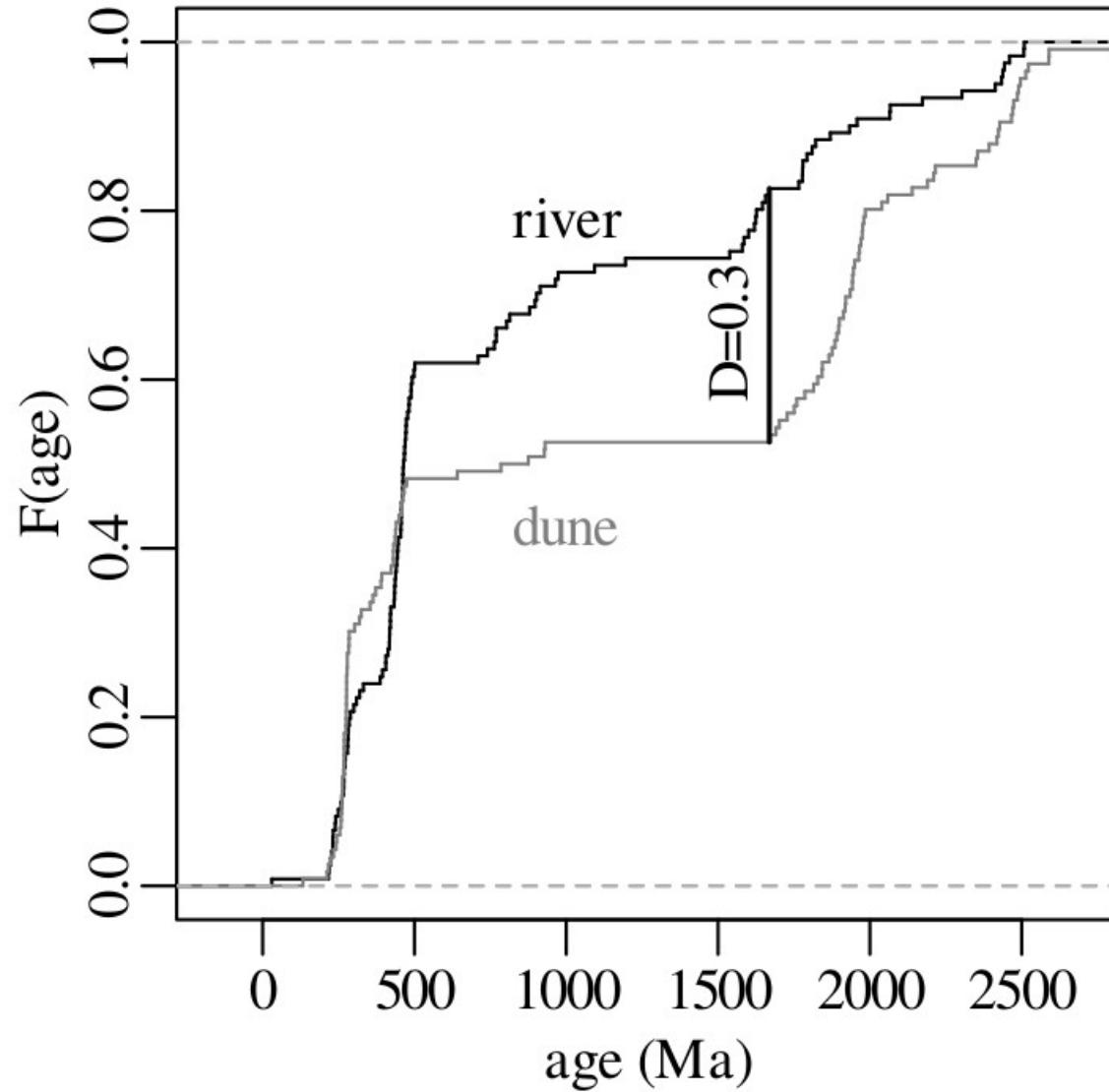
# Chapter 8

## Error propagation



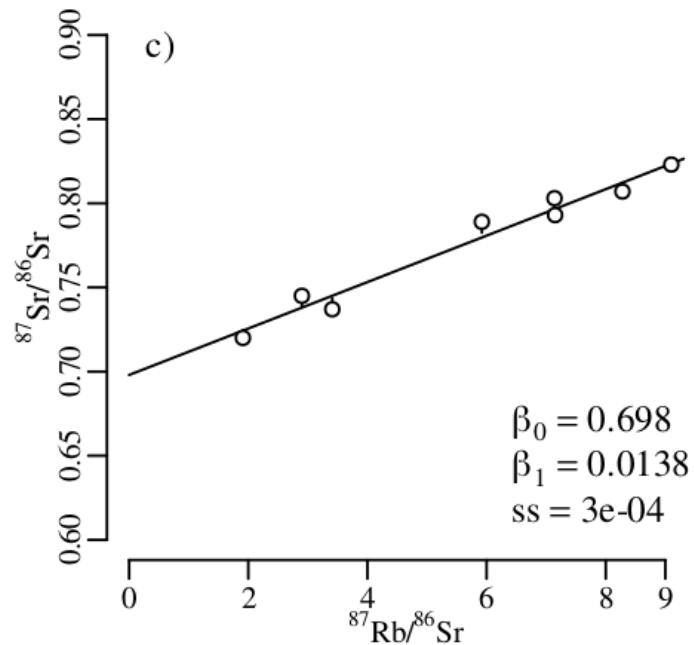
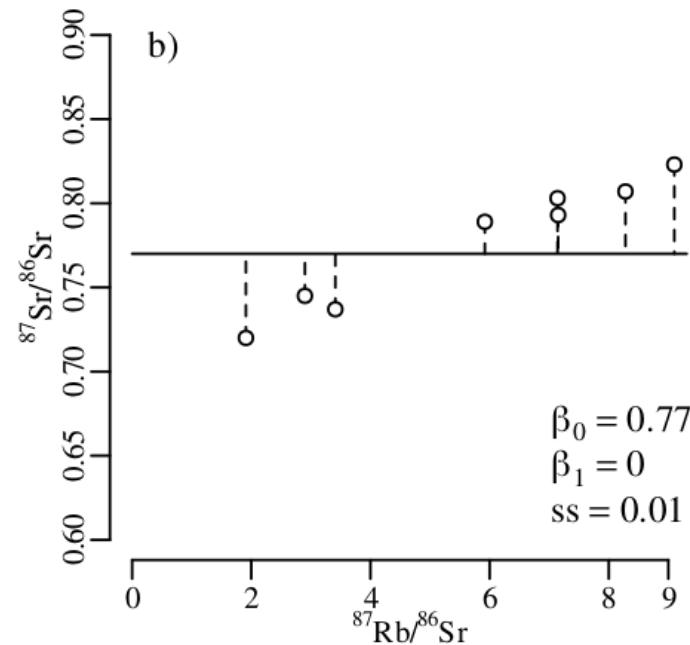
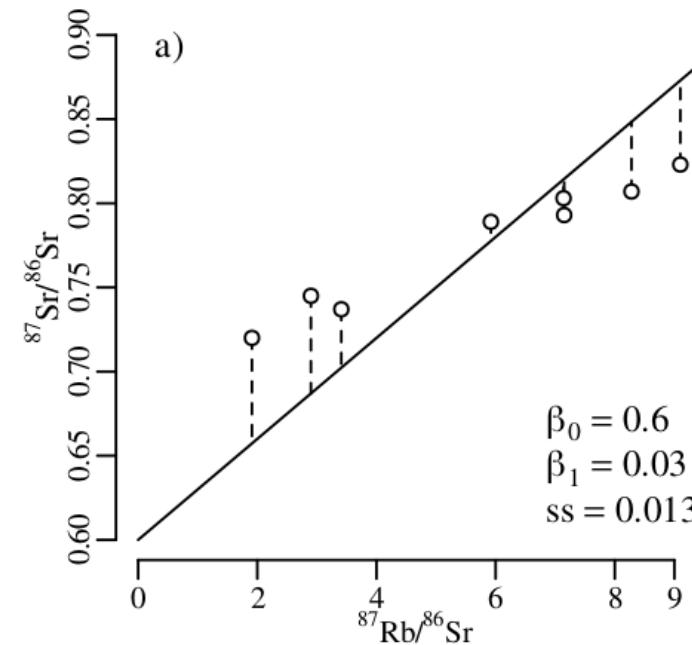
## Chapter 9

### Comparing distributions



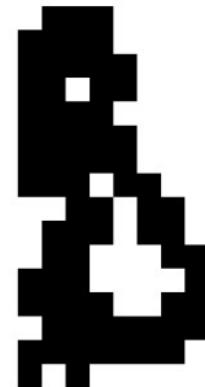
# Chapter 10

## Regression

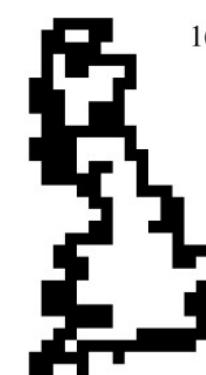


# Chapter 11

## Fractals and chaos



74/256



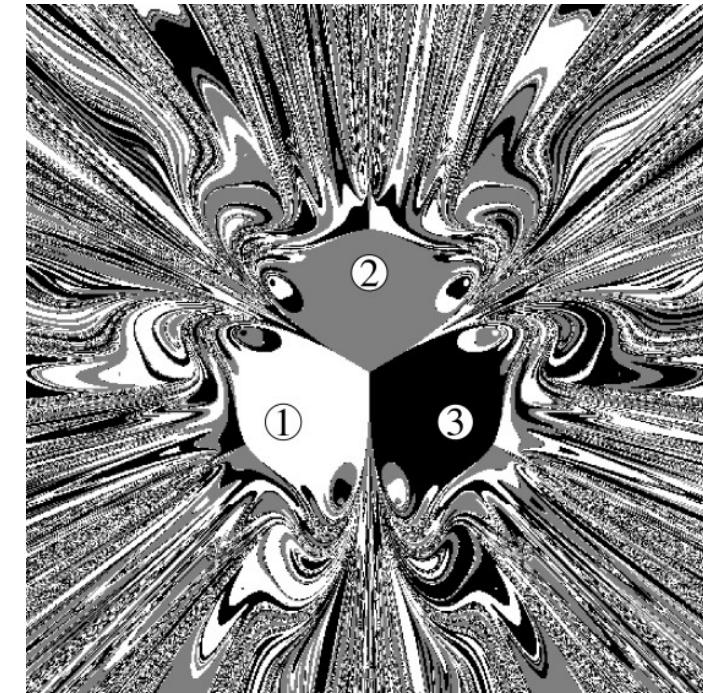
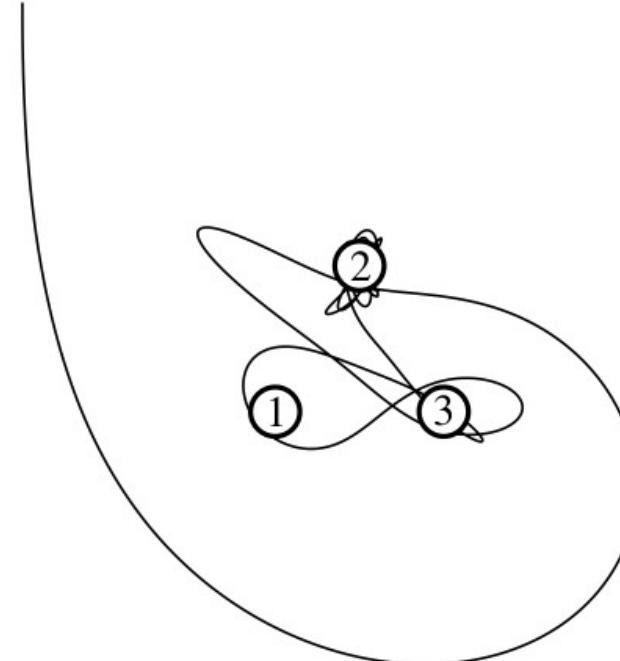
167/1024



819/16384



3058/262144

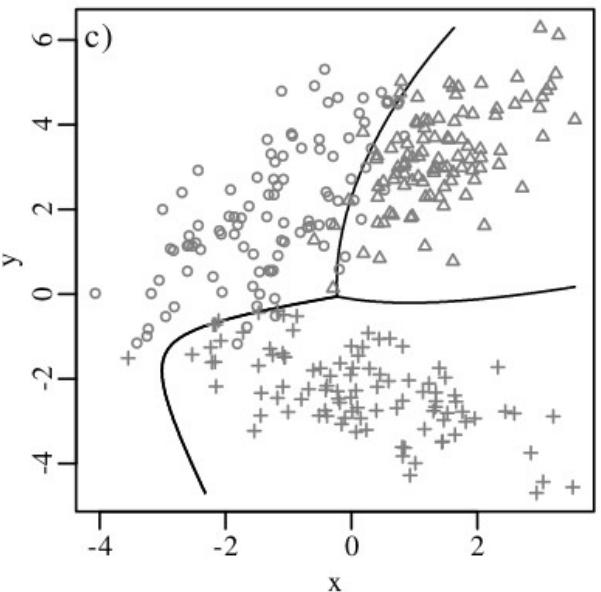
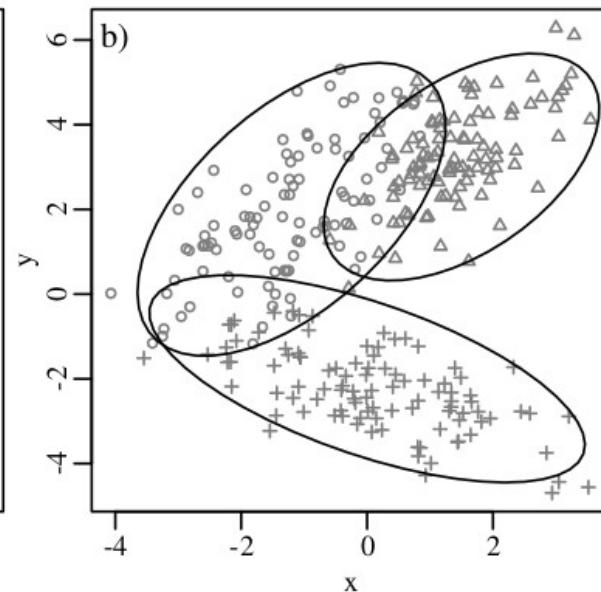
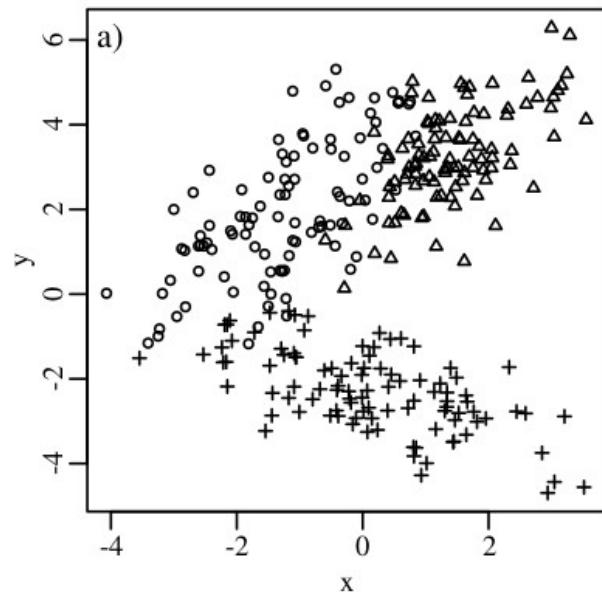
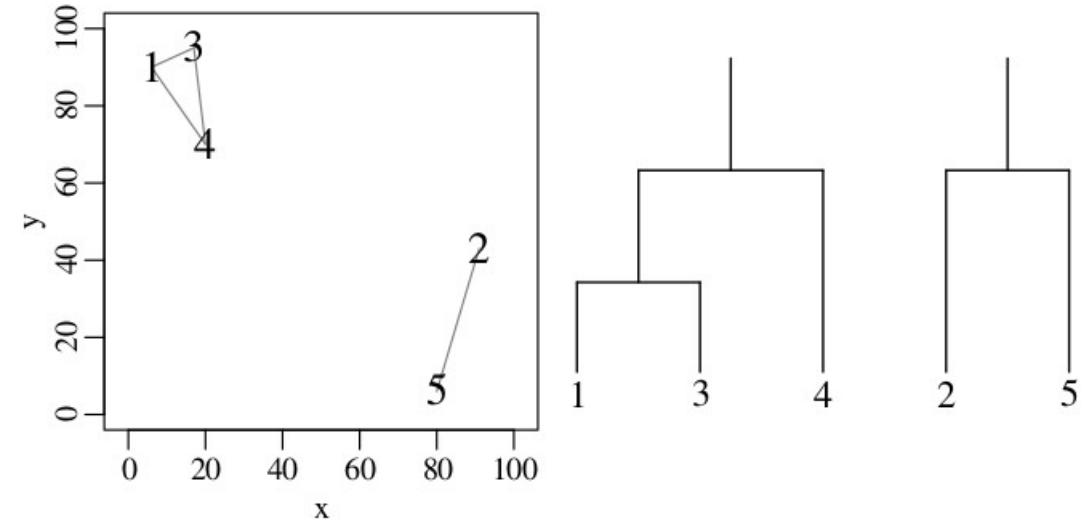


# Chapter 12

## Unsupervised learning

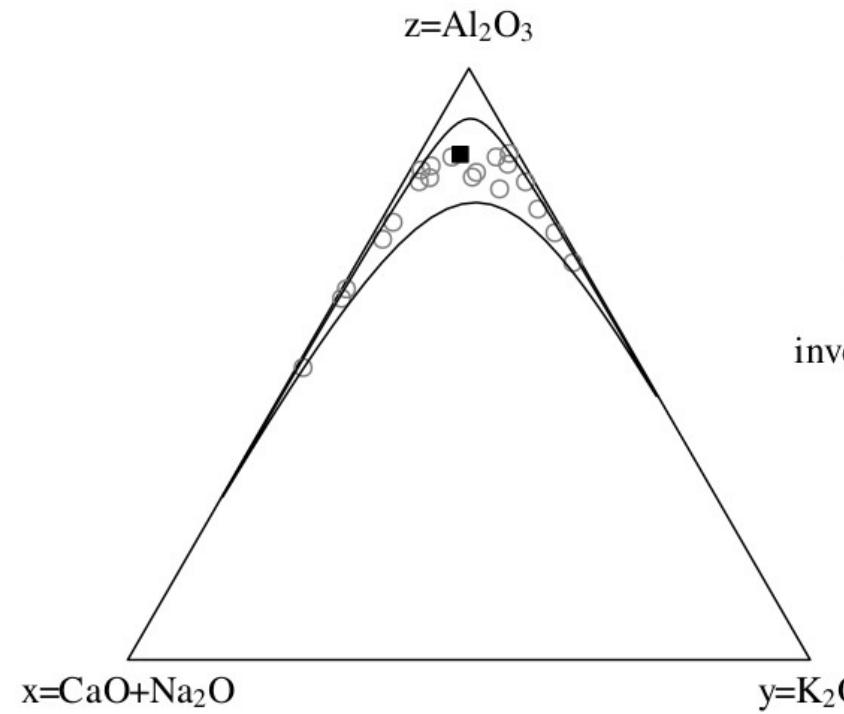
### Chapter 13

## Supervised learning

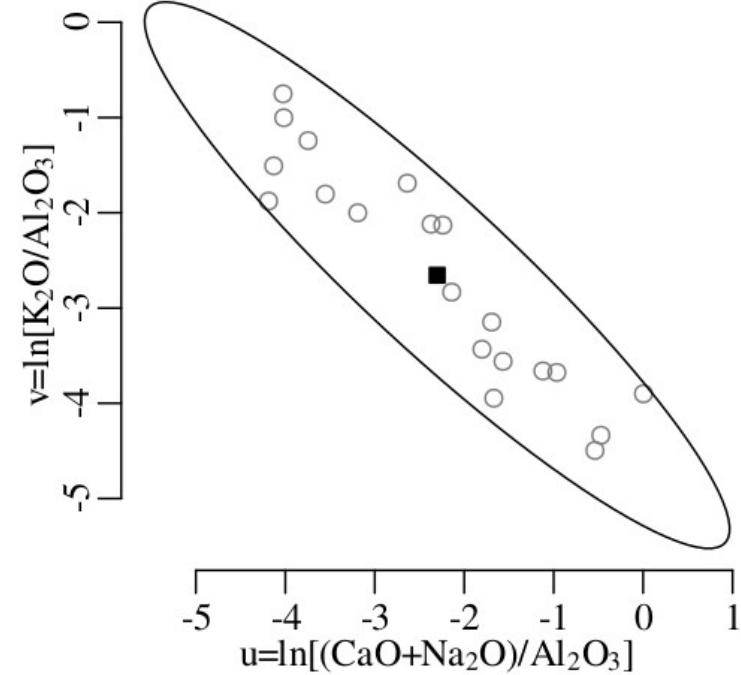


# Chapter 14

## Compositional data

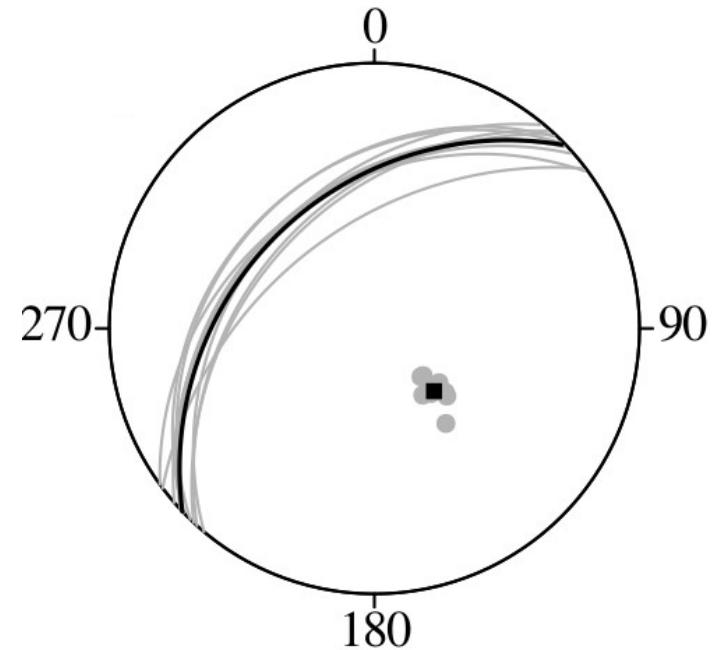
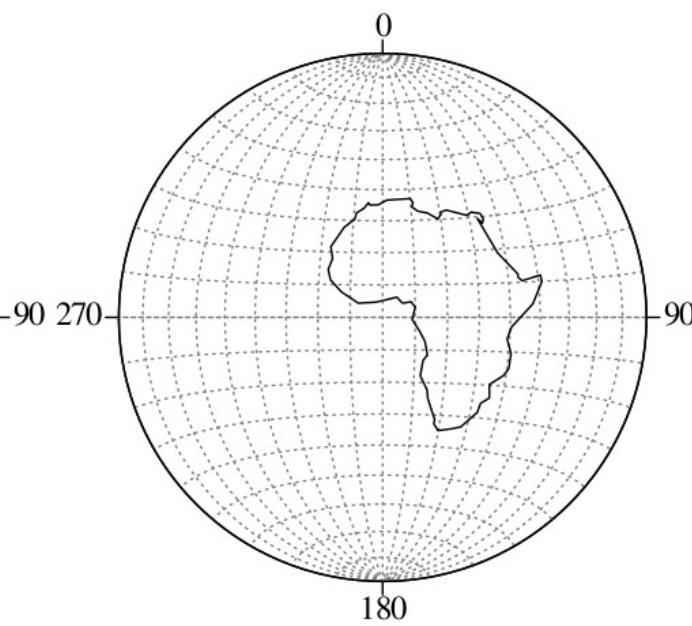
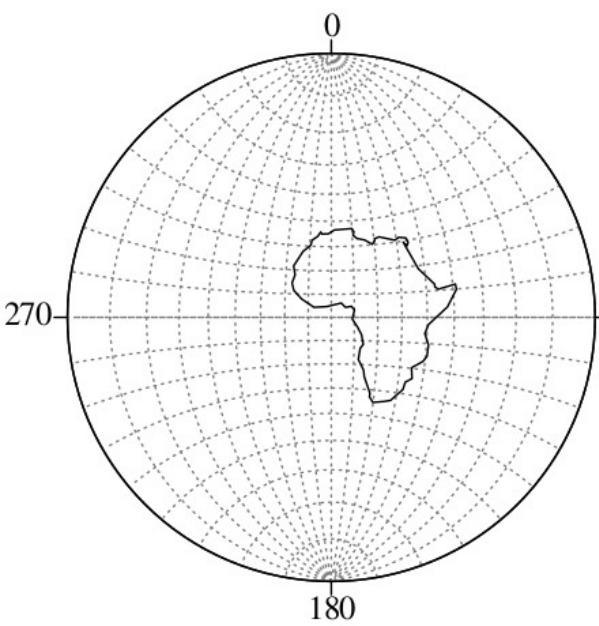
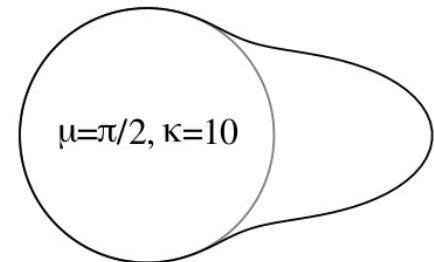
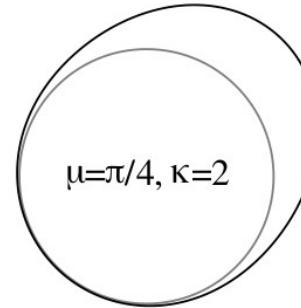
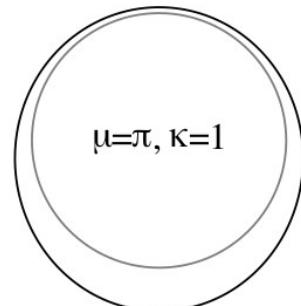
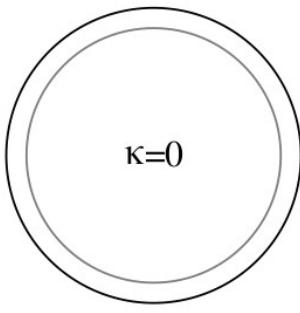


logratio transformation  
↔  
inverse logratio transformation



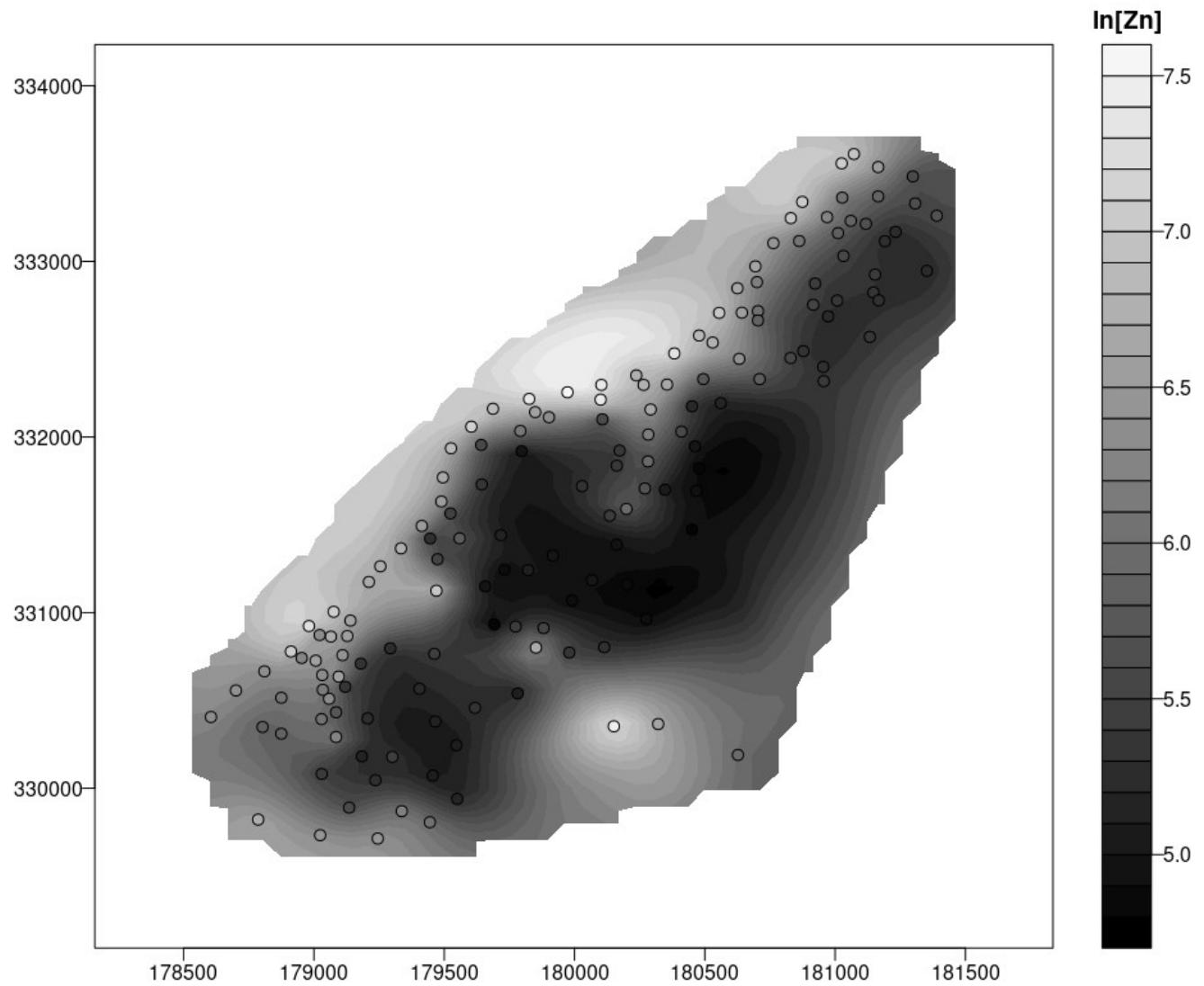
# Chapter 15

## Directional data



# Chapter 16

## Spatial data



## **17 An introduction to R**

|                                          |
|------------------------------------------|
| 17.1 The basics . . . . .                |
| 17.2 Plotting data . . . . .             |
| 17.3 Summary Statistics . . . . .        |
| 17.4 Probability . . . . .               |
| 17.5 The binomial distribution . . . . . |
| 17.6 The Poisson distribution . . . . .  |
| 17.7 The normal distribution . . . . .   |
| 17.8 Error propagation . . . . .         |
| 17.9 Comparing distributions . . . . .   |
| 17.10 Regression . . . . .               |
| 17.11 Fractals and chaos . . . . .       |
| 17.12 Unsupervised learning . . . . .    |
| 17.13 Supervised learning . . . . .      |
| 17.14 Compositional data . . . . .       |
| 17.15 Directional data . . . . .         |
| 17.16 Spatial data . . . . .             |

## **18 Exercises**

|                                          |
|------------------------------------------|
| 18.1 The basics . . . . .                |
| 18.2 Plotting data . . . . .             |
| 18.3 Summary statistics . . . . .        |
| 18.4 Probability . . . . .               |
| 18.5 The binomial distribution . . . . . |
| 18.6 The Poisson distribution . . . . .  |
| 18.7 The normal distribution . . . . .   |
| 18.8 Error propagation . . . . .         |
| 18.9 Comparing distributions . . . . .   |
| 18.10 Regression . . . . .               |
| 18.11 Fractals and chaos . . . . .       |
| 18.12 Unsupervised learning . . . . .    |
| 18.13 Supervised learning . . . . .      |
| 18.14 Compositional data . . . . .       |
| 18.15 Directional data . . . . .         |
| 18.16 Spatial data . . . . .             |

## **19 Solutions**

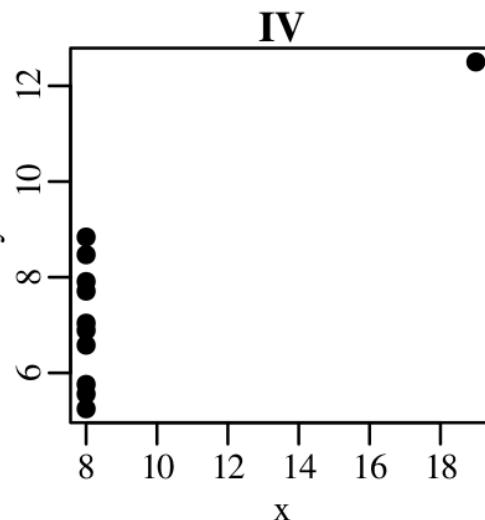
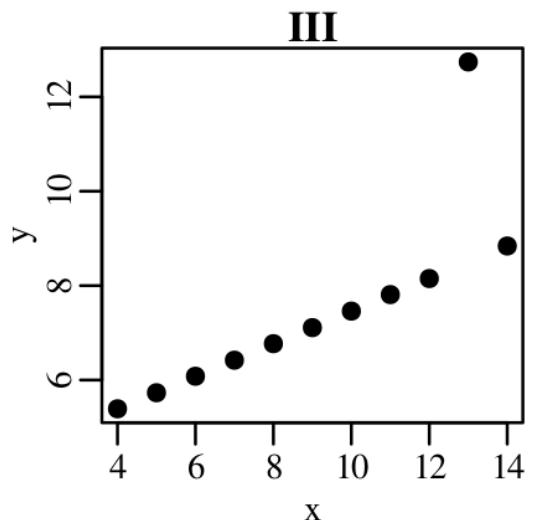
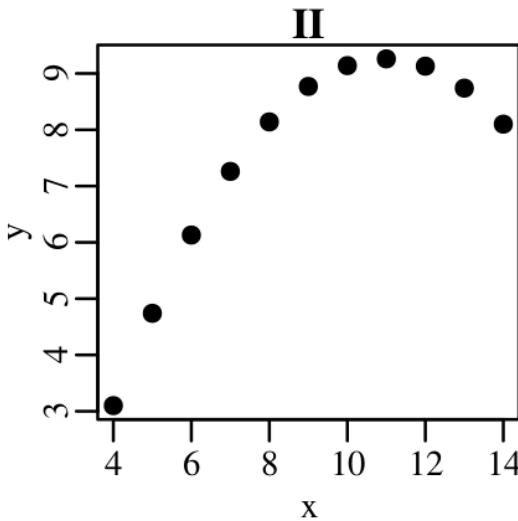
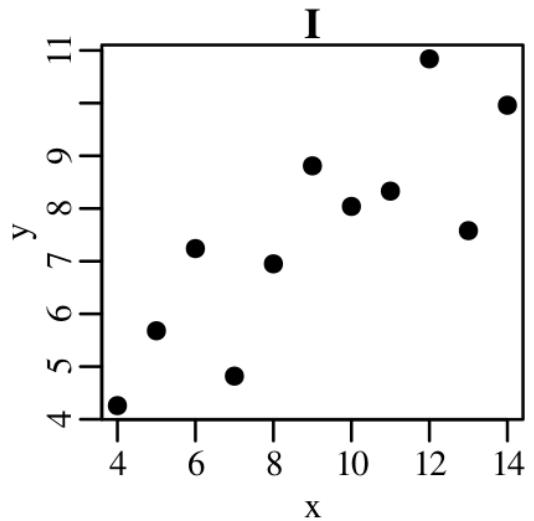
|                                          |
|------------------------------------------|
| 19.1 The basics . . . . .                |
| 19.2 Plotting data . . . . .             |
| 19.3 Summary statistics . . . . .        |
| 19.4 Probability . . . . .               |
| 19.5 The binomial distribution . . . . . |
| 19.6 The Poisson distribution . . . . .  |
| 19.7 The normal distribution . . . . .   |
| 19.8 Error propagation . . . . .         |
| 19.9 Comparing distributions . . . . .   |
| 19.10 Regression . . . . .               |
| 19.11 Fractals and chaos . . . . .       |
| 19.12 Unsupervised learning . . . . .    |
| 19.13 Supervised learning . . . . .      |
| 19.14 Compositional data . . . . .       |
| 19.15 Directional data . . . . .         |
| 19.16 Spatial data . . . . .             |



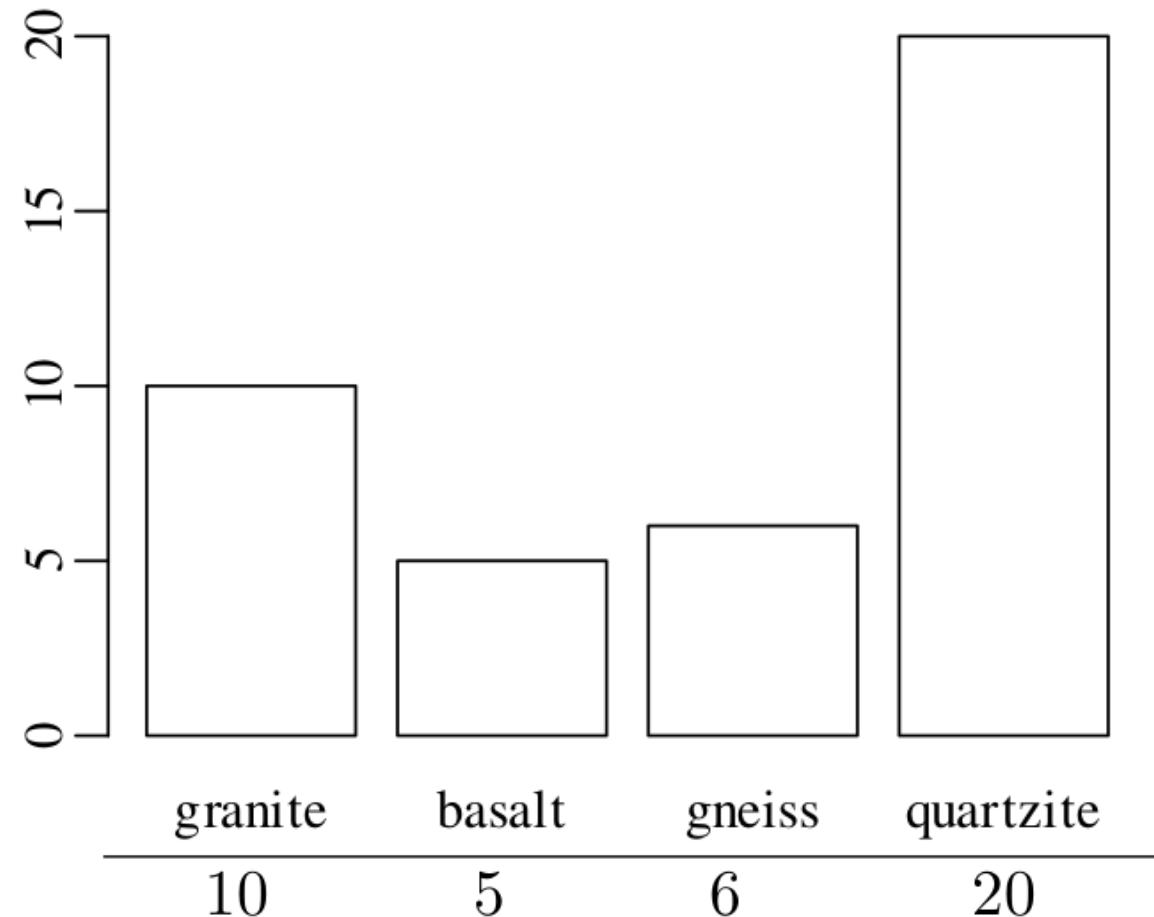
# Statistics for geoscientists

Plotting data

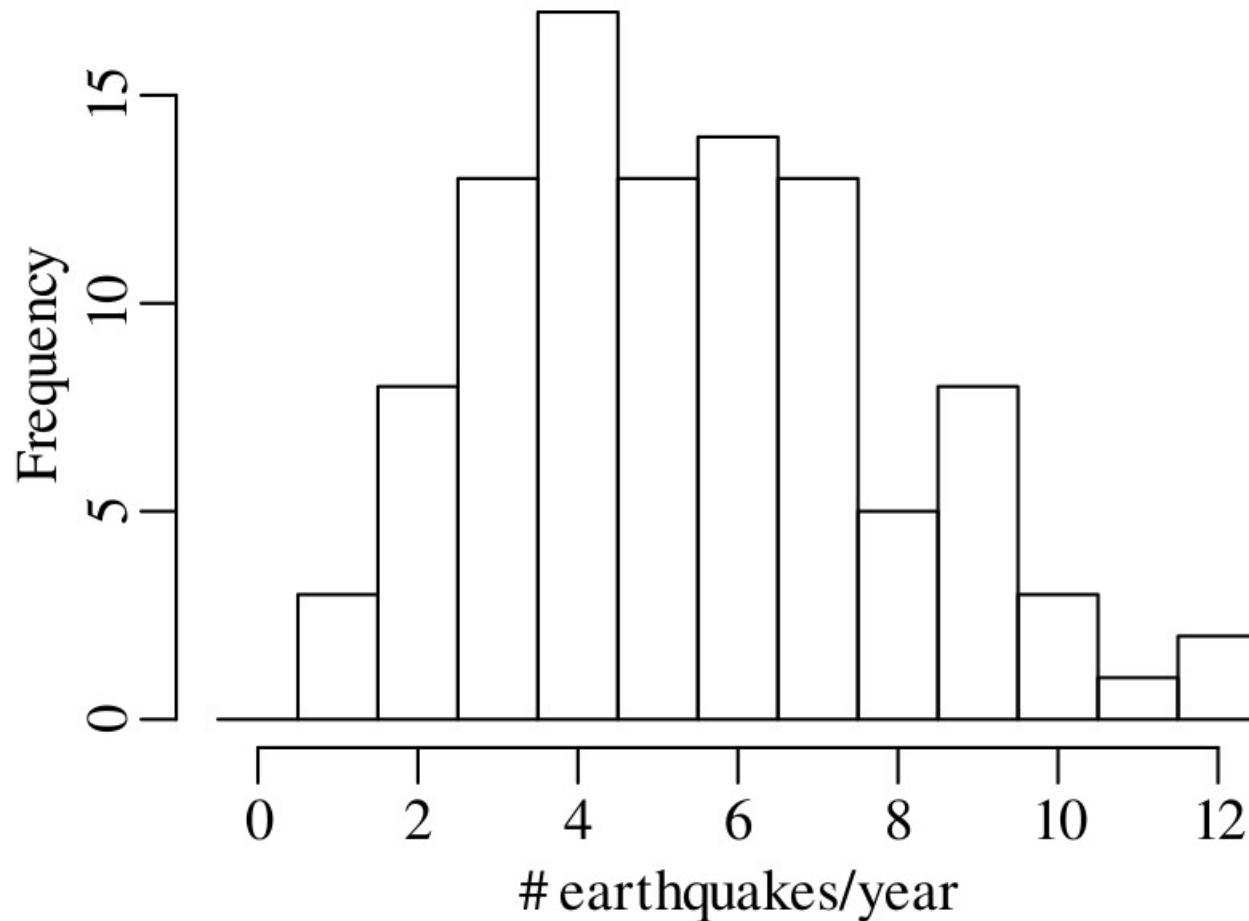
| I    | II    |      | III  |      | IV    |      |       |                                                       |
|------|-------|------|------|------|-------|------|-------|-------------------------------------------------------|
| $x$  | $y$   | $x$  | $y$  | $x$  | $y$   | $x$  | $y$   |                                                       |
| 10.0 | 8.04  | 10.0 | 9.14 | 10.0 | 7.46  | 8.0  | 6.58  | – the mean of $x$ is 9                                |
| 8.0  | 6.95  | 8.0  | 8.14 | 8.0  | 6.77  | 8.0  | 5.76  |                                                       |
| 13.0 | 7.58  | 13.0 | 8.74 | 13.0 | 12.74 | 8.0  | 7.71  | – the variance of $x$ is 11                           |
| 9.0  | 8.81  | 9.0  | 8.77 | 9.0  | 7.11  | 8.0  | 8.84  |                                                       |
| 11.0 | 8.33  | 11.0 | 9.26 | 11.0 | 7.81  | 8.0  | 8.47  | – the mean of $y$ is 7.50                             |
| 14.0 | 9.96  | 14.0 | 8.10 | 14.0 | 8.84  | 8.0  | 7.04  |                                                       |
| 6.0  | 7.24  | 6.0  | 6.13 | 6.0  | 6.08  | 8.0  | 5.25  | – the variance of $y$ is 4.125                        |
| 4.0  | 4.26  | 4.0  | 3.10 | 4.0  | 5.39  | 19.0 | 12.50 |                                                       |
| 12.0 | 10.84 | 12.0 | 9.13 | 12.0 | 8.15  | 8.0  | 5.56  | – the correlation coefficient of $x$ and $y$ is 0.816 |
| 7.0  | 4.82  | 7.0  | 7.26 | 7.0  | 6.42  | 8.0  | 7.91  |                                                       |
| 5.0  | 5.68  | 5.0  | 4.74 | 5.0  | 5.73  | 8.0  | 6.89  | – the best fit line is given by $y = 3.00 + 0.500x$   |



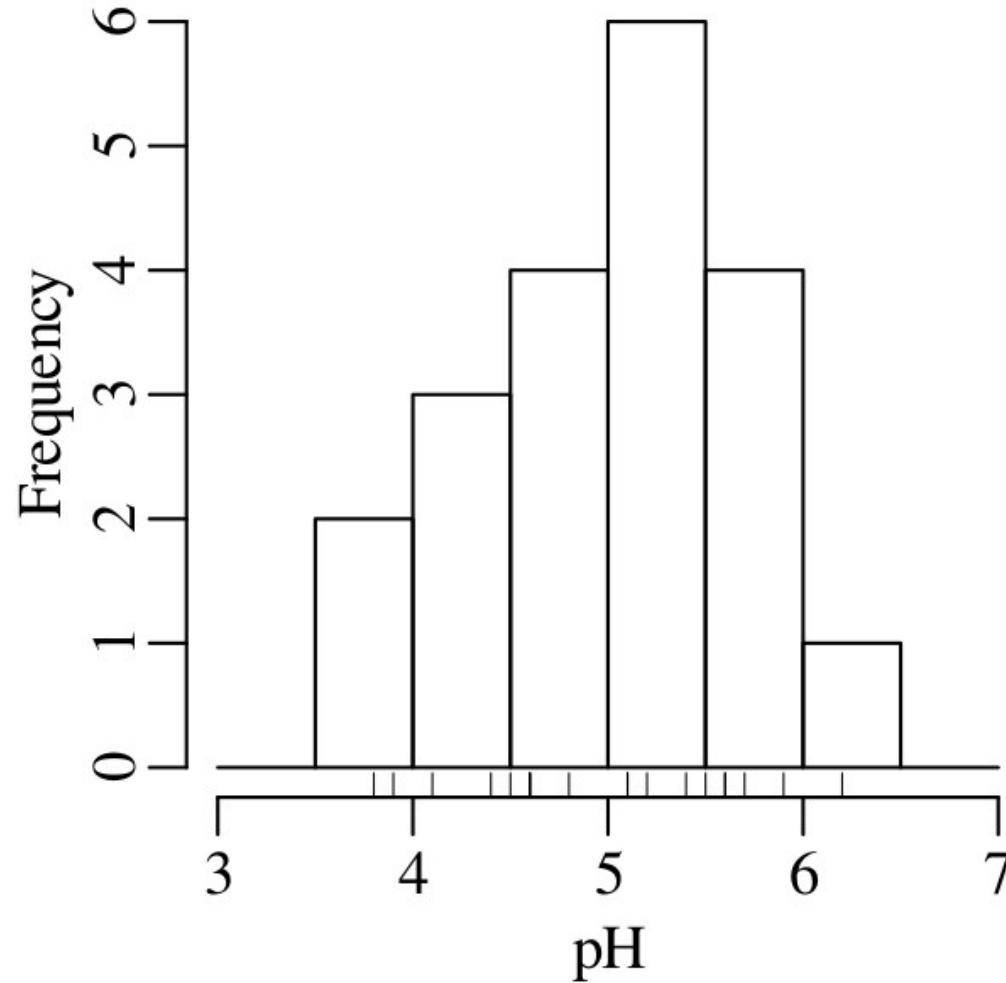
# Categorical data



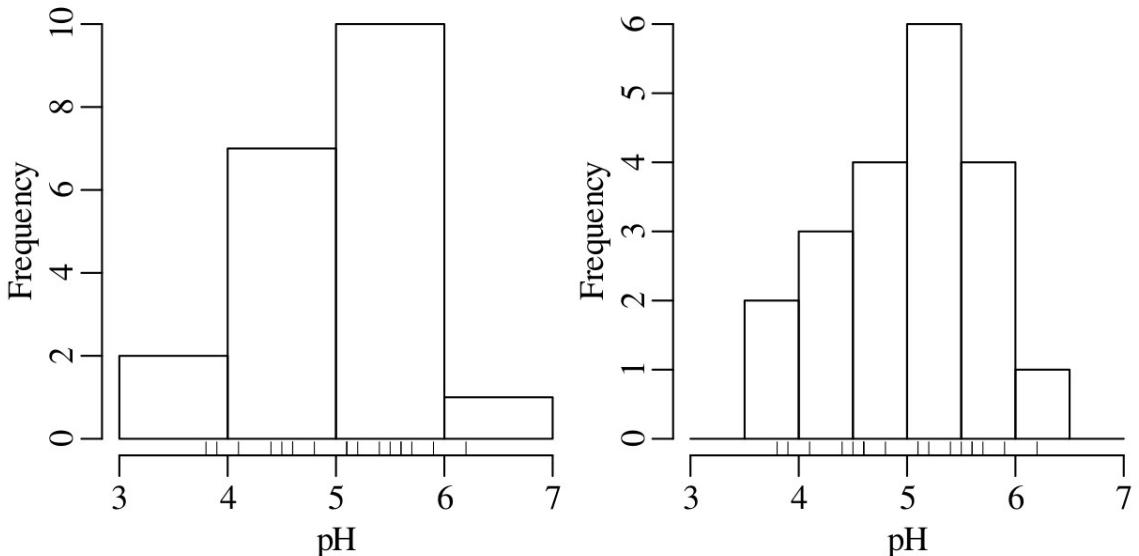
# Count data



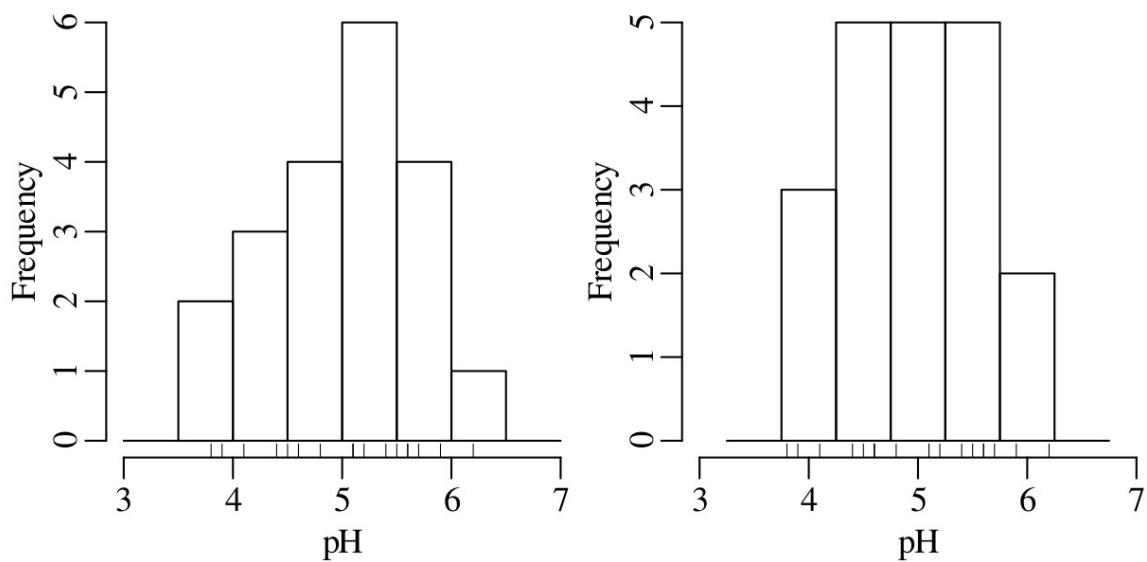
# Continuous data



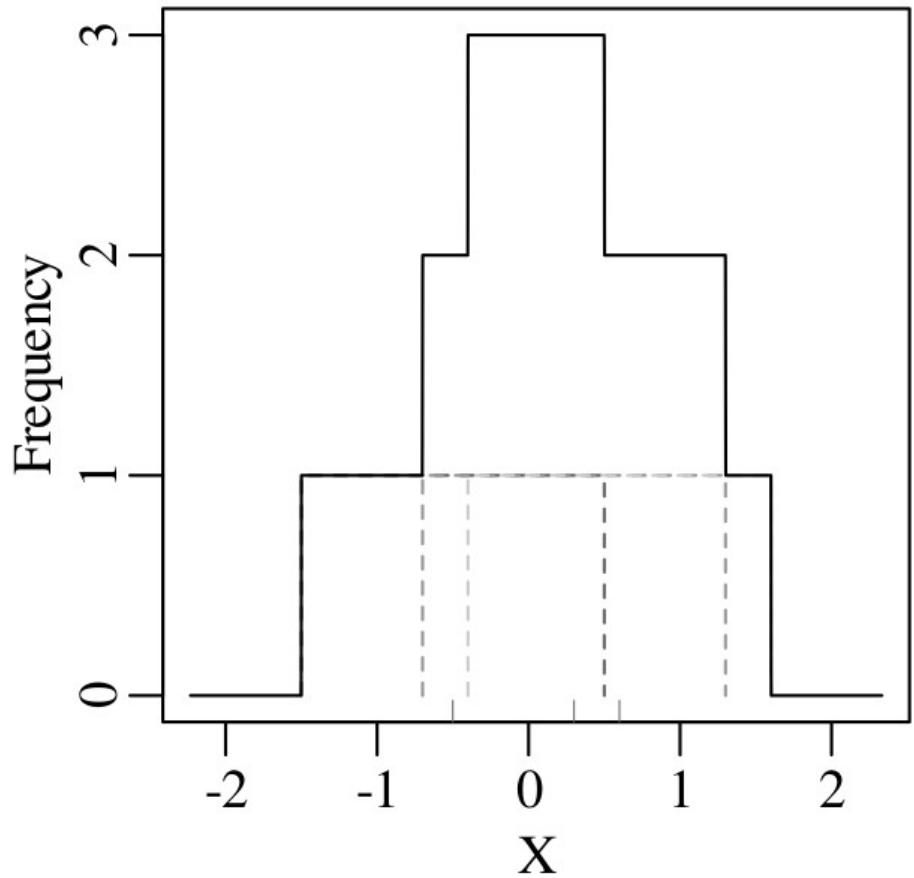
i. How many bins?



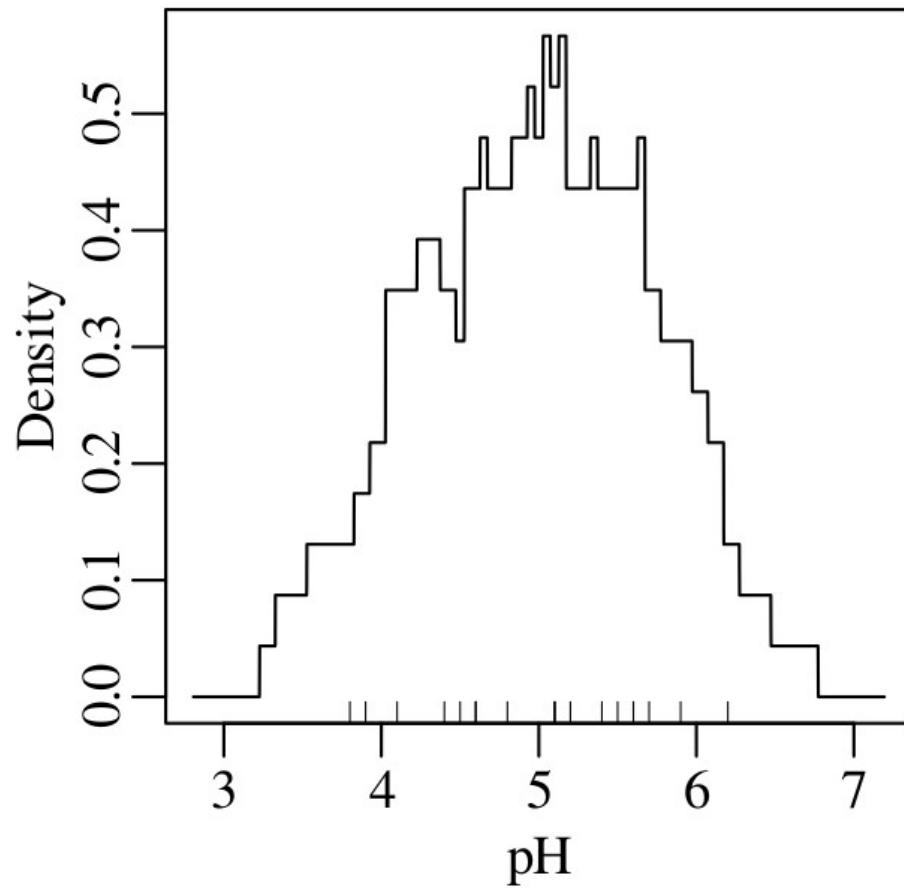
ii. Where to place the bins?



# Kernel Density Estimate

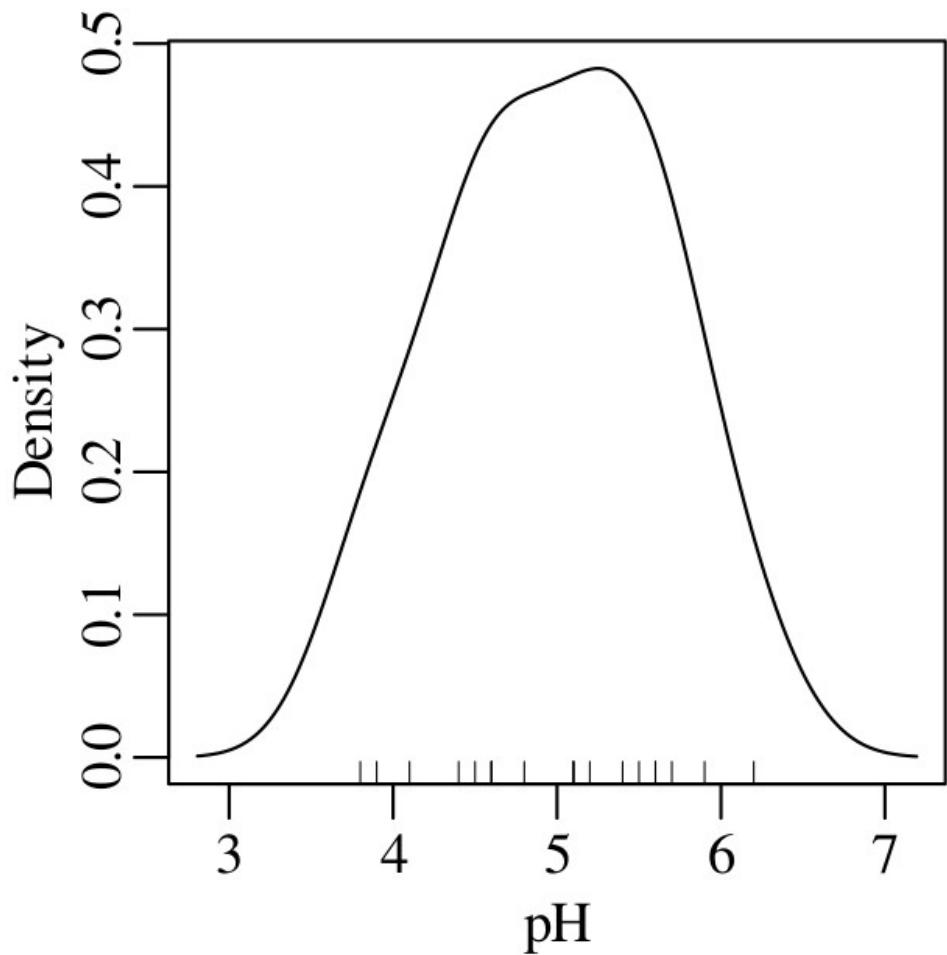
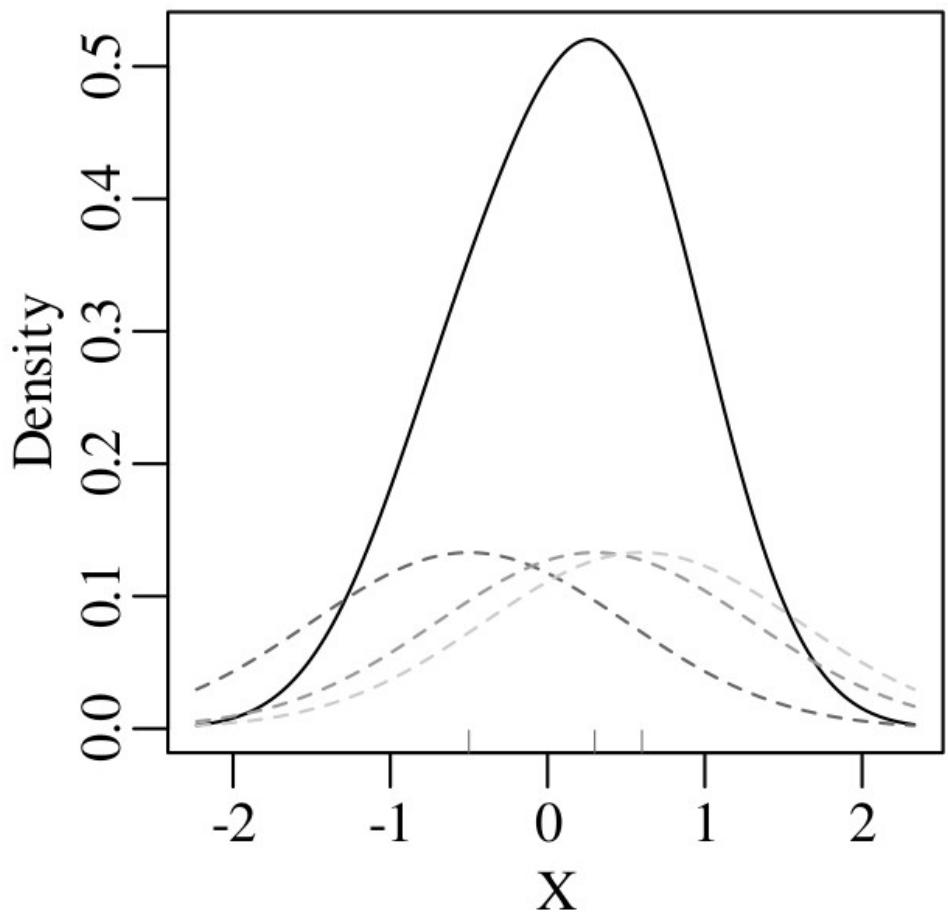


$$KDE(x) = \frac{1}{nh} \sum_{i=1}^n K\left(\frac{x - x_i}{h}\right)$$

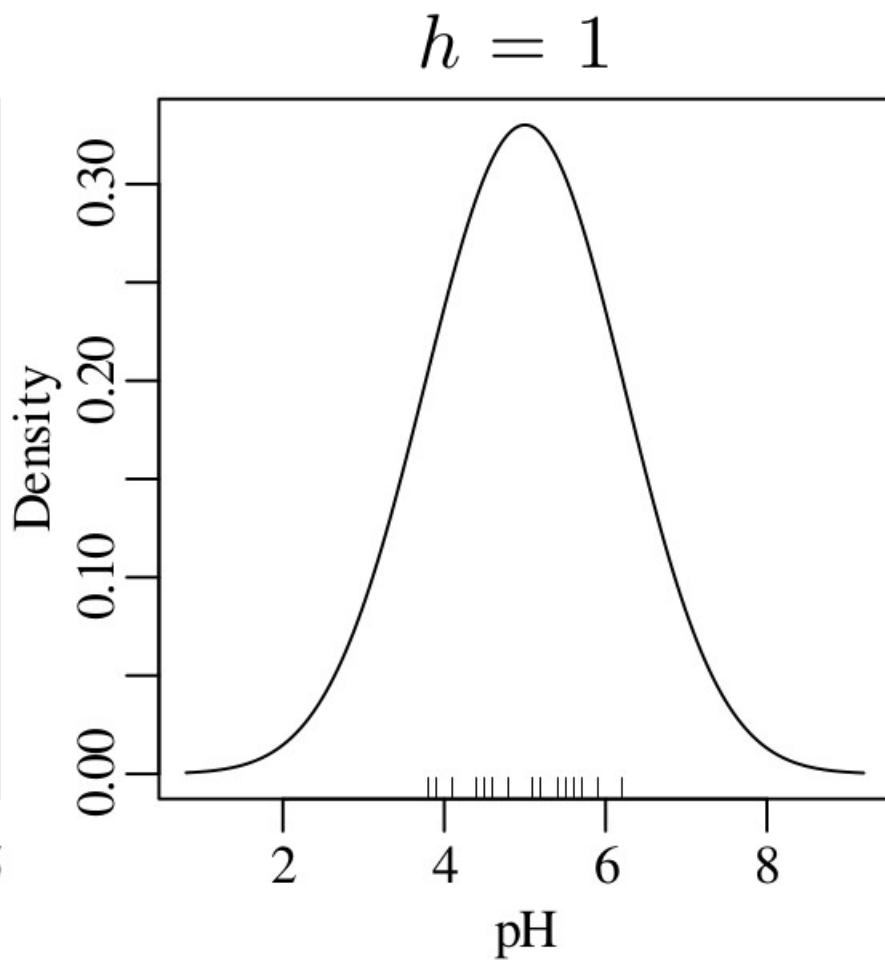
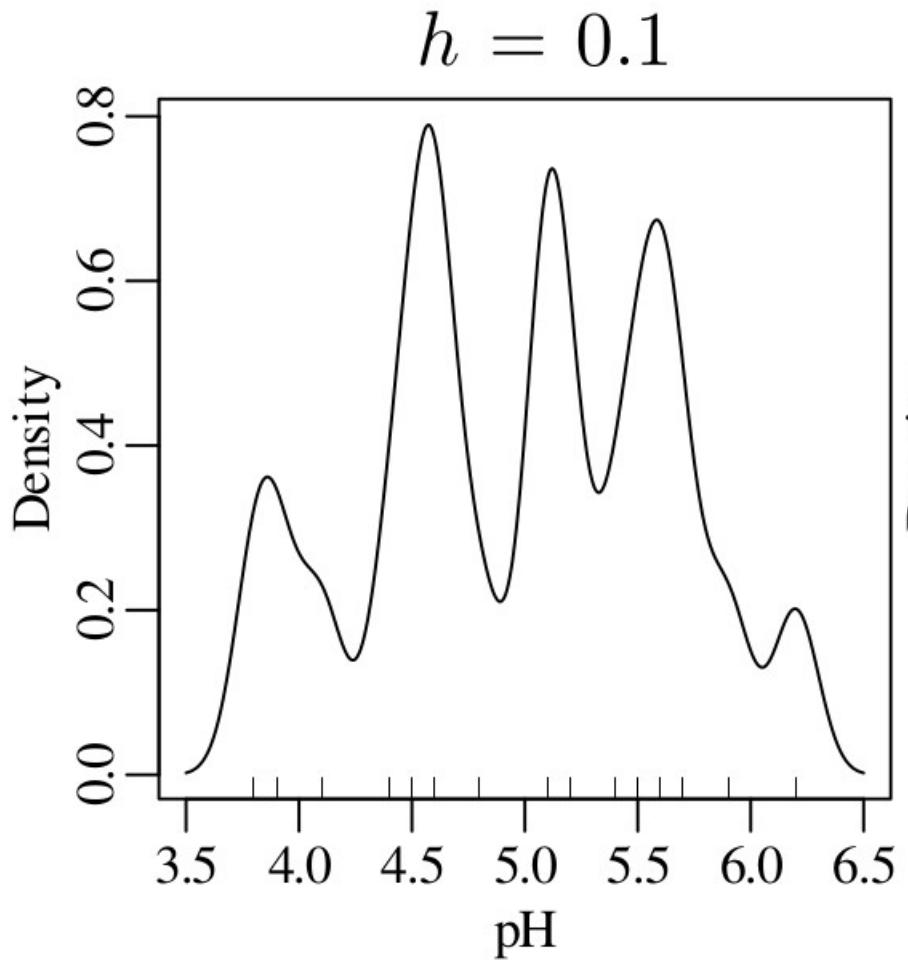


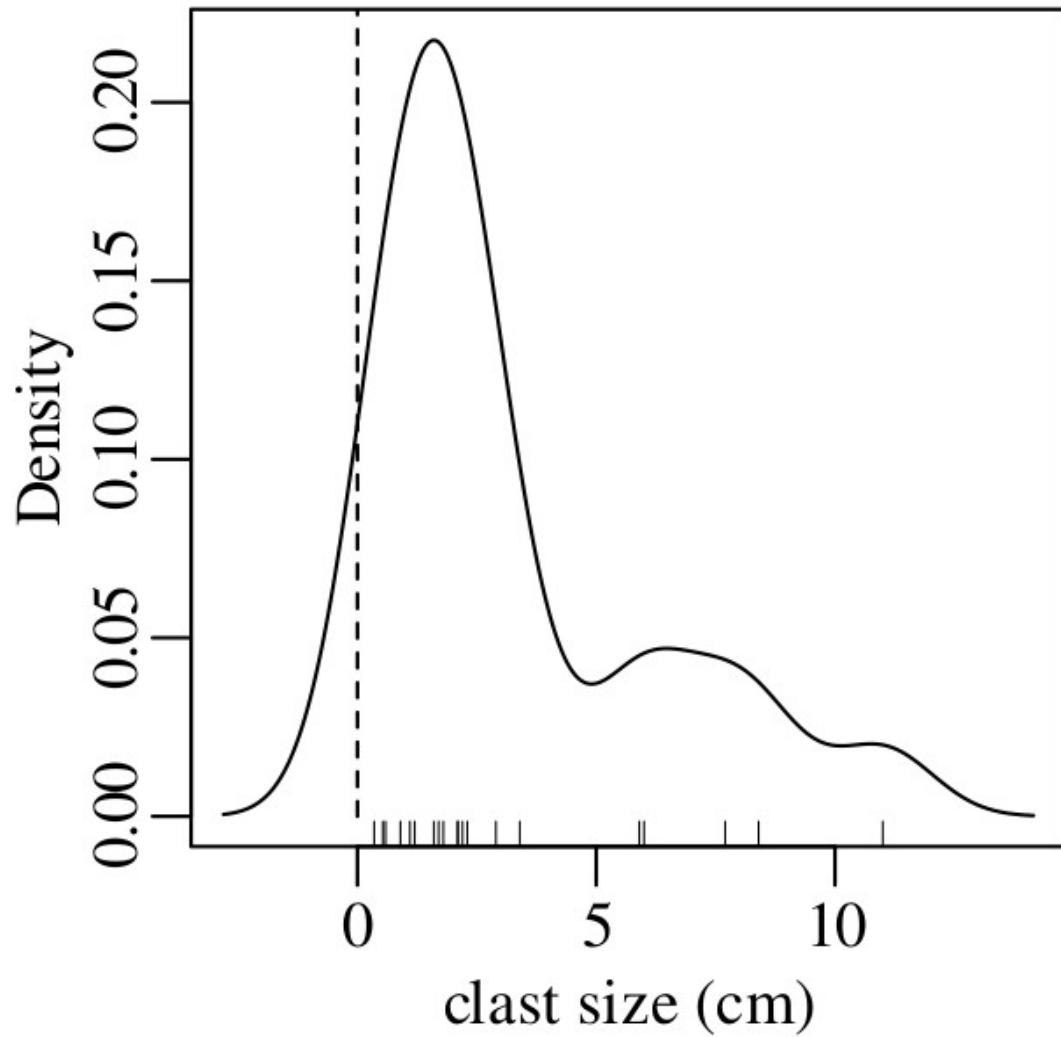
Gaussian kernel:

$$K(u) = \frac{1}{\sqrt{2\pi}} \exp\left[-\frac{u^2}{2}\right]$$

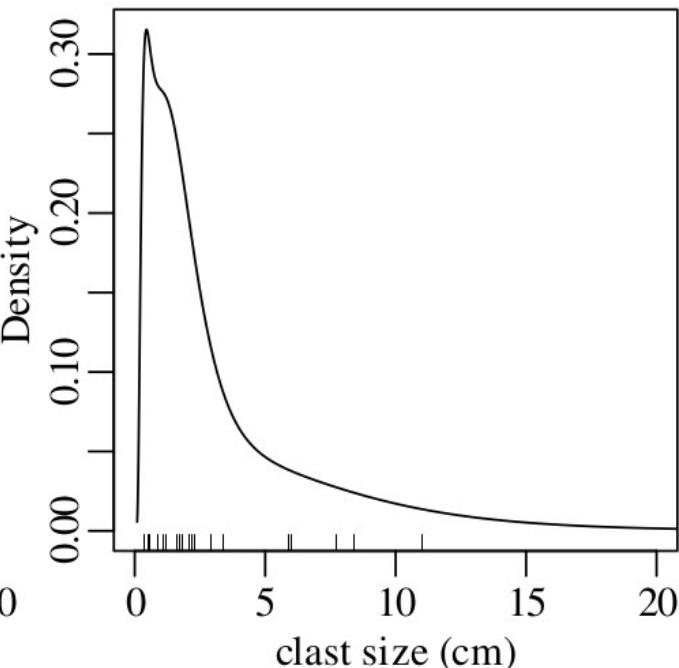
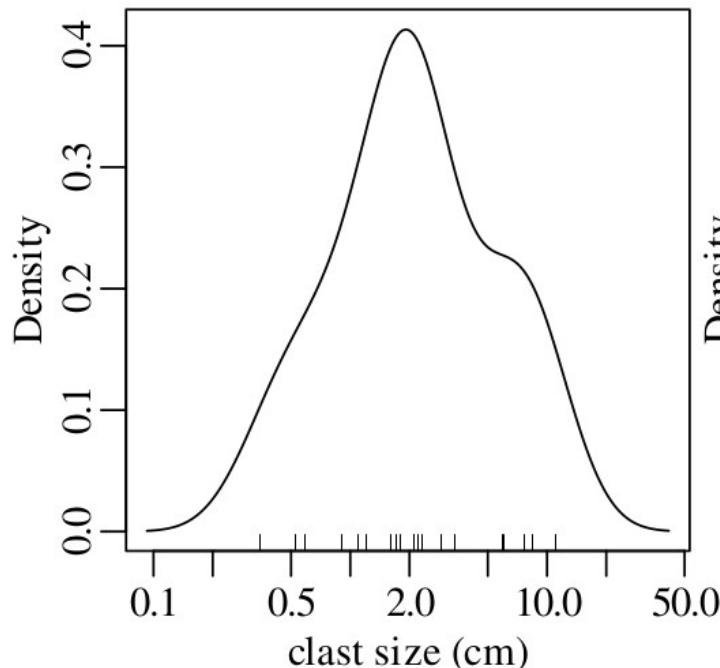
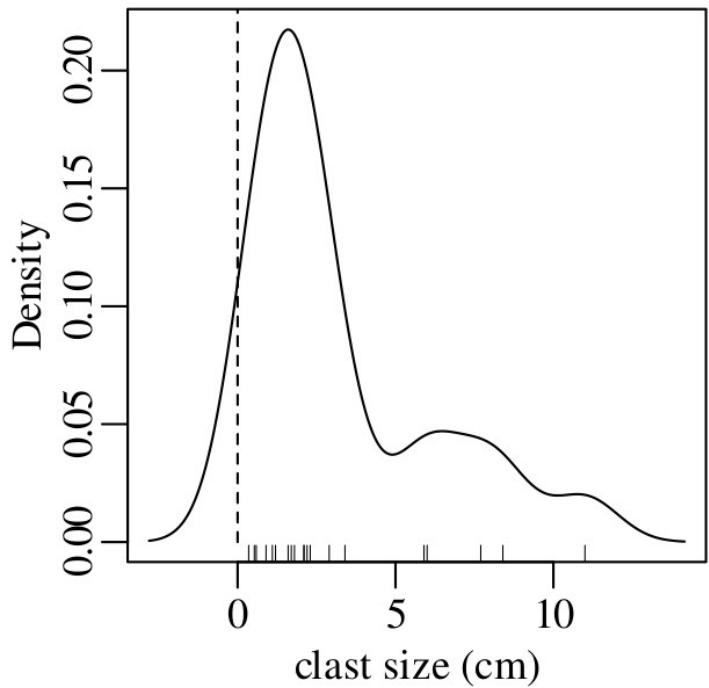


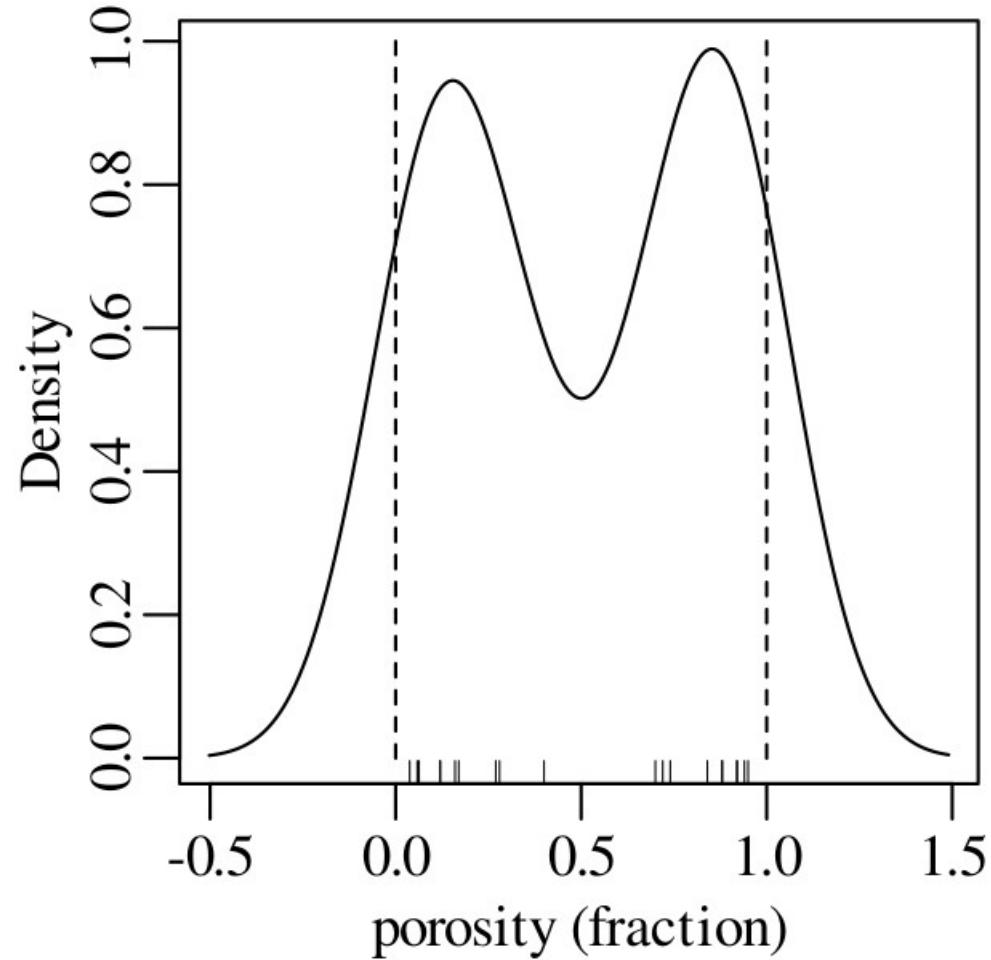
bandwidth





# logarithmic transformation

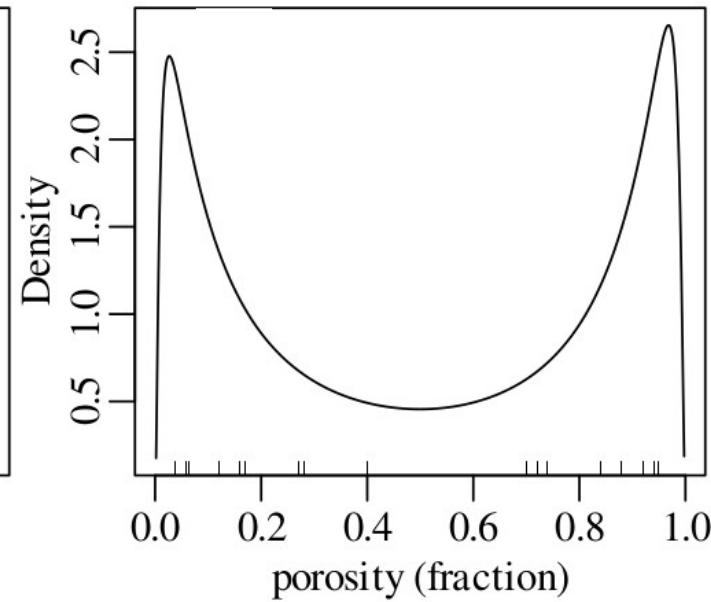
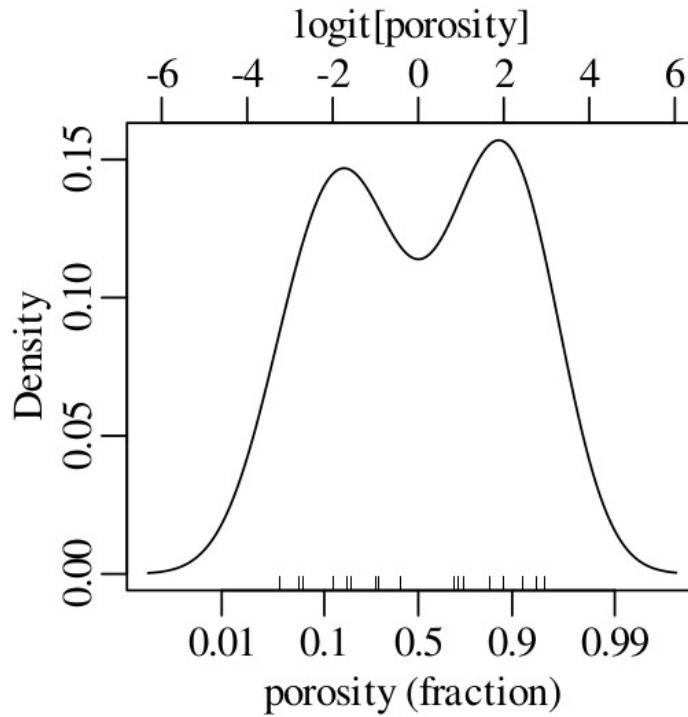
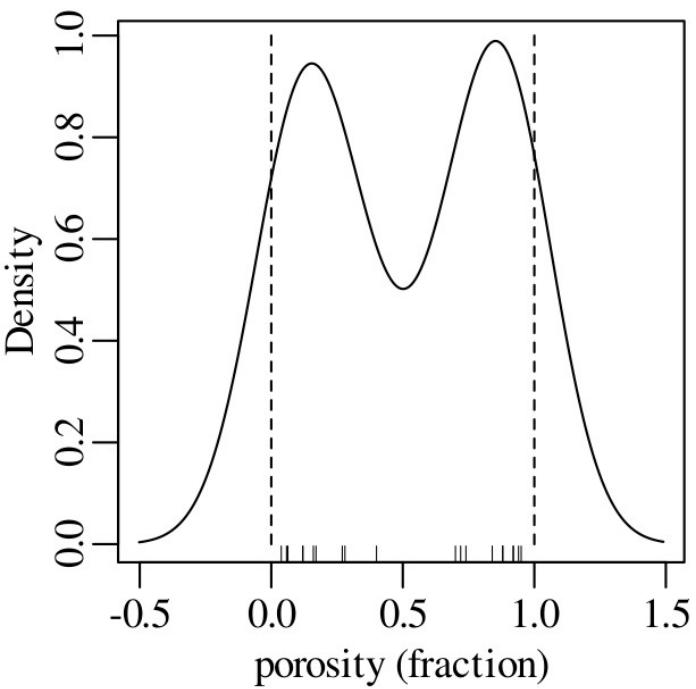




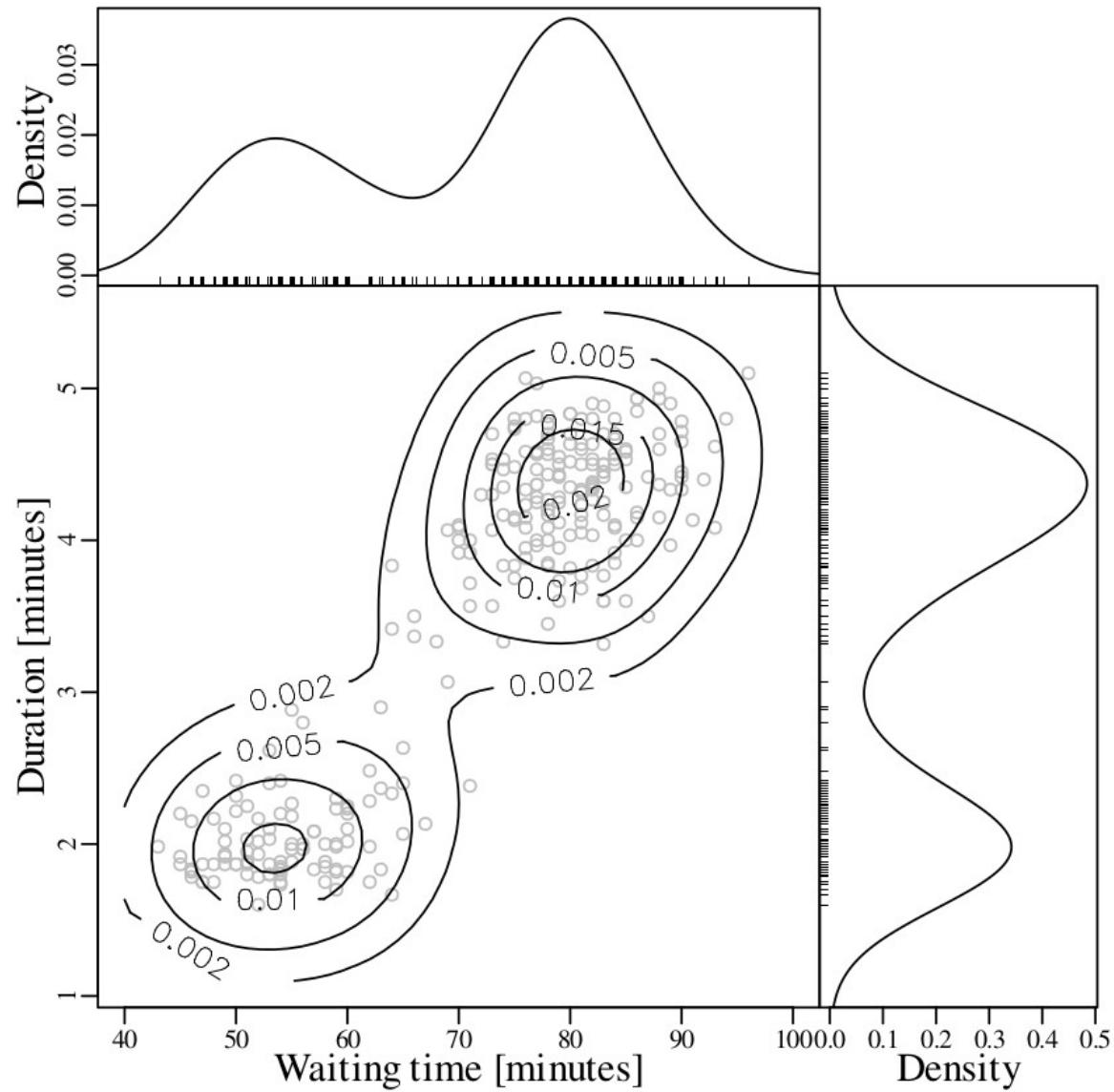
logistic transformation

$$u = \text{logit}(x) = \ln \left[ \frac{x}{1 - x} \right]$$

$$x = \text{logit}^{-1}(u) = \frac{\exp[u]}{\exp[u] + 1}$$

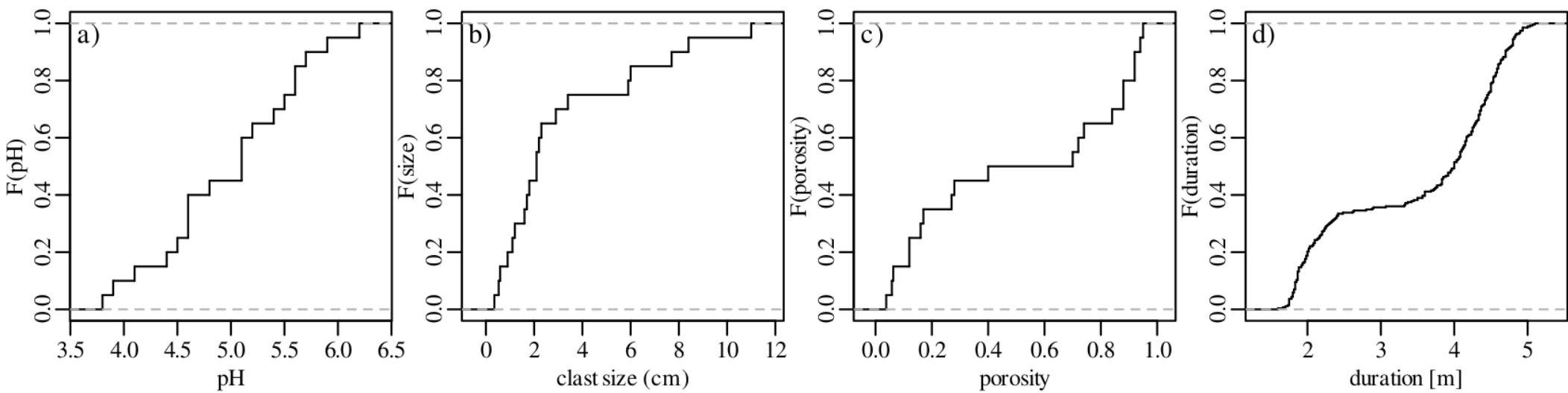


# Multivariate distributions



# Empirical cumulative distribution fuctions

$$F(x) = \sum_{i=1}^n 1(x_i < x)/n$$

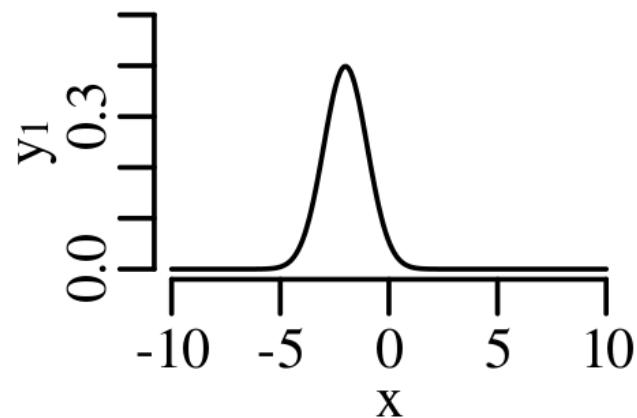


# Statistics for geoscientists

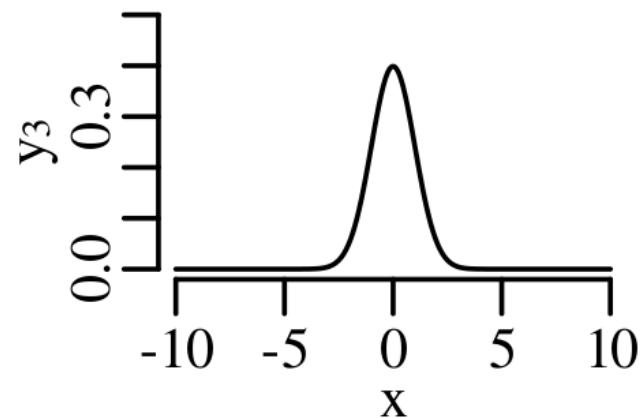
Summary statistics

| I    | II    |      | III  |      | IV    |      |       |                                                       |
|------|-------|------|------|------|-------|------|-------|-------------------------------------------------------|
| $x$  | $y$   | $x$  | $y$  | $x$  | $y$   | $x$  | $y$   |                                                       |
| 10.0 | 8.04  | 10.0 | 9.14 | 10.0 | 7.46  | 8.0  | 6.58  | – the mean of $x$ is 9                                |
| 8.0  | 6.95  | 8.0  | 8.14 | 8.0  | 6.77  | 8.0  | 5.76  |                                                       |
| 13.0 | 7.58  | 13.0 | 8.74 | 13.0 | 12.74 | 8.0  | 7.71  | – the variance of $x$ is 11                           |
| 9.0  | 8.81  | 9.0  | 8.77 | 9.0  | 7.11  | 8.0  | 8.84  |                                                       |
| 11.0 | 8.33  | 11.0 | 9.26 | 11.0 | 7.81  | 8.0  | 8.47  | – the mean of $y$ is 7.50                             |
| 14.0 | 9.96  | 14.0 | 8.10 | 14.0 | 8.84  | 8.0  | 7.04  |                                                       |
| 6.0  | 7.24  | 6.0  | 6.13 | 6.0  | 6.08  | 8.0  | 5.25  | – the variance of $y$ is 4.125                        |
| 4.0  | 4.26  | 4.0  | 3.10 | 4.0  | 5.39  | 19.0 | 12.50 |                                                       |
| 12.0 | 10.84 | 12.0 | 9.13 | 12.0 | 8.15  | 8.0  | 5.56  | – the correlation coefficient of $x$ and $y$ is 0.816 |
| 7.0  | 4.82  | 7.0  | 7.26 | 7.0  | 6.42  | 8.0  | 7.91  |                                                       |
| 5.0  | 5.68  | 5.0  | 4.74 | 5.0  | 5.73  | 8.0  | 6.89  | – the best fit line is given by $y = 3.00 + 0.500x$   |

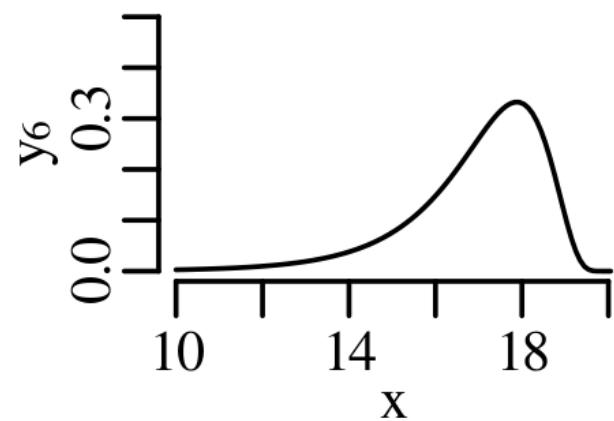
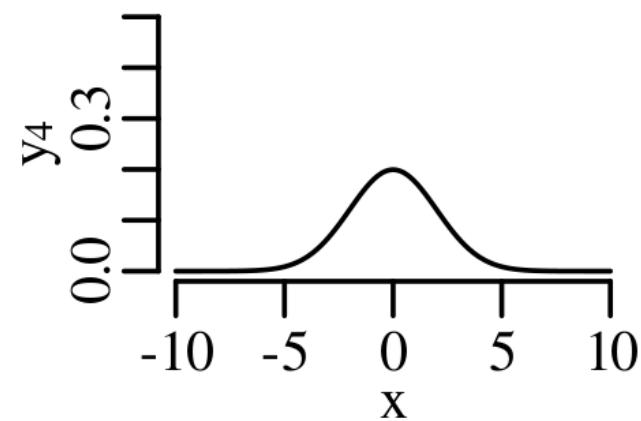
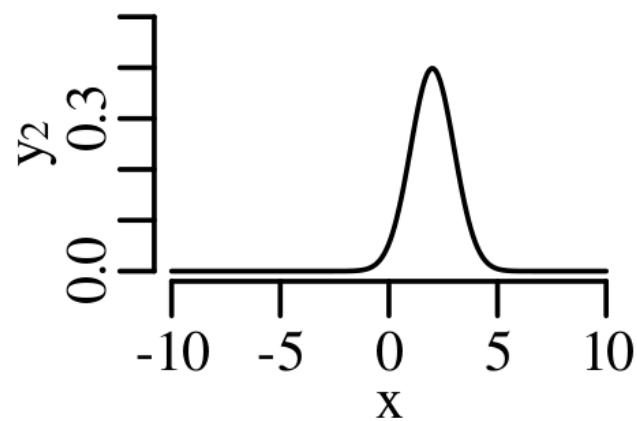
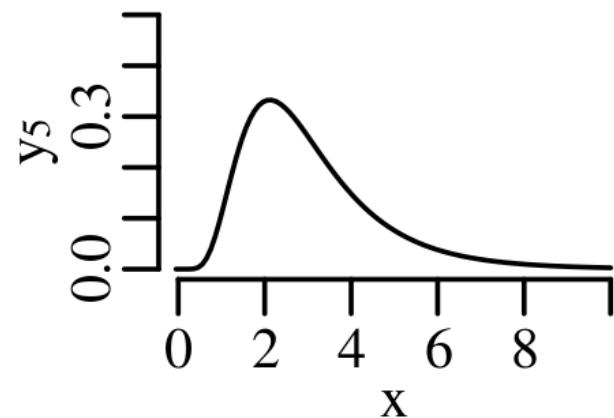
Location



Dispersion



Shape

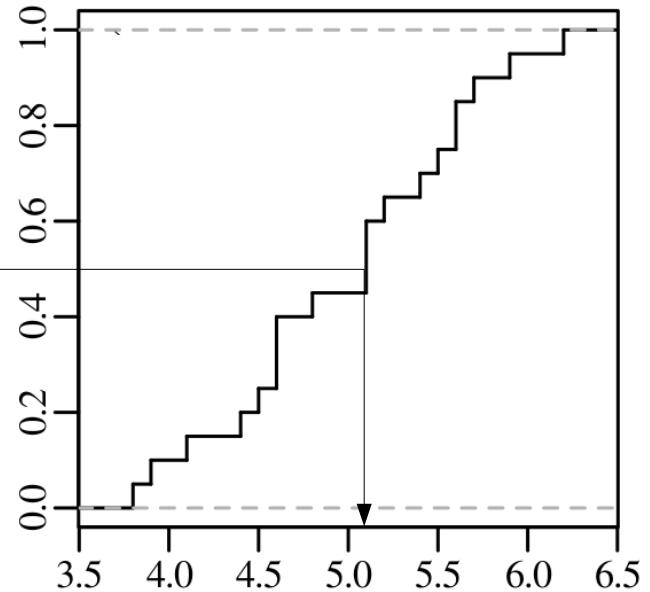
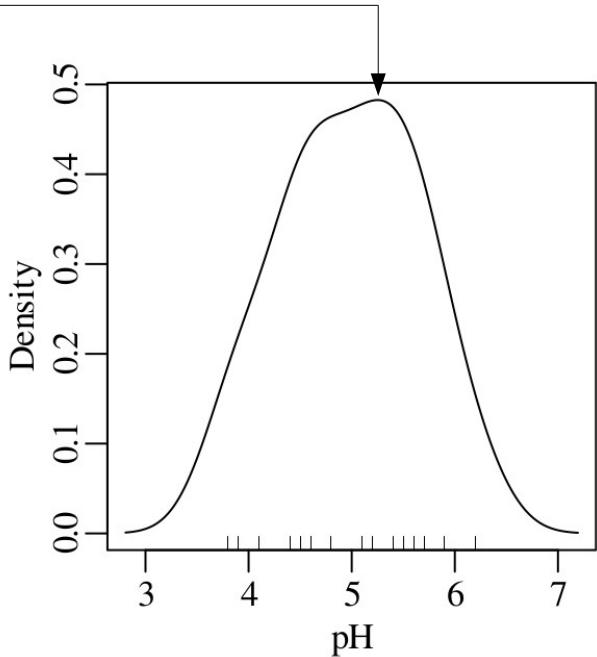


# Location

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

**Median**  $\text{med}(x) = x_1 < \dots < x_{n/2} < \dots < x_n$

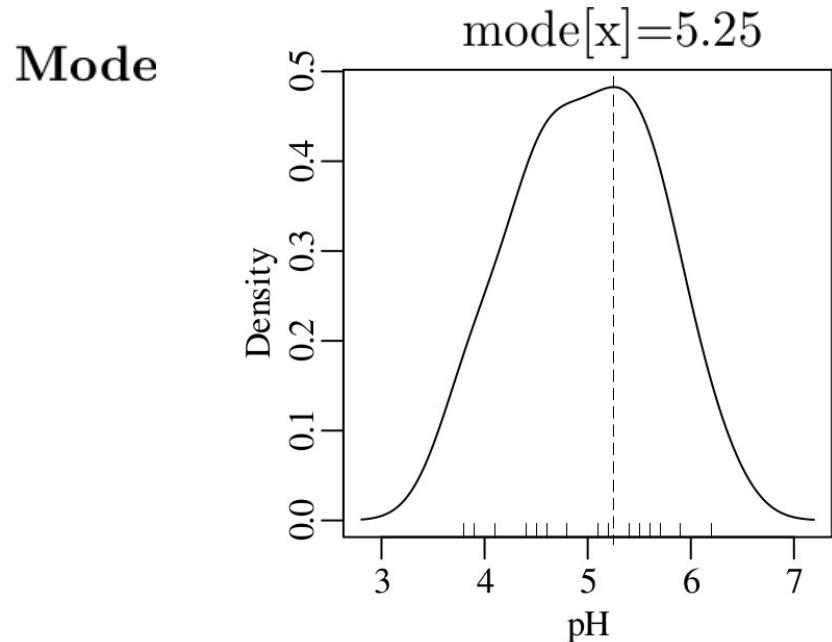
**Mode**

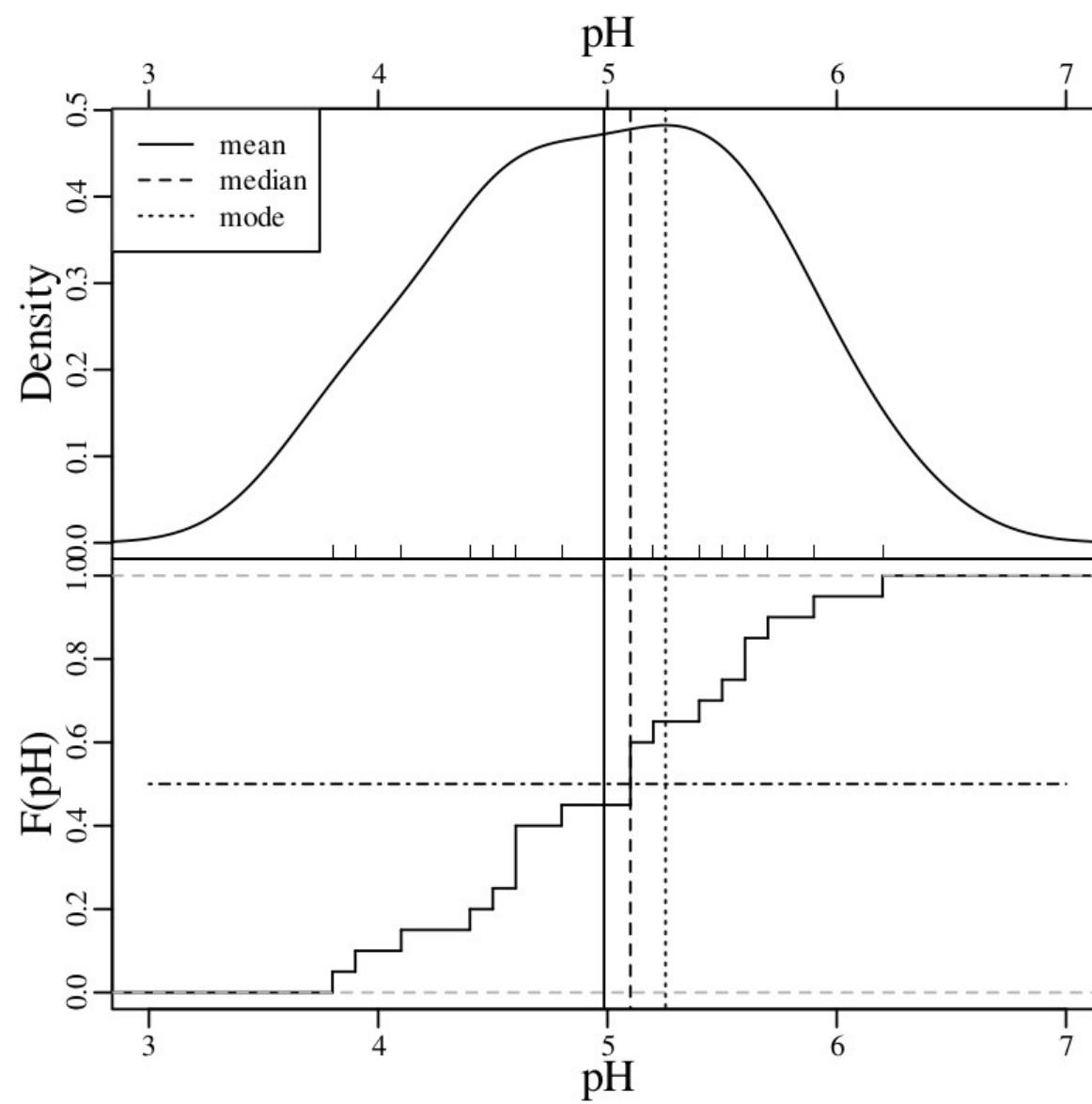


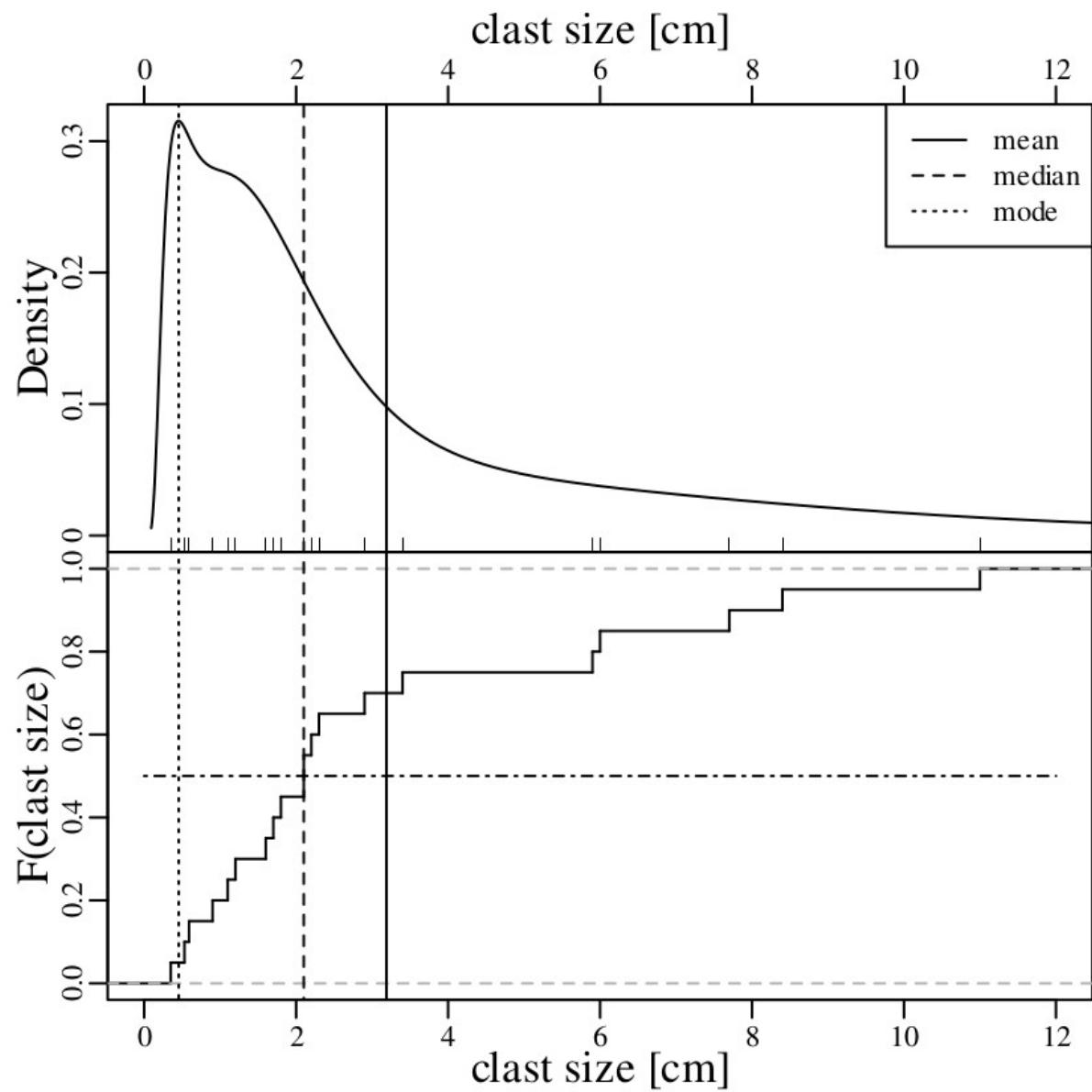
pH data    6.2, 4.4, 5.6, 5.2, 4.5, 5.4, 4.8, 5.9, 3.9, 3.8, 5.1, 4.1, 5.1, 5.5, 5.1, 4.6, 5.7, 4.6, 4.6, 5.6

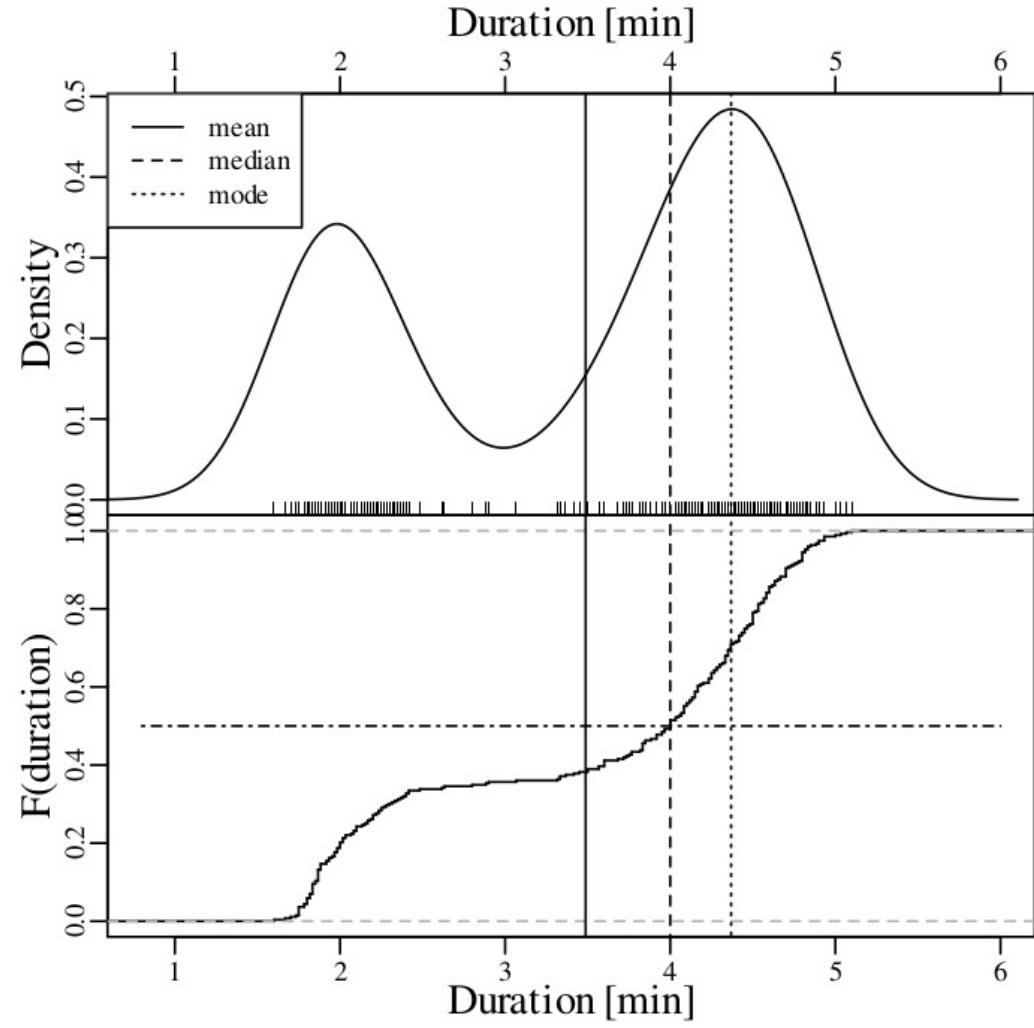
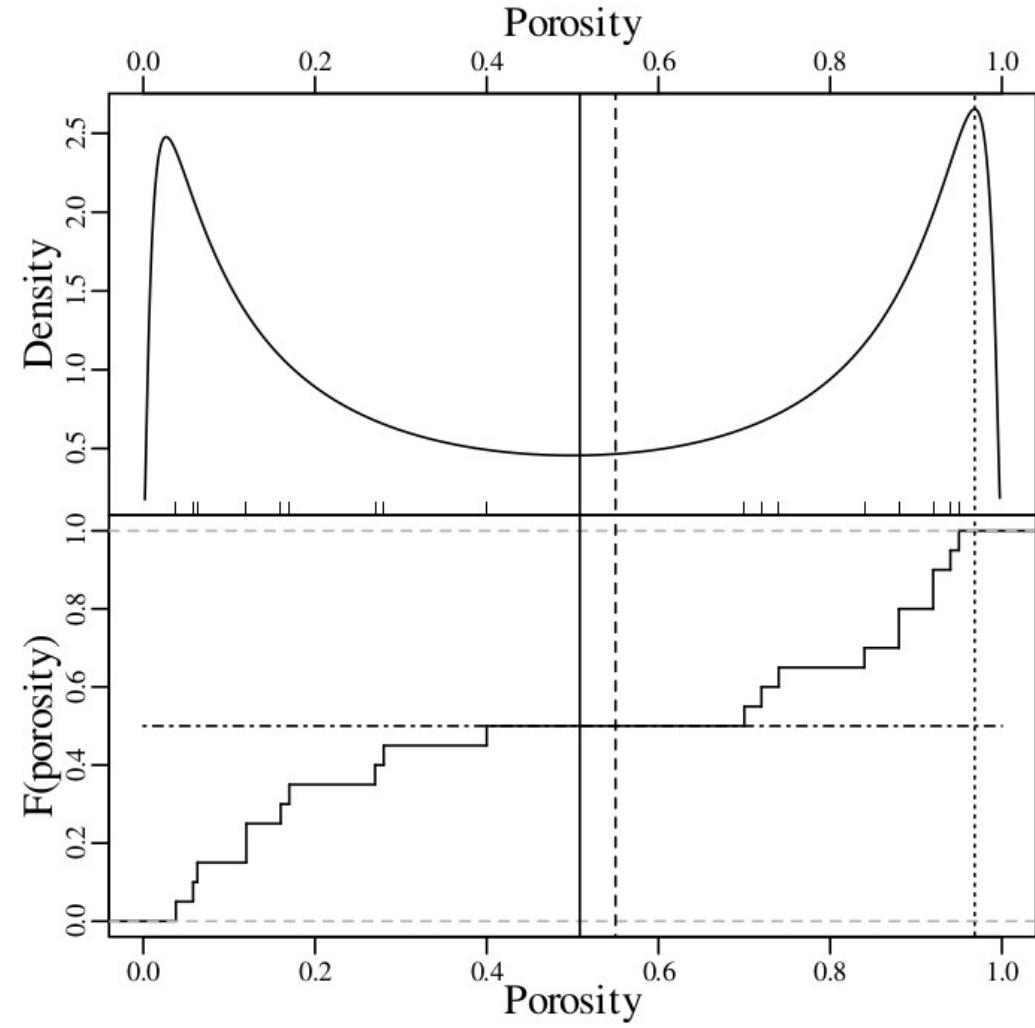
$$\text{Mean } \bar{x} = \frac{6.2 + 4.4 + 5.6 + 5.2 + 4.5 + 5.4 + 4.8 + 5.9 + 3.9 + 3.8 + 5.1 + 4.1 + 5.1 + 5.5 + 5.1 + 4.6 + 5.7 + 4.6 + 4.6 + 5.6}{20} = 5.00$$

Median    3.8, 3.9, 4.1, 4.4, 4.5, 4.6, 4.6, 4.6, 4.8, **5.1**, **5.1**, 5.1, 5.2, 5.4, 5.5, 5.6, 5.6, 5.7, 5.9, 6.2









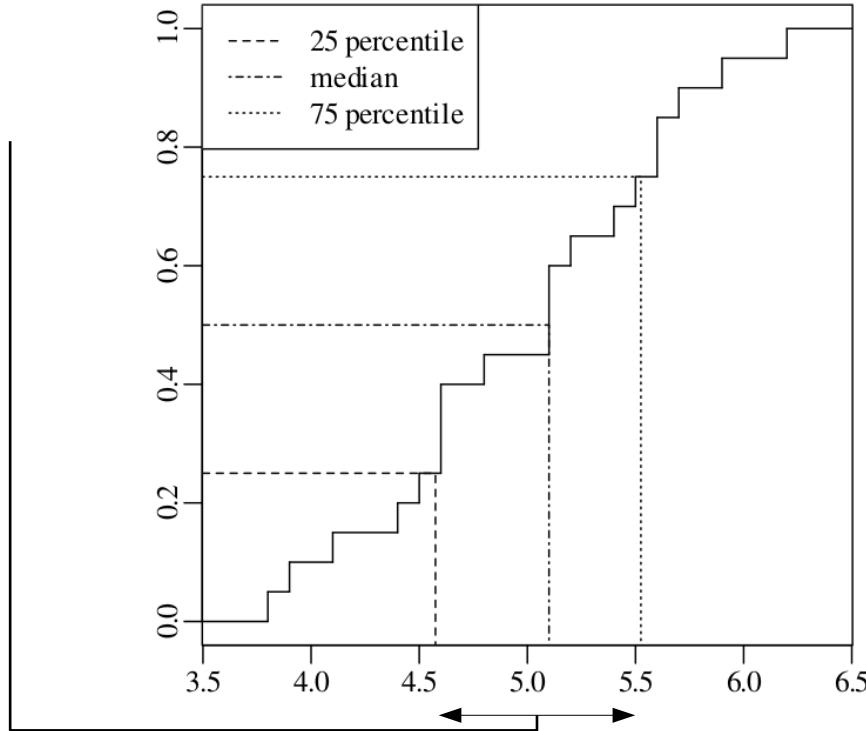
# Dispersion

## Standard deviation

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

Median Absolute Deviation  $\text{MAD} = \text{median}|x_i - \text{median}(x)|$

## Interquartile range



## Standard deviation

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

| $i$                                        | 1                              | 2    | 3   | 4    | 5    | 6   | 7    | 8   | 9    | 10   | 11   | 12   | 13   | 14  | 15   | 16   | 17  | 18   | 19   | 20  |
|--------------------------------------------|--------------------------------|------|-----|------|------|-----|------|-----|------|------|------|------|------|-----|------|------|-----|------|------|-----|
| $x_i$                                      | 6.2                            | 4.4  | 5.6 | 5.2  | 4.5  | 5.4 | 4.8  | 5.9 | 3.9  | 3.8  | 5.1  | 4.1  | 5.1  | 5.5 | 5.1  | 4.6  | 5.7 | 4.6  | 4.6  | 5.6 |
| $(x_i - \bar{x})$                          | 1.20                           | -.58 | .61 | .21  | -.49 | .42 | -.19 | .92 | -1.1 | -1.2 | .11  | -.89 | .11  | .51 | .11  | -.39 | .71 | -.39 | -.39 | .61 |
| $(x_i - \bar{x})^2$                        | 1.5                            | .34  | .38 | .046 | .24  | .17 | .034 | .84 | 1.2  | 1.4  | .013 | .78  | .013 | .27 | .013 | .15  | .51 | .15  | .15  | .38 |
| $\sum_{i=1}^{20} (x_i - \bar{x})^2 = 8.52$ | $s[x] = \sqrt{8.52/19} = 0.70$ |      |     |      |      |     |      |     |      |      |      |      |      |     |      |      |     |      |      |     |

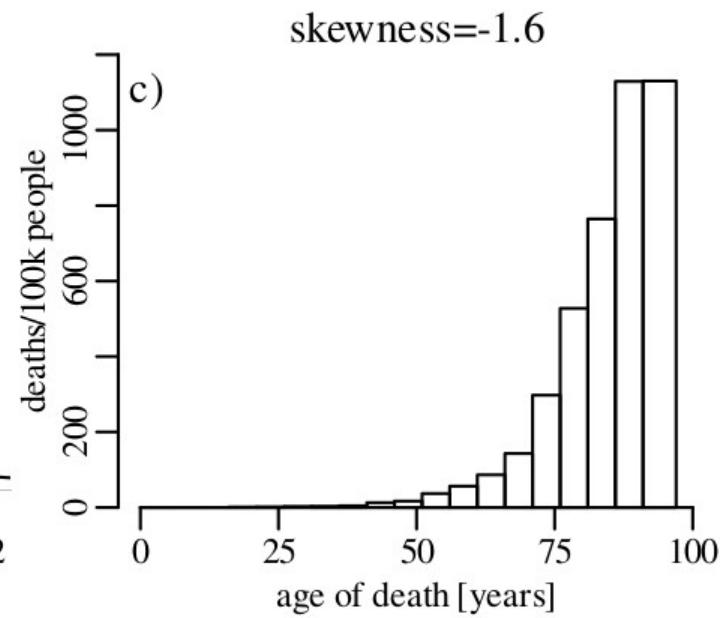
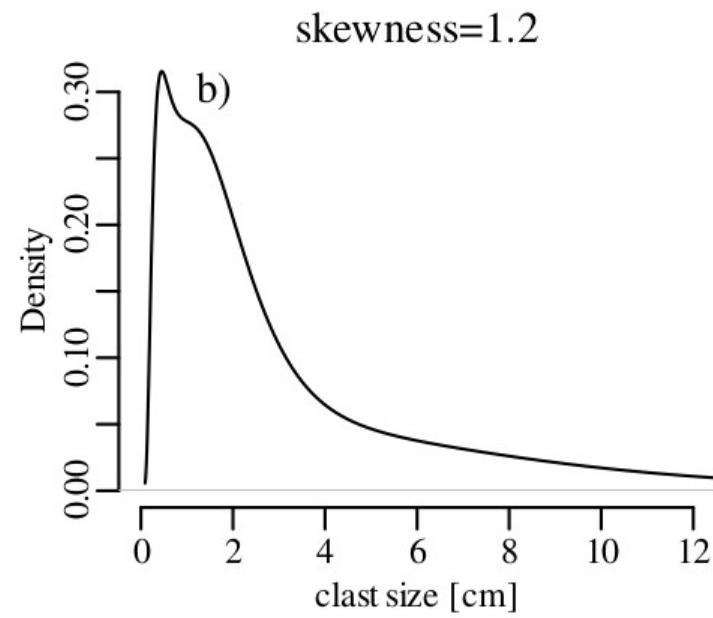
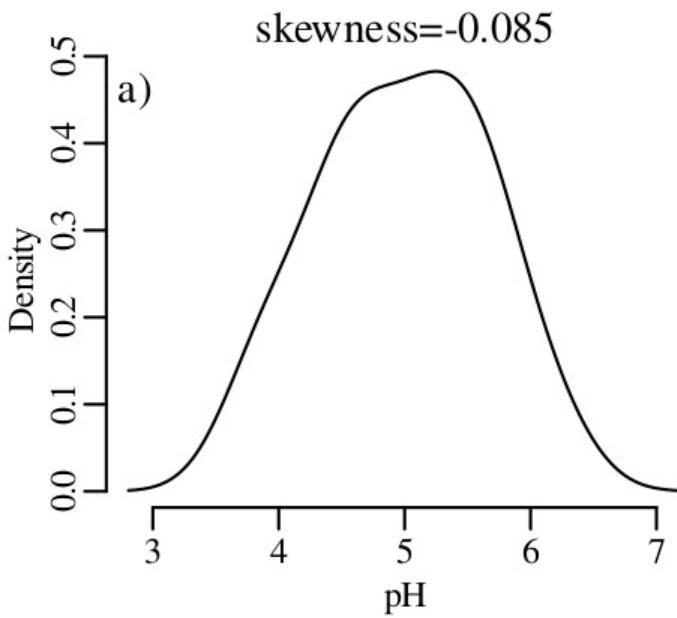
Interquartile range IQR =  $5.55 - 4.55 = 1.00$

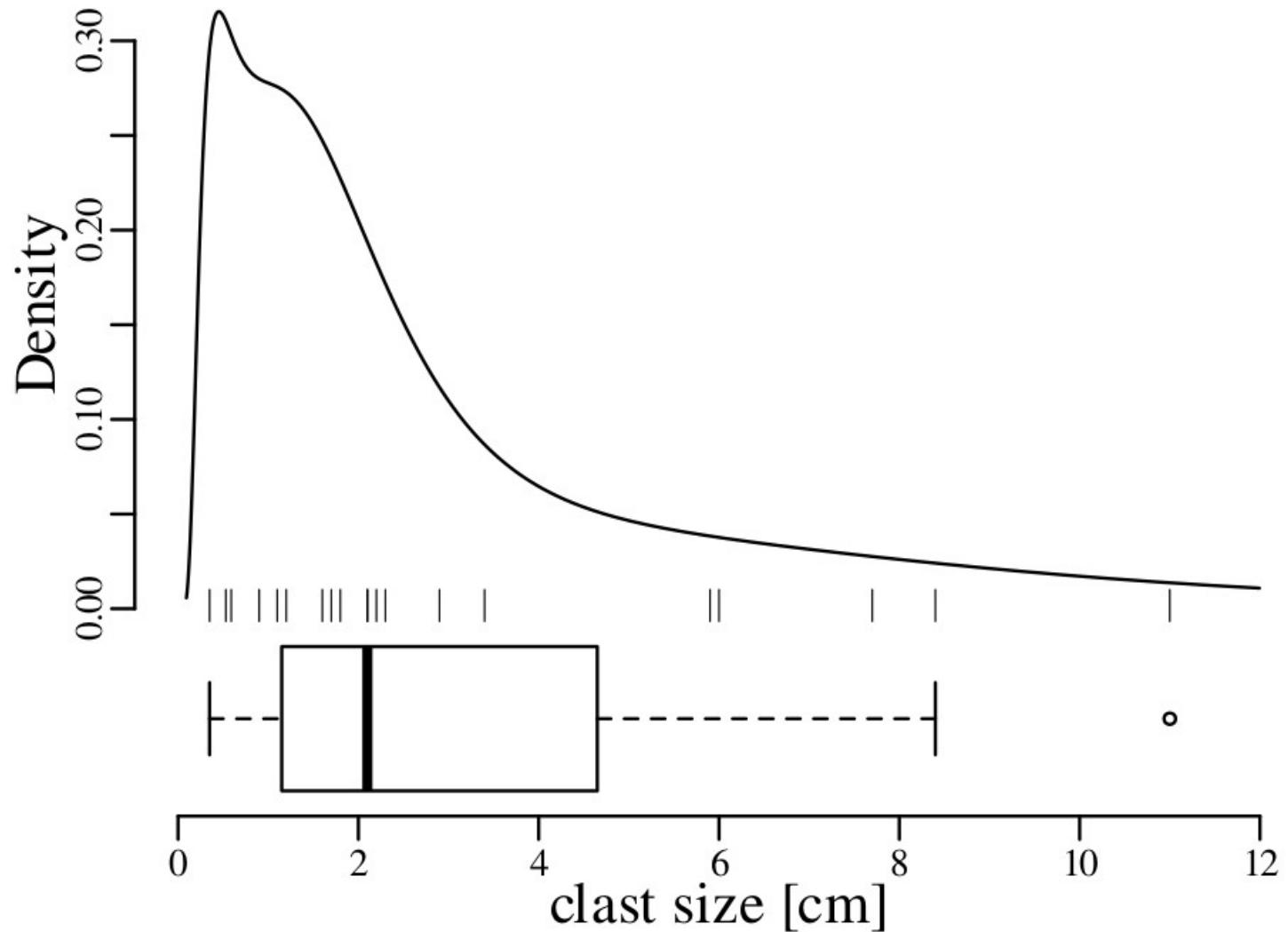
| $i$                     | 1    | 2    | 3    | 4    | 5    | 6    | 7    | 8    | 9    | 10  | 11  | 12  | 13  | 14  | 15  | 16  | 17  | 18  | 19  | 20  |
|-------------------------|------|------|------|------|------|------|------|------|------|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|
| $x_i$                   | 3.8  | 3.9  | 4.1  | 4.4  | 4.5  | 4.6  | 4.6  | 4.6  | 4.8  | 5.1 | 5.1 | 5.1 | 5.2 | 5.4 | 5.5 | 5.6 | 5.6 | 5.7 | 5.9 | 6.2 |
| $x_i - \text{med}(x)$   | -1.3 | -1.2 | -1.0 | -0.7 | -0.6 | -0.5 | -0.5 | -0.5 | -0.3 | 0.0 | 0.0 | 0.0 | 0.1 | 0.3 | 0.4 | 0.5 | 0.5 | 0.6 | 0.8 | 1.1 |
| $ x_i - \text{med}(x) $ | 1.3  | 1.2  | 1.0  | 0.7  | 0.6  | 0.5  | 0.5  | 0.5  | 0.3  | 0.0 | 0.0 | 0.0 | 0.1 | 0.3 | 0.4 | 0.5 | 0.5 | 0.6 | 0.8 | 1.1 |
| sorted                  | 0.0  | 0.0  | 0.0  | 0.1  | 0.3  | 0.3  | 0.4  | 0.5  | 0.5  | 0.5 | 0.5 | 0.6 | 0.6 | 0.7 | 0.8 | 1.0 | 1.1 | 1.2 | 1.3 |     |

Median Absolute Deviation MAD =  $\text{median}|x_i - \text{median}(x)|$

**skewness**

$$\text{skew}(x) = \frac{1}{n \cdot s[x]^3} \sum_{i=1}^n (x_i - \bar{x})^3$$





# Statistics for geoscientists

## Probability

$$P(A) = \frac{\text{the number of ways } A \text{ can occur}}{\text{the total number of outcomes}}$$

$$P(A) = \frac{\text{the number of ways } A \text{ can occur}}{\text{the total number of outcomes}}$$

$$\frac{\{H\}}{\{H\}\{T\}} = \frac{1}{2} = 0.5$$

$$P(A) = \frac{\text{the number of ways } A \text{ can occur}}{\text{the total number of outcomes}}$$

$$P(2 \times H \cap 1 \times T) = \frac{\{THH\}\{HTH\}\{HHT\}}{\{HHH\}\{THH\}\{HTH\}\{HHT\}\{TTH\}\{THT\}\{HTT\}\{TTT\}} = \frac{3}{8}$$

$$P(A) = \frac{\text{the number of ways } A \text{ can occur}}{\text{the total number of outcomes}}$$

**multiplicative rule**

$$P(\{2 \times H \cap 1 \times T\} \cap \{1 \times \begin{array}{|c|} \hline \bullet \\ \hline \end{array} \cap 1 \times \begin{array}{|c|c|} \hline \bullet & \bullet \\ \hline \bullet & \bullet \\ \hline \end{array}\}) = \frac{3}{8} \frac{1}{18} = \frac{3}{144} = 0.021$$

**additive rule**

$$P(\{2 \times H \cap 1 \times T\} \cup \{3 \times H\}) = \frac{3}{8} + \frac{1}{8} = \frac{4}{8} = 0.5$$

# Permutations

1    2    3    4    5    6    7    8    9    10    11    12    13    14    15    16    17    18    19    20

1    2    3    4    5    6    7    8    9    10    11    12    13    14    15    16    17    18    19    20

sampling with replacement    5    2    12    19    10    5    3    19    11    4

$$\overbrace{n \times n \times \dots \times n}^{k \text{ times}} = n^k$$

1    2    3    4    5    6    7    8    9    10    11    12    13    14    15    16    17    18    19    20

sampling with replacement     $\boxed{5}$     2    12     $\boxed{19}$     10     $\boxed{5}$     3     $\boxed{19}$     11    4

$$\overbrace{n \times n \times \dots \times n}^{k \text{ times}} = n^k$$

1    2    3    4    5    6    7    8    9    10    11    12    13    14    15    16    17    18    19    20

**sampling without replacement**    15    8    13    5    4    19    7    1    20    10

$$n \times (n - 1) \times (n - 2) \times \dots \times (n - k + 1) = \frac{n!}{(n - k)!}$$

*what is the probability that two students in a classroom of  $k$  celebrate their birthdays on the same day?*

*what is the probability that two students in a classroom of  $k$  celebrate their birthdays on the same day?*

$365^k$  possible combinations of birthdays (sampling with replacement)

$365!/(365-k)!$  of these combinations do not overlap (sampling without replacement)

*what is the probability that two students in a classroom of  $k$  celebrate their birthdays on the same day?*

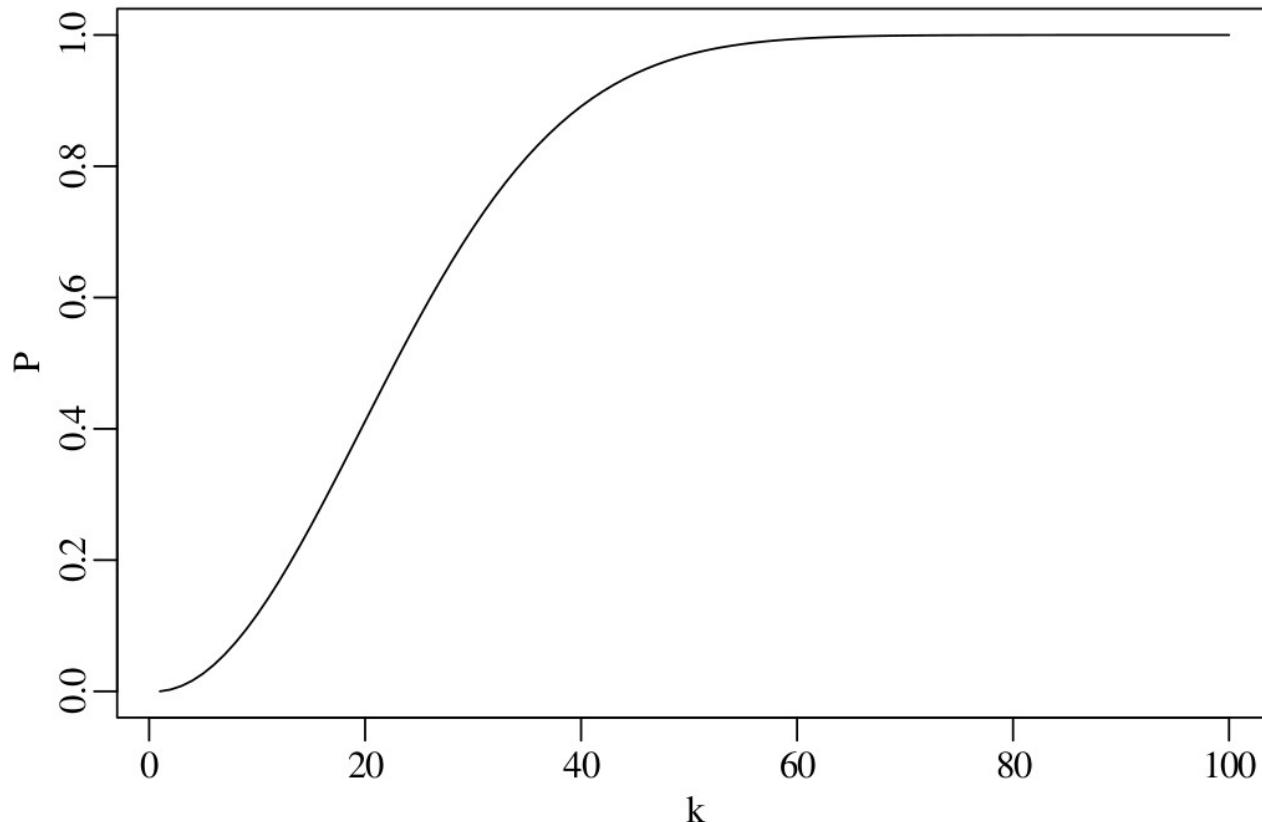
$365^k$  possible combinations of birthdays (sampling with replacement)

$365!/(365-k)!$  of these combinations do not overlap (sampling without replacement)

$$P(\text{no overlapping birthdays}) = \frac{365!}{(365 - k)!365^k}$$

$$P(> 1 \text{ overlapping birthdays}) = 1 - \frac{365!}{(365 - k)!365^k}$$

$$P(> 1 \text{ overlapping birthdays}) = 1 - \frac{365!}{(365 - k)!365^k}$$

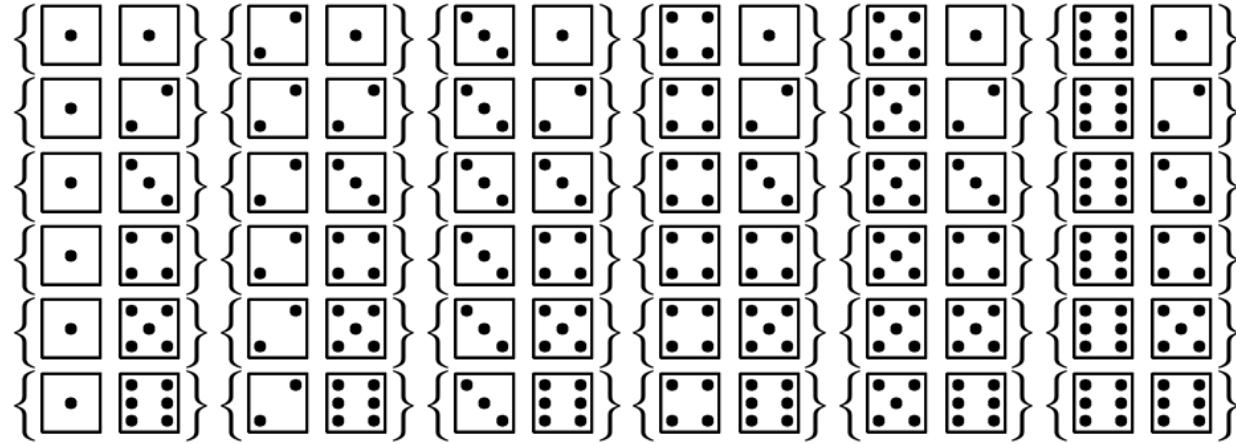


If  $k = 23$ , then  $P(> 1 \text{ overlapping birthdays}) = 0.507$

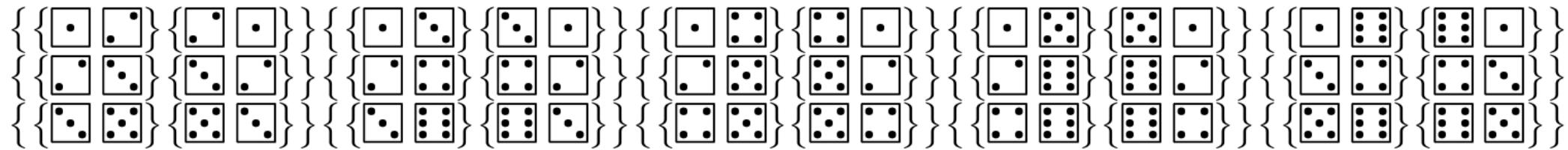
# Combinations

$\{HHH\}\{THH\}\{HTH\}\{HHT\}\{TTH\}\{THT\}\{HTT\}\{TTT\}$

duplicates:  $\{\{HTT\}\{THT\}\{HHT\}\}$  and  $\{\{THH\}\{HTH\}\{TTH\}\}$



duplicates:



$$(\# \text{ ordered samples}) = (\# \text{ unordered samples}) \times (\# \text{ ways to order the samples})$$

$$(\# \text{ unordered samples}) = \frac{(\# \text{ ordered samples})}{(\# \text{ ways to order the samples})}$$

$$= \frac{n!/(n-k)! \text{ ways to select } k \text{ objects from a collection of } n}{k! \text{ ways to order these } k \text{ objects}} = \frac{n!}{(n-k)!k!}$$

$$\binom{n}{k} = \frac{n!}{(n-k)!k!}$$

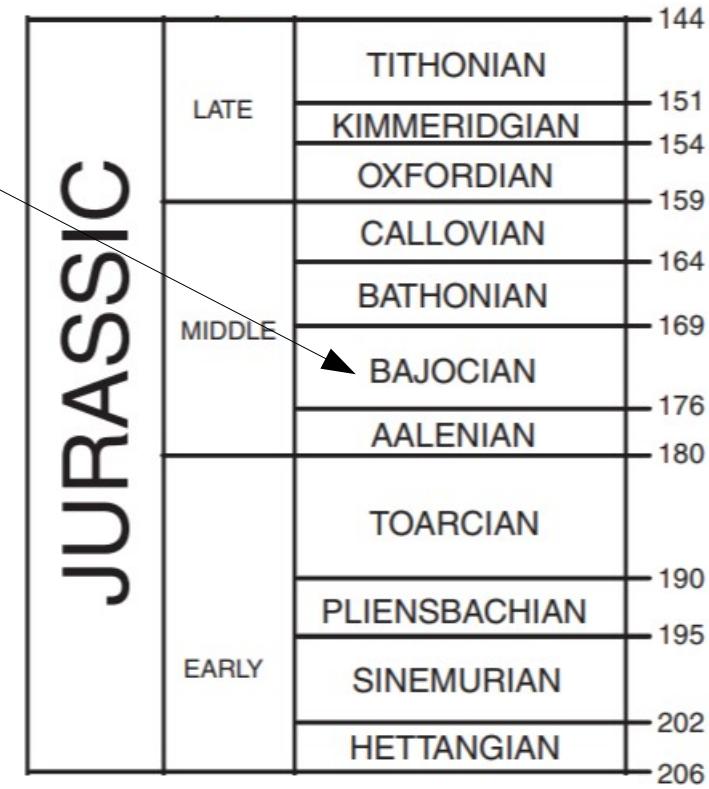
$$\binom{n}{k} = \frac{n!}{(n-k)!k!}$$

$$\{THH\}\{HTH\}\{HHT\} \hspace{10cm} \left\{\begin{array}{|c|c|}\hline \bullet & \bullet \\ \hline \bullet & \end{array}\right\} \left\{\begin{array}{|c|c|}\hline \bullet & \bullet \\ \hline \bullet & \bullet \\ \hline \end{array}\right\} \left\{\begin{array}{|c|}\hline \bullet \\ \hline \end{array}\right\}$$

$$\binom{3}{2}=\frac{3!}{1!2!}=\frac{6}{2}=3 \hspace{10cm} \binom{2}{1}=\frac{2!}{1!1!}=2$$

# Conditional probability

$P(A|B) = \text{“The conditional probability of } A \text{ given } B\text{”}$



## multiplication law

$P(A|B)$  = “The conditional probability of  $A$  given  $B$ ”

$$P(A \cap B) = P(A|B)P(B) = P(B|A)P(A) = P(B \cap A)$$

$$\left. \begin{array}{l} P(B) = 0.7 \\ P(A|B) = 0.2 \end{array} \right\} 0.7 \times 0.2 = 14\%$$

## law of total probability

$P(A|B)$  = “The conditional probability of  $A$  given  $B$ ”

$$P(A) = \sum_{i=1}^n P(A|B_i)P(B_i)$$

## law of total probability

$P(A|B)$  = “The conditional probability of  $A$  given  $B$ ”

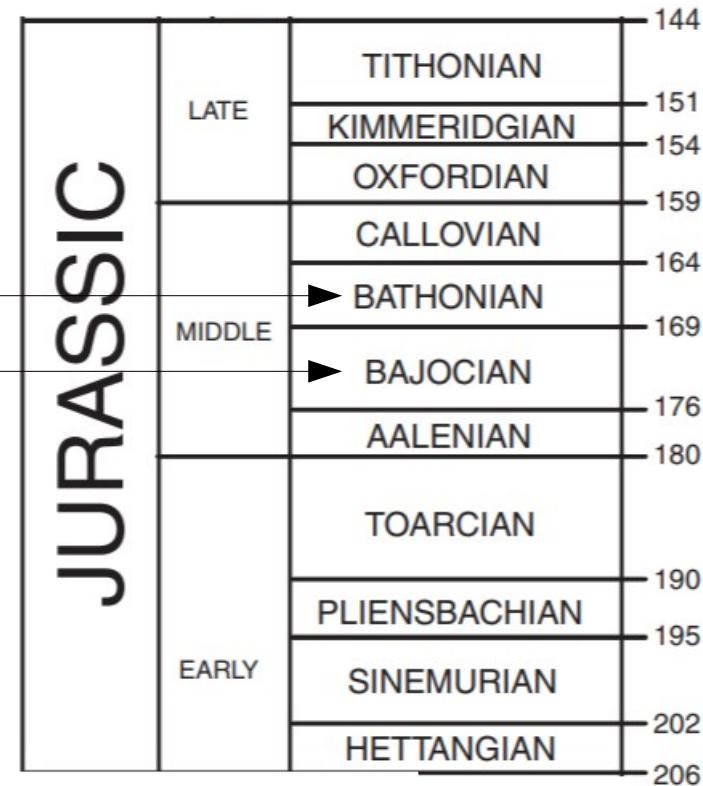
$$P(A) = \sum_{i=1}^n P(A|B_i)P(B_i)$$

$$P(B_2) = 0.3$$

$$P(B_1) = 0.7$$

$$P(A|B_2) = 0.5$$

$$P(A|B_1) = 0.2$$



$$P(A) = P(A|B_1)P(B_1) + P(A|B_2)P(B_2) = 0.2 \times 0.7 + 0.5 \times 0.3 = 0.29$$

**multiplication law**  $P(A \cap B) = P(A|B)P(B) = P(B|A)P(A) = P(B \cap A)$

**Bayes' Rule** 
$$P(B|A) = \frac{P(A|B)P(B)}{P(A)}$$

**multiplication law**  $P(A \cap B) = P(A|B)P(B) = P(B|A)P(A) = P(B \cap A)$

**law of total probability**  $P(A) = \sum_{i=1}^n P(A|B_i)P(B_i)$

**Bayes' Rule**  $P(B|A) = \frac{P(A|B)P(B)}{P(A)}$

$P(B_j|A) = \frac{P(A|B_j)P(B_j)}{\sum_{i=1}^n P(A|B_i)P(B_i)}$

**Bayes' Rule**

$$P(B_j|A) = \frac{P(A|B_j)P(B_j)}{\sum_{i=1}^n P(A|B_i)P(B_i)}$$

Suppose that we have found an ammonite fossil in the river bed. What is its likely age?

$$P(B_1|A) = \frac{P(A|B_1)P(B_1)}{P(A|B_1)P(B_1) + P(A|B_2)P(B_2)}$$

$$= \frac{0.2 \times 0.7}{0.2 \times 0.7 + 0.5 \times 0.3} = 0.48$$

# Statistics for geoscientists

The binomial distribution

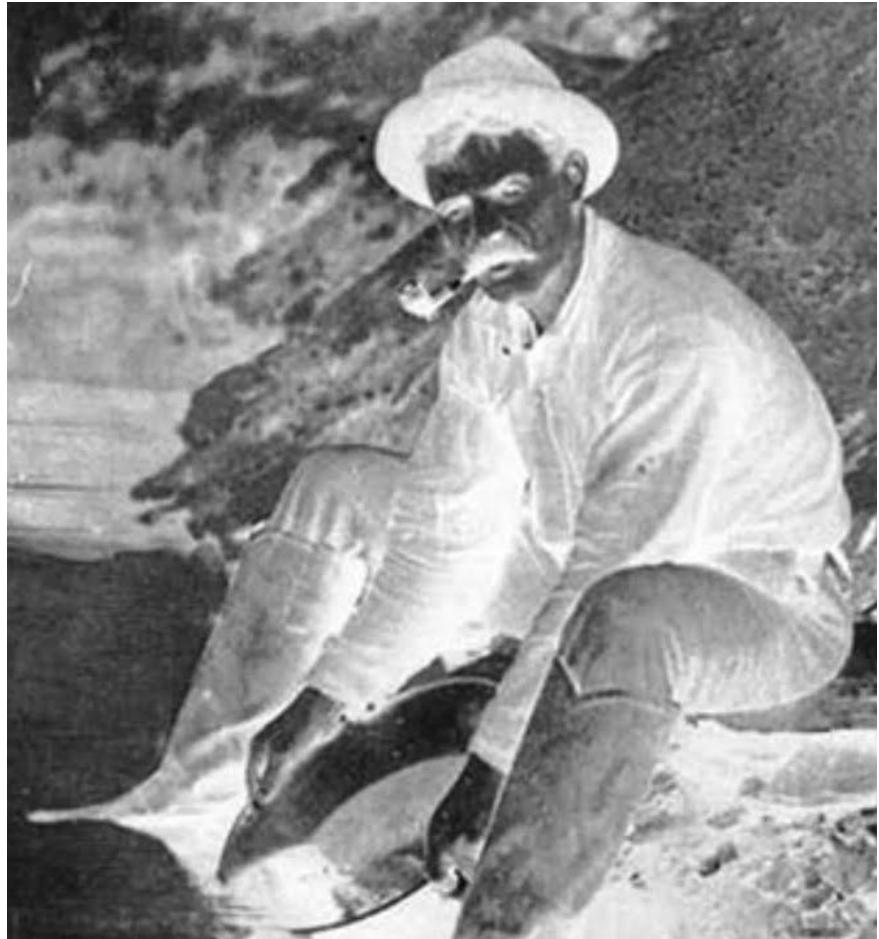
## Bernoulli variable

1. a coin may land on its head (1) or tail (0);
2. a die may land on  (1) or not (0);
3. a ‘wildcat’ exploration well may find petroleum (1) or be dry (0).

five gold claims

$$p = 2/3$$

$$P(0 \times \text{gold}) = P(00000) = (1/3)^5 = 0.0041$$



five gold claims               $p = 2/3$

$$P(0 \times \text{gold}) = P(00000) = (1/3)^5 = 0.0041$$

$$P(1 \times \text{gold}) = P(10000) + P(01000) + P(00100) + P(00010) + P(00001)$$

$$P(10000) = (2/3)(1/3)^4 = 0.0082$$

$$P(01000) = (1/3)(2/3)(1/3)^3 = 0.0082$$

$$P(00100) = (1/3)^2(2/3)(1/3)^2 = 0.0082$$

$$P(00010) = (1/3)^3(2/3)(1/3) = 0.0082$$

$$P(00001) = (1/3)^4(2/3) = 0.0082$$

$$P(1 \times \text{gold}) = \binom{5}{1} (2/3)(1/3)^4 = 5 \times 0.0082 = 0.041$$

$$\text{five gold claims} \quad p = 2/3$$

$$P(0 \times \text{gold}) = P(00000) = (1/3)^5 = 0.0041$$

$$P(1 \times \text{gold}) = \binom{5}{1} (2/3)(1/3)^4 = 5 \times 0.0082 = 0.041$$

$$P(2 \times \text{gold}) = \binom{5}{2} (2/3)^2 (1/3)^3 = 10 \times 0.016 = 0.16$$

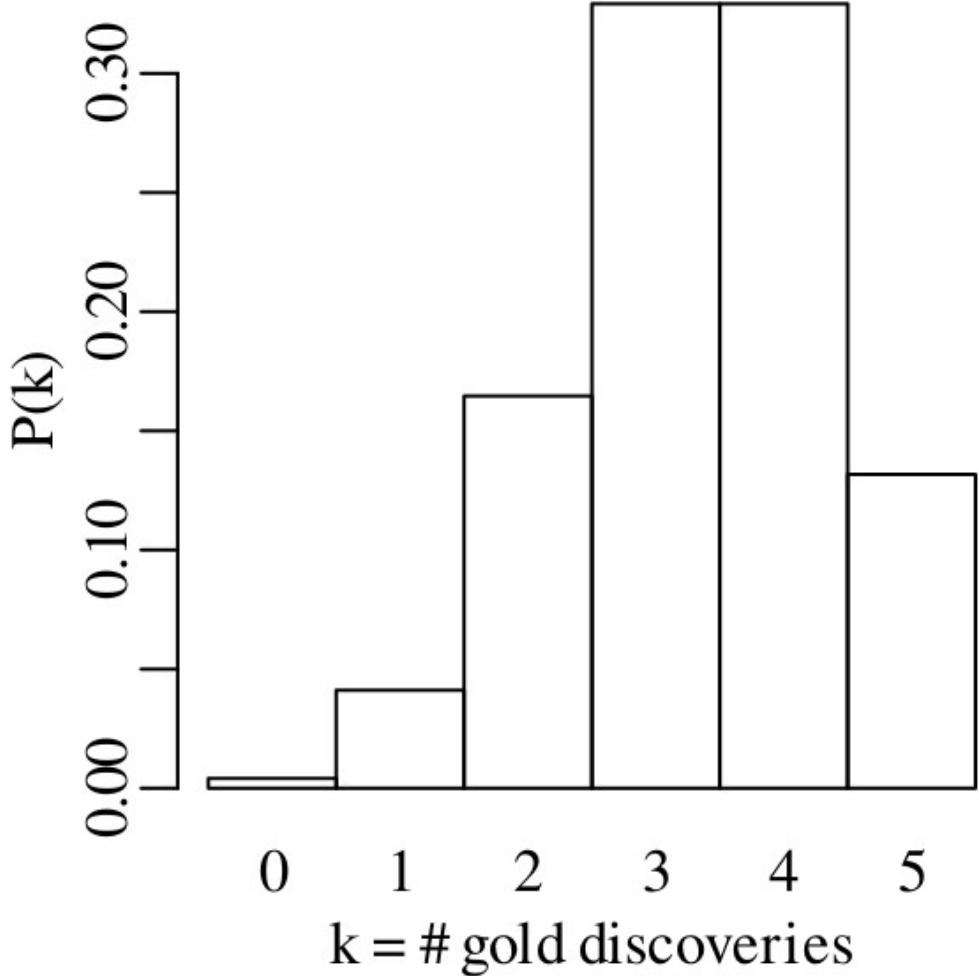
$$P(3 \times \text{gold}) = \binom{5}{3} (2/3)^3 (1/3)^2 = 10 \times 0.033 = 0.33$$

$$P(4 \times \text{gold}) = \binom{5}{4} (2/3)^4 (1/3) = 5 \times 0.066 = 0.33$$

$$P(5 \times \text{gold}) = (2/3)^5 = 0.13$$

## probability mass function (PMF)

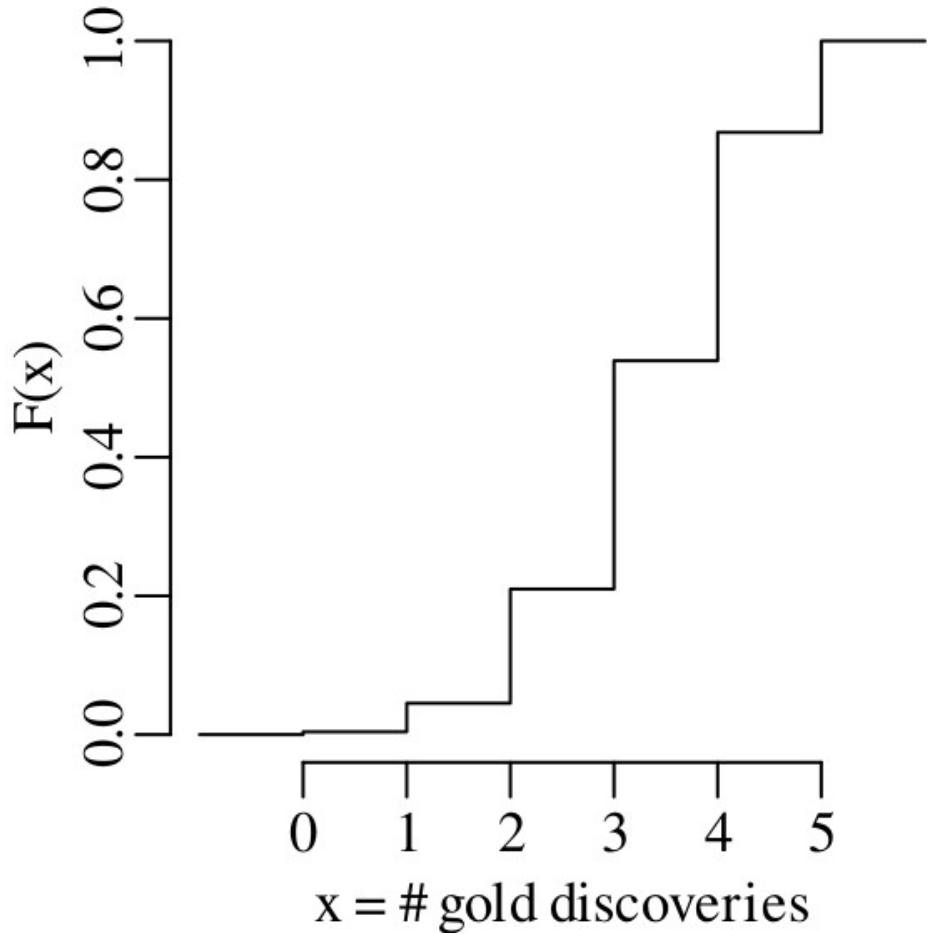
$$P(k|n, p) = \binom{n}{k} p^k (1-p)^{n-k}$$



## cumulative distribution function (CDF)

$$P(k|n, p) = \binom{n}{k} p^k (1-p)^{n-k}$$

$$F(x) = P(X \leq x)$$



# method of maximum likelihood

$$P(k|n, p) = \binom{n}{k} p^k (1-p)^{n-k}$$

$$\mathcal{L}(p|n, k) = \binom{n}{k} p^k (1-p)^{n-k}$$

$$\hat{p} = \frac{k}{n} \quad \text{e.g.} \quad \hat{p} = \frac{2}{5} = 0.4$$

# Hypothesis tests

$$\hat{p} = 2/5 \quad \longleftrightarrow \quad p = 2/3$$

1. Formulate two hypotheses:

$$H_0 \text{ (null hypothesis)} \qquad \qquad p = 2/3$$

$$H_a \text{ (alternative hypothesis):} \qquad \qquad p < 2/3$$

2. Calculate the **test statistic**  $T^-(k)$
3. Determine the **null distribution** of  $T$  under  $H_0$ .

| $k$           | 0      | 1      | 2      | 3      | 4      | 5      |
|---------------|--------|--------|--------|--------|--------|--------|
| $P(T = k)$    | 0.0041 | 0.0411 | 0.1646 | 0.3292 | 0.3292 | 0.1317 |
| $P(T \leq k)$ | 0.0041 | 0.0453 | 0.2099 | 0.5391 | 0.8683 | 1.0000 |


**p-value**

4. Choose a **significance level** ( $\alpha = 0.05$ )

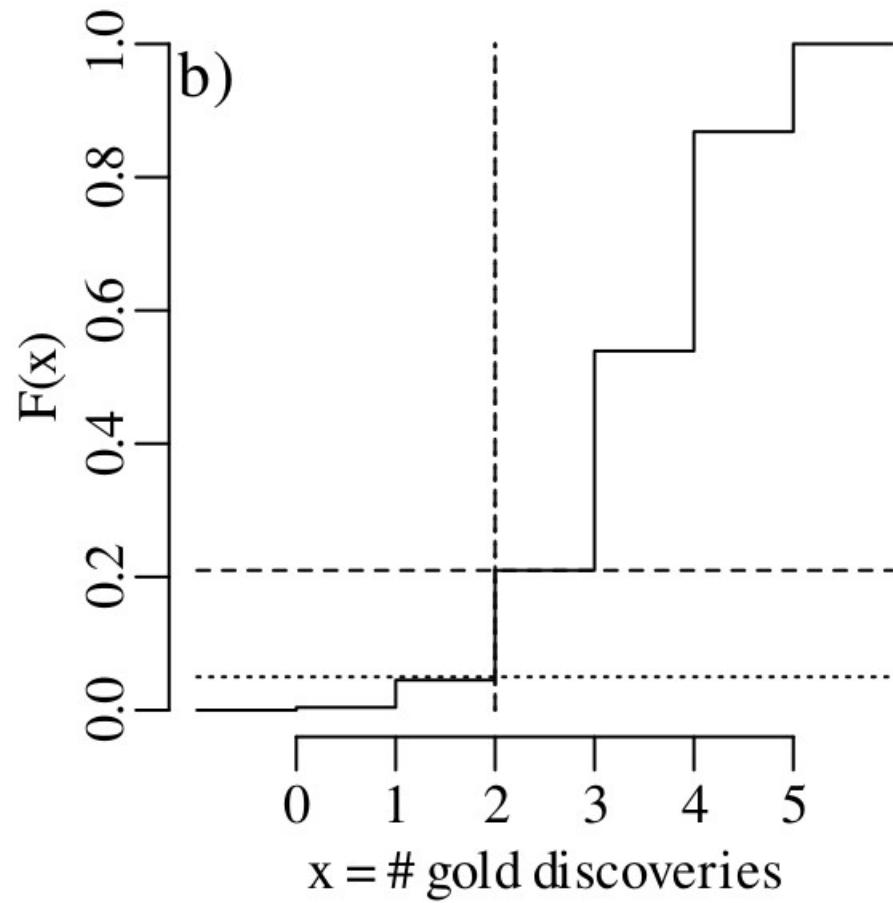
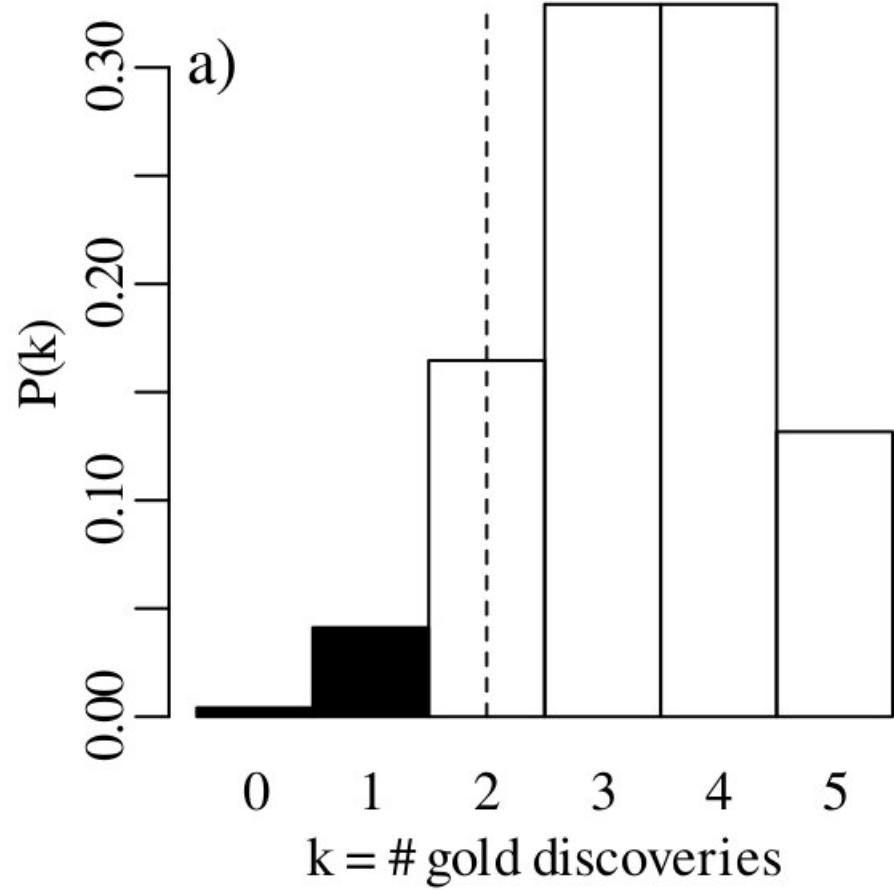
5. Mark the **rejection region**

| k             | 0             | 1             | 2             | 3      | 4      | 5      |
|---------------|---------------|---------------|---------------|--------|--------|--------|
| $P(T = k)$    | 0.0041        | 0.0411        | 0.1646        | 0.3292 | 0.3292 | 0.1317 |
| $P(T \leq k)$ | <b>0.0041</b> | <b>0.0453</b> | <i>0.2099</i> | 0.5391 | 0.8683 | 1.0000 |

rejection region     $R = \{0, 1\}$

6. Reach a **decision**  $k = 2 \notin R$

7. Alternatively,               $0.2099 > \alpha$



$$\hat{p} = 2/5 \quad \longleftrightarrow \quad p = 2/3$$

**one-sided** hypothesis test

$$H_0 \text{ (null hypothesis)} \qquad \qquad \qquad p = 2/3$$

$$H_a \text{ (alternative hypothesis):} \qquad \qquad \qquad p < 2/3$$

**two-sided** hypothesis test

$$H_0 \text{ (null hypothesis)} \qquad \qquad \qquad p = 2/3$$

$$H_a \text{ (alternative hypothesis):} \qquad \qquad \qquad p \neq 2/3$$

2. Calculate the **test statistic**  $T^-(k)$
3. Determine the **null distribution** of  $T$  under  $H_0$ .

| $k$           | 0      | 1      | 2      | 3      | 4      | 5      |
|---------------|--------|--------|--------|--------|--------|--------|
| $P(T = k)$    | 0.0041 | 0.0411 | 0.1646 | 0.3292 | 0.3292 | 0.1317 |
| $P(T \leq k)$ | 0.0041 | 0.0453 | 0.2099 | 0.5391 | 0.8683 | 1.0000 |
| $P(T \geq k)$ | 1.000  | 0.9959 | 0.9547 | 0.7901 | 0.4609 | 0.1317 |

4. Choose a **significance level** ( $\alpha = 0.05$ )

but evaluated twice at  $\alpha/2$

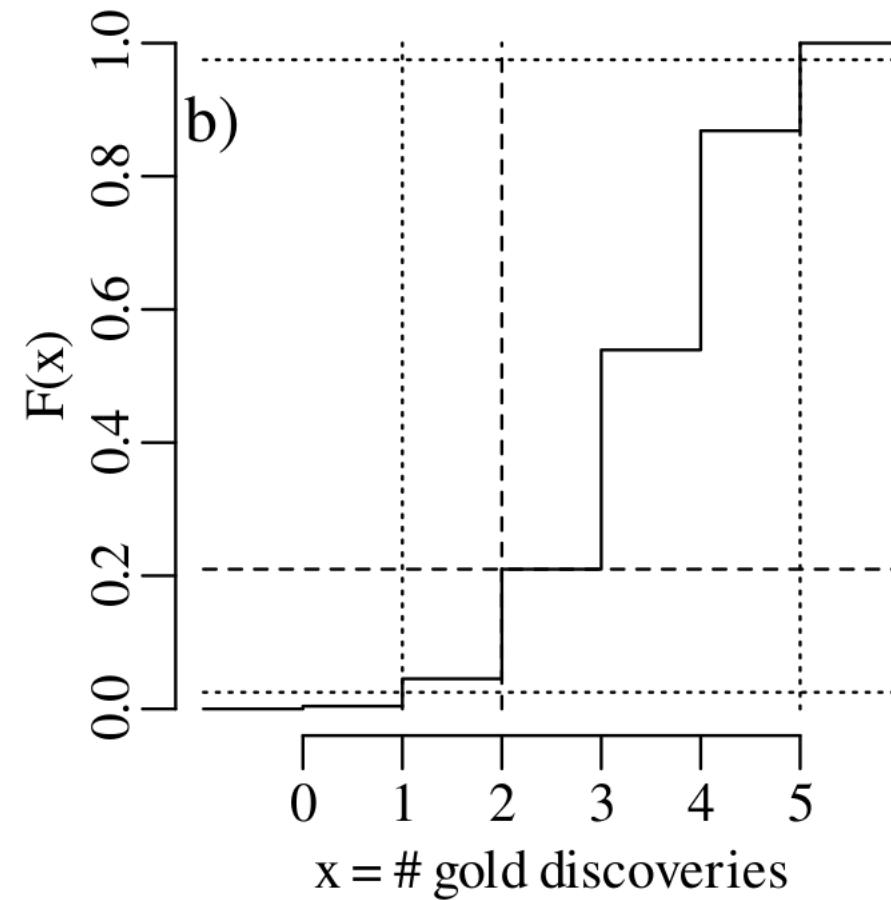
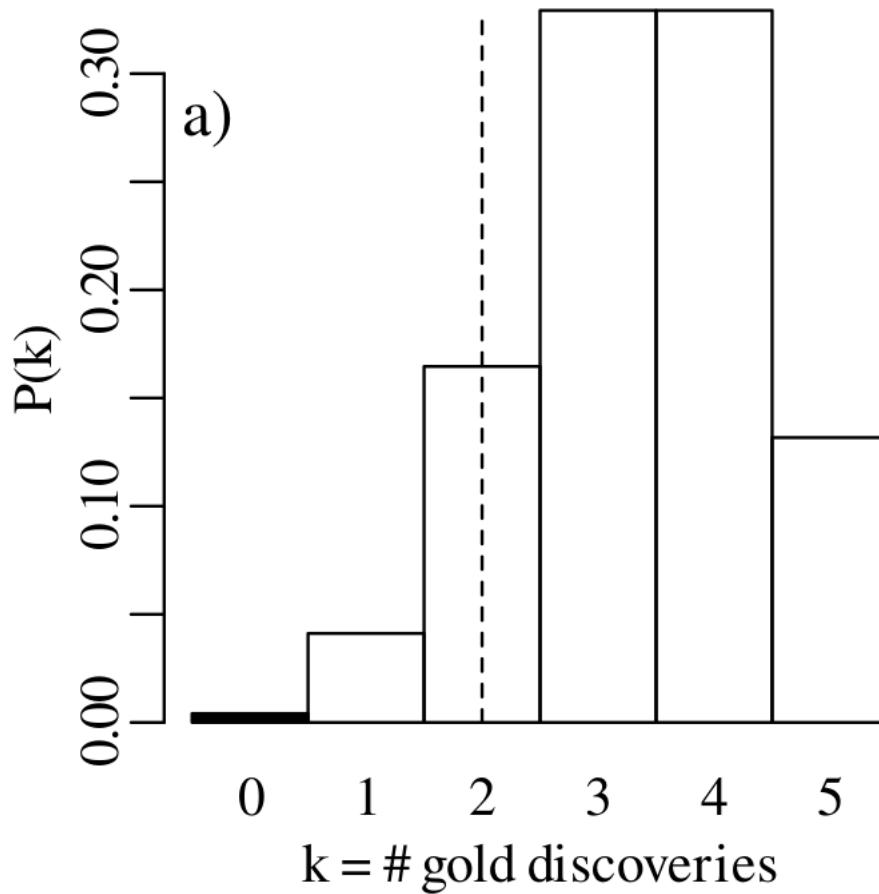
5. Mark the **rejection region**

| k             | 0             | 1      | 2             | 3      | 4      | 5      |
|---------------|---------------|--------|---------------|--------|--------|--------|
| $P(T = k)$    | 0.0041        | 0.0411 | 0.1646        | 0.3292 | 0.3292 | 0.1317 |
| $P(T \leq k)$ | <b>0.0041</b> | 0.0453 | <i>0.2099</i> | 0.5391 | 0.8683 | 1.0000 |
| $P(T \geq k)$ | 1.000         | 0.9959 | <i>0.9547</i> | 0.7901 | 0.4609 | 0.1317 |

$$R = \{0\}$$

6. Reach a **decision**  $k = 2 \notin R$

7. Alternatively,  $2 \times 0.2099 = 0.4198 > 0.05$



# Statistical power

We failed to reject  $H_0$  for  $\hat{p} = \frac{2}{5} = 0.4$ . How about  $\hat{p} = \frac{6}{15} = 0.4$ ?

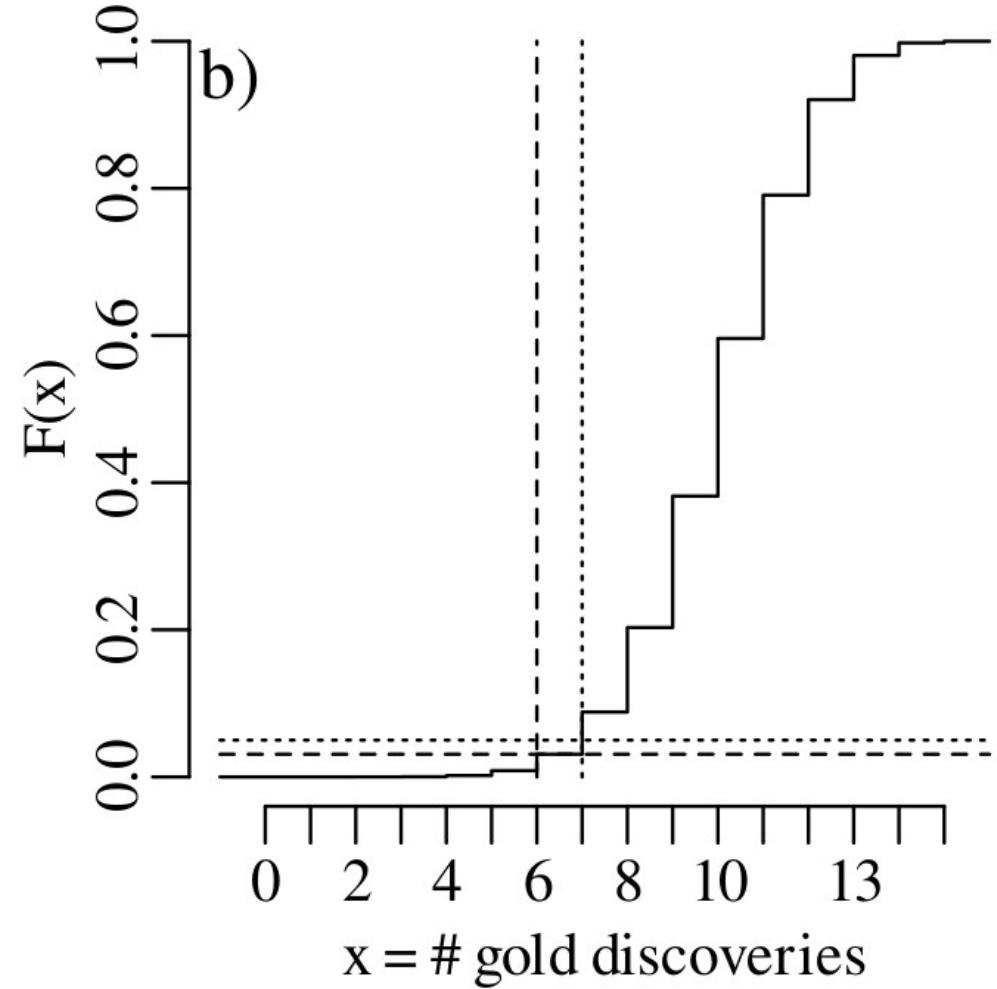
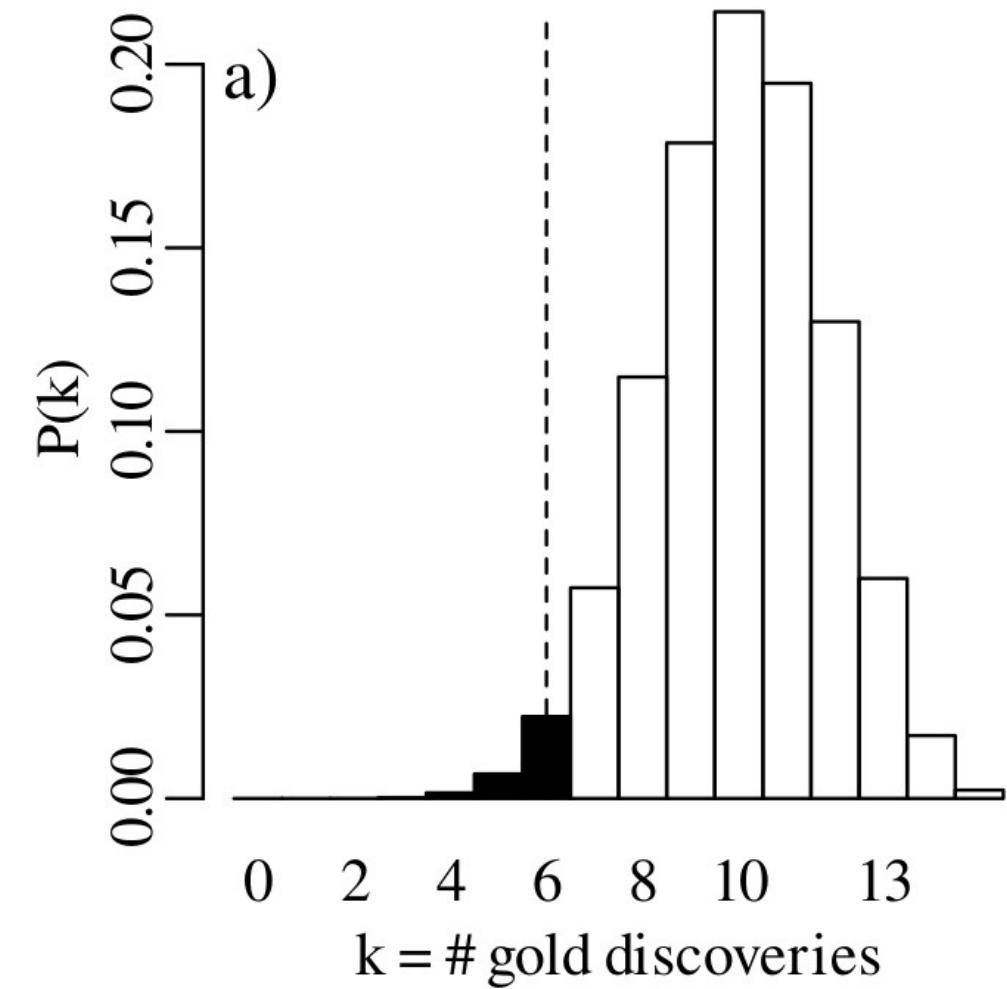
one-sided hypothesis test ( $H_0 : p = 2/3$  vs.  $H_a : p < 2/3$ )

| k             | 0                                      | 1                                      | 2                                      | 3                                      | 4             | 5             | 6             | 7      |
|---------------|----------------------------------------|----------------------------------------|----------------------------------------|----------------------------------------|---------------|---------------|---------------|--------|
| $P(T = k)$    | $7.0 \times 10^{-8}$                   | $2.1 \times 10^{-6}$                   | $2.9 \times 10^{-5}$                   | $2.5 \times 10^{-4}$                   | 0.0015        | 0.0067        | 0.0223        | 0.0574 |
| $P(T \leq k)$ | <b><math>7.0 \times 10^{-8}</math></b> | <b><math>2.2 \times 10^{-6}</math></b> | <b><math>3.1 \times 10^{-5}</math></b> | <b><math>2.8 \times 10^{-4}</math></b> | <b>0.0018</b> | <b>0.0085</b> | <b>0.0308</b> | 0.0882 |
| k             | 8                                      | 9                                      | 10                                     | 11                                     | 12            | 13            | 14            | 15     |
| $P(T = k)$    | 0.1148                                 | 0.1786                                 | 0.2143                                 | 0.1948                                 | 0.1299        | 0.0599        | 0.0171        | 0.0023 |
| $P(T \leq k)$ | 0.2030                                 | 0.3816                                 | 0.5959                                 | 0.7908                                 | 0.9206        | 0.9806        | 0.9977        | 1.0000 |

$$R = \{0, 1, 2, 3, 4, 5, 6\}$$

$$k \in R$$

$$\text{p-value} = 0.0308 < \alpha$$



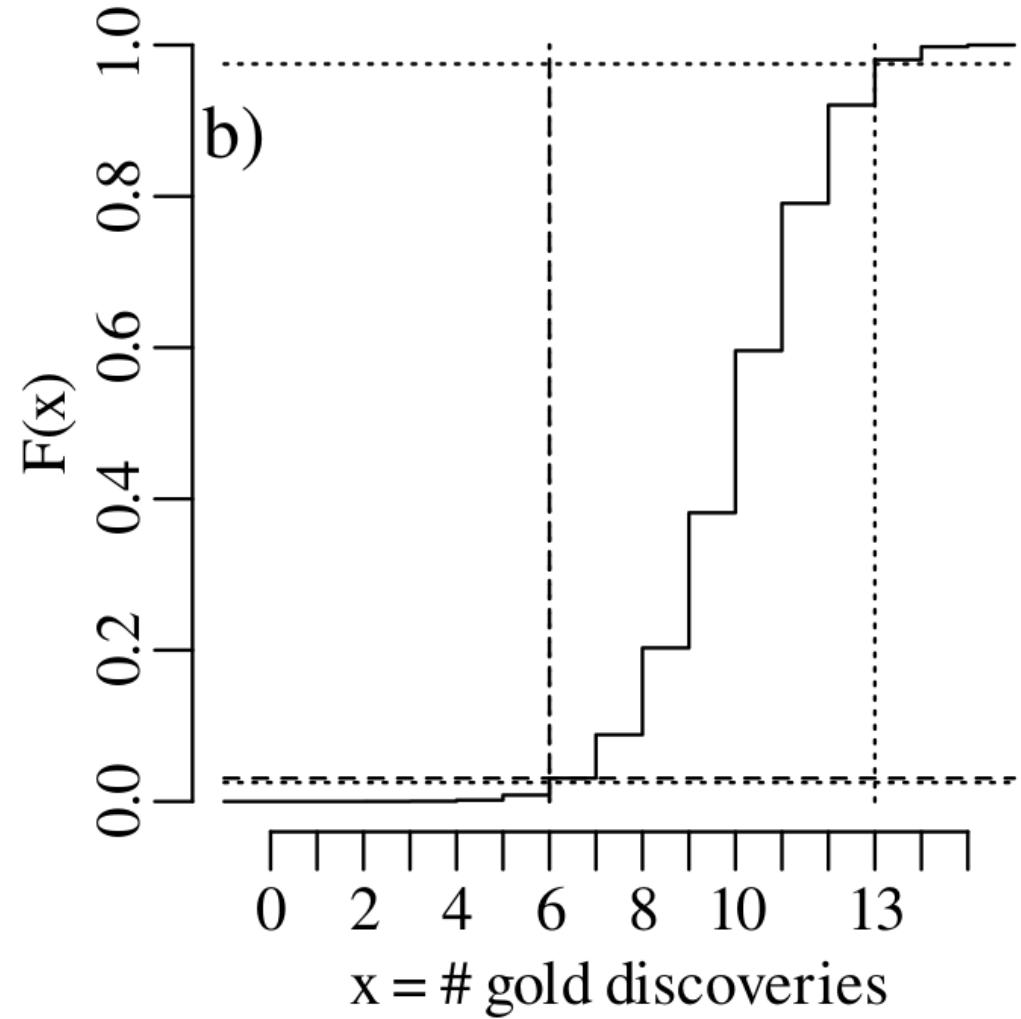
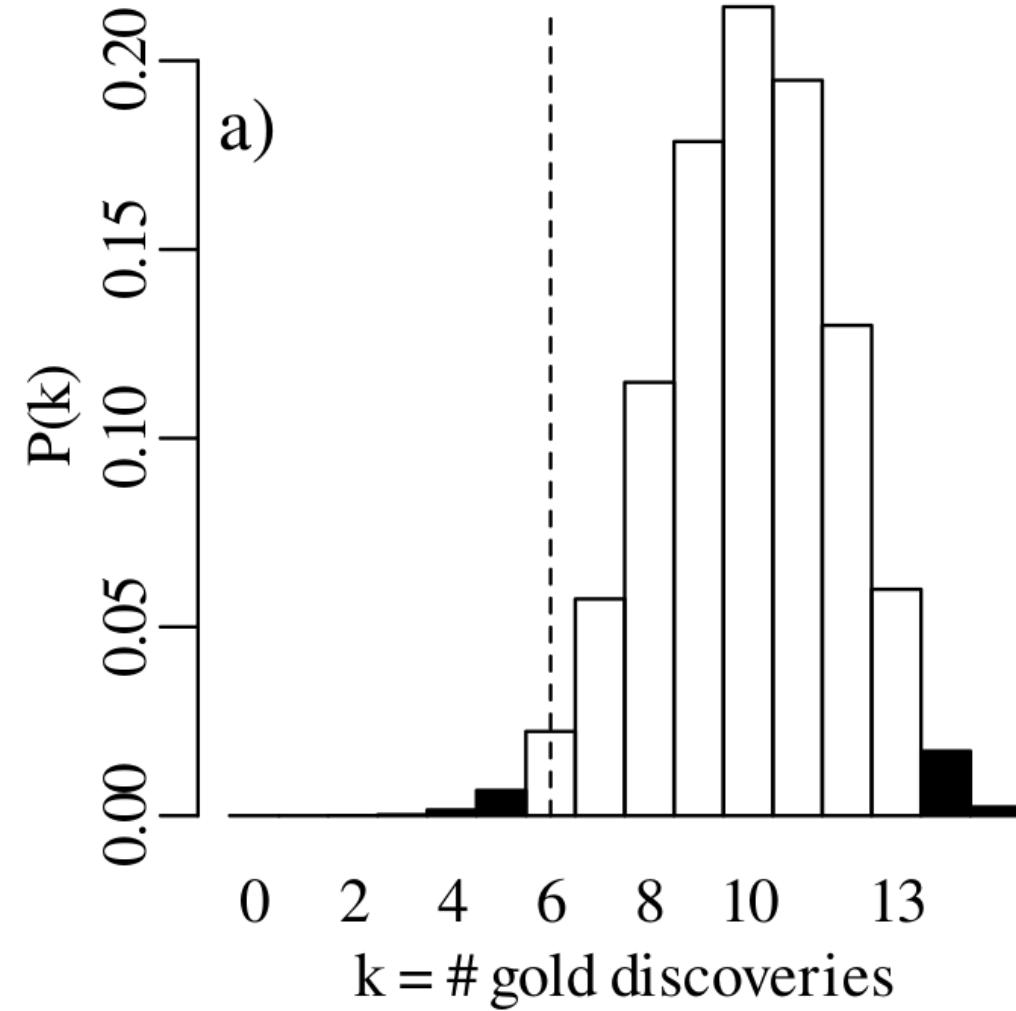
two-sided hypothesis test ( $H_0 : p = 2/3$  vs.  $H_a : p \neq 2/3$ ):

| k             | 0                                      | 1                                      | 2                                      | 3                                      | 4                        | 5             | 6             | 7             |
|---------------|----------------------------------------|----------------------------------------|----------------------------------------|----------------------------------------|--------------------------|---------------|---------------|---------------|
| $P(T = k)$    | $7.0 \times 10^{-8}$                   | $2.1 \times 10^{-6}$                   | $2.9 \times 10^{-5}$                   | $2.5 \times 10^{-4}$                   | 0.0015                   | 0.0067        | 0.0223        | 0.0574        |
| $P(T \leq k)$ | <b><math>7.0 \times 10^{-8}</math></b> | <b><math>2.2 \times 10^{-6}</math></b> | <b><math>3.1 \times 10^{-5}</math></b> | <b><math>2.8 \times 10^{-4}</math></b> | <b>0.0018</b>            | <b>0.0085</b> | <b>0.0308</b> | 0.0882        |
| $P(T \geq k)$ | 1.0000                                 | $1 - 7.0 \times 10^{-8}$               | $1 - 2.2 \times 10^{-6}$               | $1 - 3.1 \times 10^{-5}$               | $1 - 2.8 \times 10^{-4}$ | 0.9982        | 0.9915        | 0.9692        |
| k             | 8                                      | 9                                      | 10                                     | 11                                     | 12                       | 13            | 14            | 15            |
| $P(T = k)$    | 0.1148                                 | 0.1786                                 | 0.2143                                 | 0.1948                                 | 0.1299                   | 0.0599        | 0.0171        | 0.0023        |
| $P(T \leq k)$ | 0.2030                                 | 0.3816                                 | 0.5959                                 | 0.7908                                 | 0.9206                   | 0.9806        | 0.9977        | 1.0000        |
| $P(T > k)$    | 0.9118                                 | 0.7970                                 | 0.6184                                 | 0.4041                                 | 0.2092                   | 0.0794        | <b>0.0194</b> | <b>0.0023</b> |

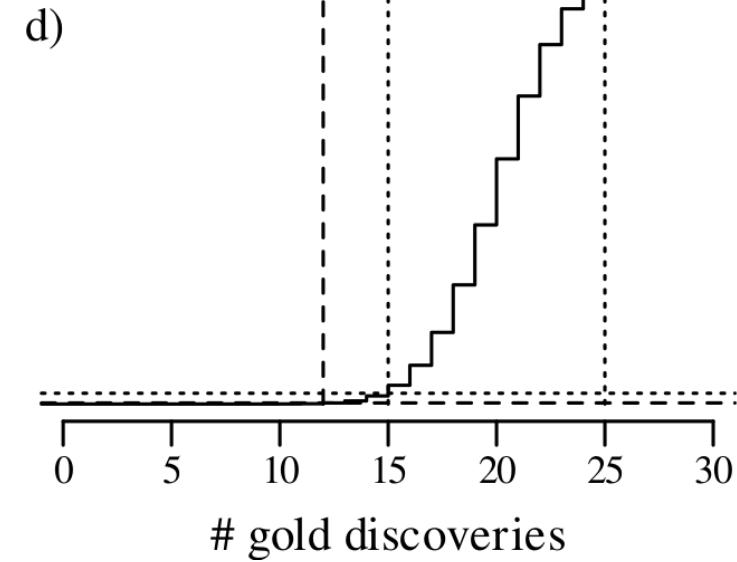
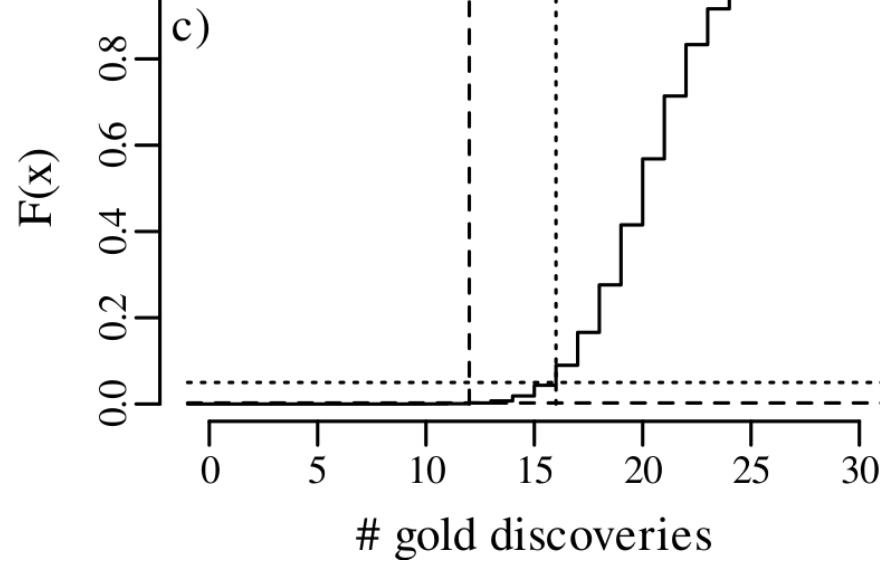
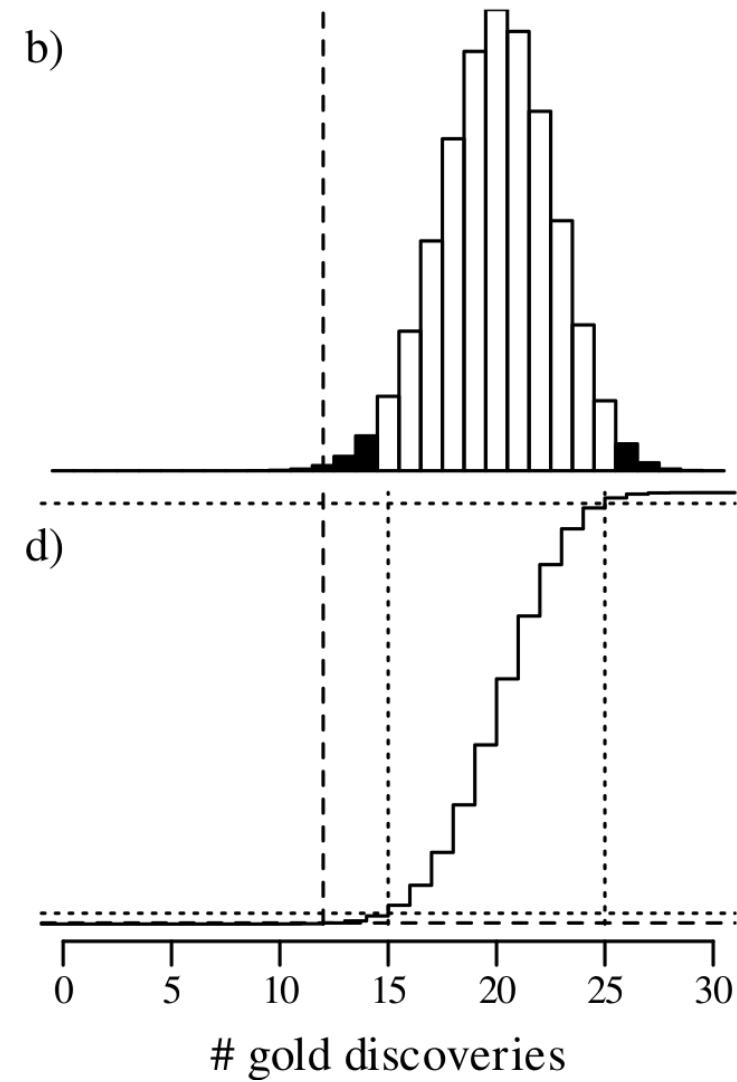
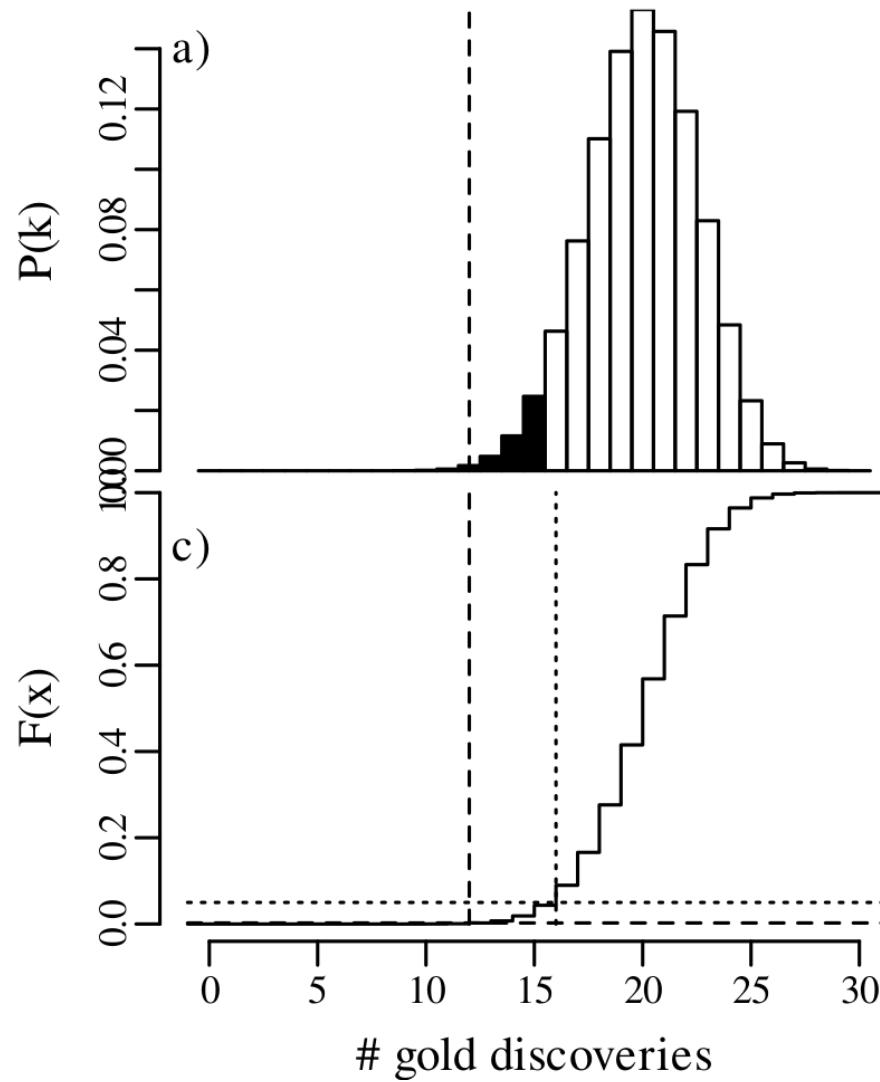
$$R = \{0, 1, 2, 3, 4, 5, 14, 15\}$$

$$k \notin R$$

$$\text{p-value} = (2 \times 0.0308 =) 0.0616 > \alpha$$



Next, how about  $\hat{p} = \frac{12}{30} = 0.4$ ?



| $H_0$ is ... | false            | true             |
|--------------|------------------|------------------|
| rejected     | correct decision | Type-I error     |
| not rejected | Type-II error    | correct decision |

| the accused is ... | guilty           | innocent         |
|--------------------|------------------|------------------|
| sentenced          | correct decision | Type-I error     |
| acquitted          | Type-II error    | correct decision |

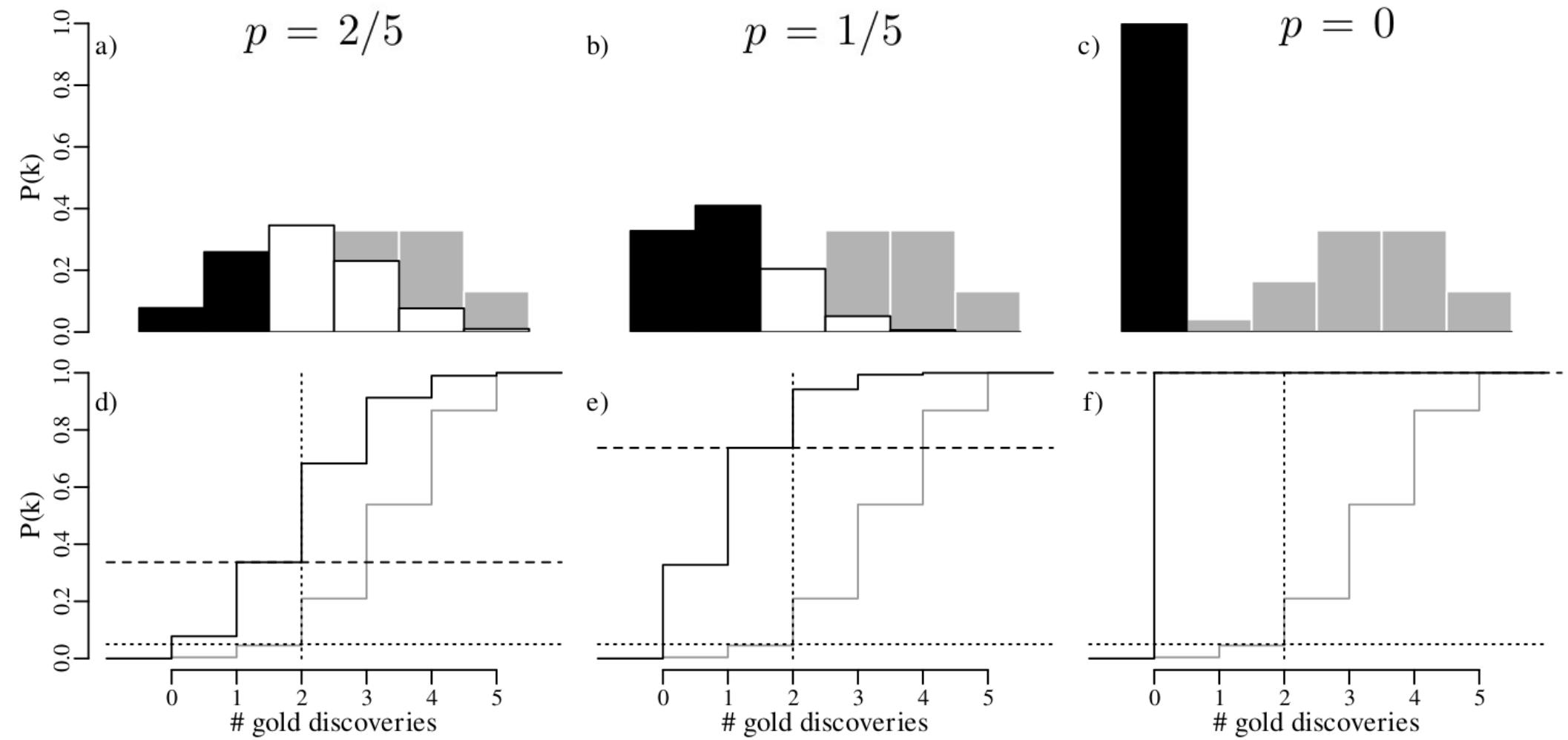
depends on the **confidence level**  $\alpha$

| $H_0$ is ... | false            | true             |
|--------------|------------------|------------------|
| rejected     | correct decision | Type-I error     |
| not rejected | Type-II error    | correct decision |

$\beta = 1 - \text{power}$

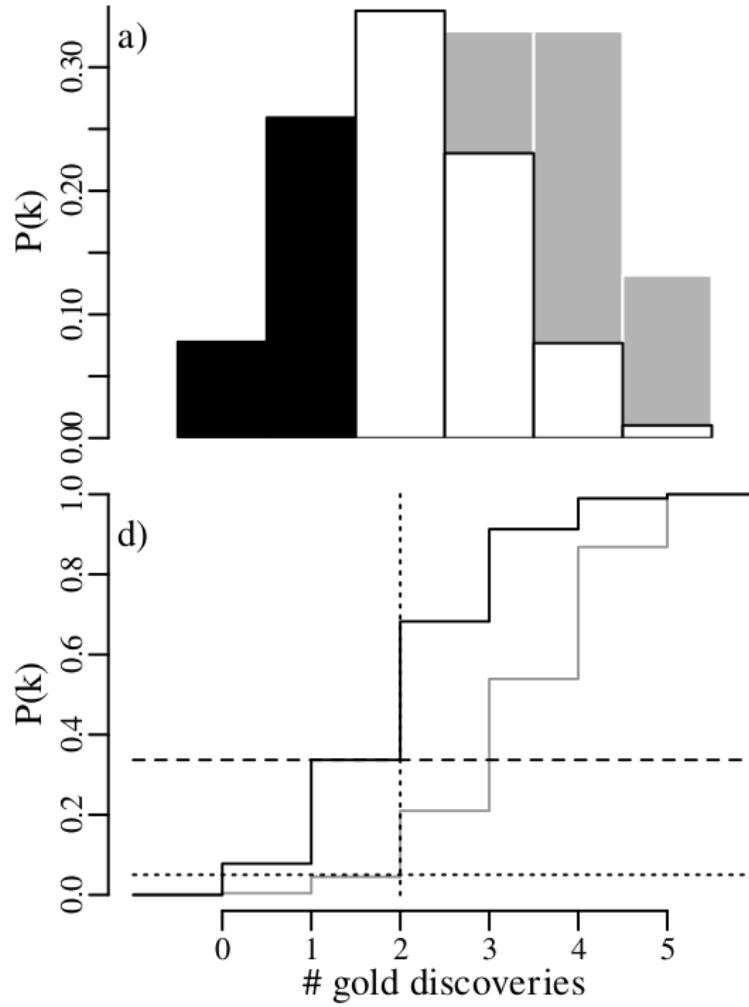
depends on 1. The **degree to which  $H_0$  is false**  
2. **Sample size**

# 1. The degree to which $H_0$ is false

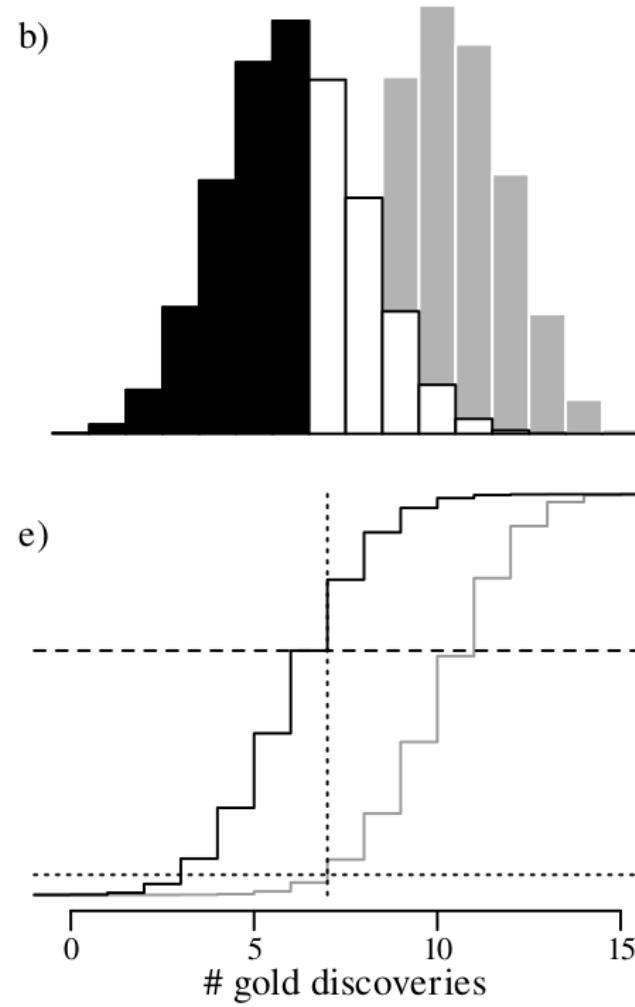


## 2. Sample size

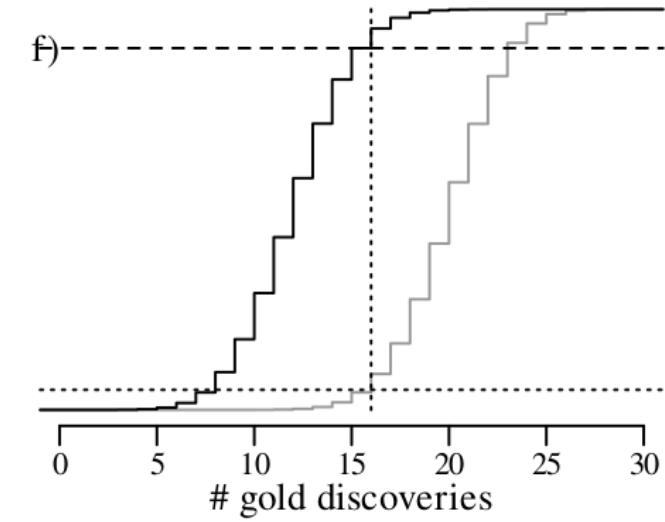
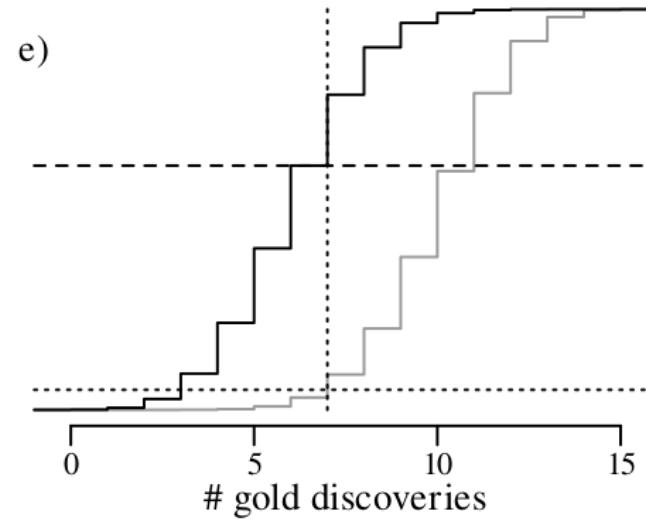
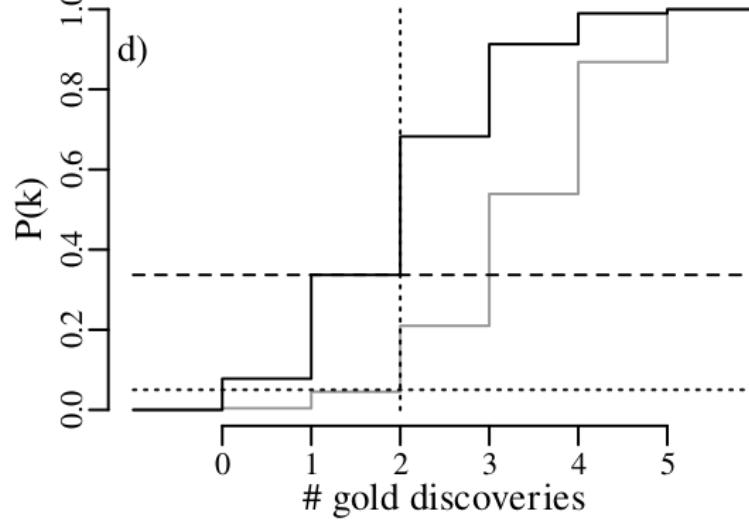
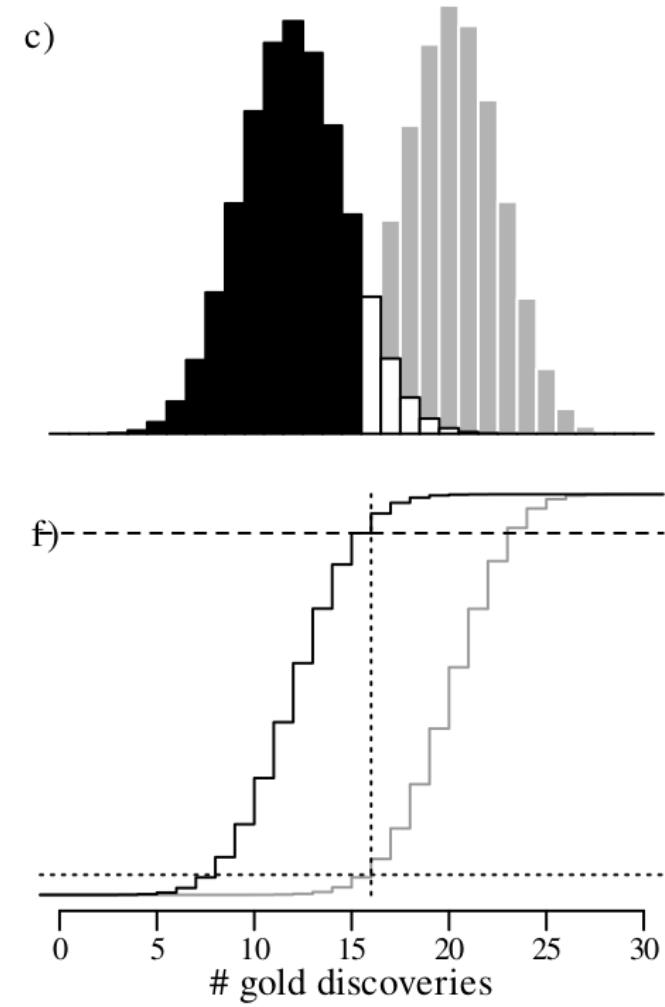
$n = 5$



$n = 15$



$n = 30$



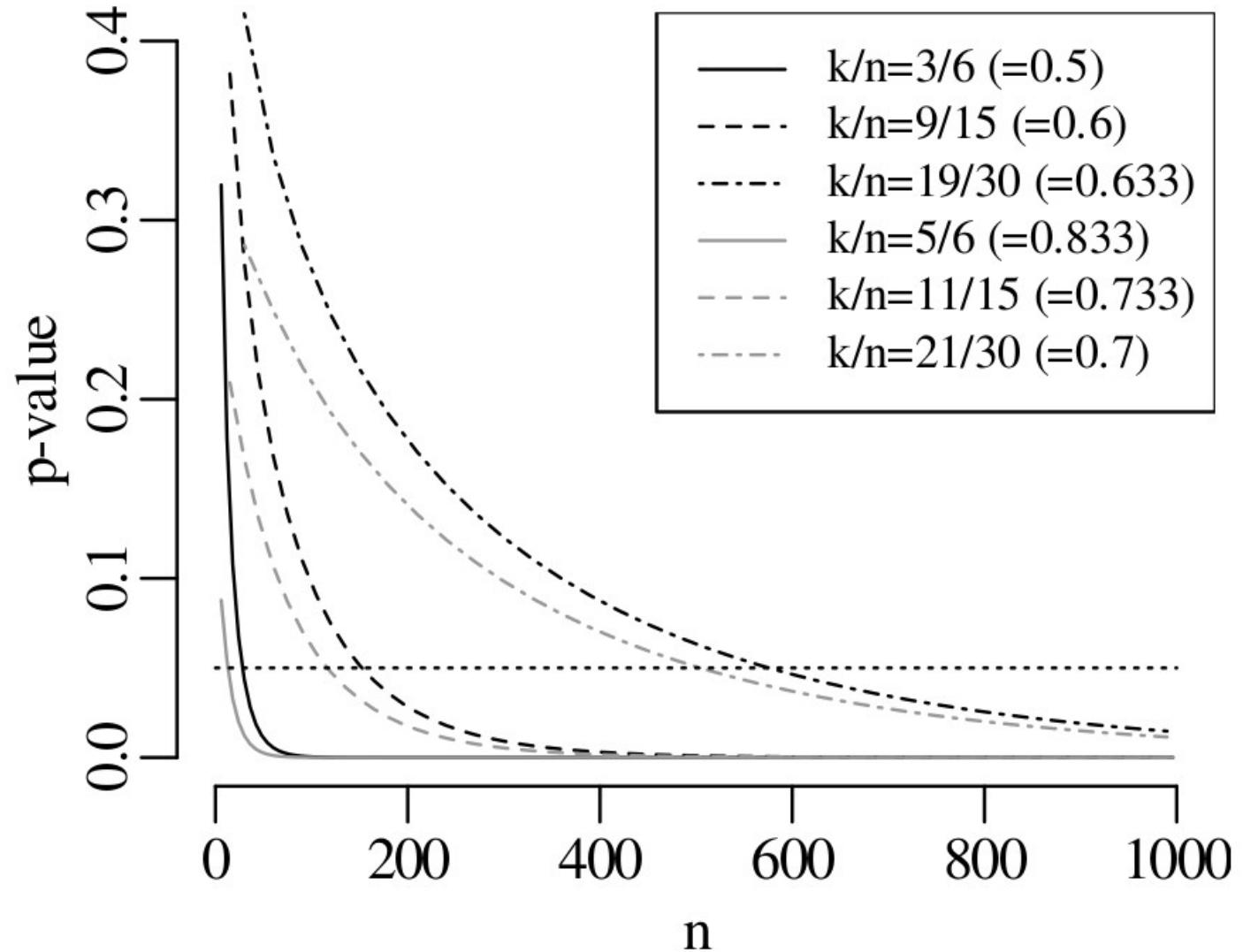
# The scientific method

1. Formulate a hypothesis.
2. Design an experiment to test said hypothesis.
3. Carry out the experiment and check to see if it matches the prediction.

1. Hypothesis: Earth's lower mantle is made of olivine.
2. Test: Study the stability of olivine at lower mantle pressures (24-136 GPa).
3. Result: Olivine is not stable at lower mantle pressures.

1. Hypothesis: Earth's lower mantle is made of perovskite.
2. Test: Study the stability of perovskite at lower mantle pressures.
3. Result: Perovskite is stable at lower mantle pressures.

$p = 2/3$  ?



# Pitfalls

*All hypotheses are wrong ... in some decimal place*

– John Tukey (paraphrased)

*All models are wrong, but some are useful*

– George Box

# Confidence intervals

Which values of  $p$  are compatible with the observation of  $k = 2$  successes out of  $n = 5$  claims?

1.  ~~$p=0?$~~

2.  $p=0.1?$

$$P(k \geq 2 | p = 0.1, n = 5) = \sum_{i=2}^5 \binom{5}{i} (0.1)^i (0.9)^{n-i} = 0.081 \geq \alpha/2$$

3.  $p=2/5? = \hat{p}$

4.  $p=2/3?$

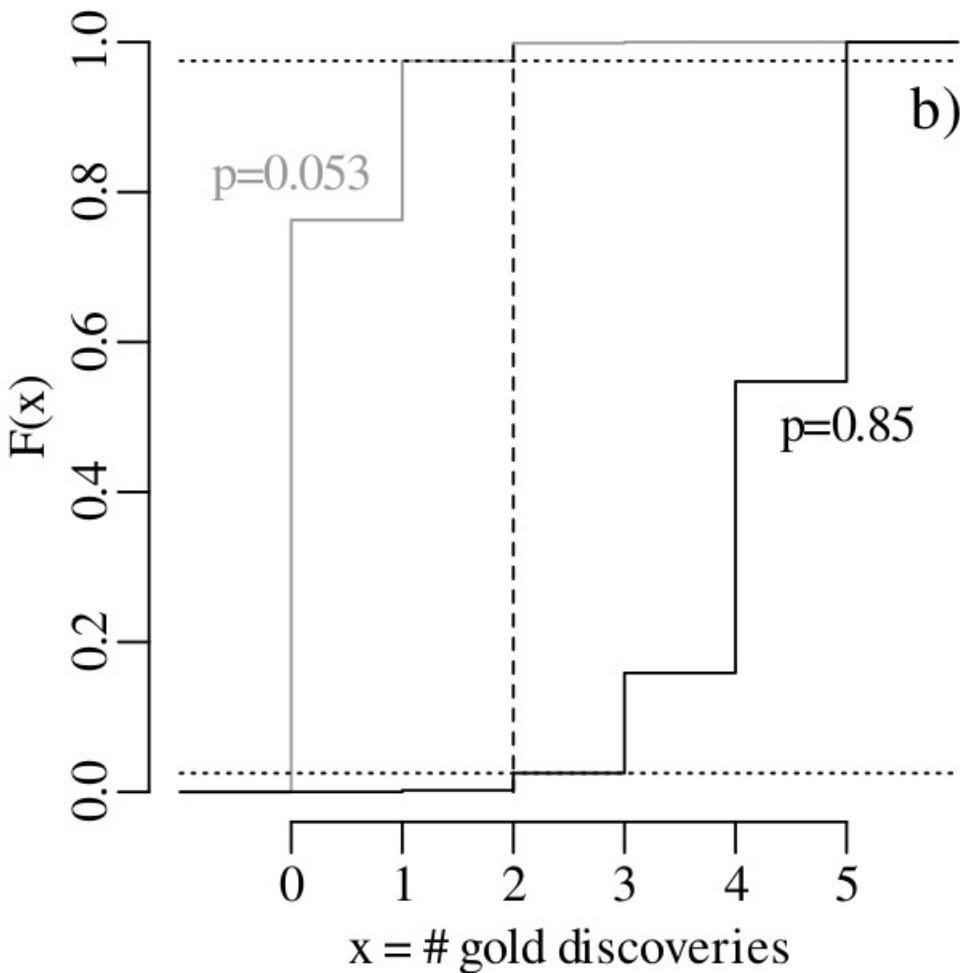
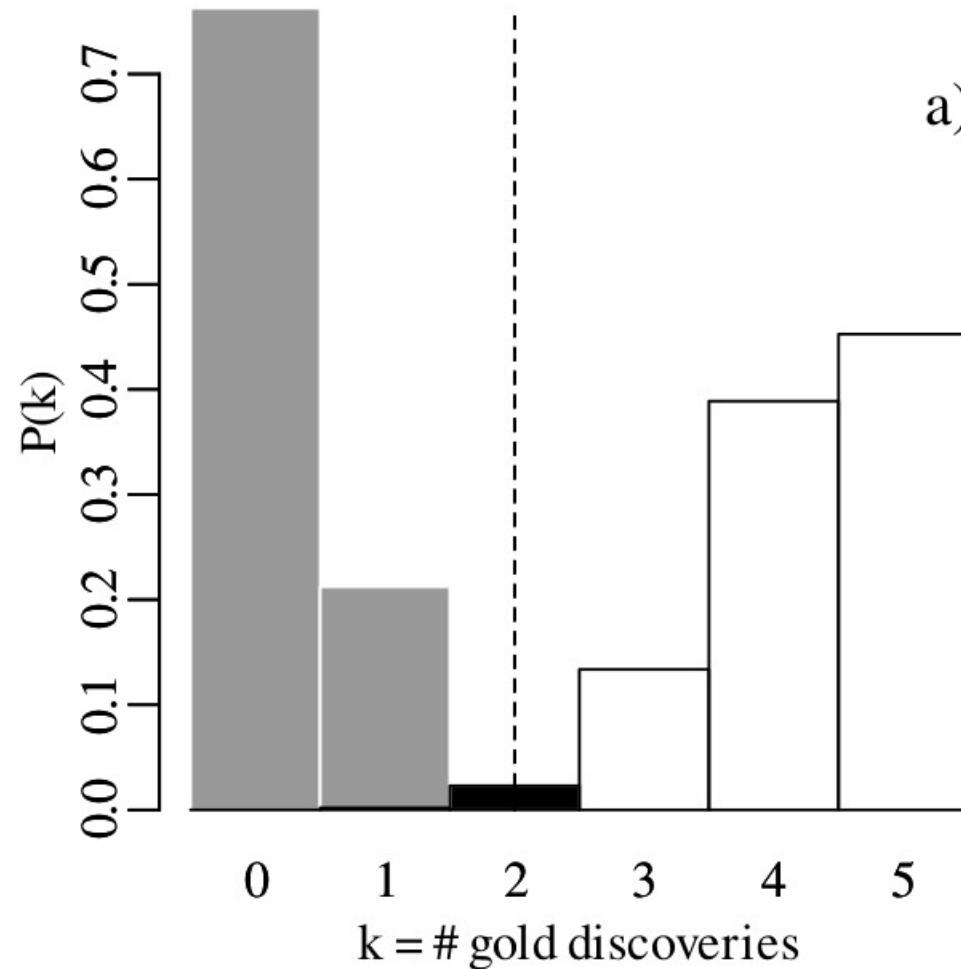
$$P(k \leq 2 | p = 2/3, n = 5) = \sum_{i=0}^2 \binom{5}{i} (2/3)^i (1/3)^{n-i} = 0.21 \geq \alpha/2$$

5.  ~~$p=0.9?$~~

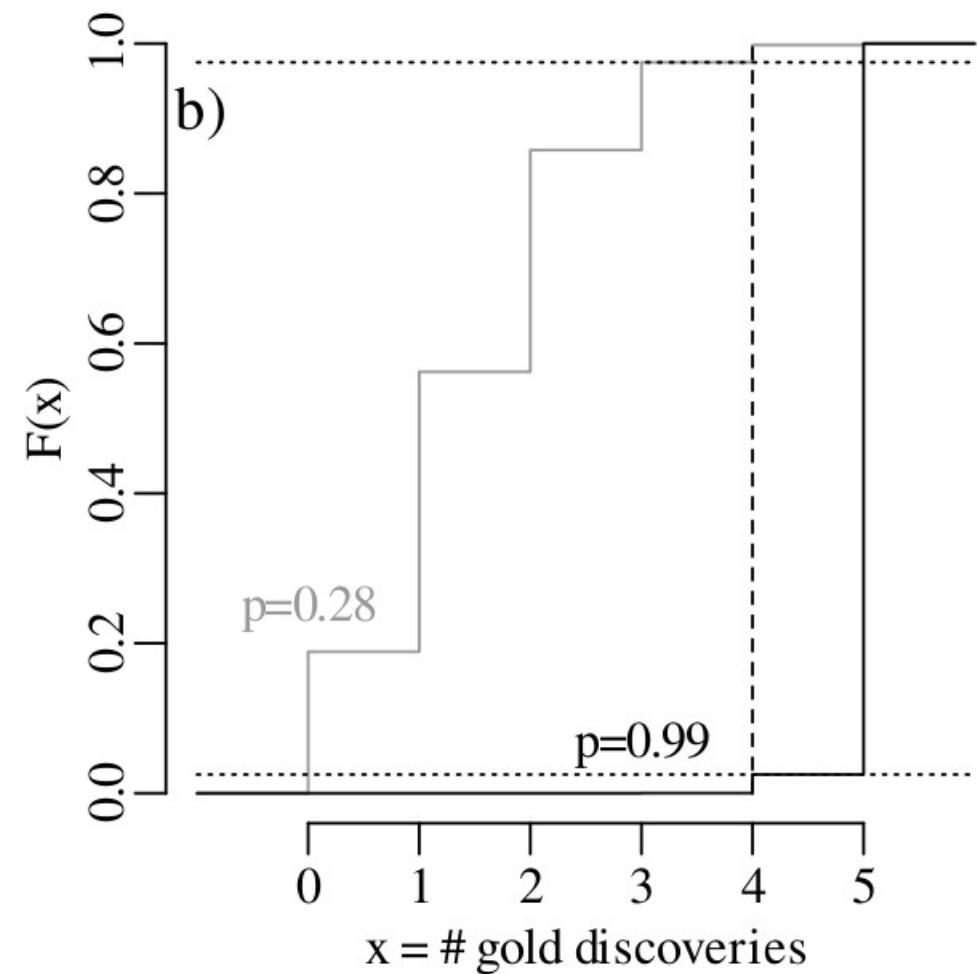
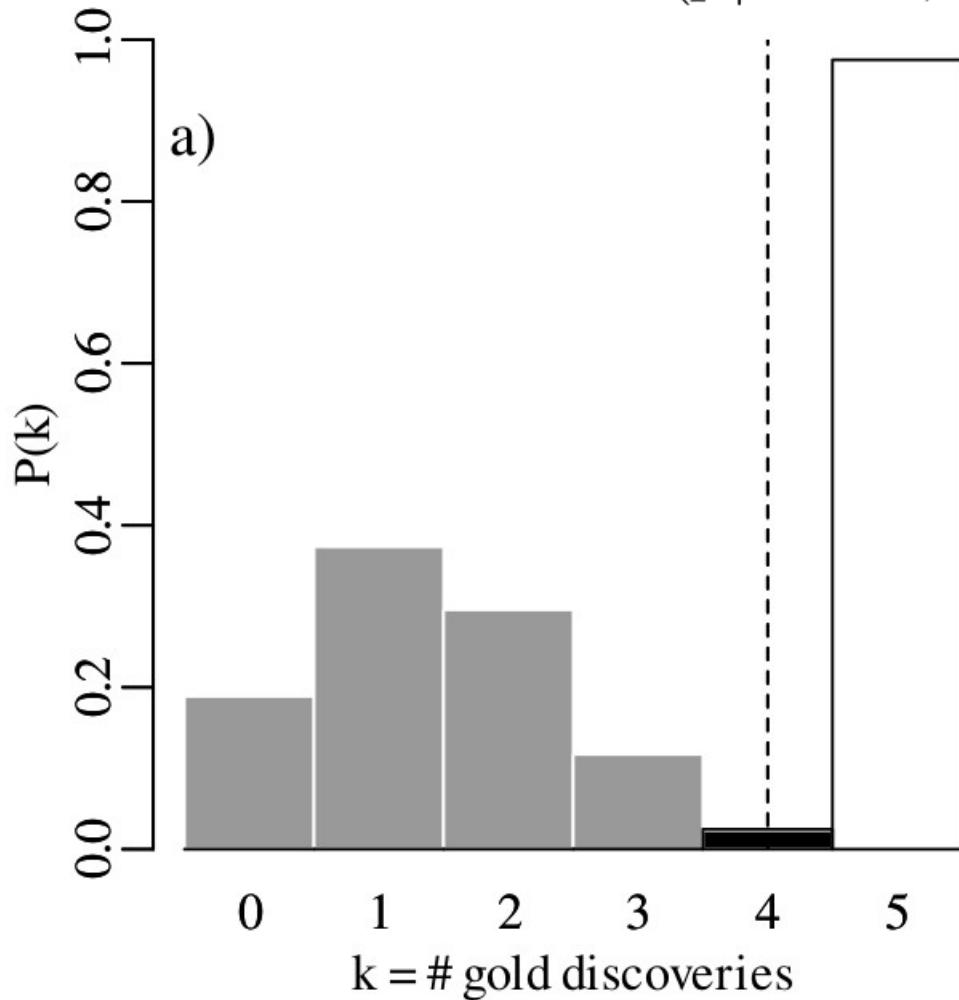
$$P(k \leq 2 | p = 0.9, n = 5) = \sum_{i=0}^2 \binom{5}{i} (0.9)^i (0.1)^{n-i} = 0.0086 \leq \alpha/2$$

6.  ~~$p=1?$~~

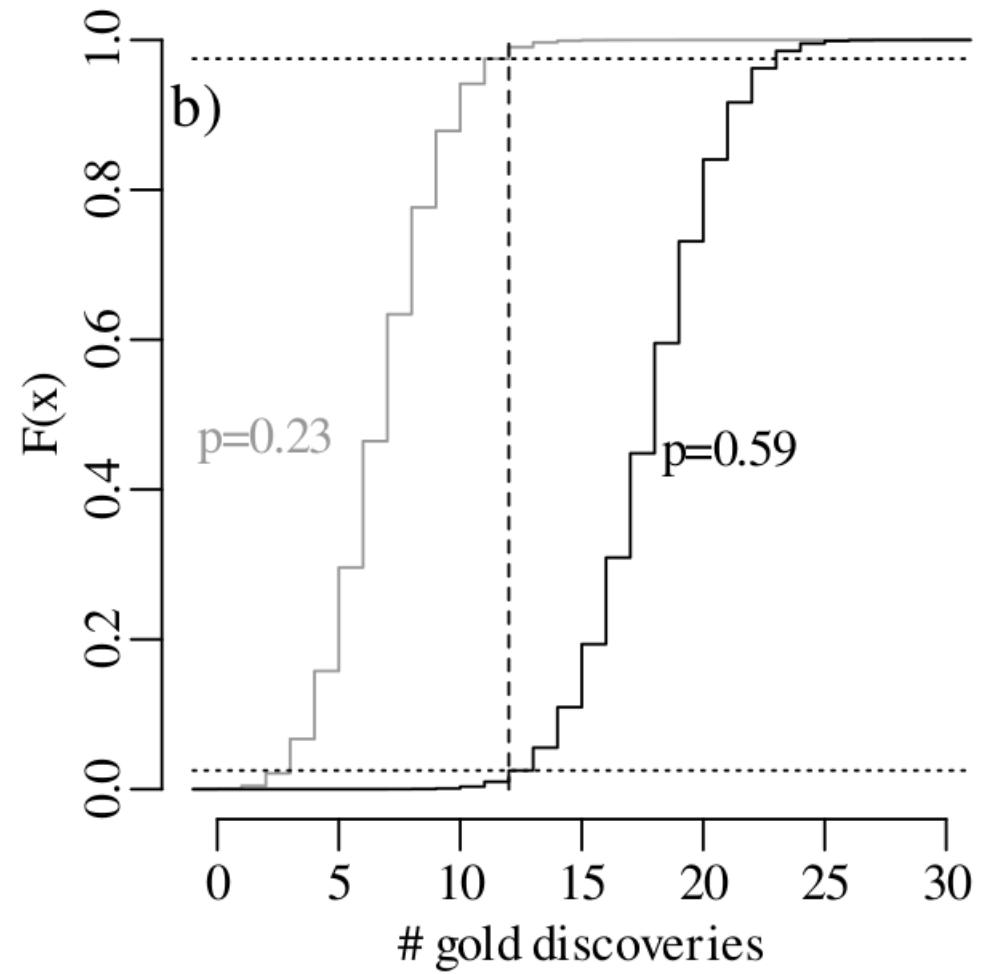
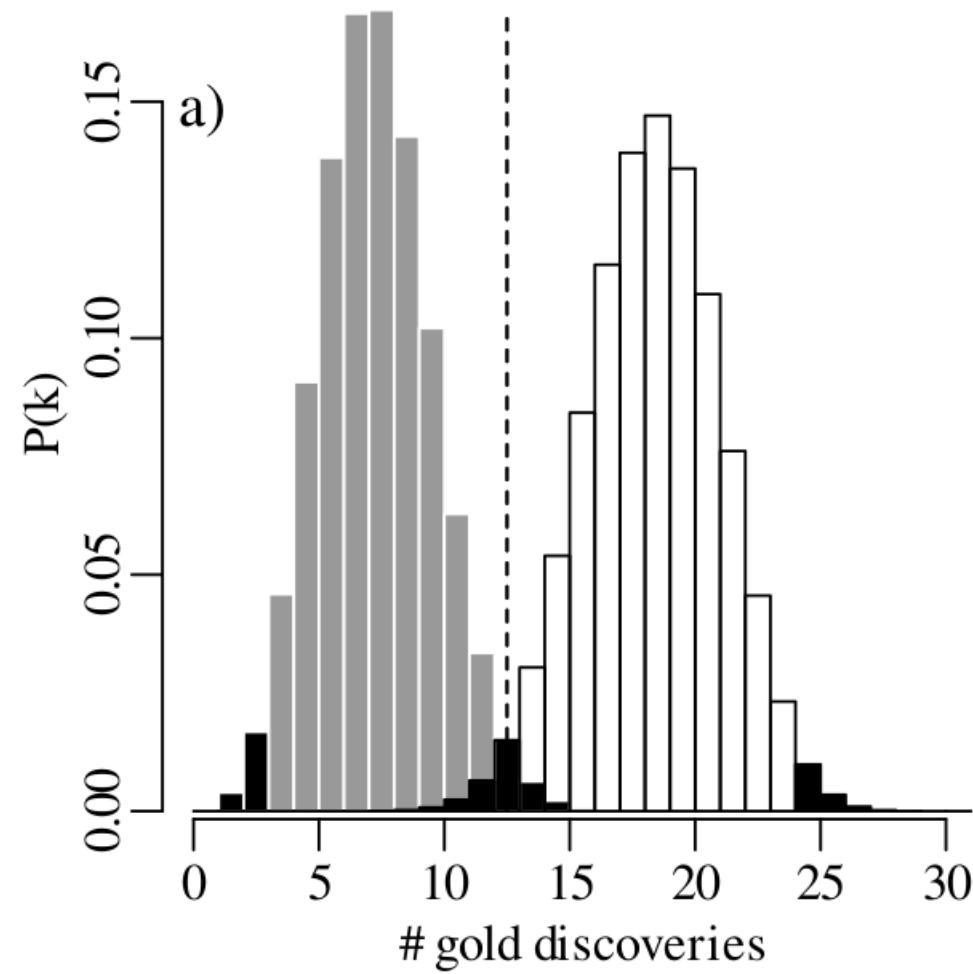
$$\text{C.I.}(p|k=2, n=5) = [0.053, 0.85]$$

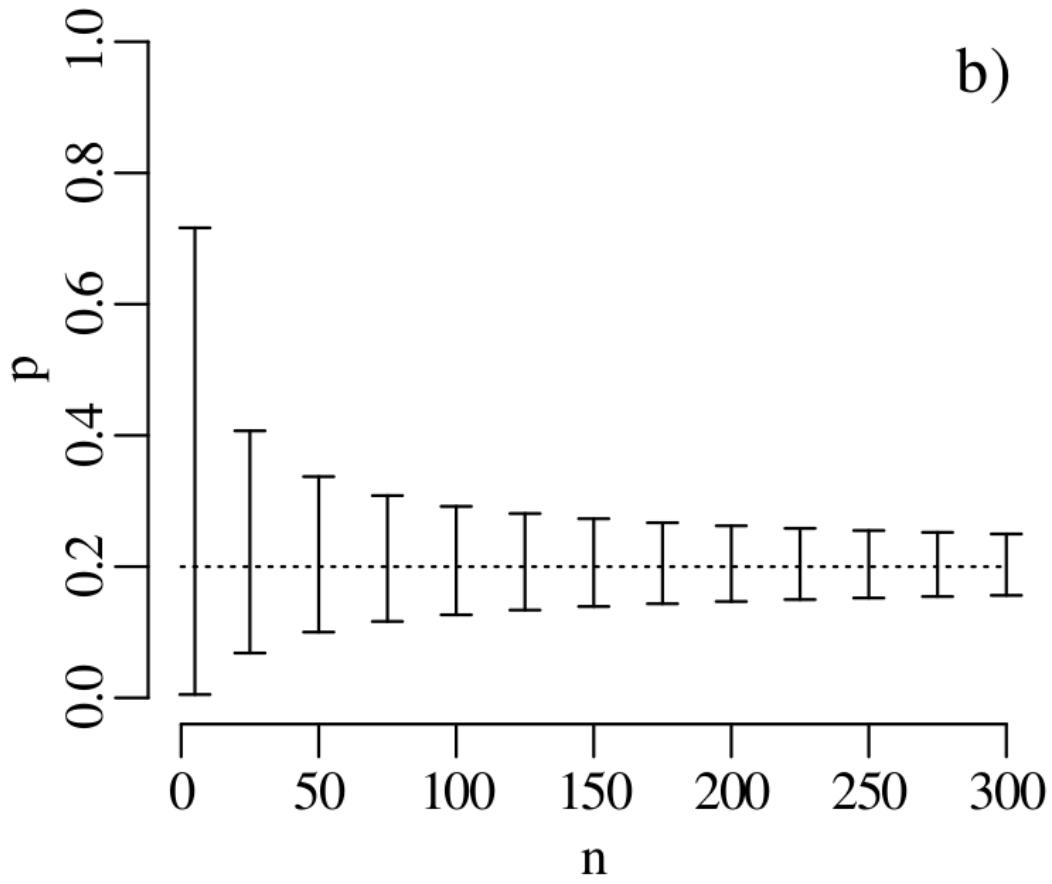
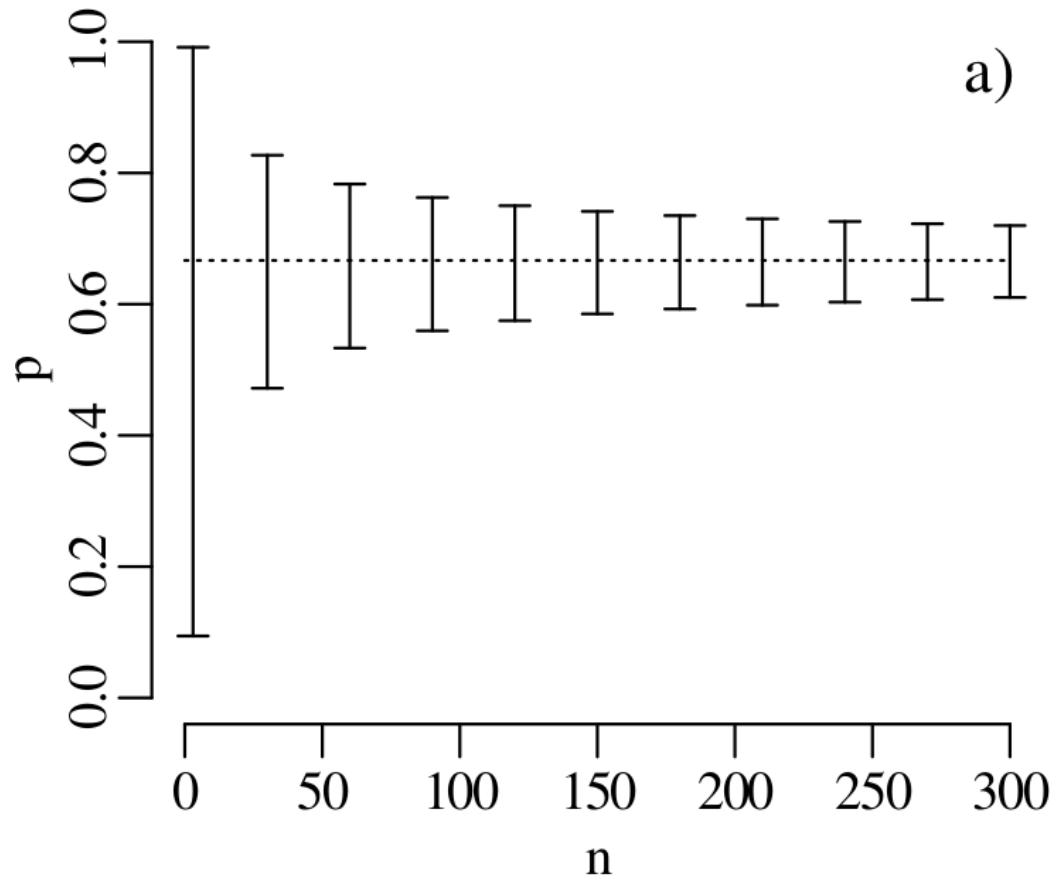


$$\text{C.I.}(p|k=4, n=5) = [0.284, 0.995]$$



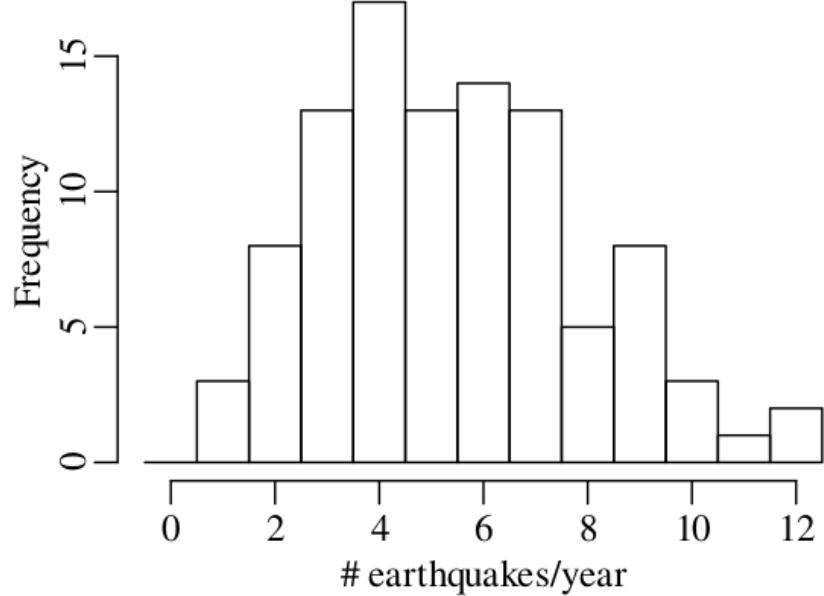
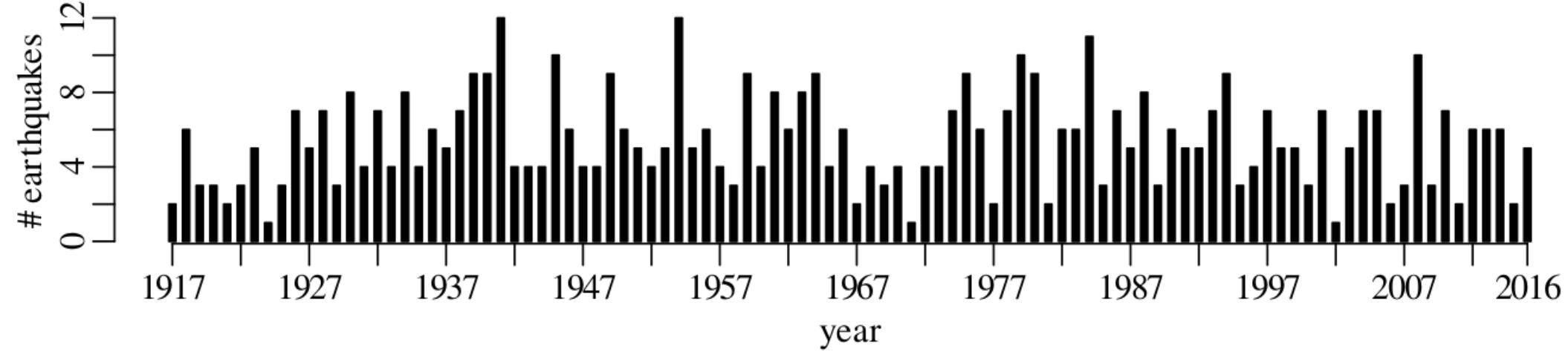
$$\text{C.I.}(p|k = 12, n = 30) = [0.23, 0.59]$$





# Statistics for geoscientists

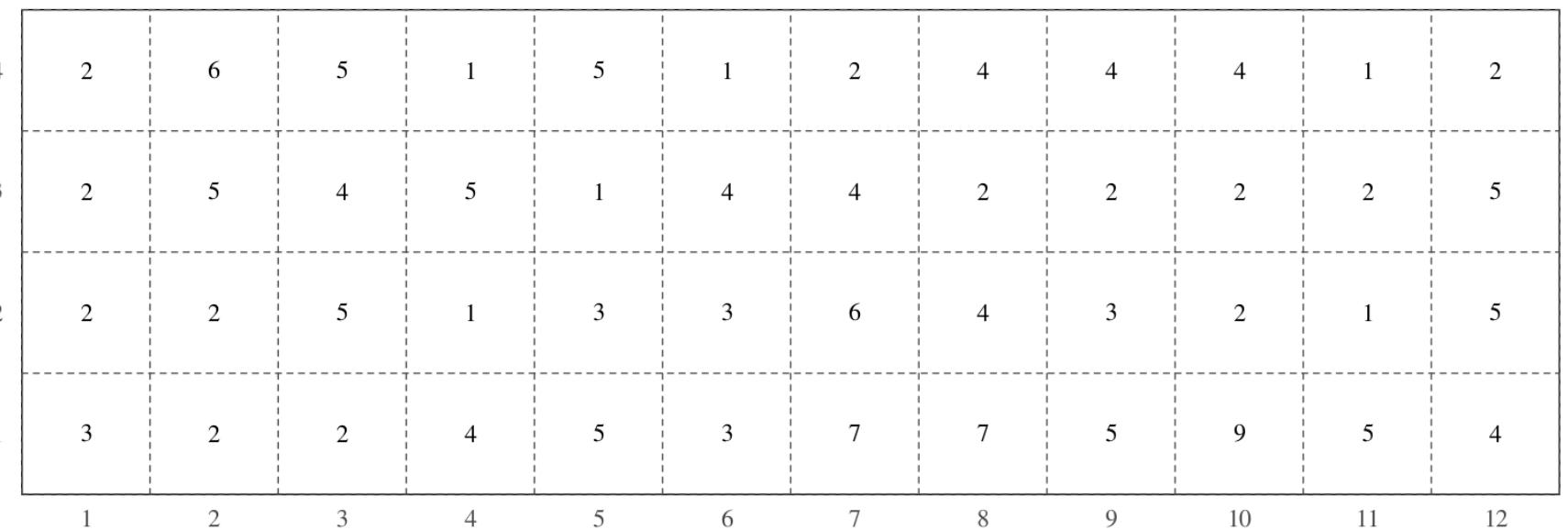
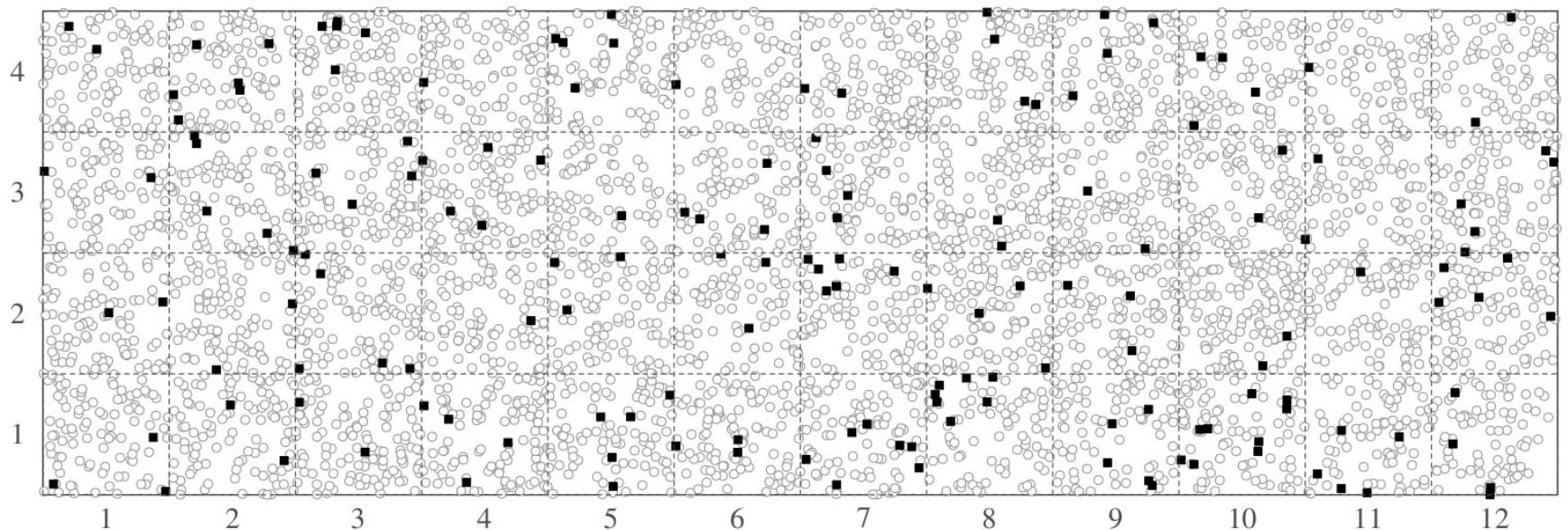
## The Poisson distribution

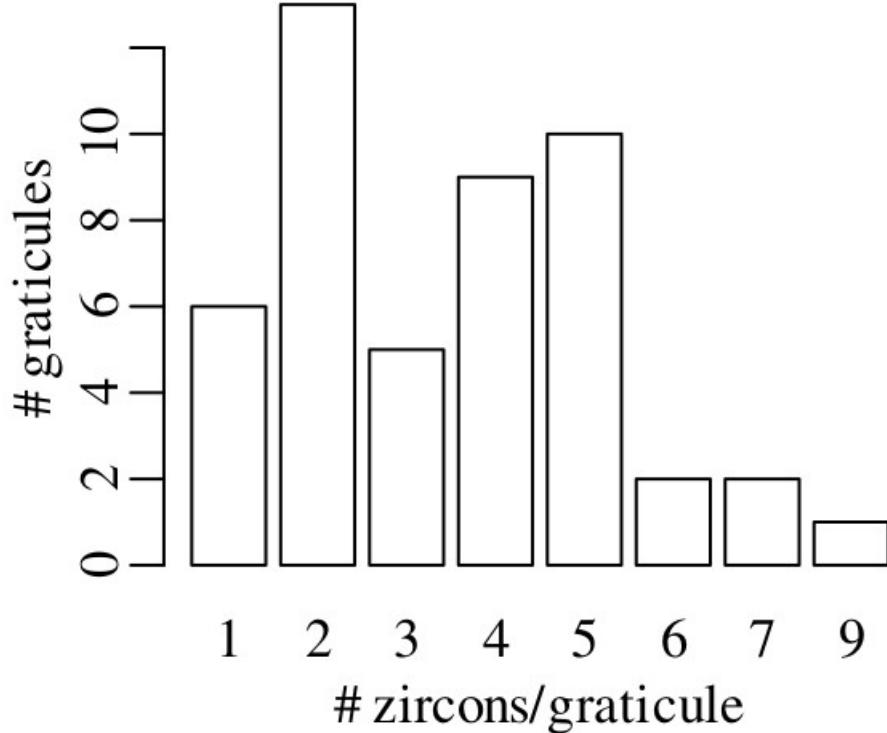


**mean: 5.43**

standard deviation: 2.50

**variance: 6.24**





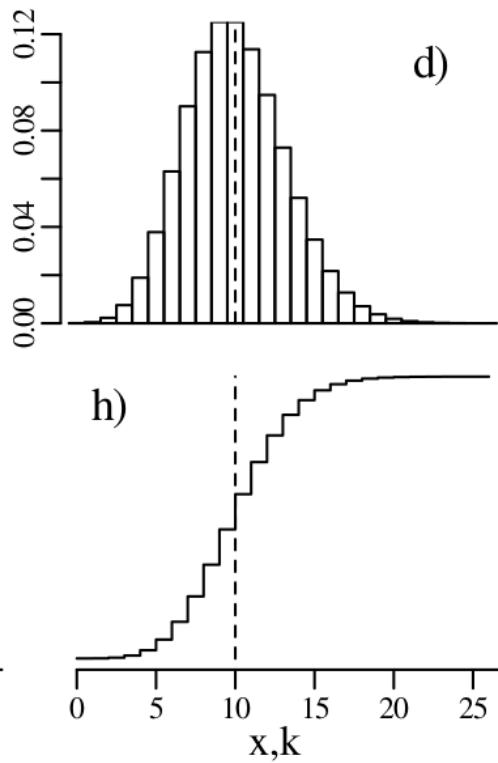
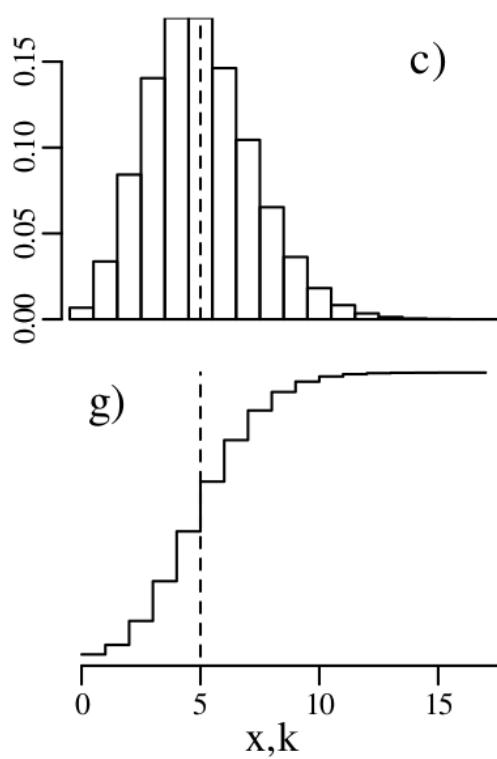
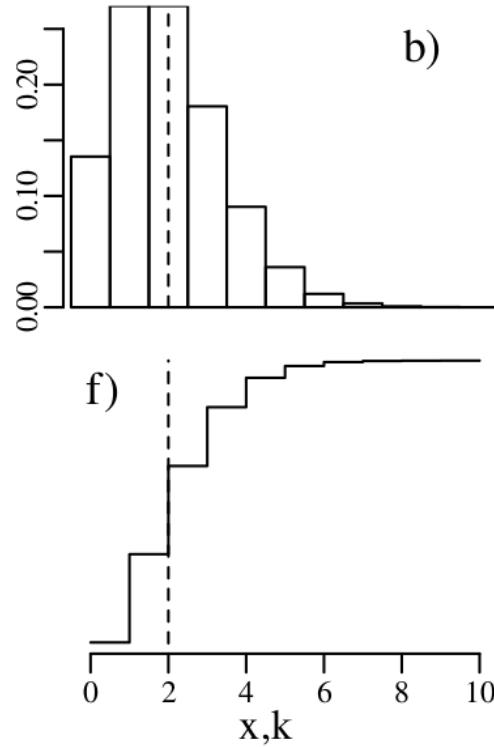
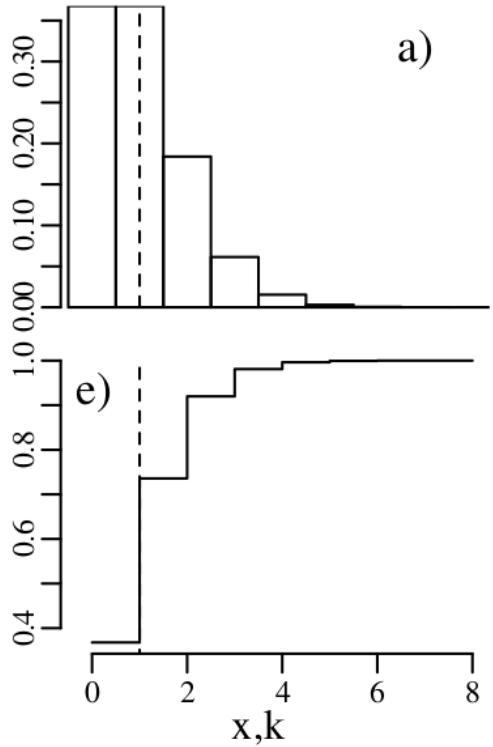
**mean: 3.50**

standard deviation: 1.85

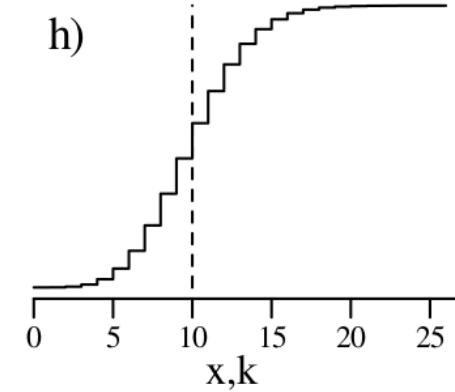
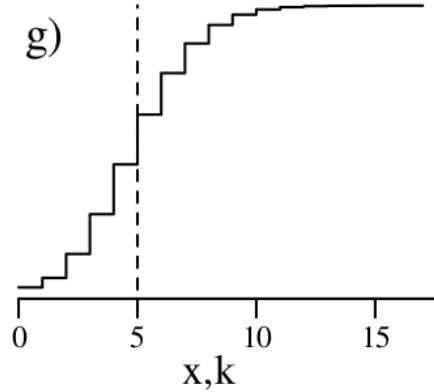
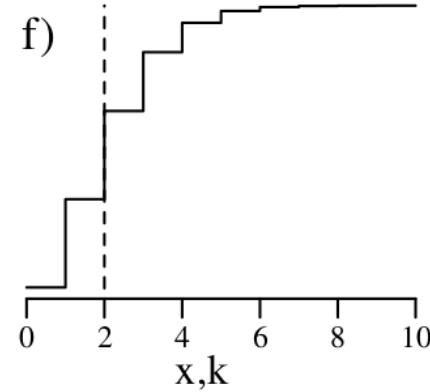
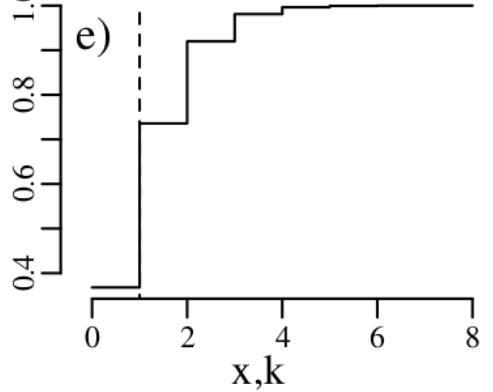
**variance: 3.40**

$$P(k|\lambda) = \frac{\lambda^k e^{-\lambda}}{k!}$$

$P(k)$



$F(x)$



$$P(k|n, p) = \binom{n}{k} p^k (1-p)^{n-k} \quad \xrightarrow{\begin{array}{l} n \rightarrow \infty \\ p \rightarrow 0 \end{array}} \quad P(k|\lambda) = \frac{\lambda^k e^{-\lambda}}{k!} \quad \text{where} \quad \lambda = np$$

| $n$           | 10     | 20     | 50    | 100   | 200   | 500    | 1000  | 2000   | 5000   | 10000  |
|---------------|--------|--------|-------|-------|-------|--------|-------|--------|--------|--------|
| $p$           | 0.5    | 0.25   | 0.1   | 0.05  | 0.025 | 0.01   | 0.005 | 0.0025 | 0.001  | 0.0005 |
| $P(k \leq 2)$ | 0.0547 | 0.0913 | 0.112 | 0.118 | 0.121 | 0.1234 | 0.124 | 0.1243 | 0.1245 | 0.1246 |

## Examples of Poisson variables

1. people killed by lightning per year;
2. mutations in DNA per generation;
3. radioactive disintegrations per unit time;
4. mass extinctions per 100 million years.

~~—earthquakes including aftershocks—~~

~~—floods per year~~

# Parameter estimation

$$P(k|\lambda) = \frac{\lambda^k e^{-\lambda}}{k!}$$

$$\mathcal{L}(\lambda|k) = \frac{\lambda^k e^{-\lambda}}{k!}$$

$$\lambda = k$$

## Hypothesis tests

$k = 9$  zircons given  $\lambda = 3.5$

1.  $H_0$  (null hypothesis)  $\lambda = 3.5$

$H_a$  (alternative hypothesis):  $\lambda > 3.5$

2. test statistic :  $k = 9$  successes

3. The null distribution

| k             | 0     | 1     | 2     | 3     | 4     | 5     | 6     | 7     | 8     | 9     | 10    |
|---------------|-------|-------|-------|-------|-------|-------|-------|-------|-------|-------|-------|
| $P(T = k)$    | 0.030 | 0.106 | 0.185 | 0.216 | 0.189 | 0.132 | 0.077 | 0.038 | 0.017 | 0.007 | 0.002 |
| $P(T \geq k)$ | 1.000 | 0.970 | 0.864 | 0.679 | 0.463 | 0.275 | 0.142 | 0.065 | 0.027 | 0.010 | 0.003 |

4. significance level  $\alpha = 0.05$

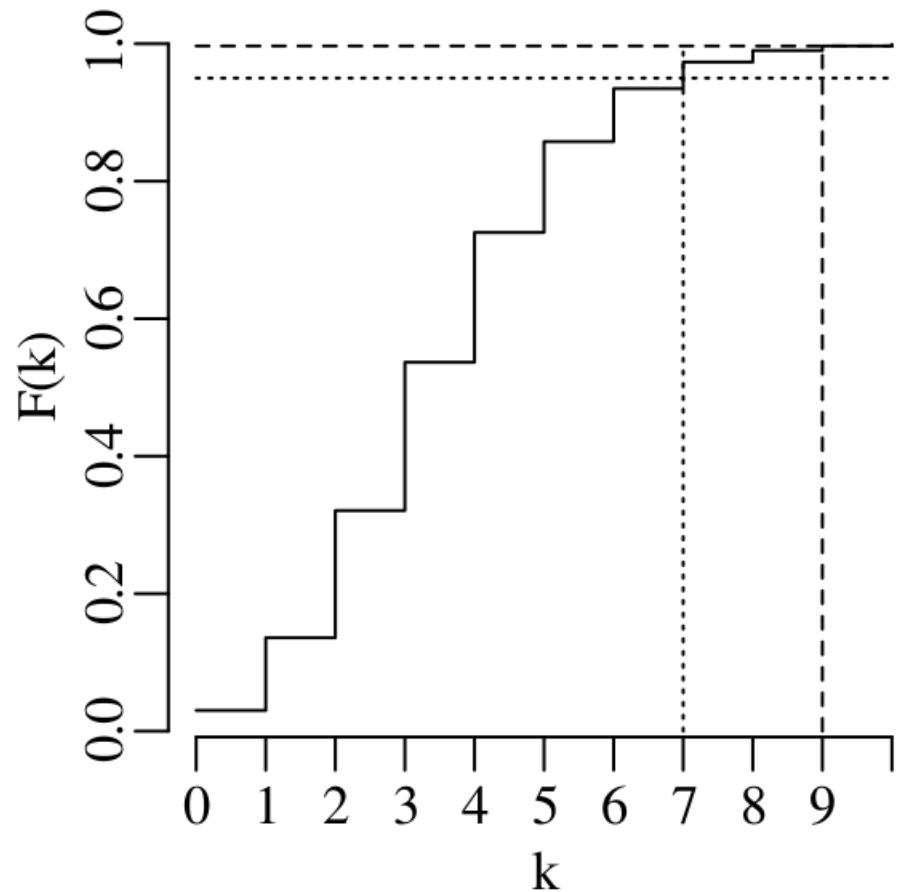
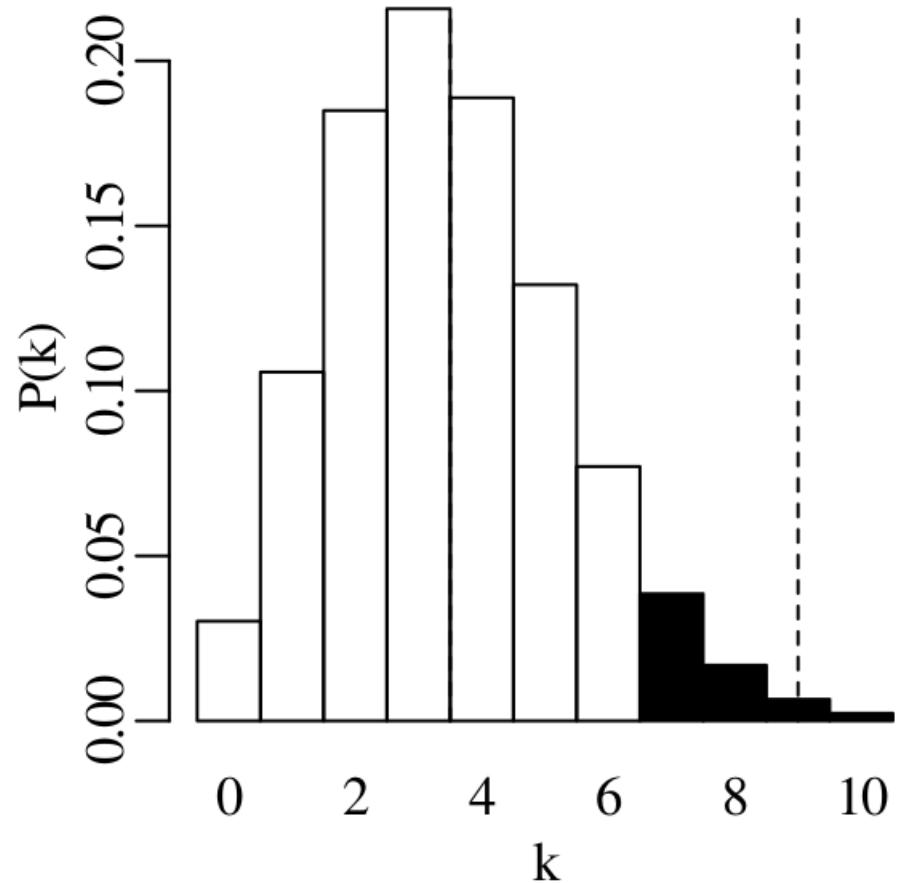
5. rejection region

| k             | 0     | 1     | 2     | 3     | 4     | 5     | 6     | 7     | 8            | 9            | 10           |
|---------------|-------|-------|-------|-------|-------|-------|-------|-------|--------------|--------------|--------------|
| $P(T = k)$    | 0.030 | 0.106 | 0.185 | 0.216 | 0.189 | 0.132 | 0.077 | 0.038 | 0.017        | 0.007        | 0.002        |
| $P(T \geq k)$ | 1.000 | 0.970 | 0.864 | 0.679 | 0.463 | 0.275 | 0.142 | 0.065 | <b>0.027</b> | <b>0.010</b> | <b>0.003</b> |

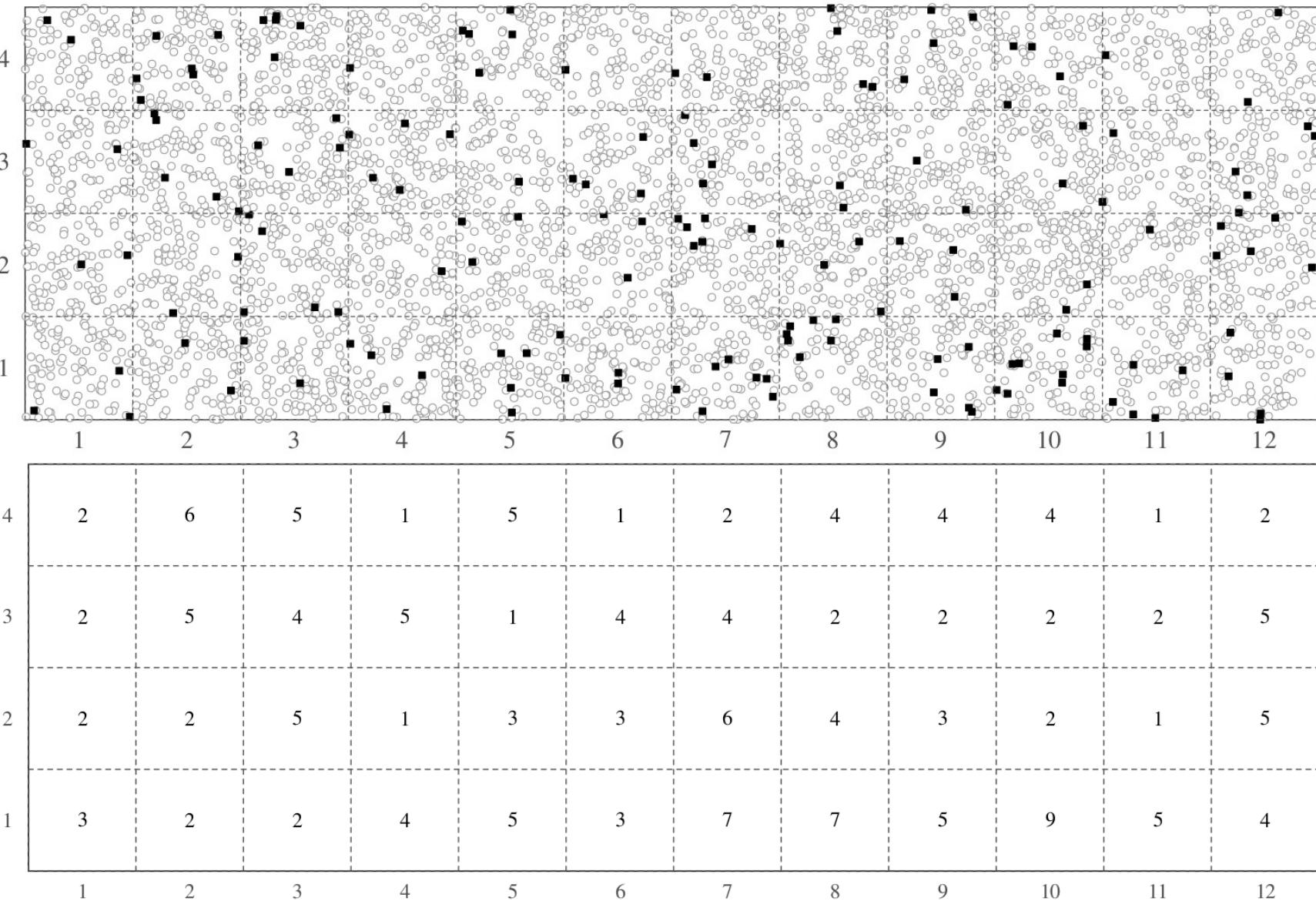
$$R = \{8, 9, 10, \dots, \infty\}$$

6.  $k \in R$

7. p-value =  $0.010 < \alpha$



# Multiple testing



Assuming that  $H_0$  is true, the probability of incurring a type-I error is:

$$1 \text{ experiment: } \alpha = 5\%$$

$$2 \text{ experiments: } 1 - (1 - \alpha)^2 = 9.75\%$$

$$3 \text{ experiments: } 1 - (1 - \alpha)^3 = 14\%$$

$$48 \text{ experiments: } 1 - (1 - \alpha)^{48} = 91.5\%$$

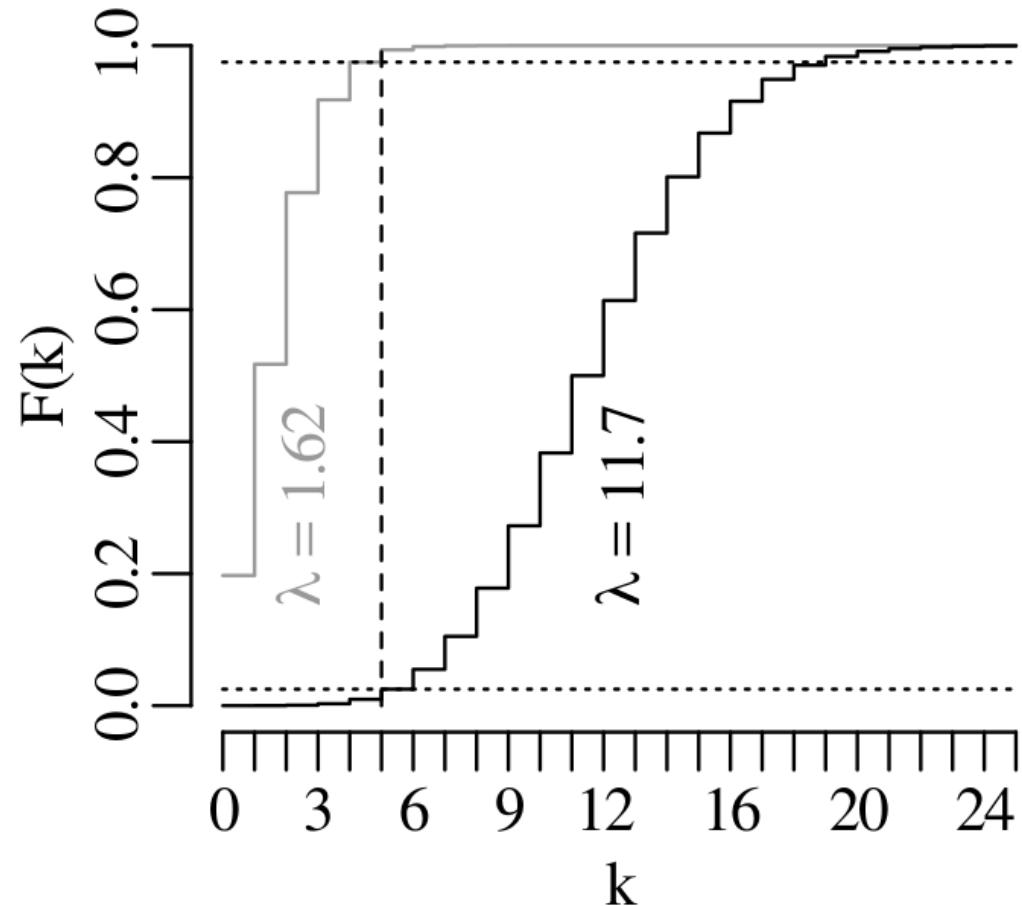
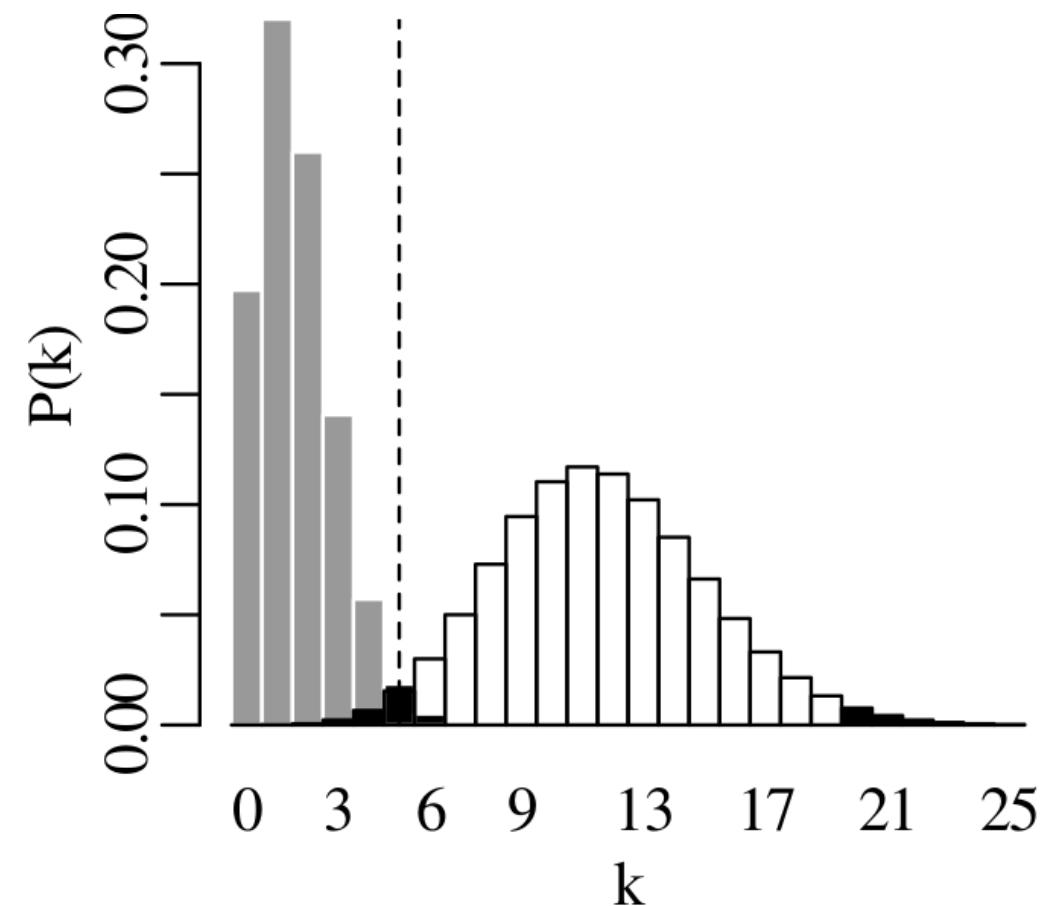
$$N \text{ experiments: } 1 - (1 - \alpha)^N\%$$

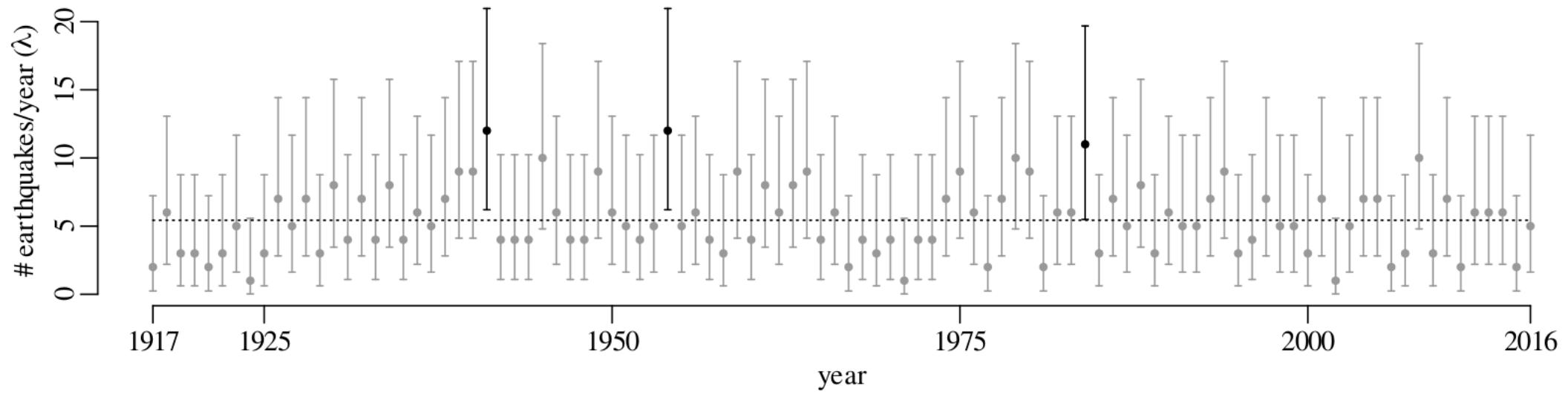
Bonferroni correction:  $\alpha/N$

# Confidence intervals

5 magnitude 5.0 or greater earthquakes occurred in the US in 2016

What is  $\lambda$ ?





# Statistics for geoscientists

The normal distribution

**the binomial distribution**

$$P(k|n,p) = \binom{n}{k} p^k (1-p)^{n-k}$$

**the Poisson distribution**

$$P(k|\lambda) = \frac{\lambda^k e^{-\lambda}}{k!}$$

**the negative binomial distribution**

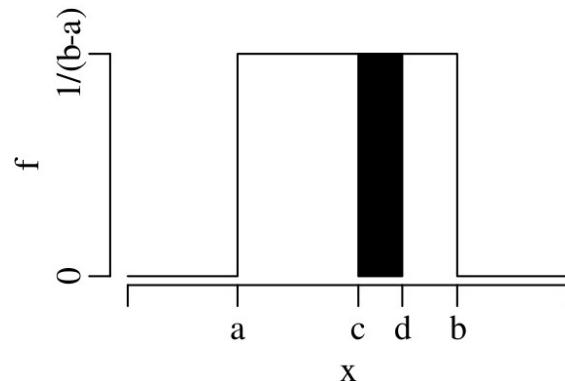
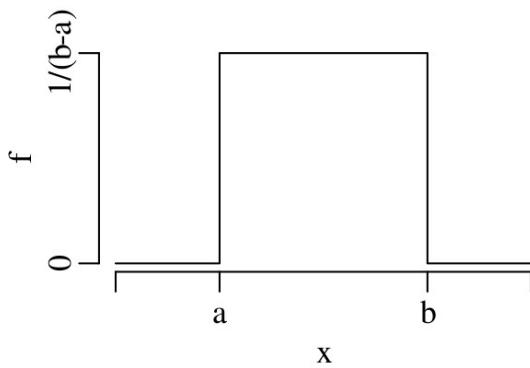
$$P(x|r,p) = \binom{r+x-1}{x} (1-p)^x p^r$$

**the multinomial distribution**

$$P(k_1, k_2, \dots, k_m | p_1, p_2, \dots, p_m) = \frac{n!}{\prod_{i=1}^m k_i!} \prod_{i=1}^m p_i^{k_i}$$

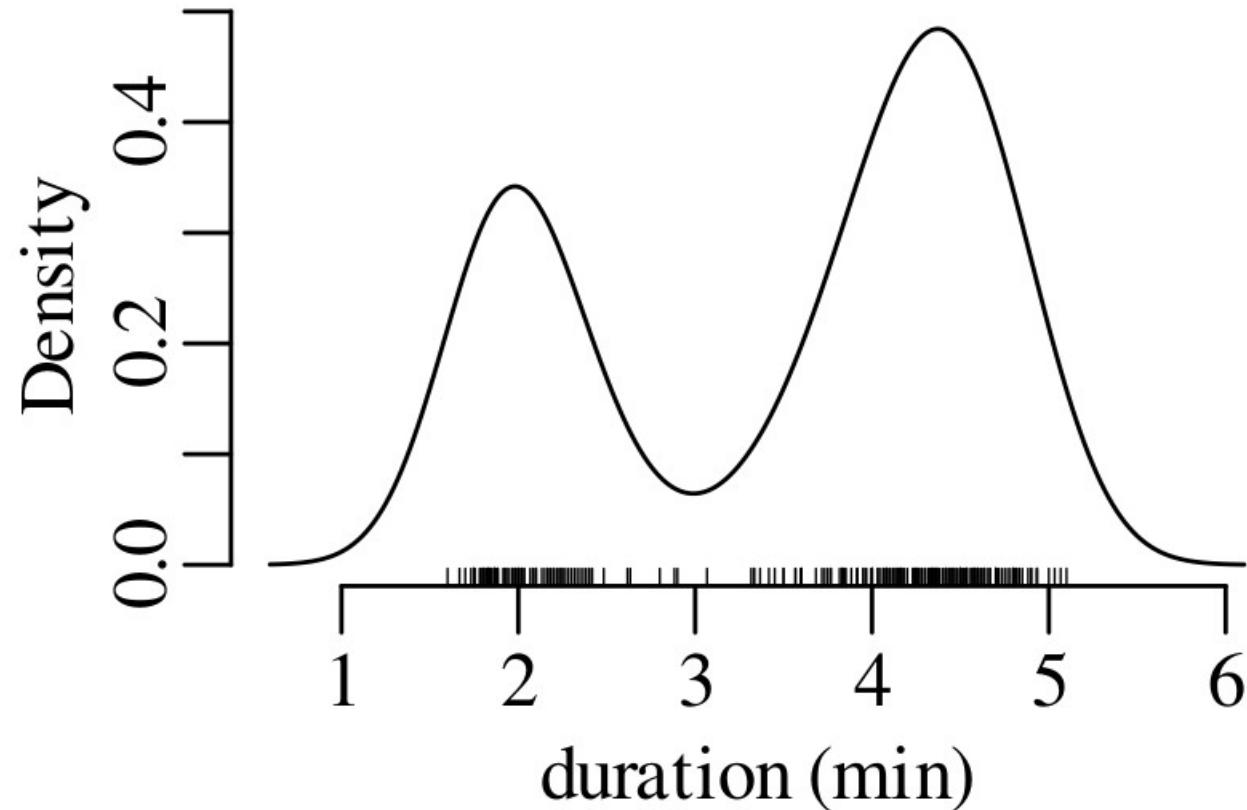
**the uniform distribution**

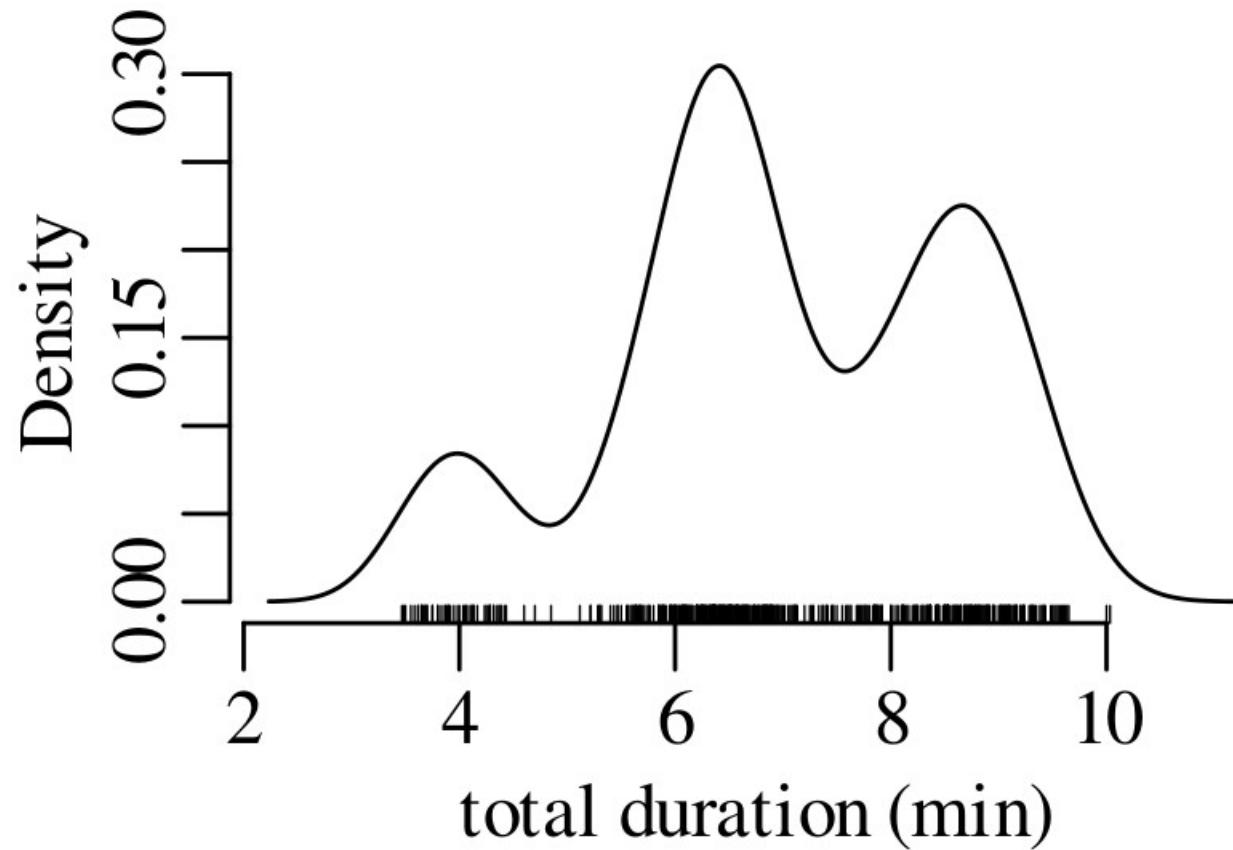
$$f(x|a,b) = \begin{cases} \frac{1}{b-a} & \text{if } a \leq x \leq b \\ 0 & \text{otherwise} \end{cases}$$

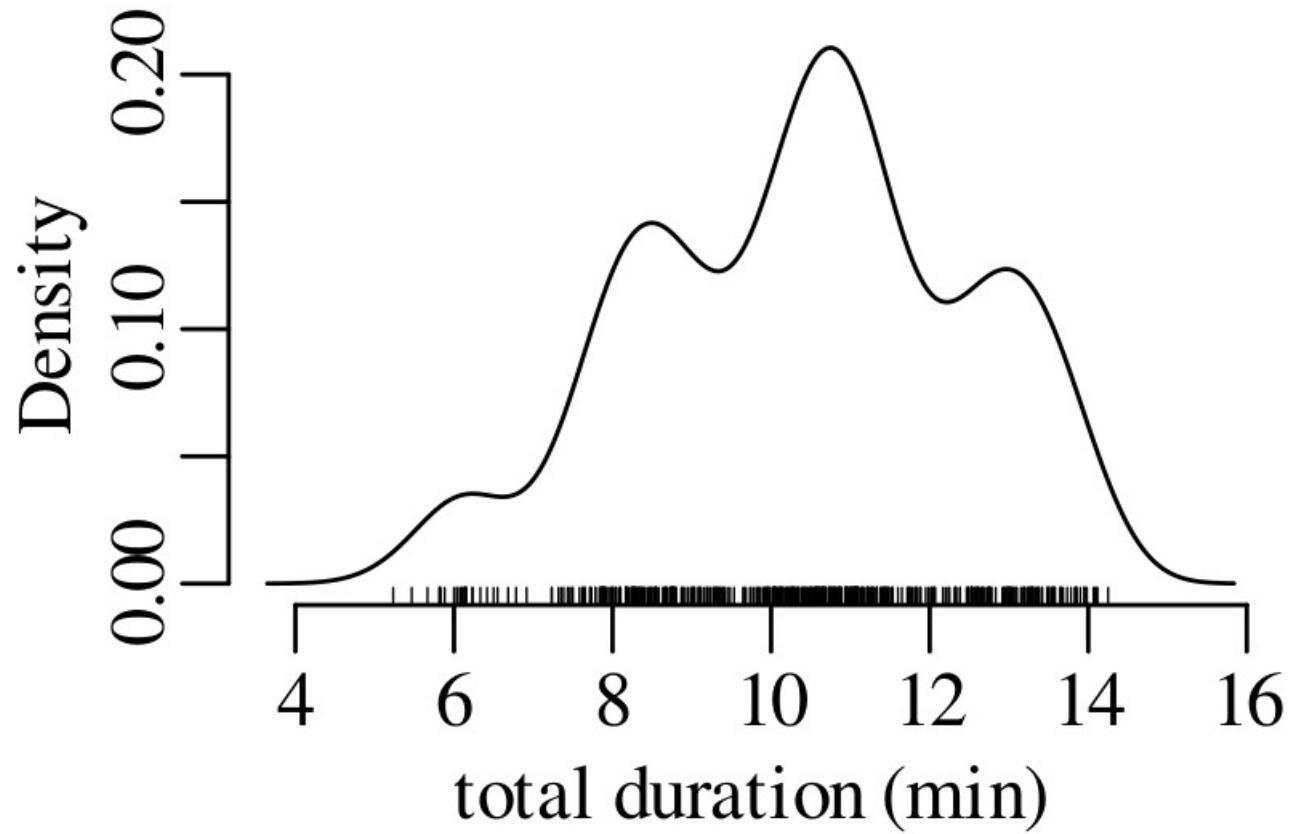


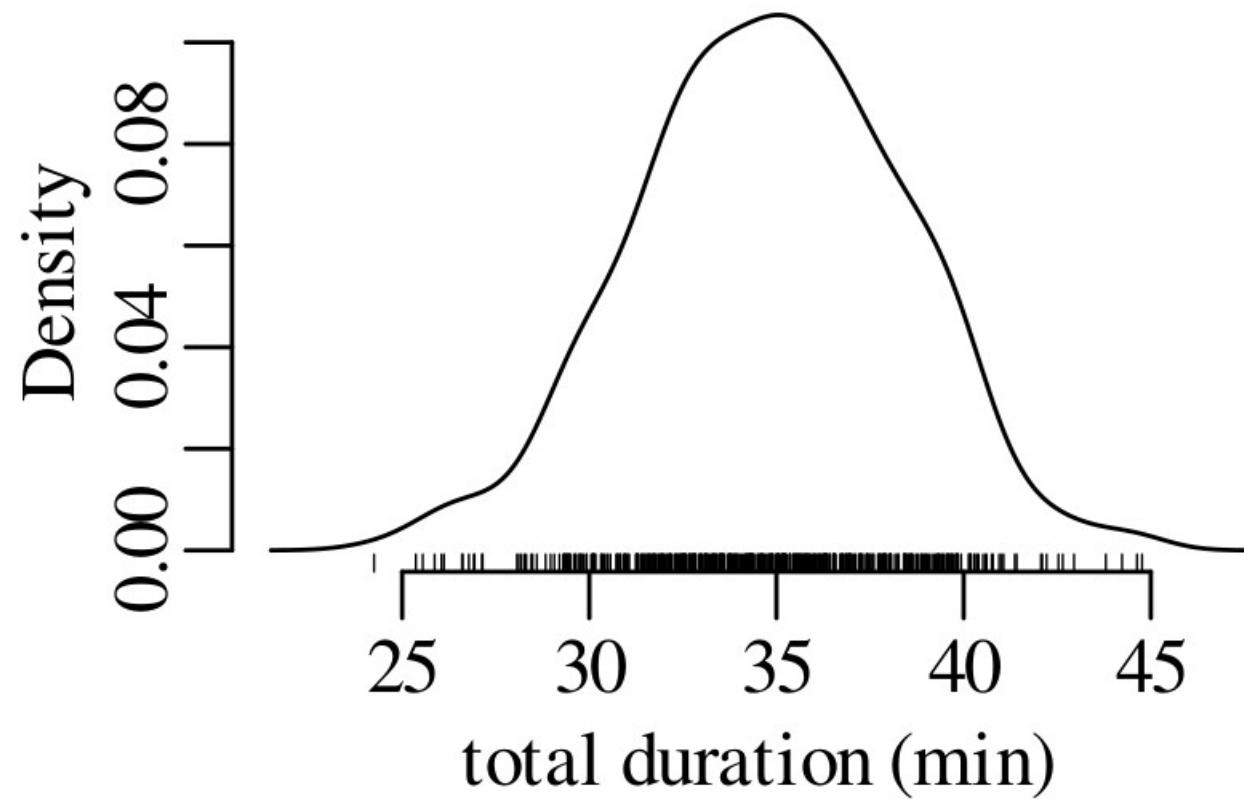
$$P(c \leq x \leq d) = \int_c^d f(x|a,b) dx$$

What is so normal about the Gaussian distribution?

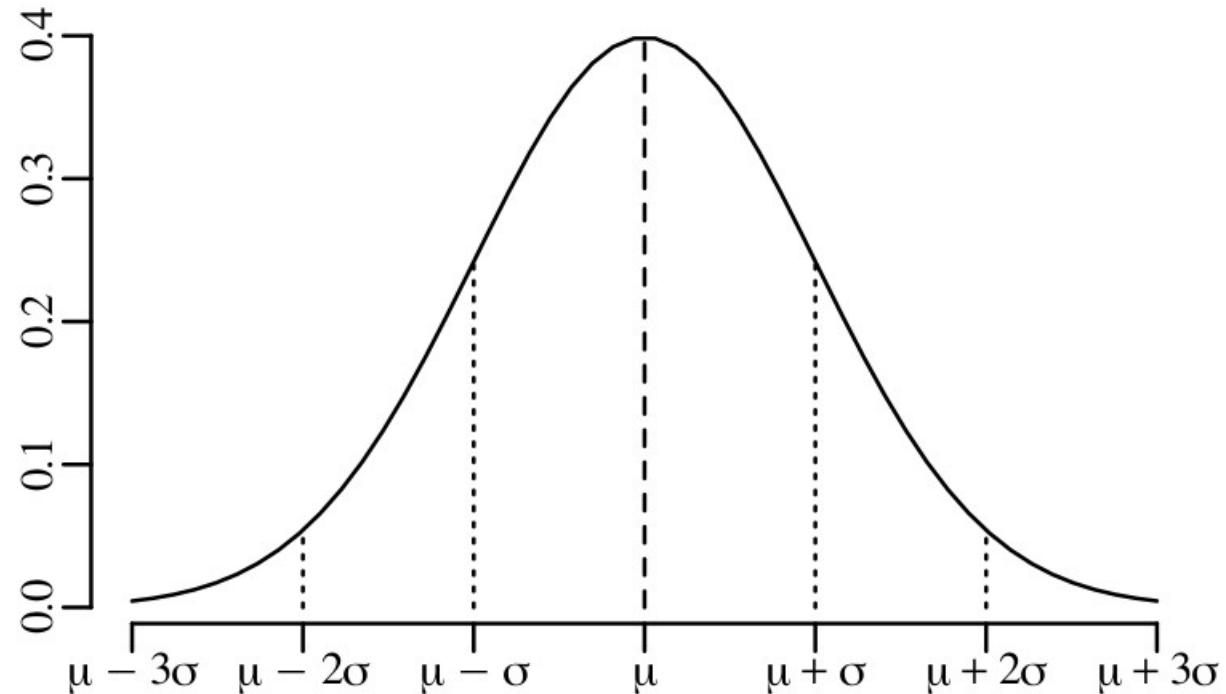


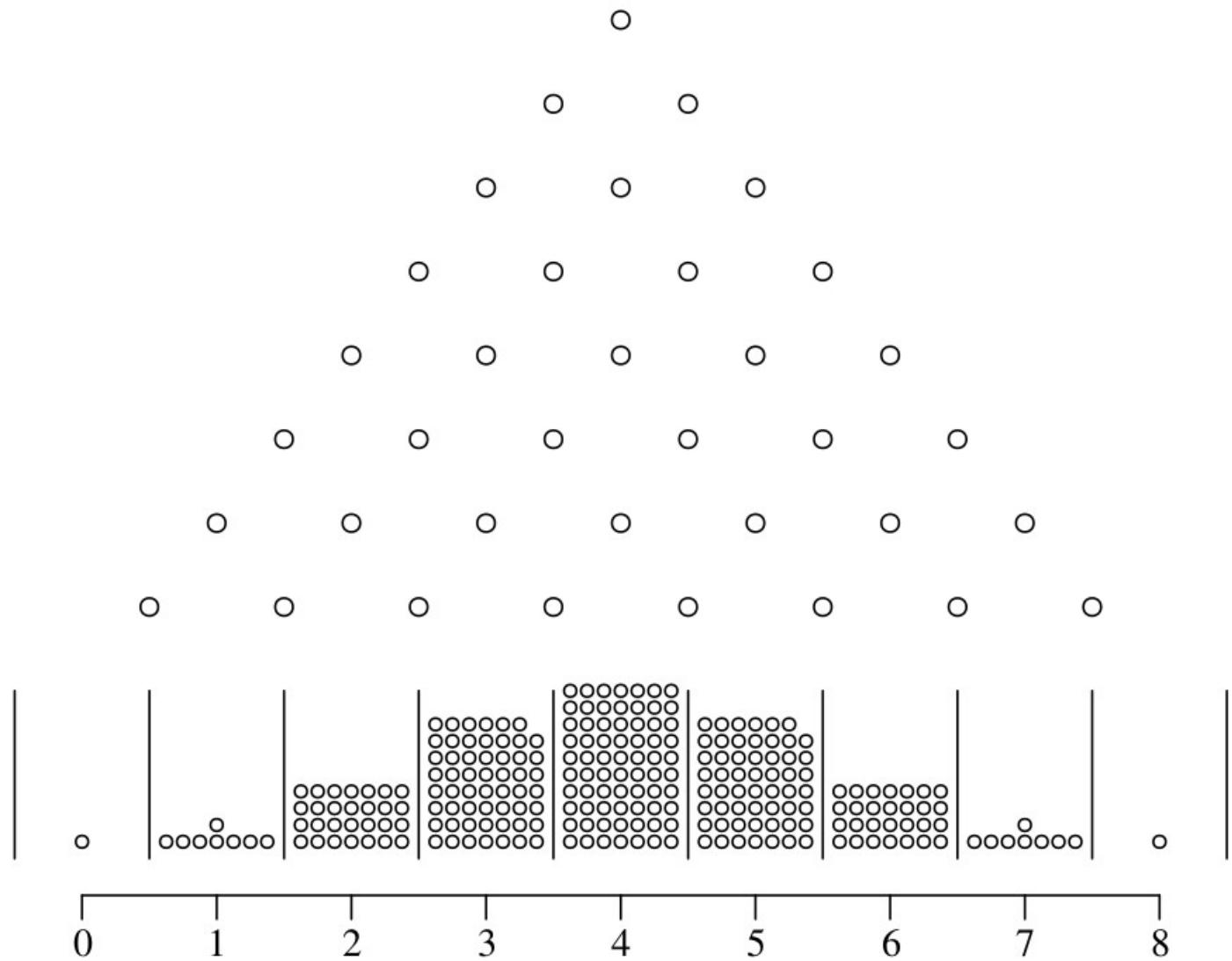


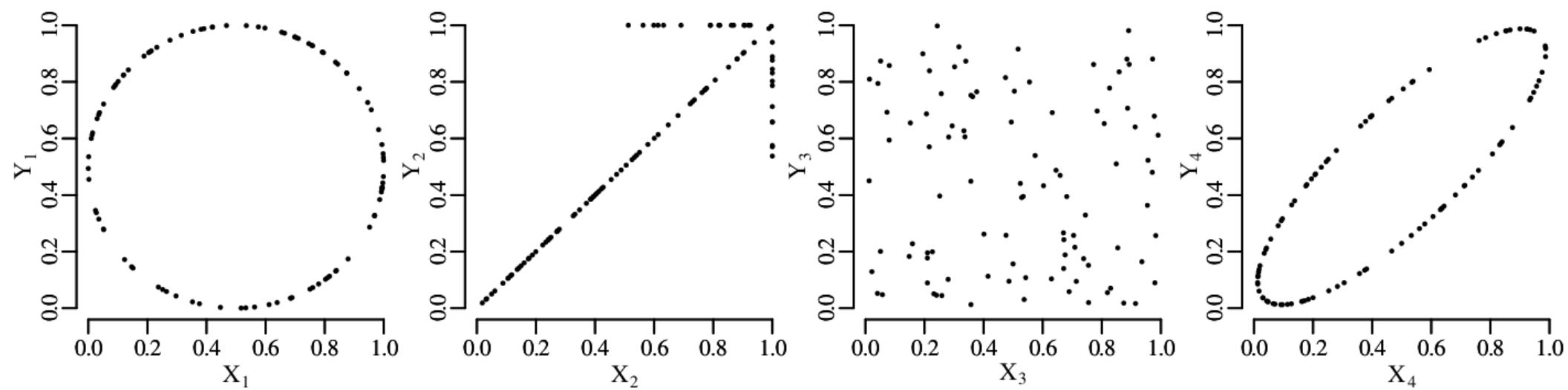
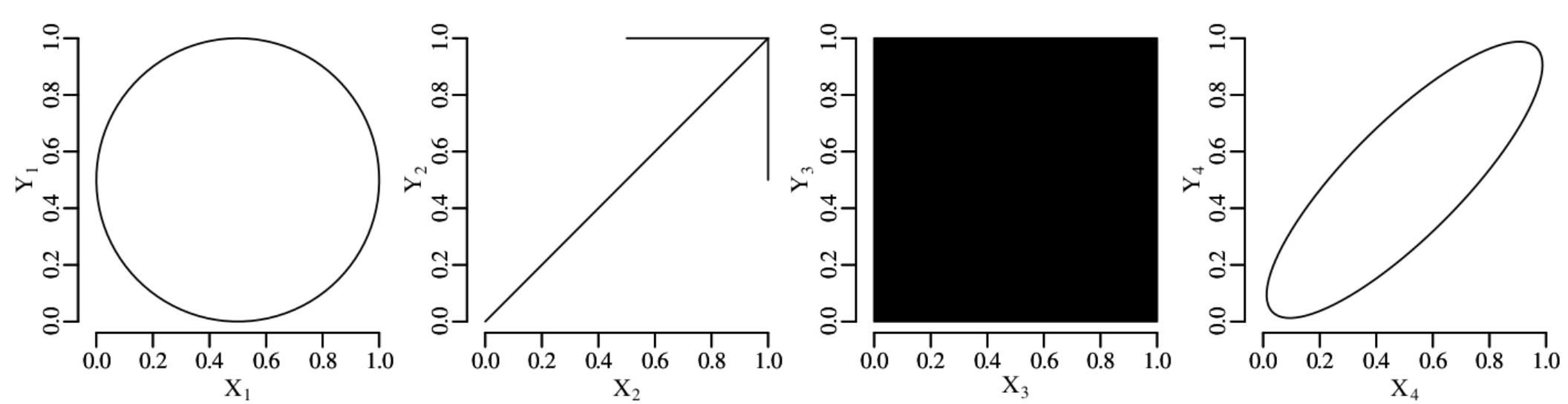


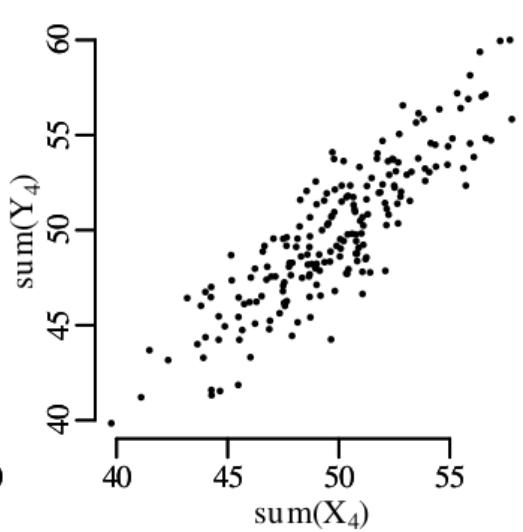
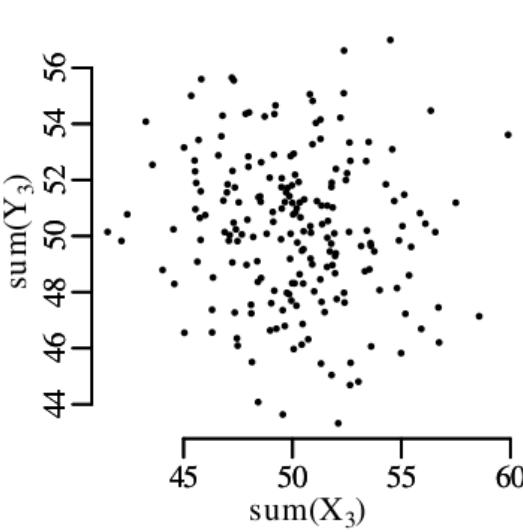
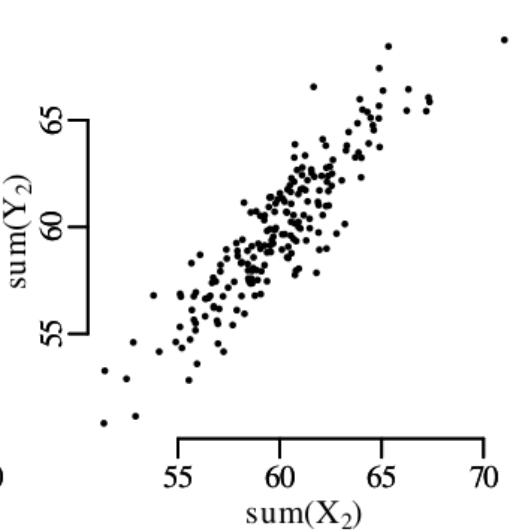
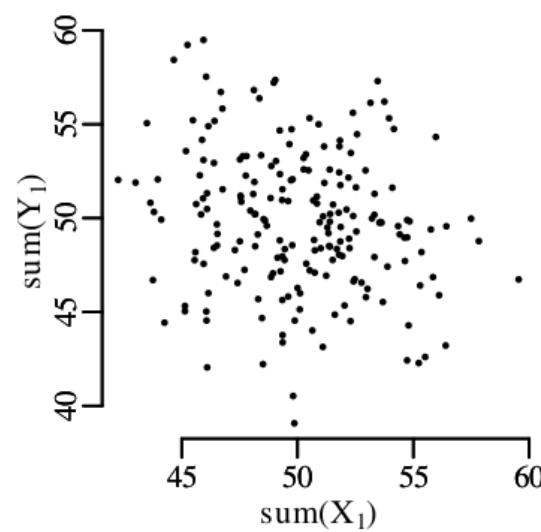
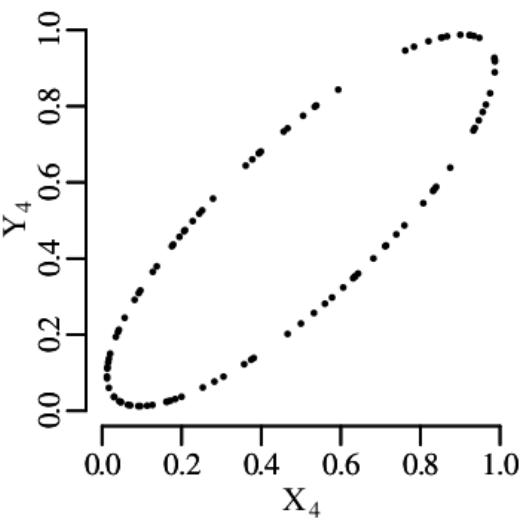
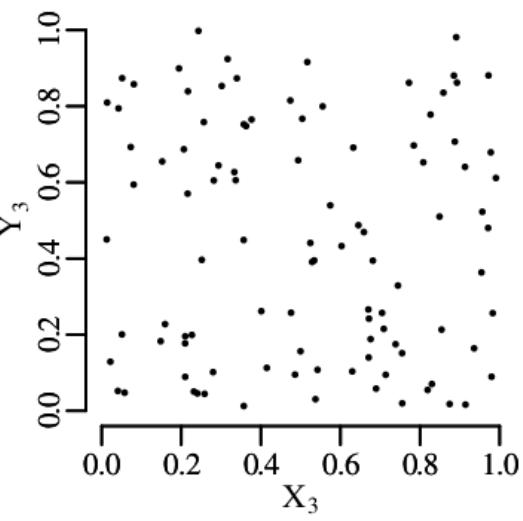
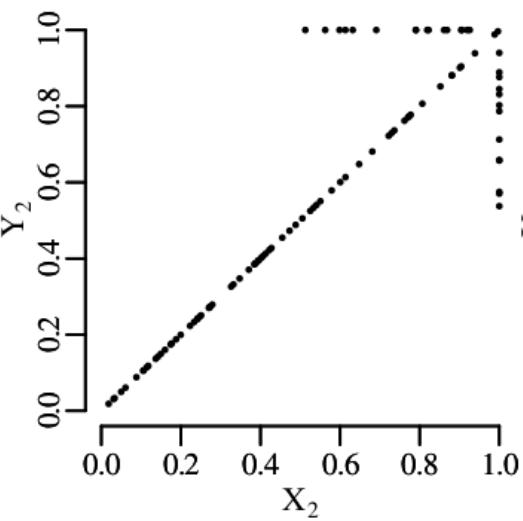
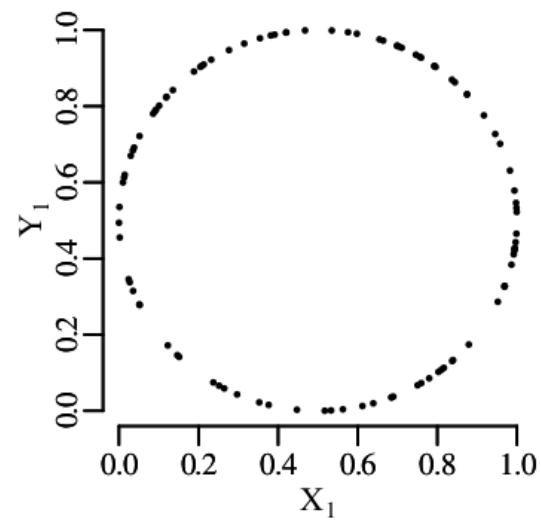


$$f(x|\mu, \sigma) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left[-\frac{(x-\mu)^2}{2\sigma^2}\right]$$

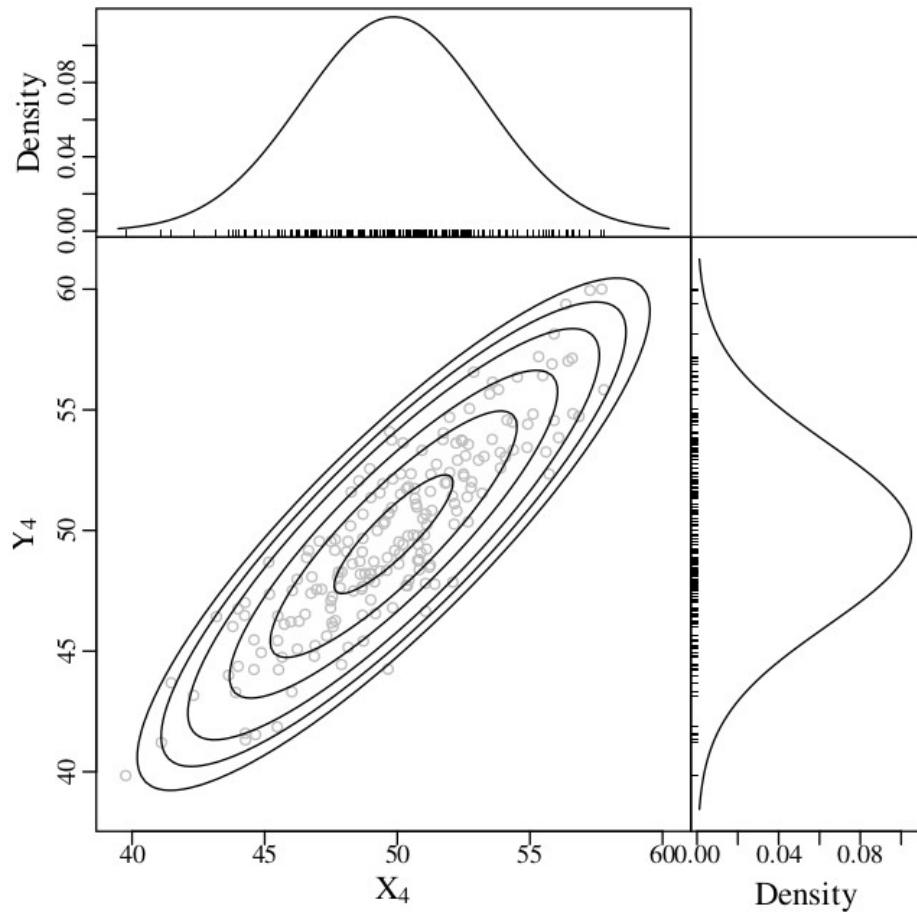




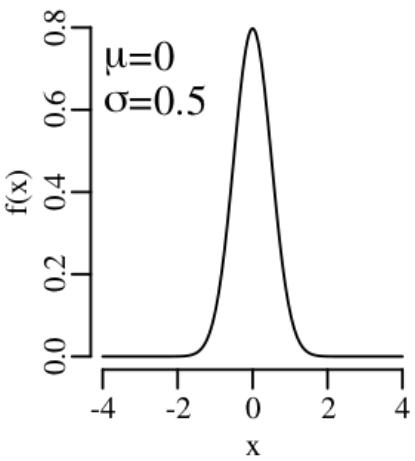
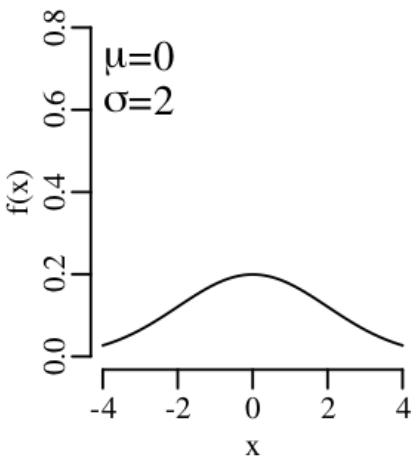
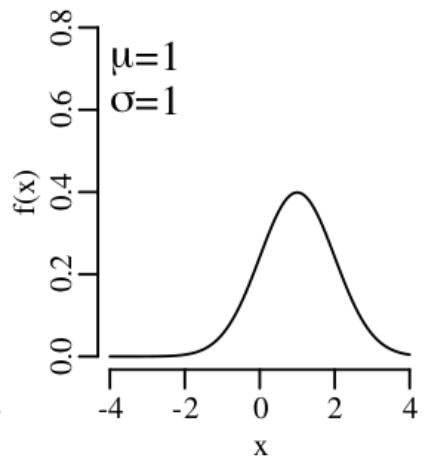
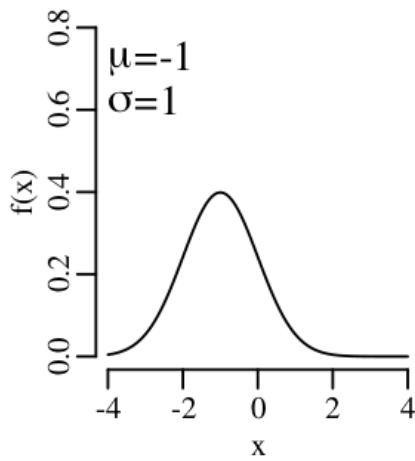
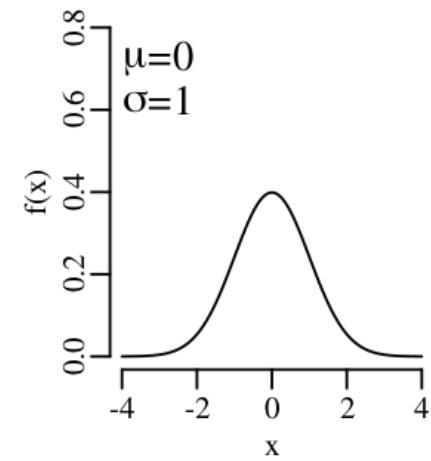




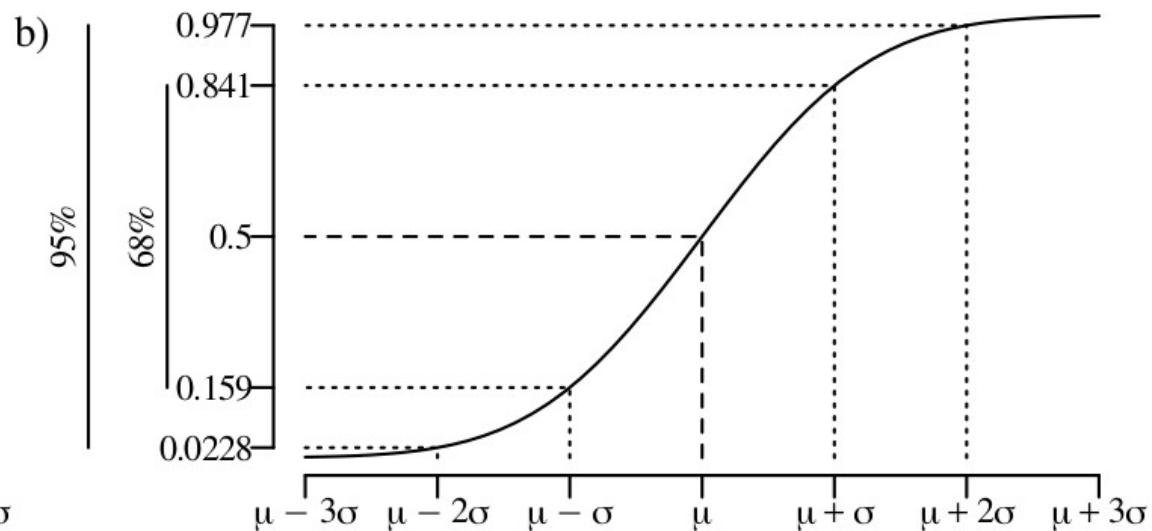
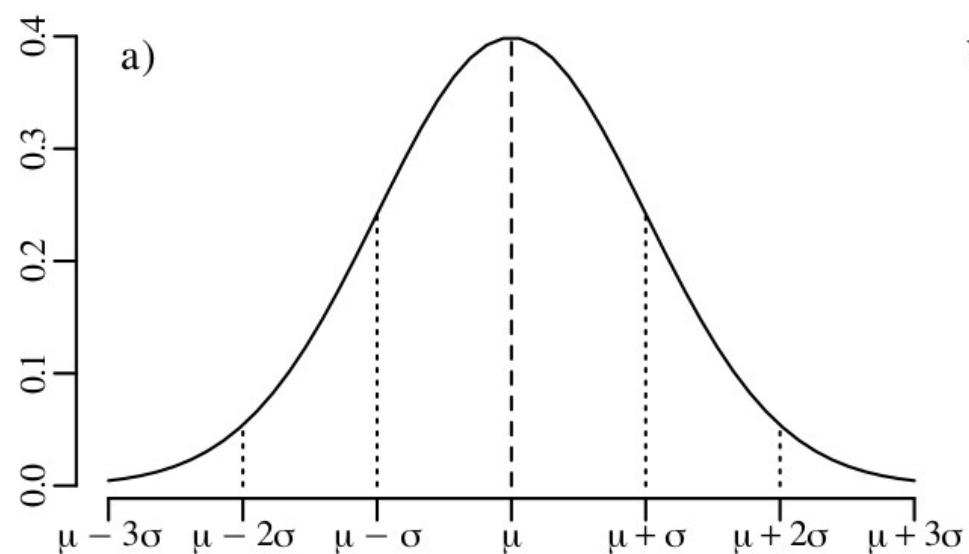
$$f(x, y | \mu_x, \mu_y, \sigma_x, \sigma_y, \sigma_{x,y}) = \frac{\exp\left(-\left[(x - \mu_x) \quad (y - \mu_y)\right] \begin{bmatrix} \sigma_x^2 & \sigma_{x,y} \\ \sigma_{x,y} & \sigma_y^2 \end{bmatrix}^{-1} \begin{bmatrix} x - \mu_x \\ y - \mu_y \end{bmatrix} / 2\right)}{2\pi \sqrt{\left|\begin{bmatrix} \sigma_x^2 & \sigma_{x,y} \\ \sigma_{x,y} & \sigma_y^2 \end{bmatrix}\right|}}$$



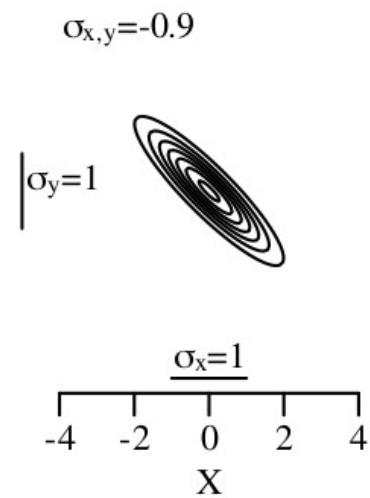
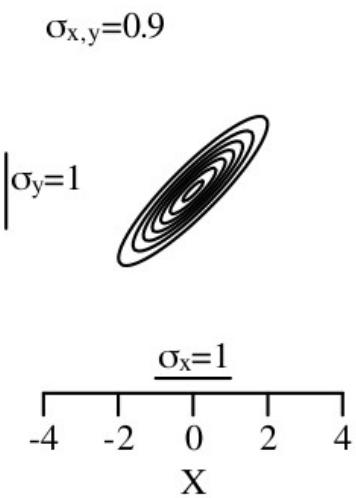
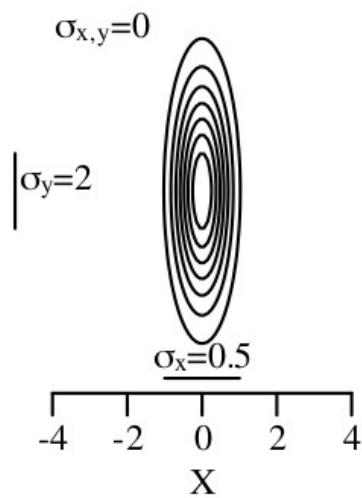
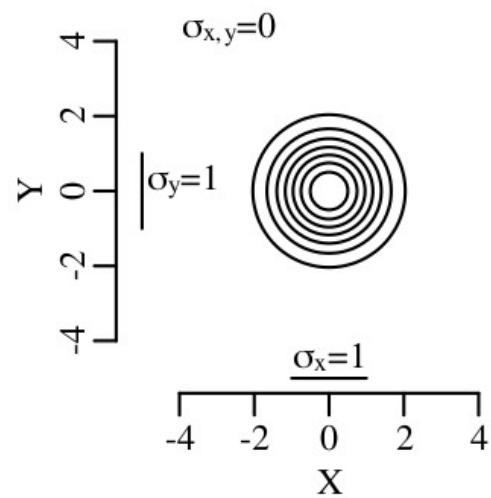
$$f(x|\mu, \sigma) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left[-\frac{(x-\mu)^2}{2\sigma^2}\right]$$



$$f(x|\mu, \sigma) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left[-\frac{(x-\mu)^2}{2\sigma^2}\right]$$



$$f(x, y | \mu_x, \mu_y, \sigma_x, \sigma_y, \sigma_{x,y}) = \frac{\exp\left(-\begin{bmatrix}(x - \mu_x) & (y - \mu_y)\end{bmatrix} \begin{bmatrix}\sigma_x^2 & \sigma_{x,y} \\ \sigma_{x,y} & \sigma_y^2\end{bmatrix}^{-1} \begin{bmatrix}x - \mu_x \\ y - \mu_y\end{bmatrix} / 2\right)}{2\pi \sqrt{\begin{vmatrix}\sigma_x^2 & \sigma_{x,y} \\ \sigma_{x,y} & \sigma_y^2\end{vmatrix}}}$$



# Parameter estimation

$$\mathcal{L}(\mu, \sigma | x_1, x_2, \dots, x_n) = \prod_{i=1}^n f(x_i | \mu, \sigma) \quad \text{with} \quad f(x | \mu, \sigma) = \frac{1}{\sigma \sqrt{2\pi}} \exp \left[ -\frac{(x - \mu)^2}{2\sigma^2} \right]$$

$$\hat{\mu} = \frac{1}{n} \sum_{i=1}^n x_i \qquad \longleftrightarrow \qquad \bar{x} = \frac{1}{n} \sum_{i=1}^n x_i$$

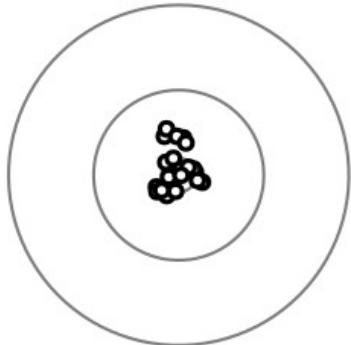
$$\hat{\sigma} = \sqrt{\frac{1}{n} \sum_{i=1}^n (x_i - \mu)^2} \qquad \longleftrightarrow \qquad s[x] = \sqrt{\frac{1}{n-1} \sum_{i=1}^n (x_i - \bar{x})^2}$$

$$\hat{\sigma}_{x,y} = \sum_{i=1}^n \frac{1}{n} (x_i - \mu_x)(y_i - \mu_y) \qquad \longleftrightarrow \qquad s[x, y] = \sum_{i=1}^n \frac{1}{n-1} (x_i - \bar{x})(y_i - \bar{y})$$

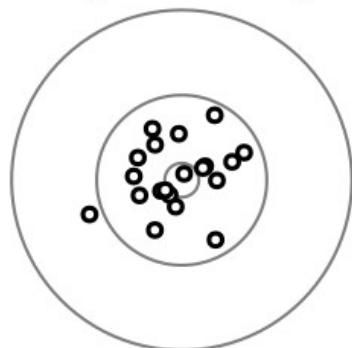
# Statistics for geoscientists

Error propagation

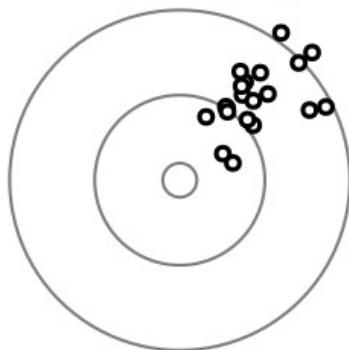
high precision  
high accuracy



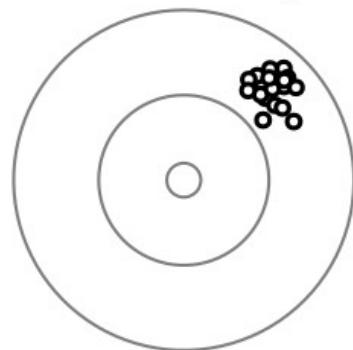
low precision  
high accuracy



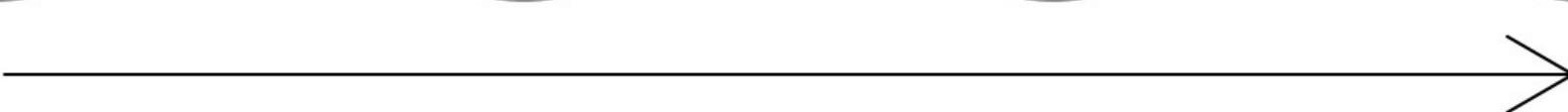
low precision  
low accuracy



high precision  
low accuracy



good



bad

$z = g(x) \rightarrow$  given  $x_i$ , for  $i = 1 \dots n$

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

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

$\bar{z} = g(\bar{x}) \longrightarrow$  What is  $s[z]$ ?

# Linear approximation

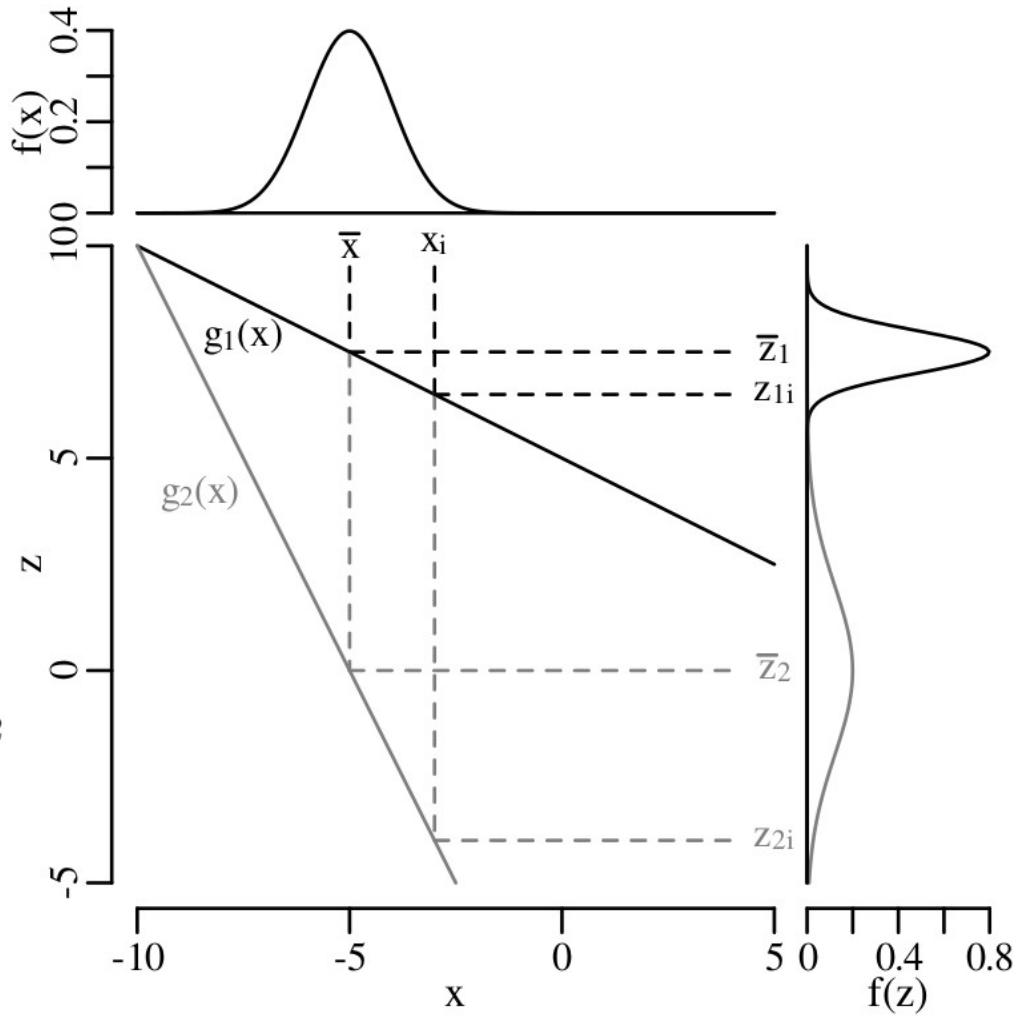
$$z = g(x)$$

$$(z_i - \bar{z}) \approx \frac{\partial z}{\partial x} (x_i - \bar{x})$$

$$s[z]^2 \equiv \frac{1}{n-1} \sum_{i=1}^n (z_i - \bar{z})^2$$

$$\begin{aligned} s[z]^2 &\approx \frac{1}{n-1} \sum_{i=1}^n \left[ (x_i - \bar{x}) \frac{\partial z}{\partial x} \right]^2 \\ &= \left[ \frac{\partial z}{\partial x} \right]^2 \left[ \frac{1}{n-1} \sum_{i=1}^n (x_i - \bar{x})^2 \right] = \left[ \frac{\partial z}{\partial x} \right]^2 s[x]^2 \end{aligned}$$

$$\Rightarrow s[z] = \left| \frac{\partial z}{\partial x} \right| s[x]$$

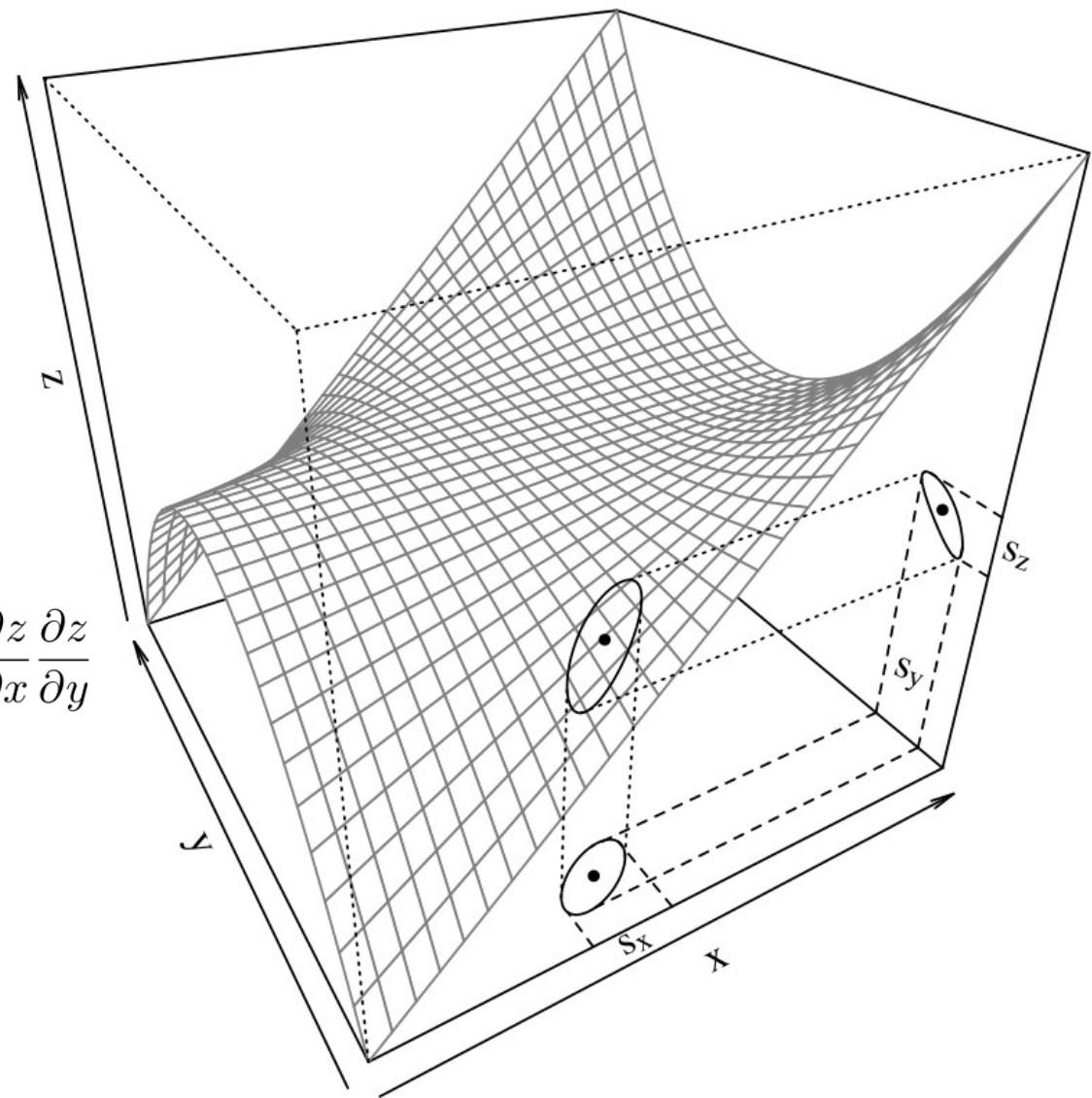


$$z = g(x, y)$$

$$s[z]^2 \approx \frac{1}{n-1} \sum_{i=1}^n \left[ (x_i - \bar{x}) \frac{\partial z}{\partial x} + (y_i - \bar{y}) \frac{\partial z}{\partial y} \right]^2$$

$$s[z]^2 \approx s[x]^2 \left( \frac{\partial z}{\partial x} \right)^2 + s[y]^2 \left( \frac{\partial z}{\partial y} \right)^2 + 2 s[x, y] \frac{\partial z}{\partial x} \frac{\partial z}{\partial y}$$

$$s[z]^2 \approx \begin{bmatrix} \frac{\partial z}{\partial x} \frac{\partial z}{\partial y} \end{bmatrix} \begin{bmatrix} s[x]^2 & s[x, y] \\ s[x, y] & s[y]^2 \end{bmatrix} \begin{bmatrix} \frac{\partial z}{\partial x} \\ \frac{\partial z}{\partial y} \end{bmatrix}$$



$$s[z]^2 \approx s[x]^2 \left( \frac{\partial z}{\partial x} \right)^2 + s[y]^2 \left( \frac{\partial z}{\partial y} \right)^2 + 2 s[x, y] \frac{\partial z}{\partial x} \frac{\partial z}{\partial y}$$

**addition**

$$z = a + bx + cy$$

$$\frac{\partial z}{\partial x} = b \text{ and } \frac{\partial z}{\partial y} = c$$

$$s[z]^2 = b^2 s[x]^2 + c^2 s[y]^2 + 2bc s[x, y]$$

$$z = x + y \longrightarrow s[z]^2 = s[x]^2 + s[y]^2$$

$$s[z]^2 \approx s[x]^2 \left( \frac{\partial z}{\partial x} \right)^2 + s[y]^2 \left( \frac{\partial z}{\partial y} \right)^2 + 2 s[x, y] \frac{\partial z}{\partial x} \frac{\partial z}{\partial y}$$

**subtraction**

$$z = ax - by$$

$$\frac{\partial z}{\partial x} = a \text{ and } \frac{\partial z}{\partial y} = -b$$

$$s[z]^2 = a^2 s[x]^2 + b^2 s[y]^2 - 2ab s[x, y]$$

$$z = x - y \longrightarrow s[z]^2 = s[x]^2 + s[y]^2$$

$$s[z]^2 \approx s[x]^2 \left( \frac{\partial z}{\partial x} \right)^2 + s[y]^2 \left( \frac{\partial z}{\partial y} \right)^2 + 2 s[x, y] \frac{\partial z}{\partial x} \frac{\partial z}{\partial y}$$

**multiplication**

$$z = axy$$

$$\frac{\partial z}{\partial x} = ay \text{ and } \frac{\partial z}{\partial y} = ax$$

$$s[z]^2 = (ay)^2 s[x]^2 + (ax)^2 s[y]^2 + 2(ay)(ax) s[x, y]$$

$$\left( \frac{s[z]}{z} \right)^2 = \left( \frac{ay}{axy} s[x] \right)^2 + \left( \frac{ax}{axy} s[y] \right)^2 + 2 \left( \frac{ay}{axy} \right) \left( \frac{ax}{axy} \right) s[x, y]$$

$$\left( \frac{s[z]}{z} \right)^2 = \left( \frac{s[x]}{x} \right)^2 + \left( \frac{s[y]}{y} \right)^2 + 2 \frac{s[x, y]}{xy}$$

**addition**       $z = a + bx + cy$        $s[z]^2 = b^2 s[x]^2 + c^2 s[y]^2 + 2bc s[x, y]$

**subtraction**       $z = ax - by$        $s[z]^2 = a^2 s[x]^2 + b^2 s[y]^2 - 2ab s[x, y]$

**multiplication**       $z = axy$        $\left(\frac{s[z]}{z}\right)^2 = \left(\frac{s[x]}{x}\right)^2 + \left(\frac{s[y]}{y}\right)^2 + 2\frac{s[x, y]}{xy}$

**division**       $z = a\frac{x}{y}$        $\left(\frac{s[z]}{z}\right)^2 = \left(\frac{s[x]}{x}\right)^2 + \left(\frac{s[y]}{y}\right)^2 - 2\frac{s[x, y]}{xy}$

**exponentiation**       $z = ae^{bx}$        $\left(\frac{s[z]}{z}\right)^2 = b^2 s[x]^2$

**logarithms**       $z = a \ln[bx]$        $s[z]^2 = a^2 \left(\frac{s[x]}{x}\right)^2$

**power**       $z = ax^b$        $\left(\frac{s[z]}{z}\right)^2 = b^2 \left(\frac{s[x]}{x}\right)^2$

# the chain rule

$$d = d_o + v_o t + g t^2$$

can be written as

$$d = x + y \quad \text{where} \quad x \equiv d_o + v_o t \quad \text{and} \quad y \equiv g t^2$$

$$\begin{aligned} s[d] &= \sqrt{s[x]^2 + s[y]^2} \\ &\quad \xrightarrow{\hspace{10cm}} s[d] = \sqrt{(v_o s[t])^2 + g^2 \left(\frac{s[t]}{t}\right)^2} \end{aligned}$$

$$z = g(x)$$

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

---

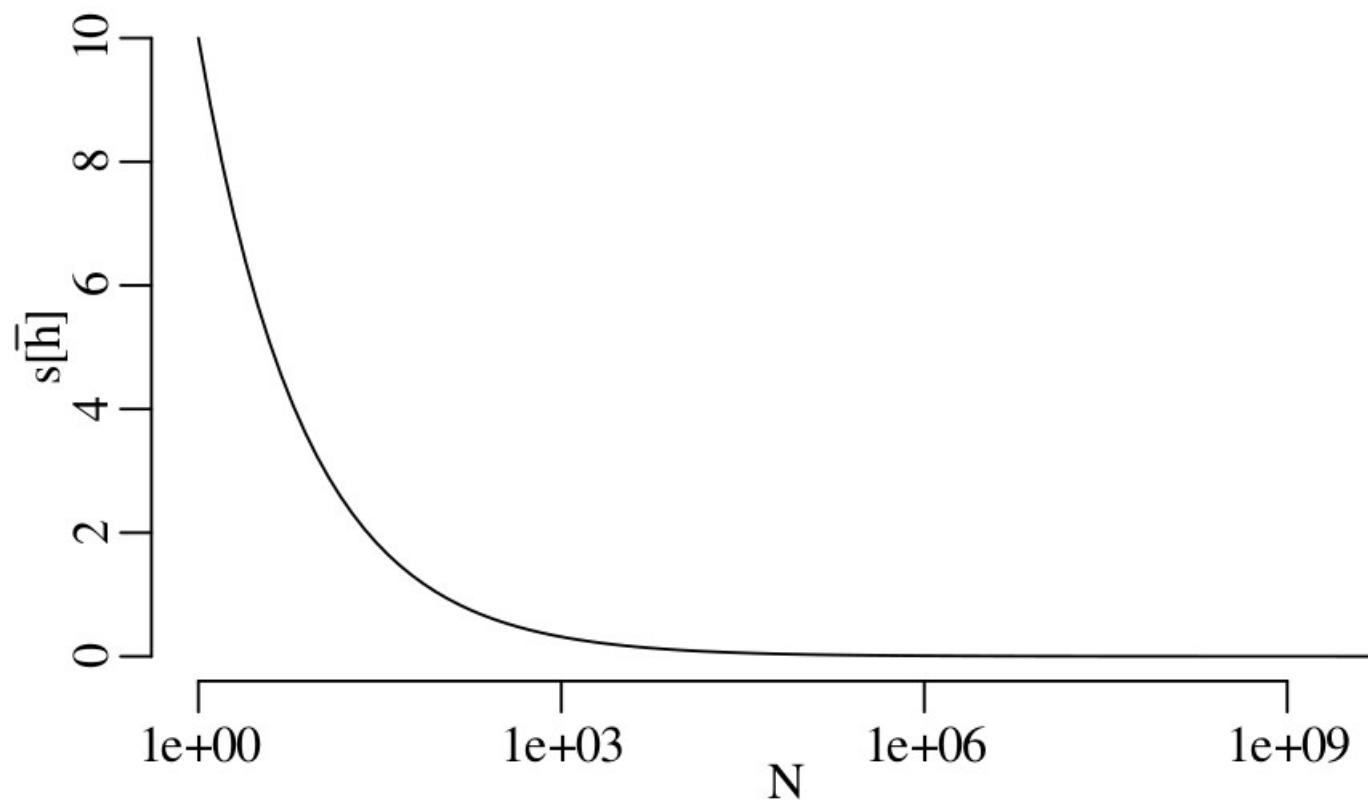
**addition**       $z = a + bx + cy$        $s[z]^2 = b^2 s[x]^2 + c^2 s[y]^2$

---

$$s[\bar{x}]^2 = \sum_{i=1}^n \left( \frac{s[x_i]}{n} \right)^2 = \left( \frac{1}{n} \right)^2 \sum_{i=1}^n s[x_i]^2$$

$$s[\bar{x}]^2 = \left( \frac{1}{n} \right)^2 \sum_{i=1}^n s[x]^2 = n \left( \frac{1}{n} \right)^2 s[x]^2$$

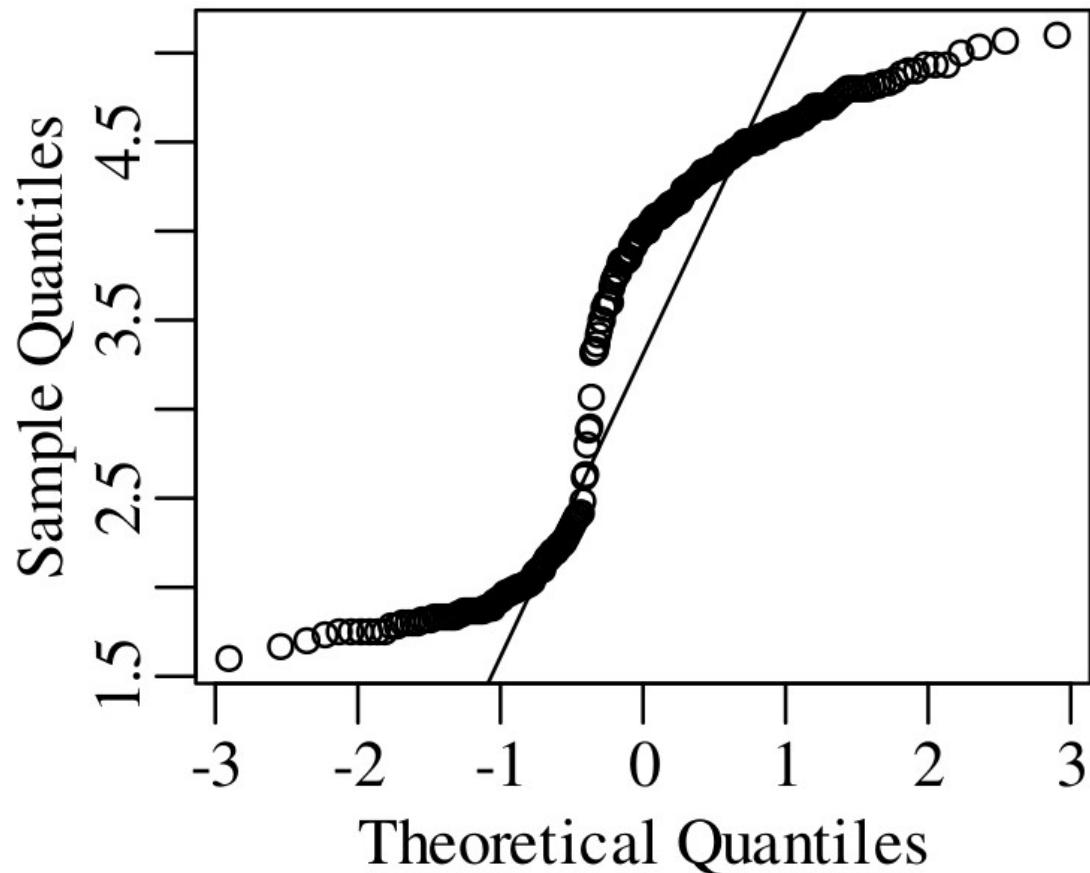
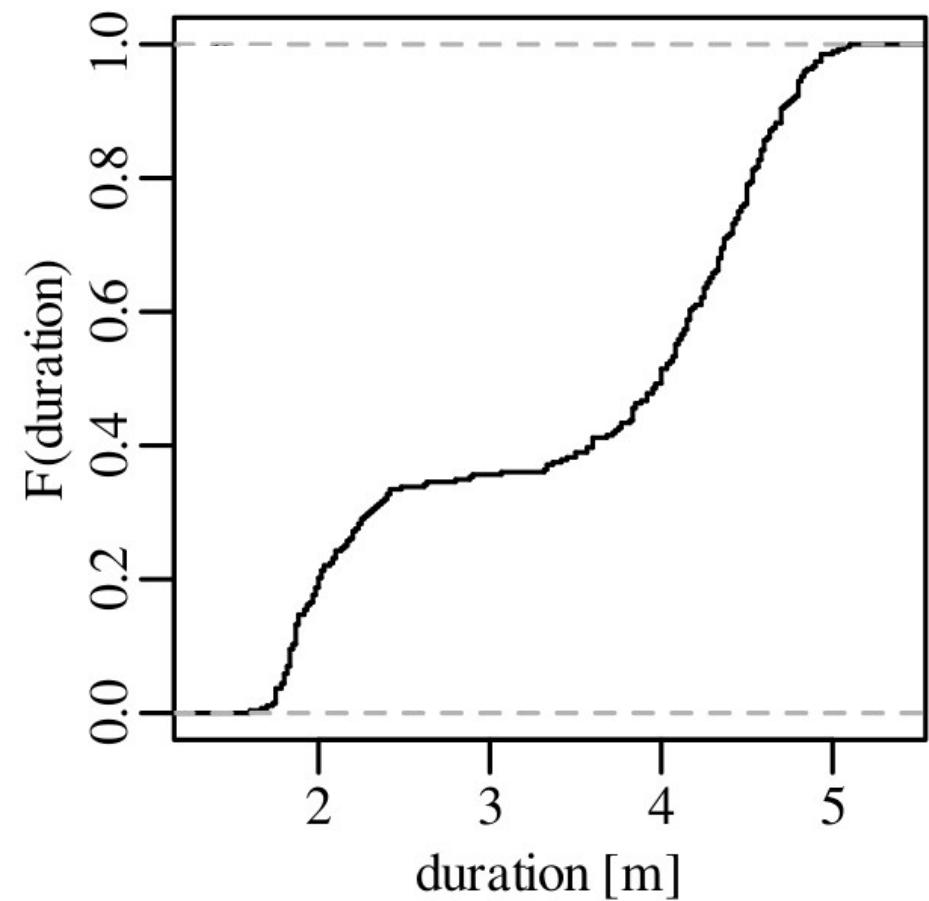
$$s[\bar{x}] = \frac{s[x]}{\sqrt{n}}$$

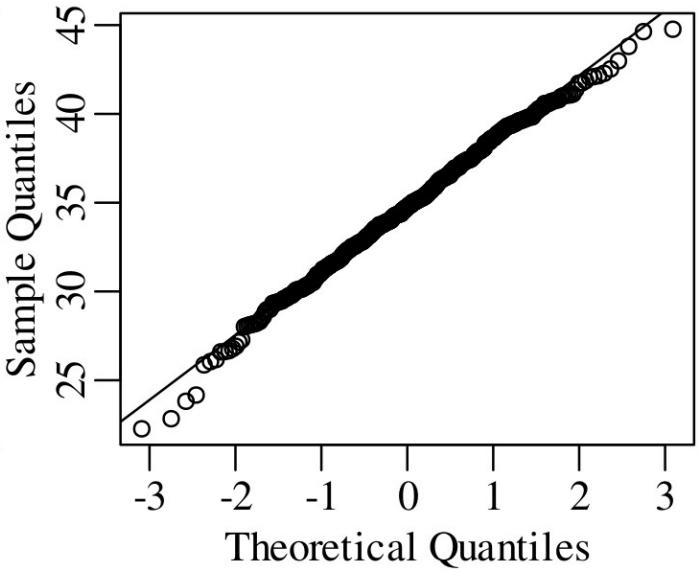
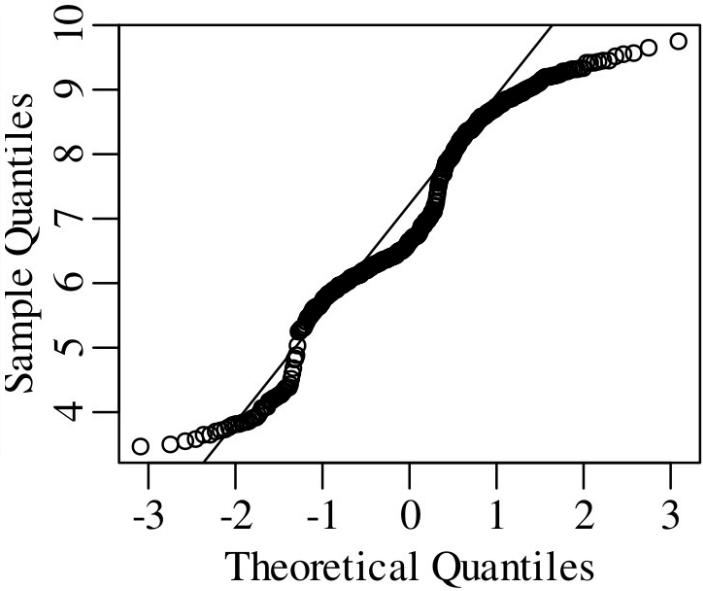
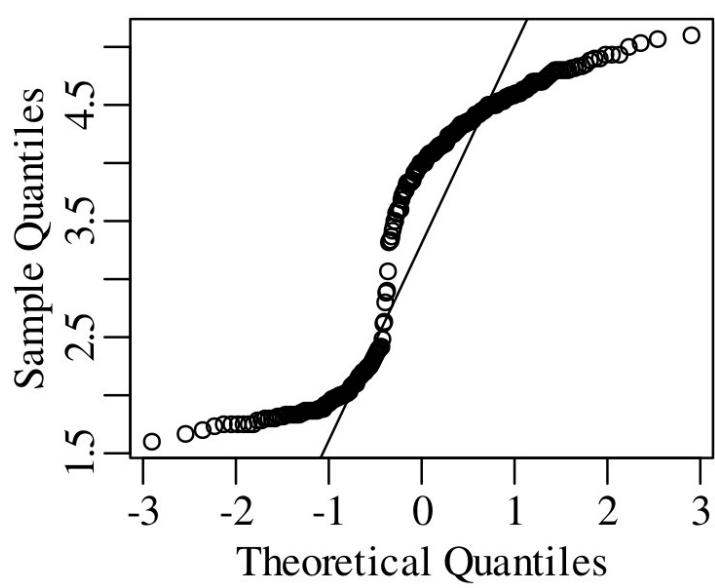


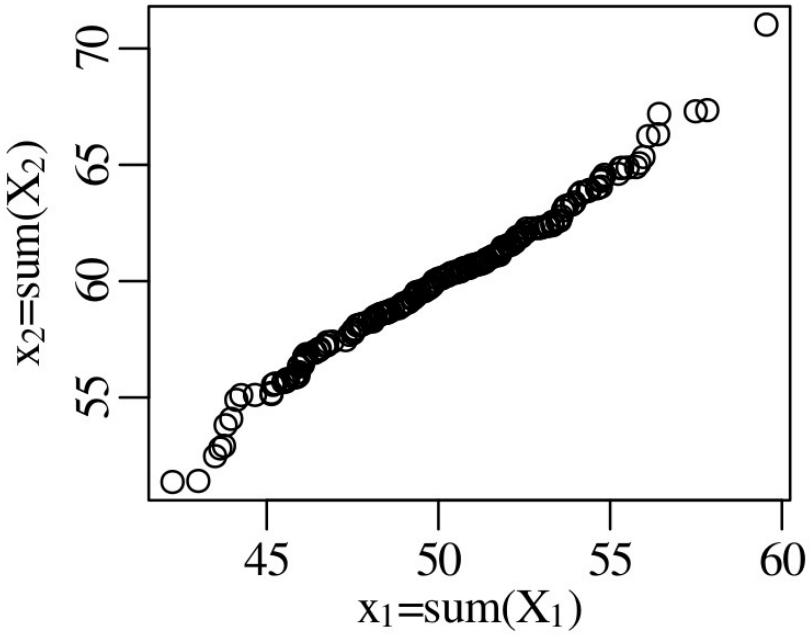
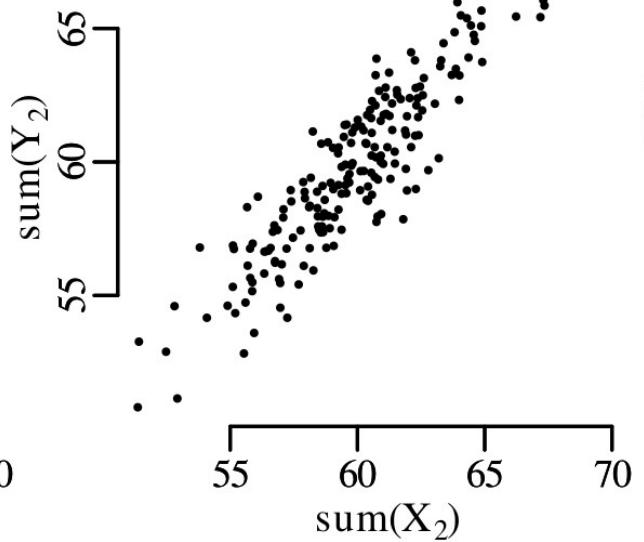
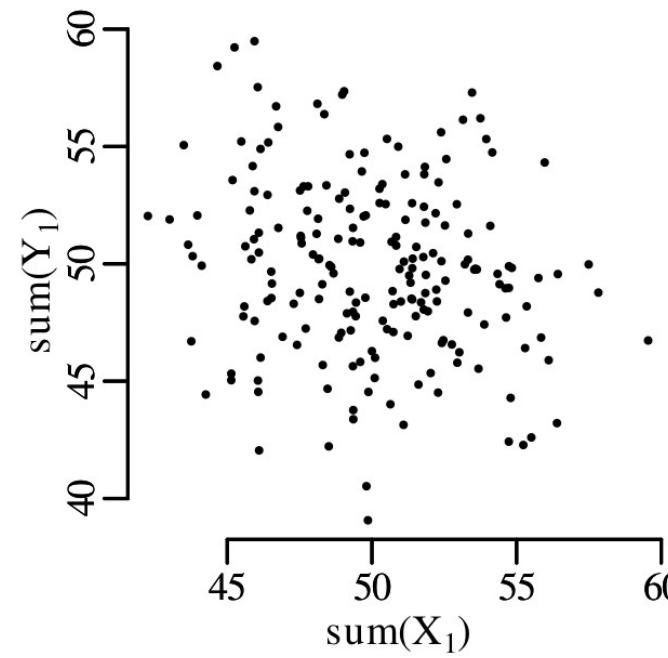
# Statistics for geoscientists

Comparing distributions

# Q-Q plots







# The t-test

| coin #                       | 1     | 2     | 3     | 4     | 5     |
|------------------------------|-------|-------|-------|-------|-------|
| density (g/cm <sup>3</sup> ) | 19.07 | 19.09 | 19.17 | 19.18 | 19.31 |

$$\bar{x} = 19.164$$

density of pure gold = 19.30 g/cm<sup>3</sup>

| coin #                       | 1     | 2     | 3     | 4     | 5     |
|------------------------------|-------|-------|-------|-------|-------|
| density (g/cm <sup>3</sup> ) | 19.07 | 19.09 | 19.17 | 19.18 | 19.31 |

$z = \frac{\bar{x} - \mu}{\sigma / \sqrt{n}}$  follows a **standard normal distribution**

$t = \frac{\bar{x} - \mu}{s[\bar{x}] / \sqrt{n}}$  follows a **Student t-distribution**  
with  $(n - 1)$  **degrees of freedom**

1.  $H_0$  (null hypothesis)  $\mu = 19.30$

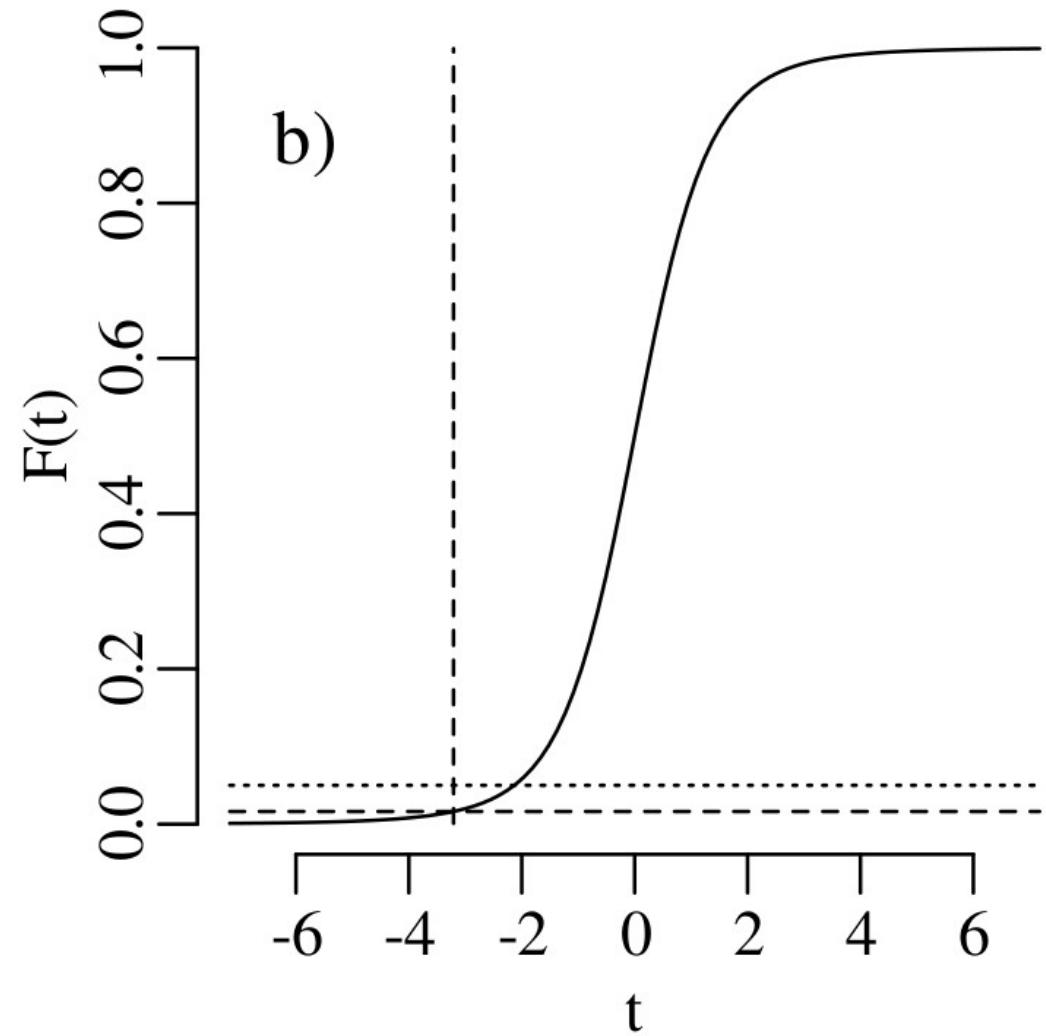
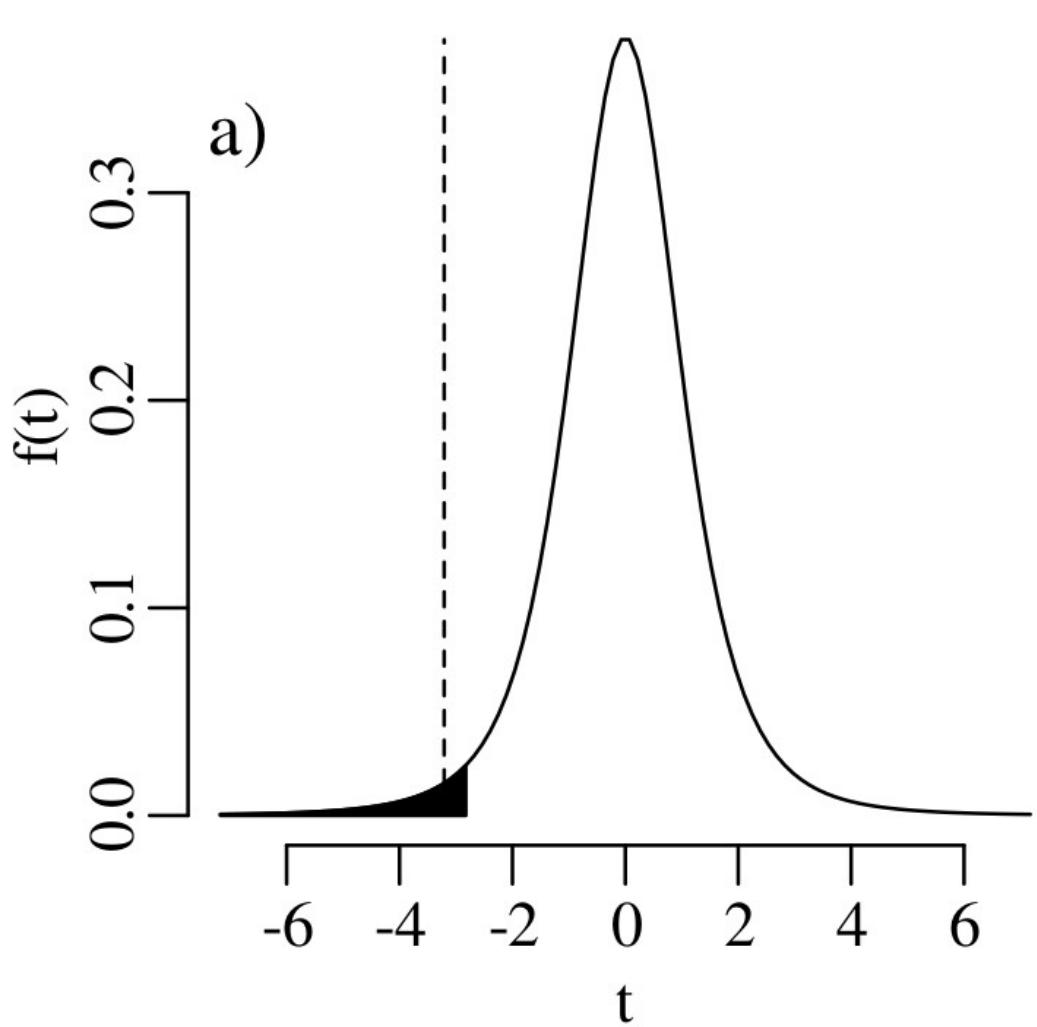
$H_a$  (alternative hypothesis):  $\mu < 19.30$

2.  $t = \frac{\bar{x} - \mu_0}{s[x]/\sqrt{n}}$  where  $\mu_0 = 19.30$   $\bar{x} = 19.164$   
 $s[x] = 0.0948$  so that  $t = -3.2091$   
 $n = 5$

|    |               |                                                                     |
|----|---------------|---------------------------------------------------------------------|
| 3. | $t$           | -3.70 -3.2091 -2.80 -2.10 -1.50 -0.74 0.00 0.74 1.50 2.10 2.80 3.70 |
|    | $P(T \leq t)$ | 0.01 0.0163 0.025 0.05 0.1 0.25 0.5 0.75 0.9 0.95 0.975 0.99        |

4.  $\alpha = 0.05$

|    |               |                                                                     |
|----|---------------|---------------------------------------------------------------------|
| 5. | $t$           | -3.70 -3.2091 -2.80 -2.10 -1.50 -0.74 0.00 0.74 1.50 2.10 2.80 3.70 |
|    | $P(T \leq t)$ | 0.01 0.0163 0.025 0.05 0.1 0.25 0.5 0.75 0.9 0.95 0.975 0.99        |



## two sample t-test

| coin #                               | 1     | 2     | 3     | 4     | 5     |
|--------------------------------------|-------|-------|-------|-------|-------|
| density (1 <sup>st</sup> collection) | 19.07 | 19.09 | 19.17 | 19.18 | 19.31 |
| density (2 <sup>nd</sup> collection) | 19.17 | 19.30 | 19.31 | 19.32 |       |

$$\bar{x}_1 = 19.164 \longleftrightarrow \bar{x}_2 = 19.275$$

$$1. \quad H_0 \text{ (null hypothesis)} \quad \mu_1 = \mu_2$$

$$H_a \text{ (alternative hypothesis):} \quad \mu_1 \neq \mu_2$$

$$2. \quad t = \frac{\bar{x}_1 - \bar{x}_2}{s_p \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} \quad \text{where} \quad s_p = \sqrt{\frac{(n_1 - 1)s[x_1]^2 + (n_2 - 1)s[x_2]^2}{n_1 + n_2 - 2}}$$
$$= -2.014$$

3. Student t-distribution with  $(n_1 + n_2 - 2)$  degrees of freedom

|               |       |       |        |       |       |       |      |      |      |      |       |      |
|---------------|-------|-------|--------|-------|-------|-------|------|------|------|------|-------|------|
| $t$           | -3.00 | -2.40 | -2.014 | -1.90 | -1.40 | -0.71 | 0.00 | 0.71 | 1.40 | 1.90 | 2.40  | 3.00 |
| $P(T \leq t)$ | 0.01  | 0.025 | 0.042  | 0.05  | 0.1   | 0.25  | 0.5  | 0.75 | 0.9  | 0.95 | 0.975 | 0.99 |
| $P(T \geq t)$ | 0.99  | 0.975 | 0.958  | 0.95  | 0.9   | 0.75  | 0.5  | 0.25 | 0.1  | 0.05 | 0.025 | 0.01 |

4.  $\alpha = 0.05$

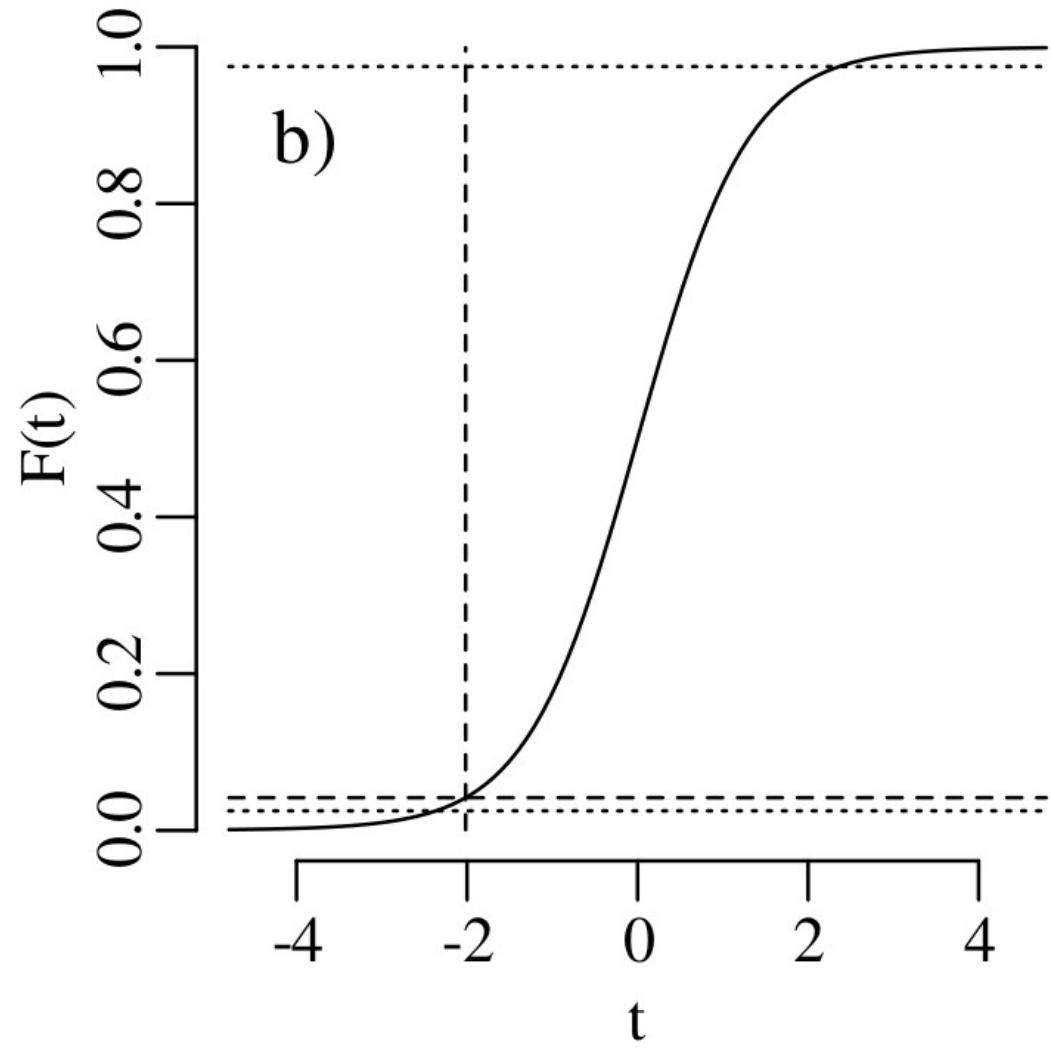
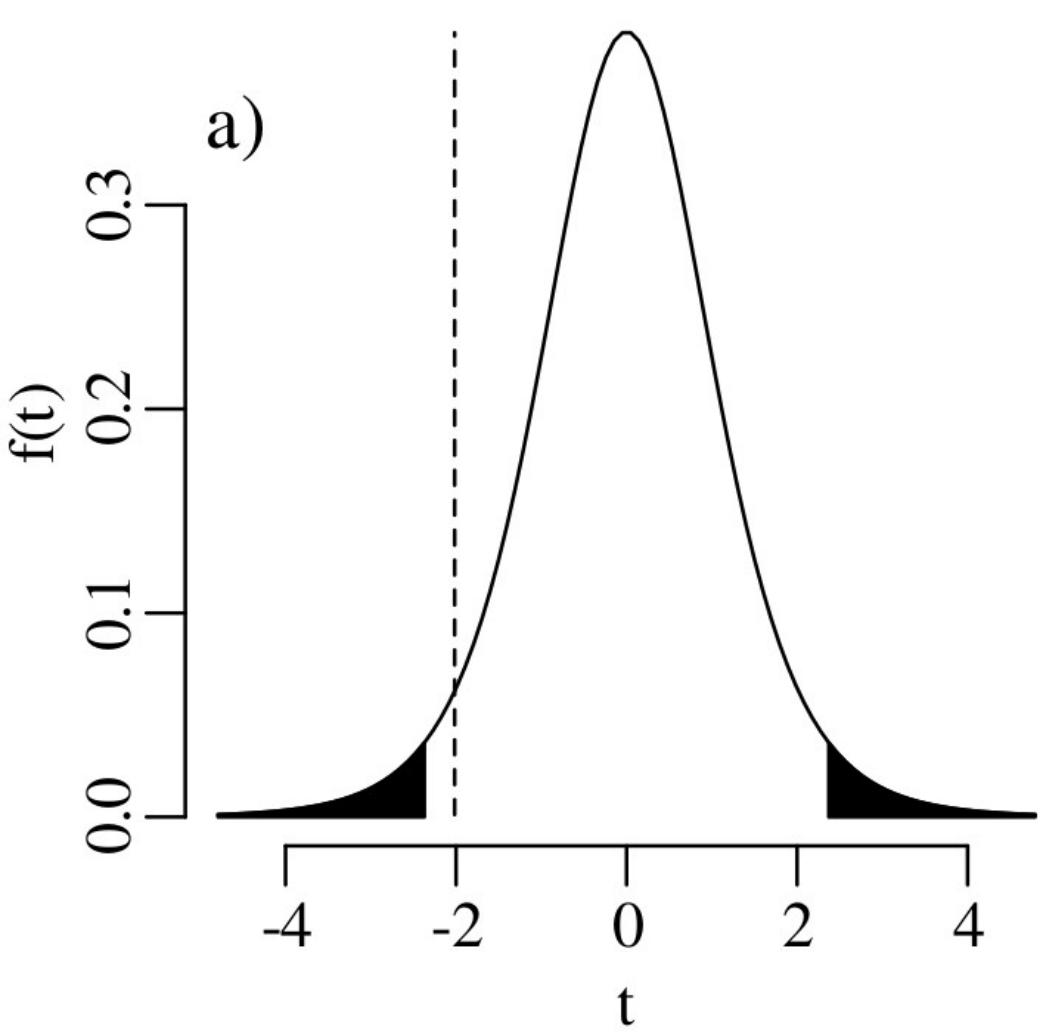
5.

| $t$           | <b>-3.00</b> | <b>-2.40</b> | <b>-2.014</b> | -1.90 | -1.40 | -0.71 | 0.00 | 0.71 | 1.40 | 1.90 | <b>2.40</b>  | <b>3.00</b> |
|---------------|--------------|--------------|---------------|-------|-------|-------|------|------|------|------|--------------|-------------|
| $P(T \leq t)$ | <b>0.01</b>  | <b>0.025</b> | <b>0.042</b>  | 0.05  | 0.1   | 0.25  | 0.5  | 0.75 | 0.9  | 0.95 | 0.975        | 0.99        |
| $P(T \geq t)$ | 0.99         | 0.975        | 0.958         | 0.95  | 0.9   | 0.75  | 0.5  | 0.25 | 0.1  | 0.05 | <b>0.025</b> | <b>0.01</b> |

$$R = \{t < -2.40 \text{ and } t > 2.40\}$$

6.  $t = -2.014 \notin R$

7. p-value = 0.084 ( $= 2 \times 0.042$ )  $> 0.05$



# Confidence intervals

$$t = \frac{\bar{x} - \mu}{s[x]/\sqrt{n}}$$

$$t_{df,\alpha/2} \leq t \leq t_{df,1-\alpha/2}$$

$$t_{df,\alpha/2} \leq \frac{\bar{x} - \mu}{s[x]/\sqrt{n}} \leq t_{df,1-\alpha/2}$$

$$\bar{x} - t_{df,\alpha/2} \frac{s[x]}{\sqrt{n}} \geq \mu \geq \bar{x} - t_{df,1-\alpha/2} \frac{s[x]}{\sqrt{n}}$$

with  $t_{df,1-\alpha/2} = -t_{df,\alpha/2}$

$$\bar{x} + t_{df,\alpha/2} \frac{s[x]}{\sqrt{n}} \leq \mu \leq \bar{x} - t_{df,\alpha/2} \frac{s[x]}{\sqrt{n}}$$

$\rightarrow \mu \in \left\{ \bar{x} \pm t_{df,\alpha/2} \frac{s[x]}{\sqrt{n}} \right\}$

$$\mu \in \left\{ \bar{x} \pm t_{df,\alpha/2} \frac{s[x]}{\sqrt{n}} \right\}$$

| coin #                       | 1     | 2     | 3     | 4     | 5     |
|------------------------------|-------|-------|-------|-------|-------|
| density (g/cm <sup>3</sup> ) | 19.07 | 19.09 | 19.17 | 19.18 | 19.31 |

$$\bar{x} = 19.164$$

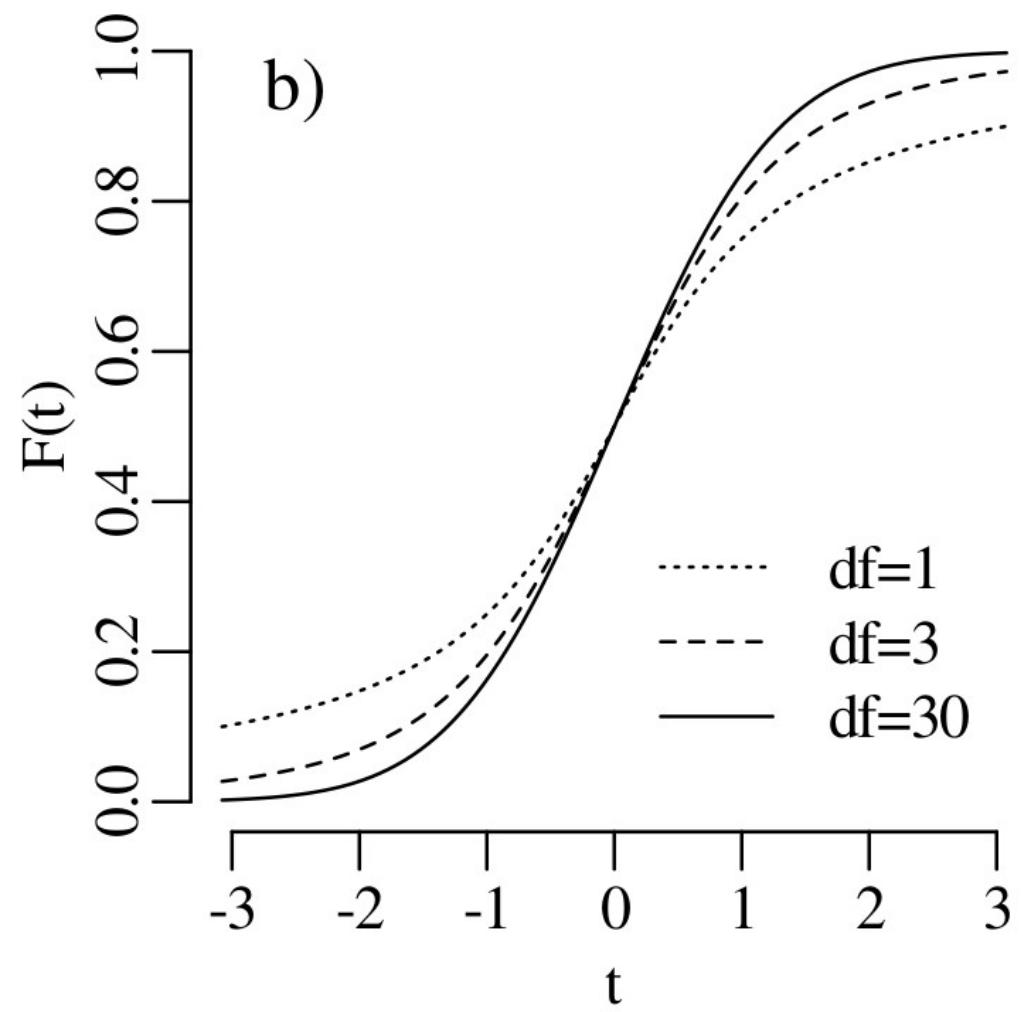
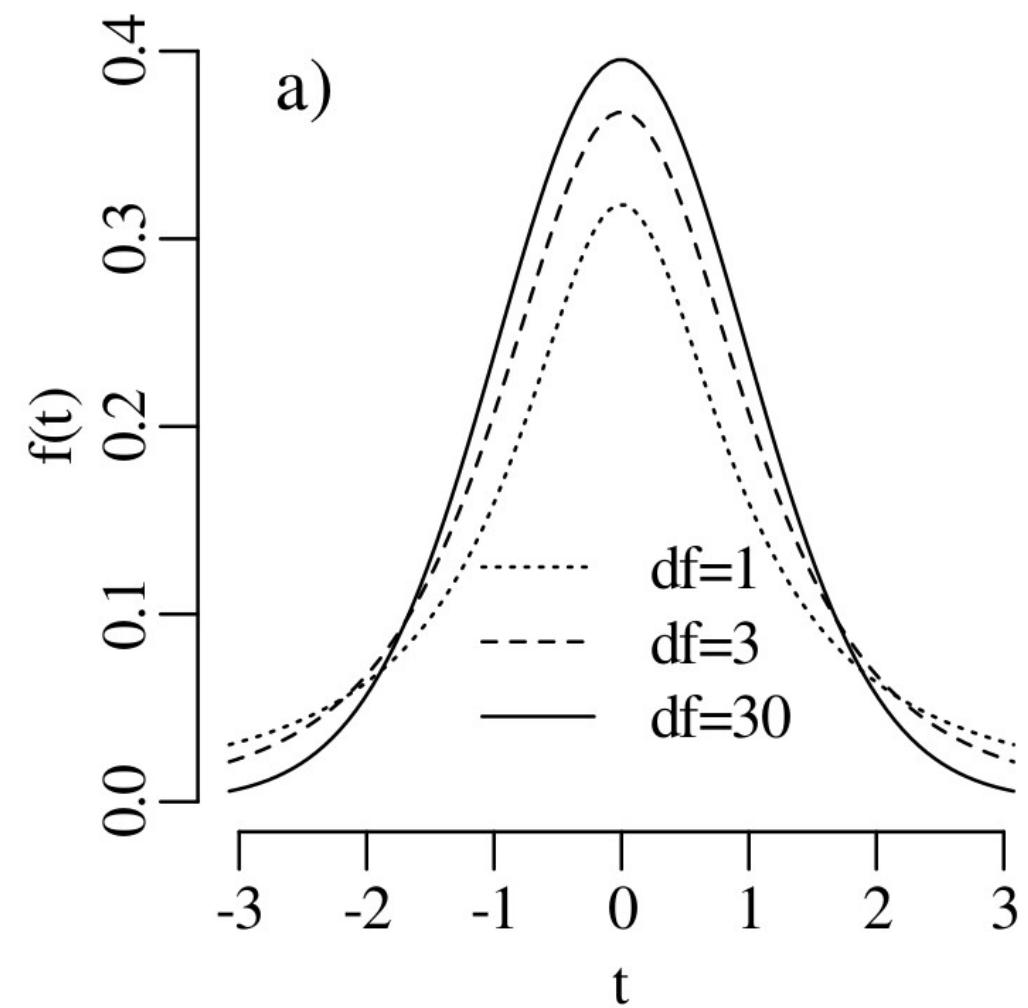
$$s[x] = 0.0948 \qquad \qquad \mu = 19.16 \pm 0.12 \text{ g/cm}^3$$

$$df = 4$$

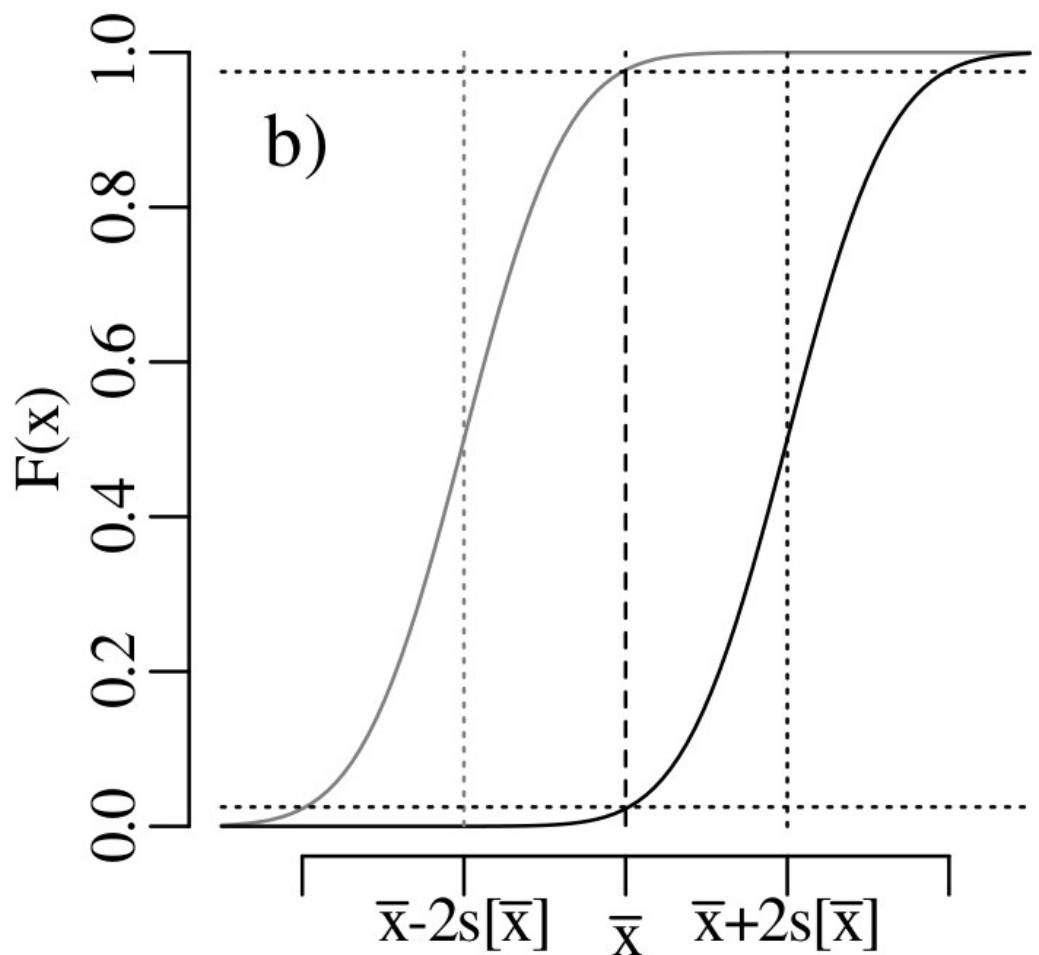
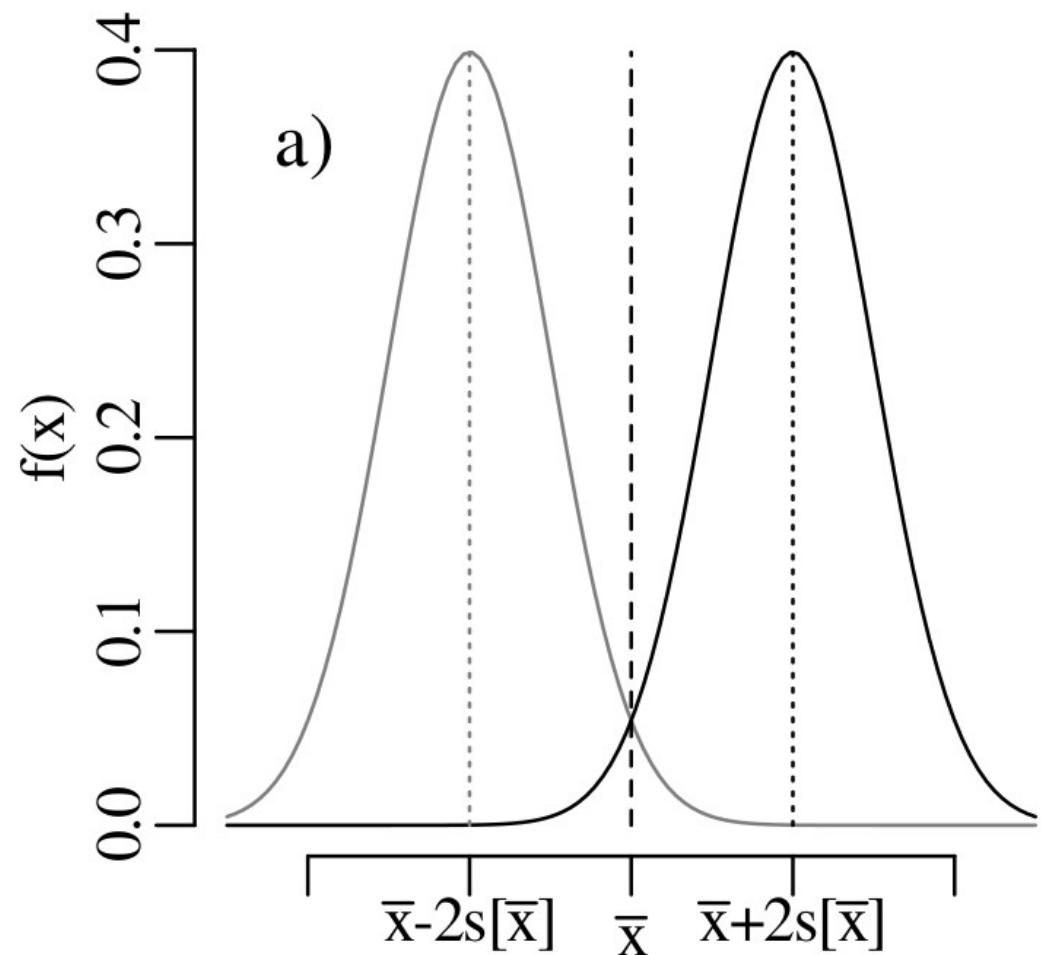
$$t_{4,0.025} = -2.776$$

| $df$           | 1      | 2     | 3     | 4     | 5     | 6     | 7     | 8     | 9     | 10    | 30    | 100   | 1000         |
|----------------|--------|-------|-------|-------|-------|-------|-------|-------|-------|-------|-------|-------|--------------|
| $t_{df,0.975}$ | 12.710 | 4.303 | 3.182 | 2.776 | 2.571 | 2.447 | 2.365 | 2.306 | 2.262 | 2.228 | 2.042 | 1.984 | <b>1.962</b> |

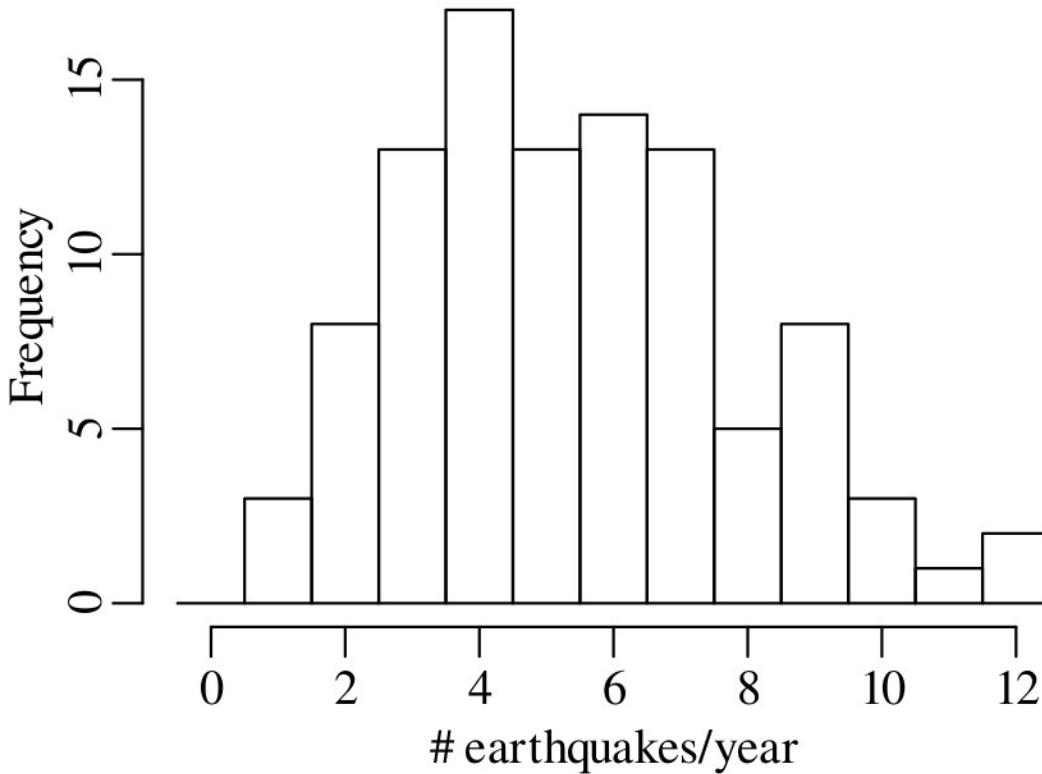
$$\mu \in \left\{ \bar{x} \pm t_{df,\alpha/2} \frac{s[\bar{x}]}{\sqrt{n}} \right\} \xrightarrow{} \mu \in \{ \bar{x} \pm 2s[\bar{x}] \}$$



$$\mu \in \left\{ \bar{x} \pm t_{df,\alpha/2} \frac{s[x]}{\sqrt{n}} \right\} \longrightarrow \mu \in \{ \bar{x} \pm 2s[\bar{x}] \}$$



# The $\chi^2$ -test



**mean: 5.43**

standard deviation: 2.50

**variance: 6.24**

|                                |                                |
|--------------------------------|--------------------------------|
| number of earthquakes per year | 0 1 2 3 4 5 6 7 8 9 10 11 12   |
| number of years                | 0 3 8 13 17 13 14 13 5 8 3 1 2 |

observed

|                                |   |   |   |    |    |    |    |    |   |   |    |    |    |
|--------------------------------|---|---|---|----|----|----|----|----|---|---|----|----|----|
| number of earthquakes per year | 0 | 1 | 2 | 3  | 4  | 5  | 6  | 7  | 8 | 9 | 10 | 11 | 12 |
| number of years                | 0 | 3 | 8 | 13 | 17 | 13 | 14 | 13 | 5 | 8 | 3  | 1  | 2  |

|                                |          |    |    |    |    |    |   |   |           |
|--------------------------------|----------|----|----|----|----|----|---|---|-----------|
| number of earthquakes per year | $\leq 2$ | 3  | 4  | 5  | 6  | 7  | 8 | 9 | $\geq 10$ |
| number of years                | 11       | 13 | 17 | 13 | 14 | 13 | 5 | 8 | 6         |

expected

$$P(k|\lambda) = \frac{\lambda^k e^{-\lambda}}{k!}$$

|                                        |          |      |      |      |      |      |      |      |           |
|----------------------------------------|----------|------|------|------|------|------|------|------|-----------|
| number of earthquakes per year ( $k$ ) | $\leq 2$ | 3    | 4    | 5    | 6    | 7    | 8    | 9    | $\geq 10$ |
| $N \times P(k \lambda = 5.43)$         | 9.28     | 11.7 | 15.9 | 17.2 | 15.6 | 12.1 | 8.22 | 4.96 | 5.02      |

$$\chi^2 = \sum_{i=1}^n \frac{(O_i - E_i)^2}{E_i}$$

1.  $H_0$  (**null hypothesis**): the earthquake data follow a Poisson distribution

$H_a$  (**alternative hypothesis**): the earthquake data do not follow a Poisson distribution

$$2. \chi^2 = \frac{(11 - 9.28)^2}{9.28} + \frac{(13 - 11.7)^2}{11.7} + \frac{(17 - 15.9)^2}{15.9} + \frac{(13 - 17.2)^2}{17.2} + \\ \frac{(14 - 15.6)^2}{15.6} + \frac{(13 - 12.1)^2}{12.1} + \frac{(5 - 8.22)^2}{8.22} + \frac{(8 - 4.96)^2}{4.96} + \frac{(6 - 5.02)^2}{5.02} = 5.14$$

3.  $\chi^2$  distribution with  $n - 2$  degrees of freedom

|                    |      |       |      |      |      |       |      |      |      |      |       |      |
|--------------------|------|-------|------|------|------|-------|------|------|------|------|-------|------|
| $\chi^2$           | 1.24 | 1.69  | 2.17 | 2.83 | 4.25 | 5.14  | 6.35 | 9.04 | 12.0 | 14.1 | 16.0  | 18.5 |
| $P(X \leq \chi^2)$ | 0.01 | 0.025 | 0.05 | 0.1  | 0.25 | 0.36  | 0.5  | 0.75 | 0.9  | 0.95 | 0.975 | 0.99 |
| $P(X \geq \chi^2)$ | 0.99 | 0.975 | 0.95 | 0.9  | 0.75 | 0.743 | 0.5  | 0.25 | 0.1  | 0.05 | 0.025 | 0.01 |

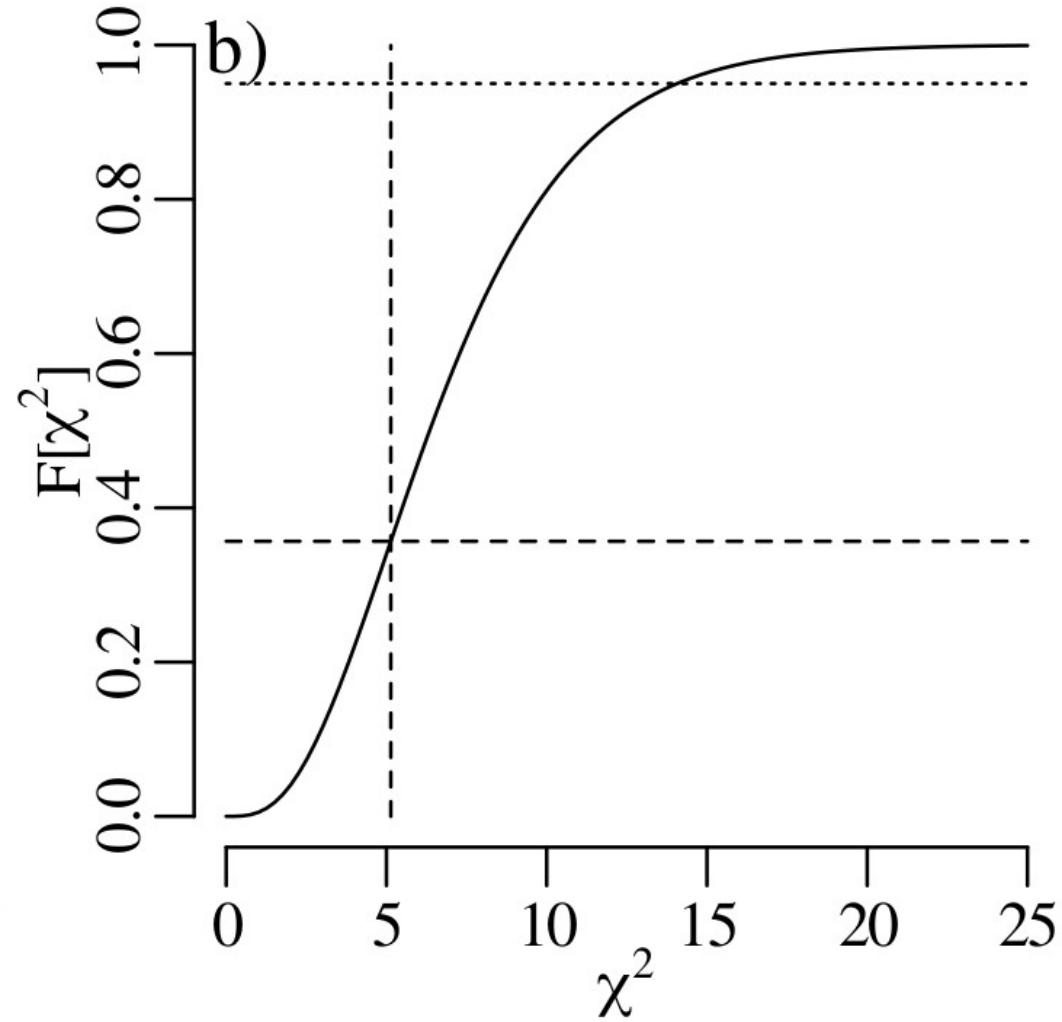
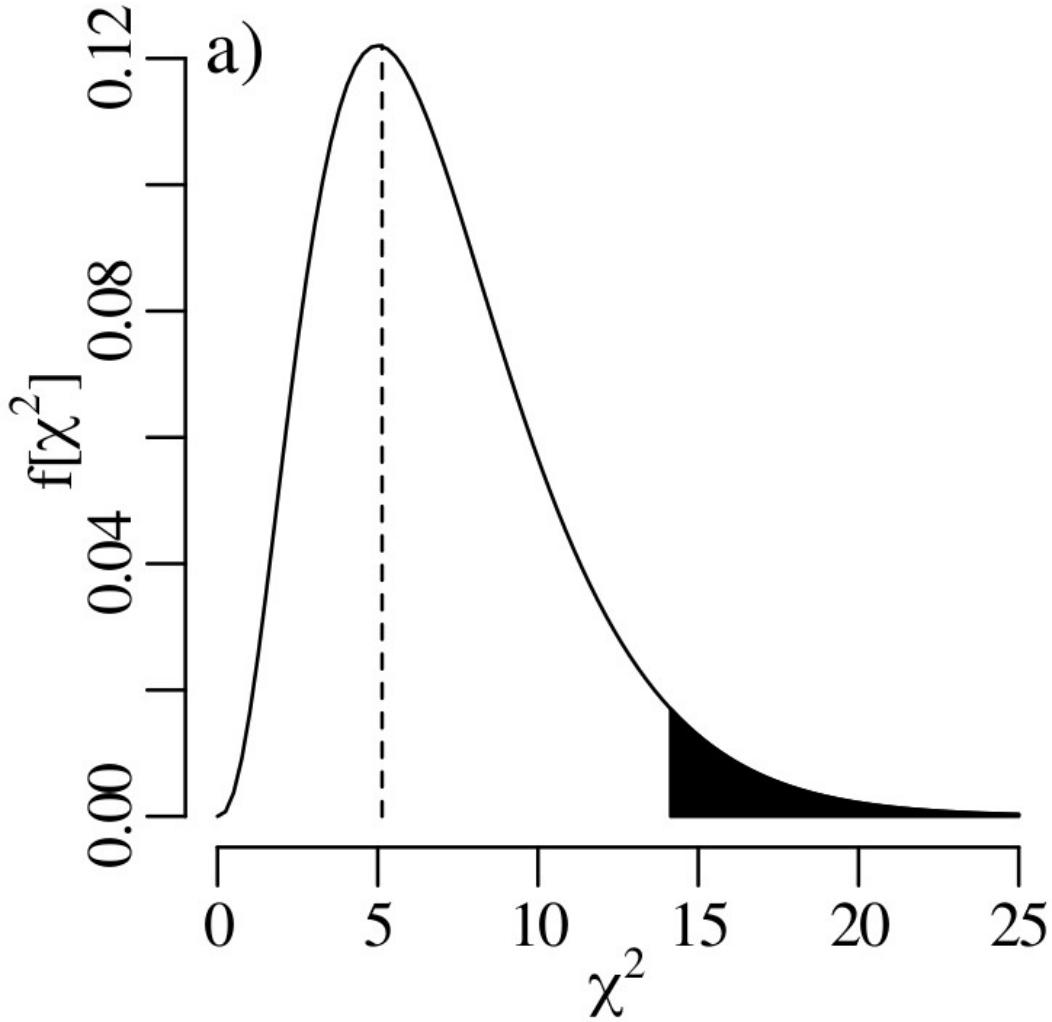
4.  $\alpha = 0.05$

|    |                    |                                              |                                      |
|----|--------------------|----------------------------------------------|--------------------------------------|
| 5. | $\chi^2$           | 1.24 1.69 2.17 2.83 4.25 5.14 6.35 9.04 12.0 | <b>14.1</b> 16.0 18.5                |
|    | $P(X \geq \chi^2)$ | 0.99 0.975 0.95 0.9 0.75 0.743 0.5 0.25 0.1  | <b>0.05</b> <b>0.025</b> <b>0.01</b> |

$$R = \{\chi^2 \geq 14.1\}$$

6.  $\chi^2 = 5.14 \notin R$

7. p-value = 0.743 > 0.05



# Comparing two or more samples

| lithology | granite | basalt | gneiss | quartzite |
|-----------|---------|--------|--------|-----------|
| sample A  | 10      | 5      | 6      | 20        |
| sample B  | 25      | 12     | 10     | 35        |

| lithology  | granite | basalt | gneiss | quartzite | row sum |
|------------|---------|--------|--------|-----------|---------|
| sample A   | 10      | 5      | 6      | 20        | 41      |
| sample B   | 25      | 12     | 10     | 35        | 82      |
| column sum | 35      | 17     | 16     | 55        | 123     |

expected counts of bin  $(i, j) = \frac{(\text{sum of row } i) \times (\text{sum of column } j)}{(\text{sum of all the cells})}$

| lithology | granite | basalt | gneiss | quartzite |
|-----------|---------|--------|--------|-----------|
| sample A  | 11.7    | 5.67   | 5.33   | 18.3      |
| sample B  | 23.3    | 11.30  | 10.70  | 36.7      |

1.  $H_0$  (**null hypothesis**): samples A and B have the same composition

$H_a$  (**alternative hypothesis**): samples A and B do not have the same composition

$$2. \chi^2 = \frac{(10 - 11.7)^2}{11.7} + \frac{(5 - 5.67)^2}{5.67} + \frac{(6 - 5.33)^2}{5.33} + \frac{(20 - 18.3)^2}{18.3} + \\ \frac{(25 - 23.3)^2}{23.3} + \frac{(12 - 11.30)^2}{11.30} + \frac{(10 - 10.70)^2}{10.70} + \frac{(35 - 36.7)^2}{36.7} = 0.86$$

3.  $\chi^2$ -distribution with  $(2-1) \times (4-1) = 3$  degrees of freedom

| $\chi^2$           | 0.115 | 0.216 | 0.352 | 0.584 | 0.86  | 1.21 | 2.37 | 4.11 | 6.25 | 7.81 | 9.35  | 11.3 |
|--------------------|-------|-------|-------|-------|-------|------|------|------|------|------|-------|------|
| $P(X \leq \chi^2)$ | 0.01  | 0.025 | 0.05  | 0.1   | 0.157 | 0.25 | 0.5  | 0.75 | 0.9  | 0.95 | 0.975 | 0.99 |
| $P(X \geq \chi^2)$ | 0.99  | 0.975 | 0.95  | 0.9   | 0.843 | 0.75 | 0.5  | 0.25 | 0.1  | 0.05 | 0.025 | 0.01 |

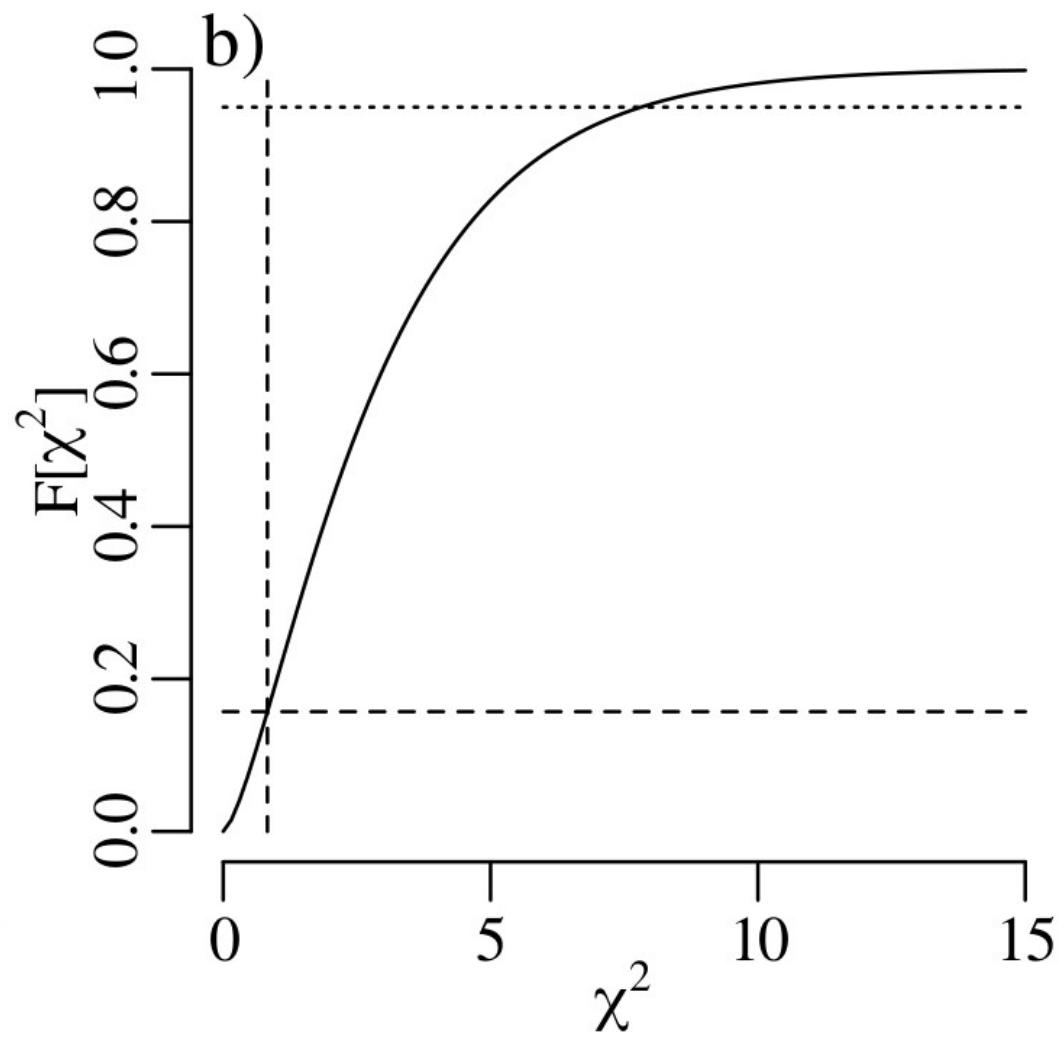
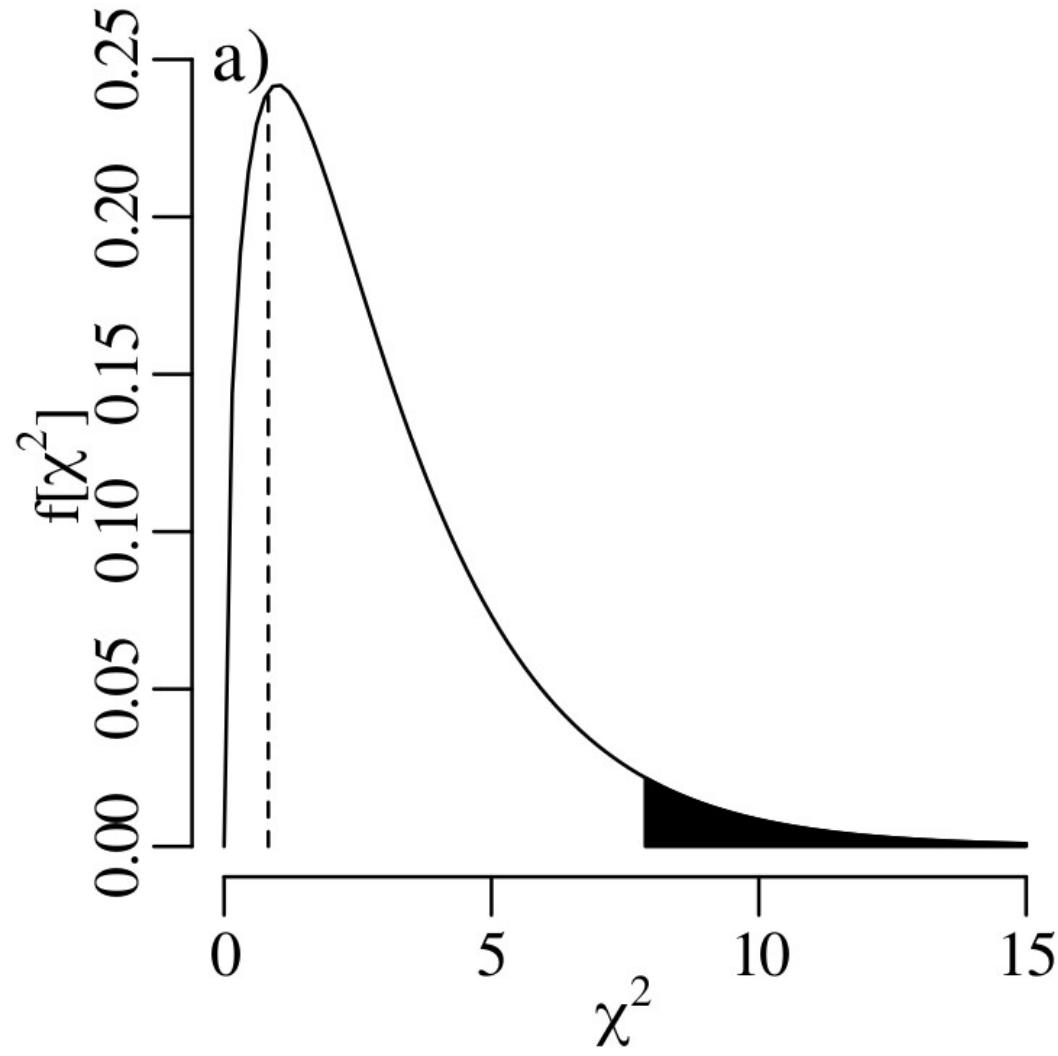
4.  $\alpha = 0.05$

|    |                    |  |       |       |       |       |       |      |      |      |      |             |              |      |
|----|--------------------|--|-------|-------|-------|-------|-------|------|------|------|------|-------------|--------------|------|
| 5. | $\chi^2$           |  | 0.115 | 0.216 | 0.352 | 0.584 | 0.86  | 1.21 | 2.37 | 4.11 | 6.25 | <b>7.81</b> | <b>9.35</b>  | 11.3 |
|    | $P(X \geq \chi^2)$ |  | 0.99  | 0.975 | 0.95  | 0.9   | 0.843 | 0.75 | 0.5  | 0.25 | 0.1  | <b>0.05</b> | <b>0.025</b> | 0.01 |

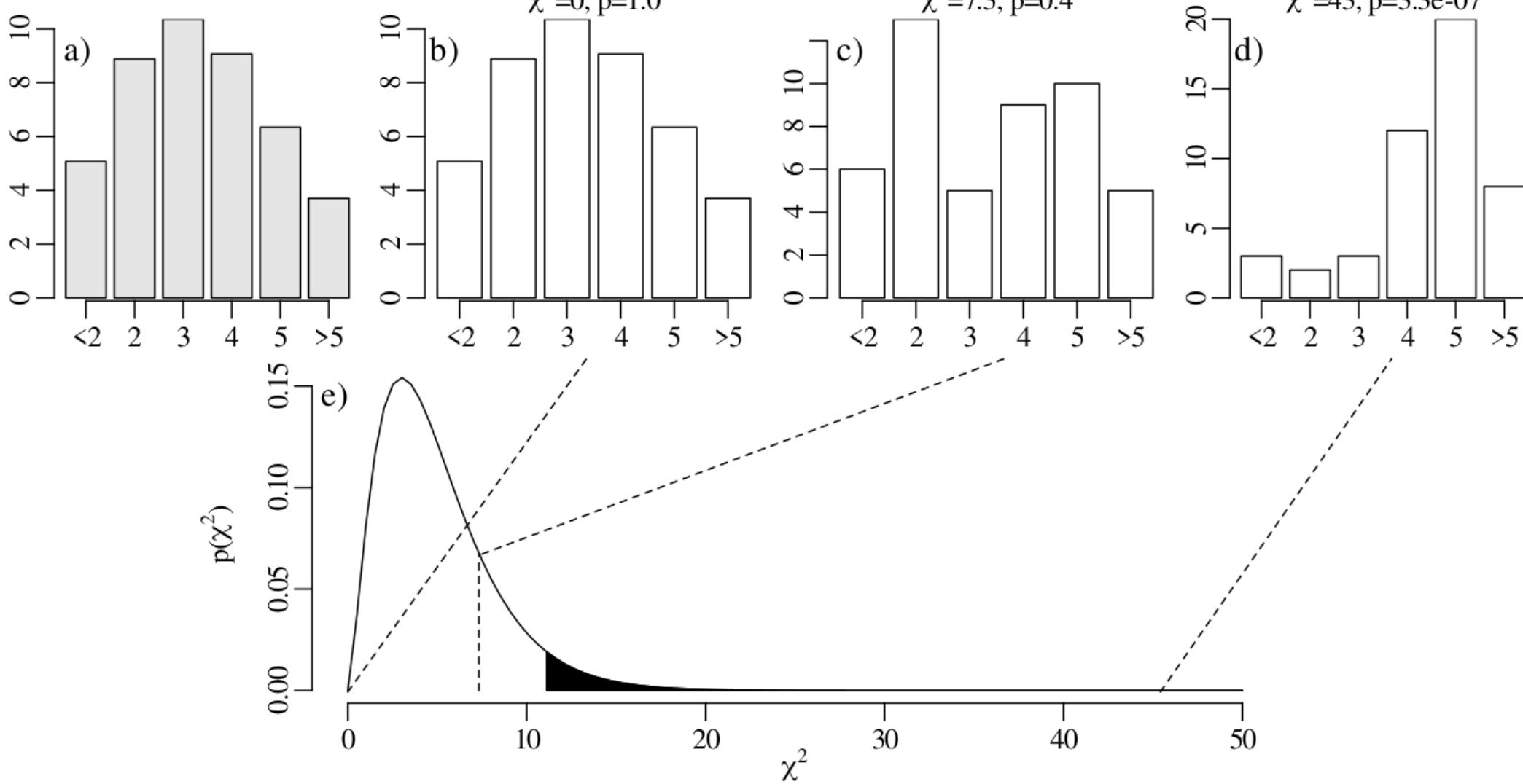
$$R = \{\chi^2 > 7.81\}$$

6.  $\chi^2 = 0.86 \notin R$

7. p-value = 0.843 > 0.05



# Cherry picking (Type-I errors revisited)



## Effect size (Type-II errors revisited)

| lithology | quartz | plagioclase | alkali feldspar | lithics |
|-----------|--------|-------------|-----------------|---------|
| sample A  | 29544  | 14424       | 13706           | 47864   |
| sample B  | 29454  | 14788       | 13948           | 47311   |

| lithology  | quartz | plagioclase | alkali feldspar | lithics | row sum |
|------------|--------|-------------|-----------------|---------|---------|
| sample A   | 29544  | 14424       | 13706           | 47864   | 105538  |
| sample B   | 29454  | 14788       | 13948           | 47311   | 105501  |
| column sum | 58998  | 29212       | 27654           | 95175   | 211039  |

1.  $H_0$  (**null hypothesis**): samples A and B have identical compositions

$H_a$  (**alternative hypothesis**): samples A and B have different compositions

| lithology | quartz | plagioclase | alkali feldspar | lithics |                 |
|-----------|--------|-------------|-----------------|---------|-----------------|
| sample A  | 29504  | 14609       | 13829           | 47596   | $\chi^2 = 10.0$ |
| sample B  | 29494  | 14603       | 13825           | 47579   |                 |

3.  $\chi^2$ -distribution 3 degrees of freedom

| $\chi^2$           | 0.115 | 0.216 | 0.352 | 0.584 | 1.21 | 2.37 | 4.11 | 6.25 | 7.81 | 9.35  | 10.0   | 11.3  |
|--------------------|-------|-------|-------|-------|------|------|------|------|------|-------|--------|-------|
| $P(X \leq \chi^2)$ | 0.01  | 0.025 | 0.05  | 0.1   | 0.25 | 0.5  | 0.75 | 0.9  | 0.95 | 0.975 | 0.9814 | 0.99  |
| $P(X \geq \chi^2)$ | 0.99  | 0.975 | 0.95  | 0.9   | 0.75 | 0.5  | 0.25 | 0.1  | 0.05 | 0.025 | 0.0186 | 0.010 |

4.  $\alpha = 0.05$

5.

| $\chi^2$           | 0.115 | 0.216 | 0.352 | 0.584 | 1.21 | 2.37 | 4.11 | 6.25 | <b>7.81</b> | <b>9.35</b>  | <b>10.0</b>   | 11.3  |
|--------------------|-------|-------|-------|-------|------|------|------|------|-------------|--------------|---------------|-------|
| $P(X \geq \chi^2)$ | 0.99  | 0.975 | 0.95  | 0.9   | 0.75 | 0.5  | 0.25 | 0.1  | <b>0.05</b> | <b>0.025</b> | <b>0.0186</b> | 0.010 |

$$R = \{\chi^2 > 7.81\}$$

6.  $\chi^2 = 10.0 \in R$

7. p-value = 0.0186 <  $\alpha$

depends on the **confidence level**  $\alpha$

| $H_0$ is ... | false            | true             |
|--------------|------------------|------------------|
| rejected     | correct decision | Type-I error     |
| not rejected | Type-II error    | correct decision |

$\beta = 1 - \text{power}$

depends on 1. The degree to which  $H_0$  is false

2. Sample size

depends on the **confidence level**  $\alpha$

| $H_0$ is ... | false            | true             |
|--------------|------------------|------------------|
| rejected     | correct decision | Type-I error     |
| not rejected | Type-II error    | correct decision |

$\beta = 1 - \text{power}$

- depends on
1. The degree to which  $H_0$  is false
  2. Sample size

# effect size

$$\chi^2 = \sum_{i=1}^n \frac{(O_i - E_i)^2}{E_i}$$

$$w = \sqrt{\sum_{i=1}^n \frac{(o_i - e_i)^2}{e_i}}$$

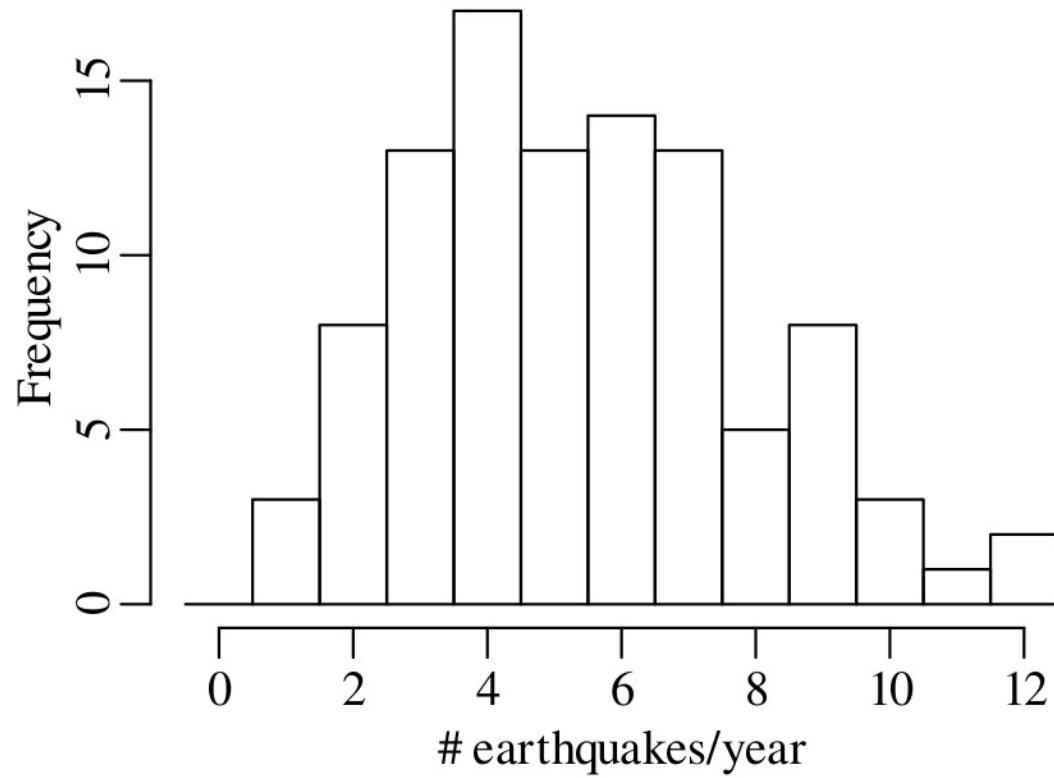
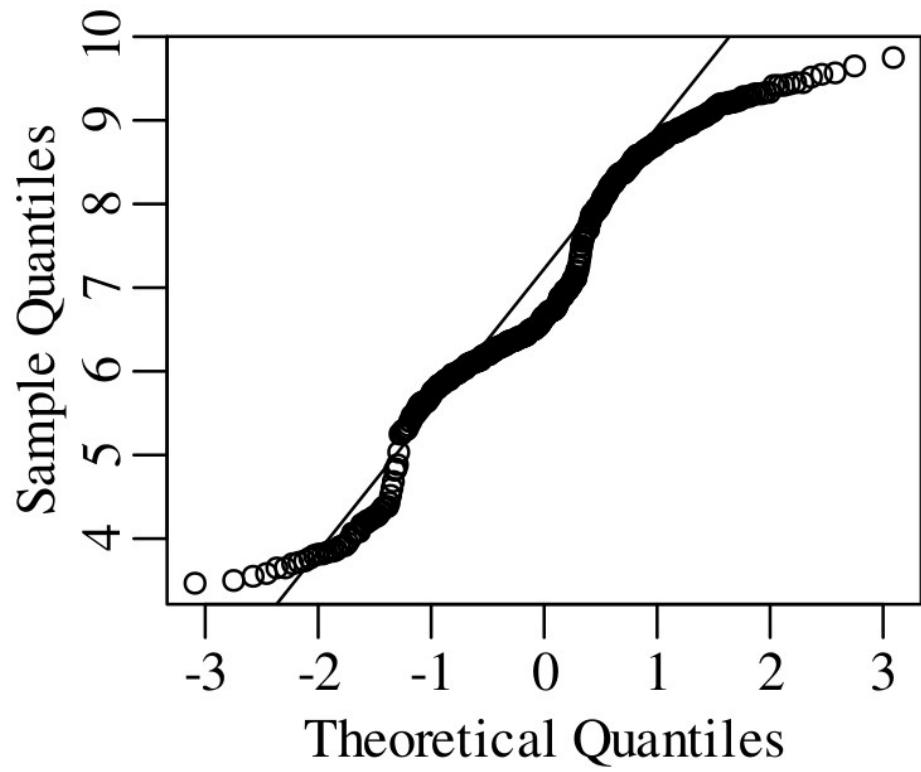
where  $o_i = O_i/N$  and  $e_i = E_i/N$

with  $N = \sum_i^n O_i = \sum_i^n E_i$

| effect size | small | medium | large |
|-------------|-------|--------|-------|
| w           | 0.1   | 0.3    | 0.5   |

| effect size |           | small  | medium      | large           |         |
|-------------|-----------|--------|-------------|-----------------|---------|
|             | w         | 0.1    | 0.3         | 0.5             |         |
| $o =$       | lithology | quartz | plagioclase | alkali feldspar | lithics |
|             | sample A  | 0.1400 | 0.06835     | 0.06495         | 0.2268  |
|             | sample B  | 0.1396 | 0.07007     | 0.06609         | 0.2242  |
| $e =$       | lithology | quartz | plagioclase | alkali feldspar | lithics |
|             | sample A  | 0.1398 | 0.06922     | 0.06553         | 0.2255  |
|             | sample B  | 0.1398 | 0.06920     | 0.06551         | 0.2255  |

$$w = 0.00688$$



# Non-parametric tests

## 1. Wilcoxon test

| coin #                               | 1     | 2     | 3     | 4     | 5     |
|--------------------------------------|-------|-------|-------|-------|-------|
| density (1 <sup>st</sup> collection) | 19.07 | 19.09 | 19.17 | 19.18 | 19.31 |
| density (2 <sup>nd</sup> collection) | 19.17 | 19.30 | 19.31 | 19.32 |       |

| sample | 1     | 2     | 3     | 4     | 5     |
|--------|-------|-------|-------|-------|-------|
| $A$    | $A_1$ | $A_2$ | $A_3$ | $A_4$ | $A_5$ |
| $B$    | $B_1$ | $B_2$ | $B_3$ | $B_4$ |       |

| sample | 1     | 2     | 3     | 4     | 5     |
|--------|-------|-------|-------|-------|-------|
| A      | $A_1$ | $A_2$ | $A_3$ | $A_4$ | $A_5$ |
| B      | $B_1$ | $B_2$ | $B_3$ | $B_4$ |       |

suppose that

$$A_4 < A_1 < A_2 < B_3 < A_5 < B_1 < B_4 < A_3 < B_2$$

then

| rank  | 1     | 2     | 3     | 4     | 5     | 6     | 7     | 8     | 9     |
|-------|-------|-------|-------|-------|-------|-------|-------|-------|-------|
| value | $A_4$ | $A_1$ | $A_2$ | $B_3$ | $A_5$ | $B_1$ | $B_4$ | $A_3$ | $B_2$ |

$$W = 4 + 6 + 7 + 9 = 26$$

e.g.  $W = 10$

|                |       |       |       |       |       |       |       |       |       |
|----------------|-------|-------|-------|-------|-------|-------|-------|-------|-------|
| arrangement 1: | $B_1$ | $B_2$ | $B_3$ | $B_4$ | $A_1$ | $A_2$ | $A_3$ | $A_4$ | $A_5$ |
| arrangement 2: | $B_2$ | $B_1$ | $B_3$ | $B_4$ | $A_1$ | $A_2$ | $A_3$ | $A_4$ | $A_5$ |
| arrangement 2: | $B_1$ | $B_2$ | $B_3$ | $B_4$ | $A_2$ | $A_1$ | $A_3$ | $A_4$ | $A_5$ |
| etc.           |       |       |       |       |       |       |       |       |       |

$$P(W = 10 | n_A = 5, n_B = 4) = \frac{4!5!}{9!} = 0.00794$$

| coin #                                     | 1            | 2            | 3            | 4            | 5     |
|--------------------------------------------|--------------|--------------|--------------|--------------|-------|
| density (1 <sup>st</sup> collection)       | 19.07        | 19.09        | 19.17        | 19.18        | 19.31 |
| <b>density (2<sup>nd</sup> collection)</b> | <b>19.17</b> | <b>19.30</b> | <b>19.31</b> | <b>19.32</b> |       |

1.  $H_0$  (null hypothesis)  $\text{median}(\text{sample 1}) = \text{median}(\text{sample 2})$

$H_a$  (alternative hypothesis):  $\text{median}(\text{sample 1}) \neq \text{median}(\text{sample 2})$

| rank    | 1     | 2     | 3.5   | <b>3.5</b>   | 5     | <b>6</b>     | 7.5          | <b>7.5</b> | <b>9</b>     |
|---------|-------|-------|-------|--------------|-------|--------------|--------------|------------|--------------|
| density | 19.07 | 19.09 | 19.17 | <b>19.17</b> | 19.18 | <b>19.30</b> | <b>19.31</b> | 19.31      | <b>19.32</b> |

$$W = 3.5 + 6 + 7.5 + 9 = 26$$

3.

|               |      |       |      |     |      |     |      |     |      |       |       |
|---------------|------|-------|------|-----|------|-----|------|-----|------|-------|-------|
| $W$           | 10   | 11    | 12   | 14  | 16   | 19  | 22   | 24  | 26   | 27    | 28    |
| $P(w \leq W)$ | 0.01 | 0.025 | 0.05 | 0.1 | 0.25 | 0.5 | 0.75 | 0.9 | 0.95 | 0.975 | 0.99  |
| $P(w \geq W)$ | 0.99 | 0.975 | 0.95 | 0.9 | 0.75 | 0.5 | 0.25 | 0.1 | 0.05 | 0.025 | 0.010 |

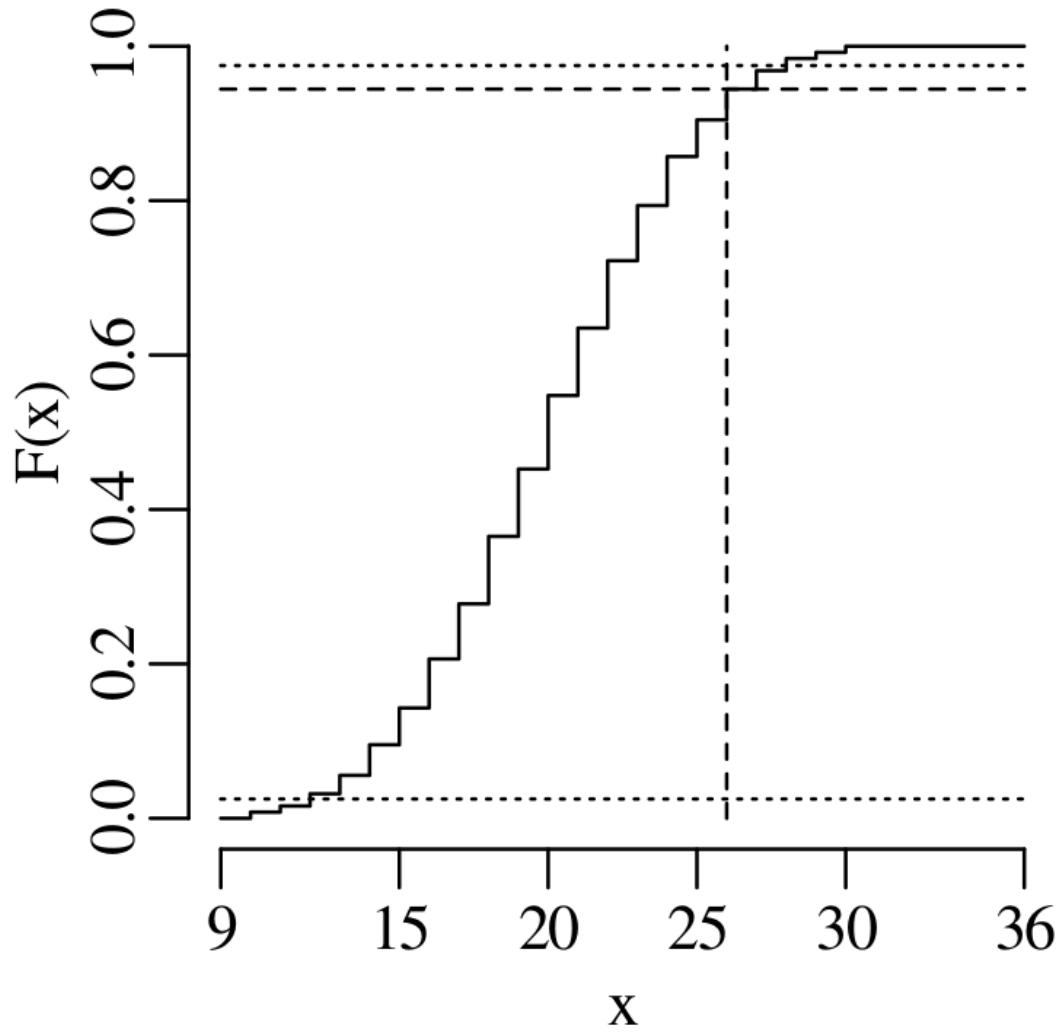
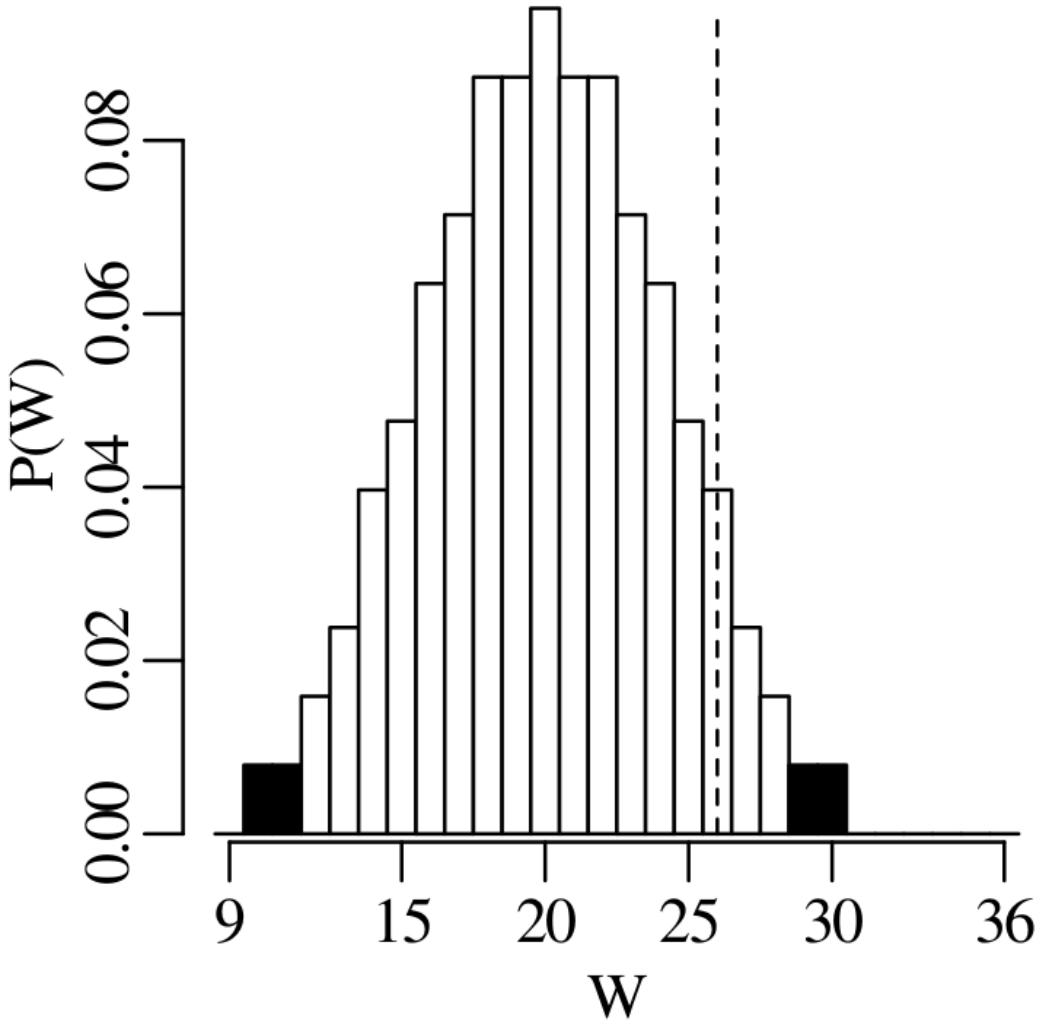
4.  $\alpha = 0.05$

|    |               |      |       |      |     |      |     |      |     |      |       |             |
|----|---------------|------|-------|------|-----|------|-----|------|-----|------|-------|-------------|
| 5. | $W$           | 10   | 11    | 12   | 14  | 16   | 19  | 22   | 24  | 26   | 27    | <b>28</b>   |
|    | $P(w \leq W)$ | 0.01 | 0.025 | 0.05 | 0.1 | 0.25 | 0.5 | 0.75 | 0.9 | 0.95 | 0.975 | 0.99        |
|    | $P(w \geq W)$ | 0.99 | 0.975 | 0.95 | 0.9 | 0.75 | 0.5 | 0.25 | 0.1 | 0.05 | 0.025 | <b>0.01</b> |

$$R = \{ W = 10 \text{ and } W > 27 \}$$

6.  $W = 26 \notin R$

7. p-value = 0.10 >  $\alpha$



## 2. Kolmogorov-Smirnov test

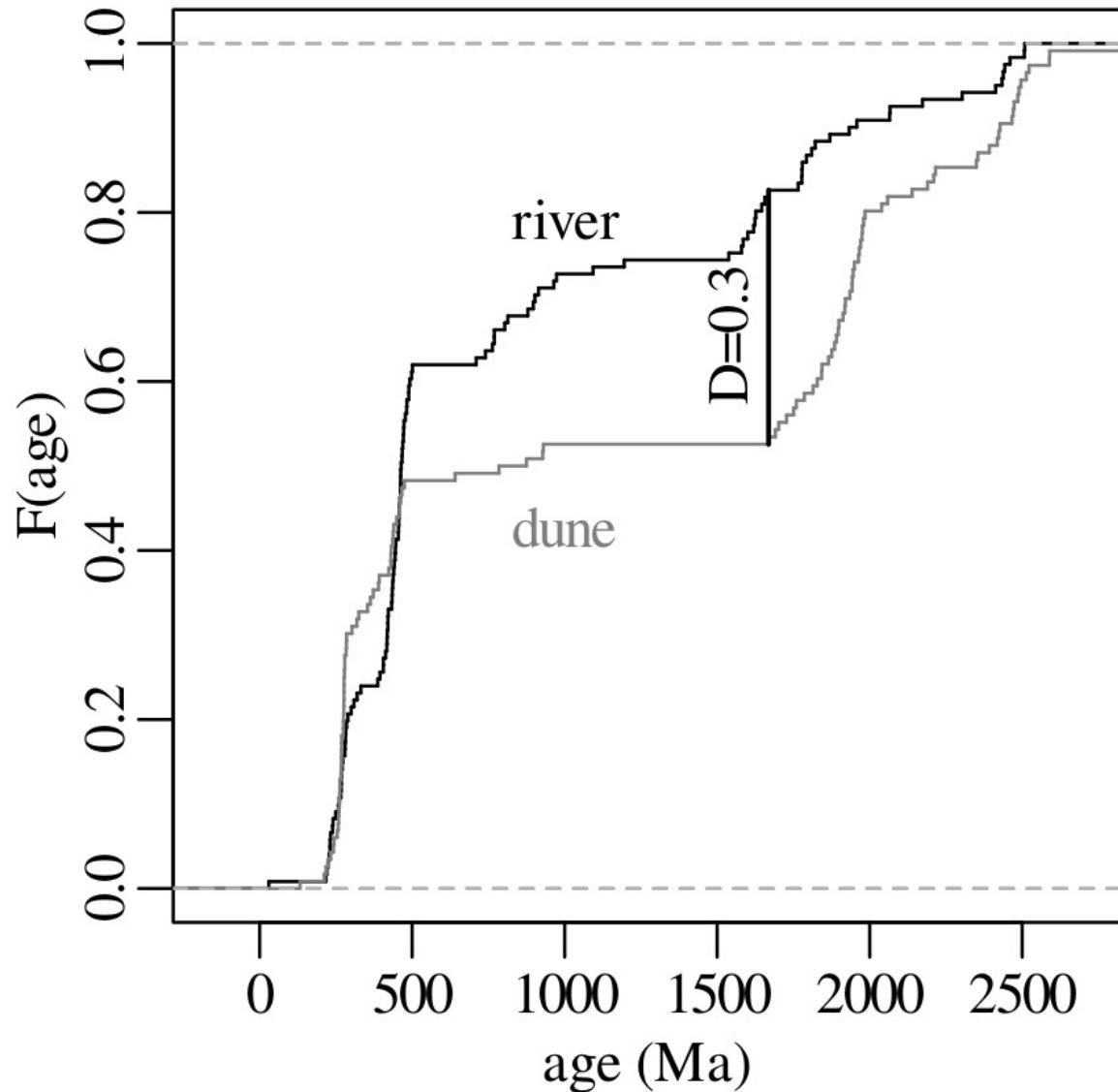
Given

$$x = \{x_1, x_2, \dots, x_n\}$$

and

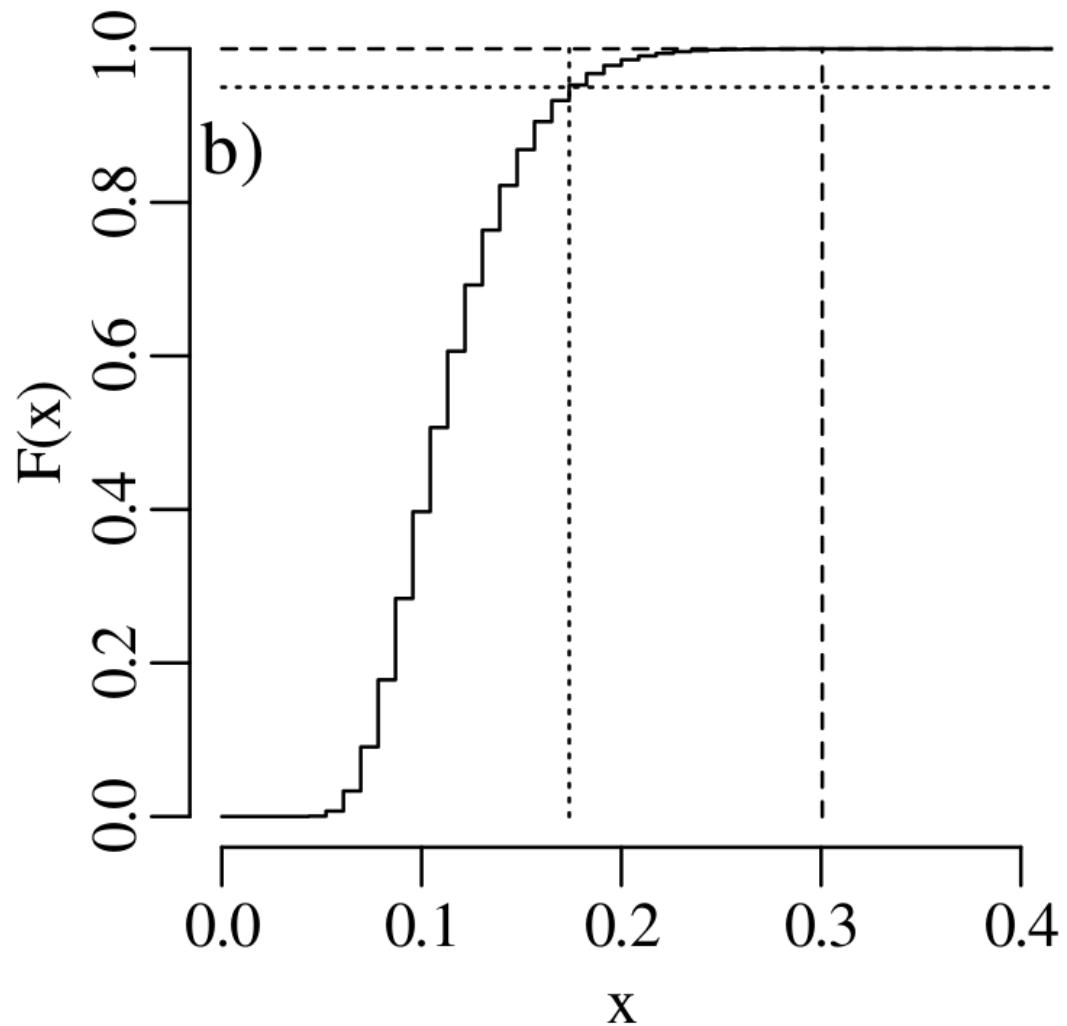
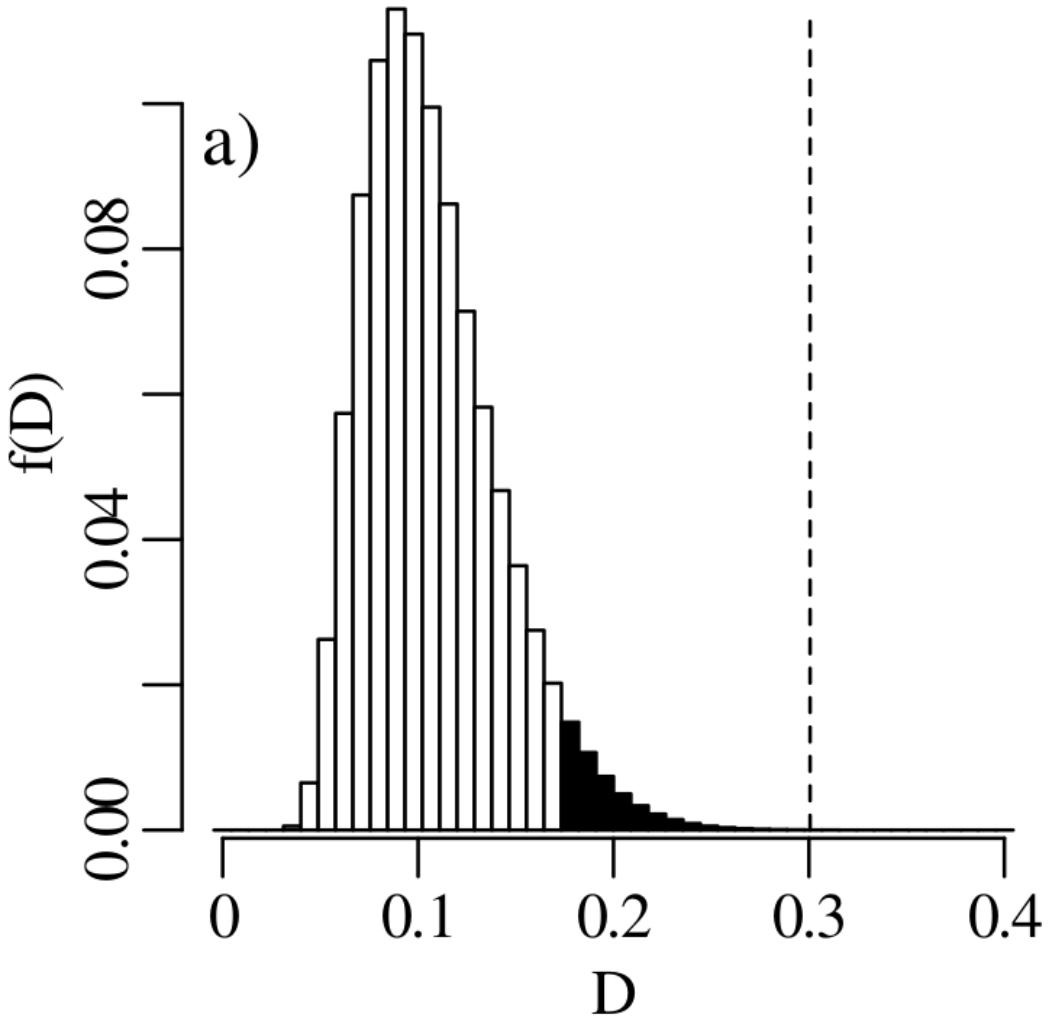
$$y = \{y_1, y_2, \dots, y_m\}$$

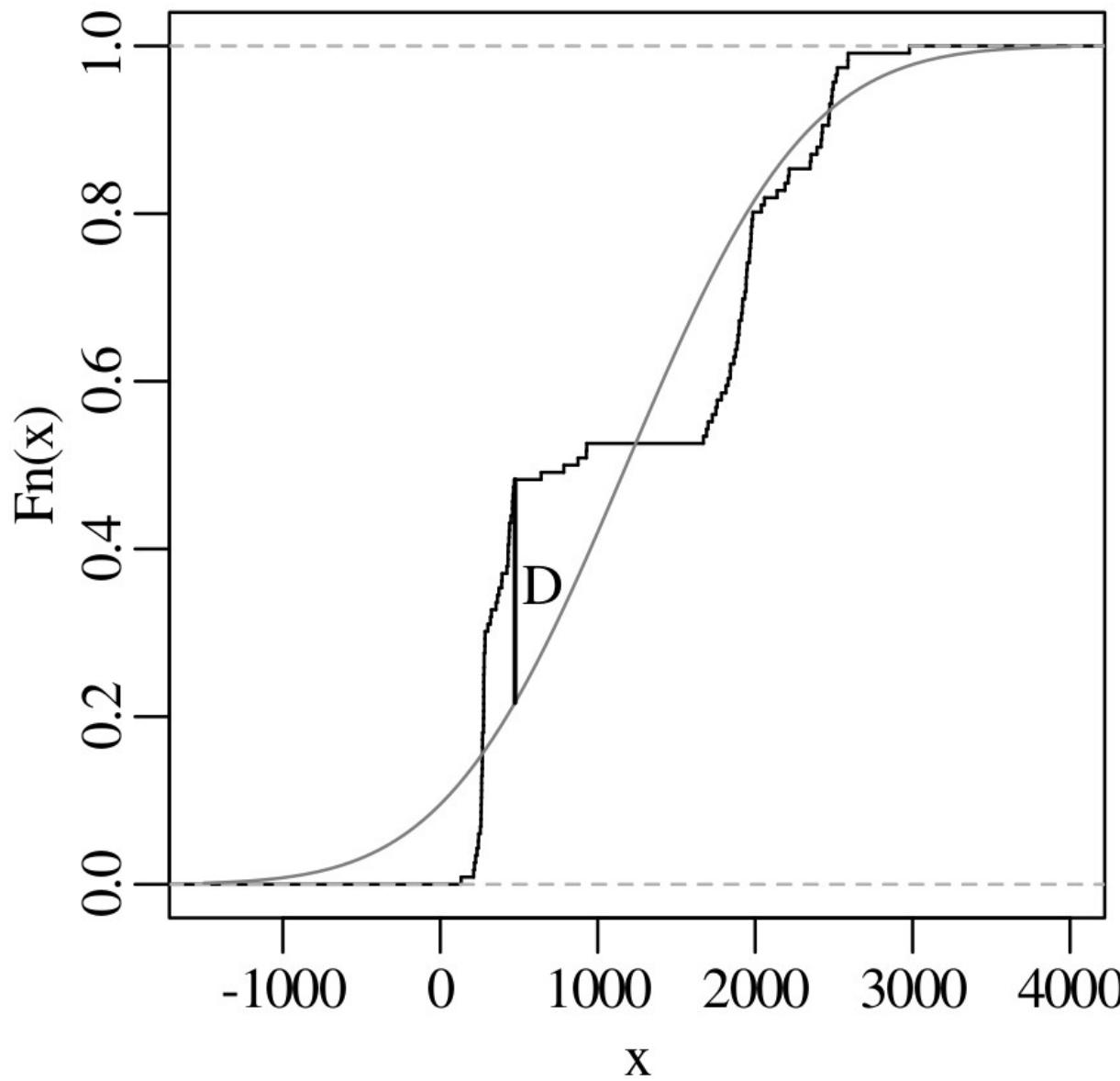
$$D = \max_z |F_x(z) - F_y(z)|$$



- $H_0$  (**null hypothesis**): samples 1 and 2 were drawn from the same distribution  
 $H_a$  (**alternative hypothesis**): samples 1 and 2 were drawn from different distributions
- $D = 0.30$
- | $D$           | 0.061 | 0.061 | 0.070 | 0.078 | 0.087 | 0.104 | 0.130 | 0.157 | 0.174 | 0.191 | 0.209 | 0.301                |
|---------------|-------|-------|-------|-------|-------|-------|-------|-------|-------|-------|-------|----------------------|
| $P(d \leq D)$ | 0.01  | 0.025 | 0.05  | 0.1   | 0.25  | 0.5   | 0.75  | 0.9   | 0.95  | 0.975 | 0.99  | 0.99996              |
| $P(d \geq D)$ | 0.99  | 0.975 | 0.95  | 0.9   | 0.75  | 0.5   | 0.25  | 0.1   | 0.05  | 0.025 | 0.010 | $4.5 \times 10^{-5}$ |
- $\alpha = 0.05$
- | $D$           | 0.061 | 0.061 | 0.070 | 0.078 | 0.087 | 0.104 | 0.130 | 0.157 | <b>0.174</b> | <b>0.191</b> | <b>0.209</b> | 0.301                |
|---------------|-------|-------|-------|-------|-------|-------|-------|-------|--------------|--------------|--------------|----------------------|
| $P(d \geq D)$ | 0.99  | 0.975 | 0.95  | 0.9   | 0.75  | 0.5   | 0.25  | 0.1   | <b>0.05</b>  | <b>0.025</b> | <b>0.010</b> | $4.5 \times 10^{-5}$ |

 $R = \{D > 0.174\}$
- $D = 0.301 \in R$
- p-value =  $4.5 \times 10^{-5} < \alpha$





# Statistics for geoscientists

## Regression

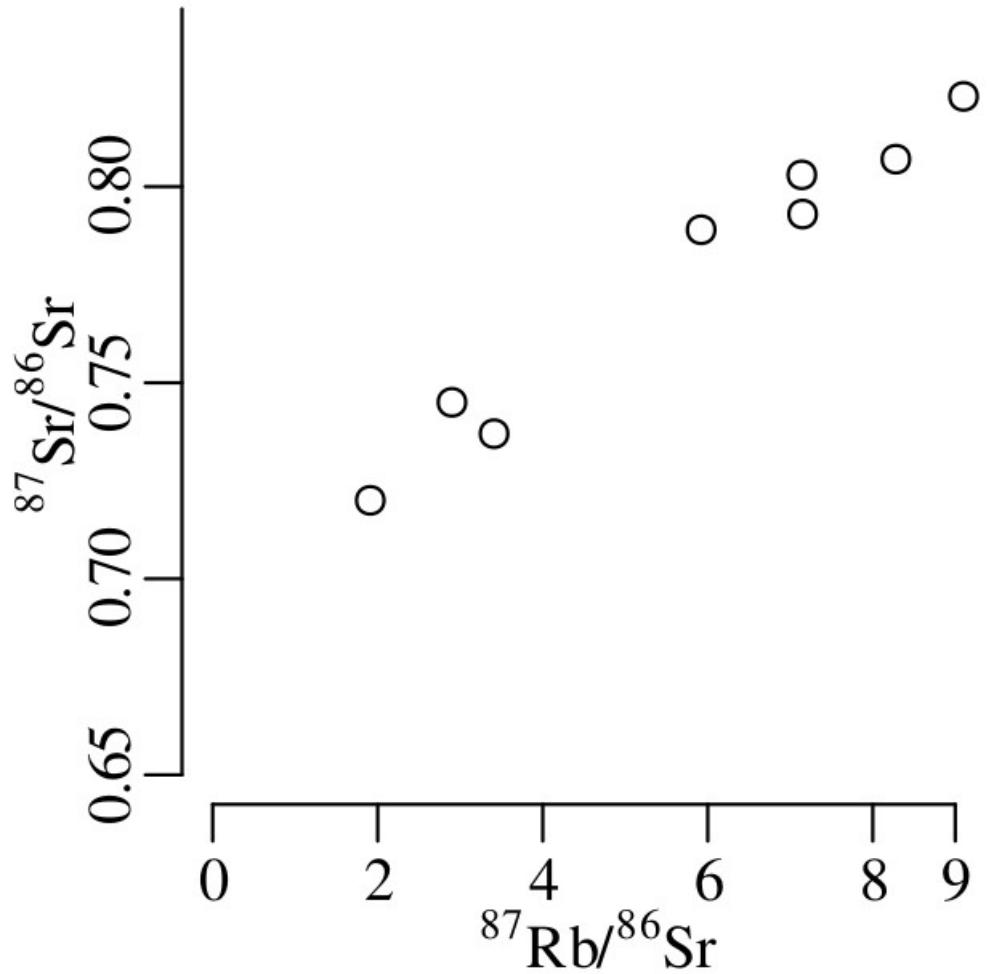
$$\left[ \frac{^{87}Sr}{^{86}Sr} \right] = \left[ \frac{^{87}Sr}{^{86}Sr} \right]_0 + \left[ \frac{^{87}Rb}{^{86}Sr} \right] (e^{\lambda t} - 1)$$

$$y_i = \beta_0 + \beta_1 x_i$$

| $i$                       | 1     | 2     | 3     | 4     | 5     | 6     | 7     | 8     |
|---------------------------|-------|-------|-------|-------|-------|-------|-------|-------|
| $[^{87}Rb/^{86}Sr] = x_i$ | 2.90  | 7.14  | 9.10  | 3.41  | 1.91  | 7.15  | 5.92  | 8.28  |
| $[^{87}Sr/^{86}Sr] = y_i$ | 0.745 | 0.803 | 0.823 | 0.737 | 0.720 | 0.793 | 0.789 | 0.807 |

$$\left[ \frac{^{87}Sr}{^{86}Sr} \right] = \left[ \frac{^{87}Sr}{^{86}Sr} \right]_o + \left[ \frac{^{87}Rb}{^{86}Sr} \right] (e^{\lambda t} - 1)$$

$$y_i = \beta_0 + \beta_1 x_i + \epsilon_i$$



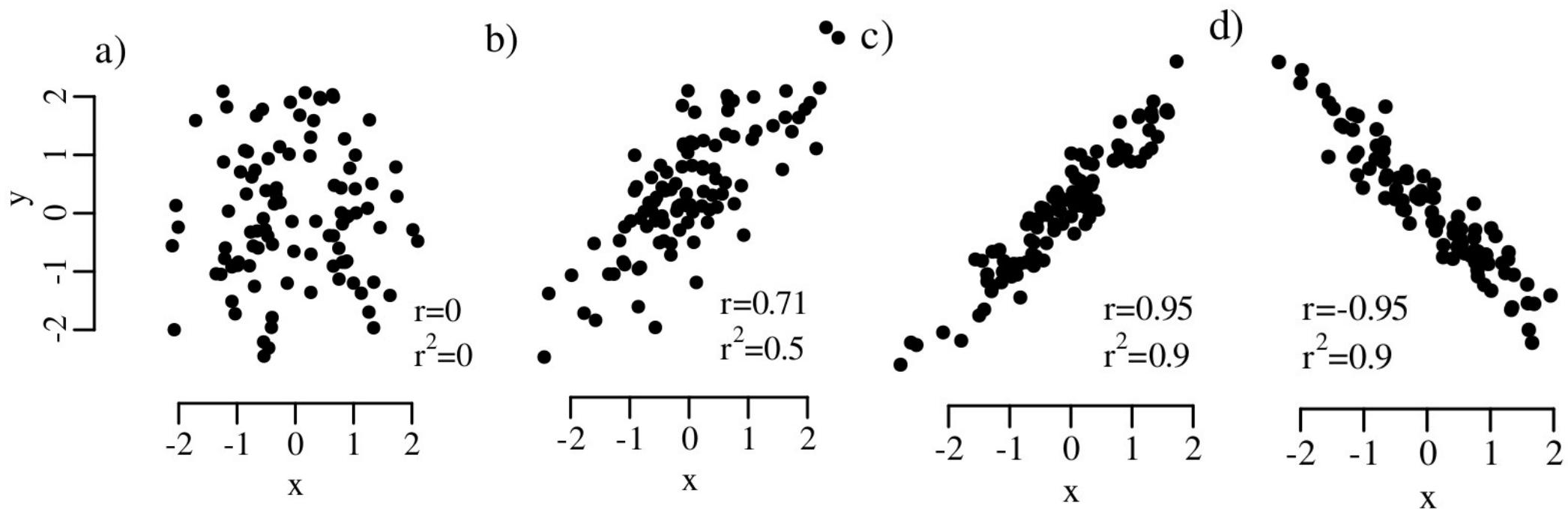
$$\rho = \frac{\sigma_{x,y}}{\sigma_x\sigma_y}$$

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

$$s[x,y] = \sum_{i=1}^n \frac{1}{n-1}(x_i - \bar{x})(y_i - \bar{y})$$

$$r=\frac{s[x,y]}{s[x]s[y]}$$

$$r = \frac{s[x, y]}{s[x]s[y]}$$



$$t = \frac{r\sqrt{n-2}}{\sqrt{1-r^2}}$$

$$t = \frac{r\sqrt{n-2}}{\sqrt{1-r^2}}$$

1.  $H_0$  (null hypothesis)  $\rho = 0$

$H_a$  (alternative hypothesis):  $\rho \neq 0$

2.  $t = \frac{0.985\sqrt{8-2}}{\sqrt{1-0.985^2}} = 13.98$

3.

|               |       |       |       |       |       |     |      |      |      |       |       |           |
|---------------|-------|-------|-------|-------|-------|-----|------|------|------|-------|-------|-----------|
| $t$           | -3.10 | -2.40 | -1.90 | -1.40 | -0.72 | 0   | 0.72 | 1.40 | 1.90 | 2.40  | 3.10  | 13.98     |
| $P(t \leq T)$ | 0.01  | 0.025 | 0.05  | 0.1   | 0.25  | 0.5 | 0.75 | 0.9  | 0.95 | 0.975 | 0.99  | 0.9999958 |
| $P(t \geq T)$ | 0.99  | 0.975 | 0.95  | 0.9   | 0.75  | 0.5 | 0.25 | 0.1  | 0.05 | 0.025 | 0.010 | 0.0000042 |

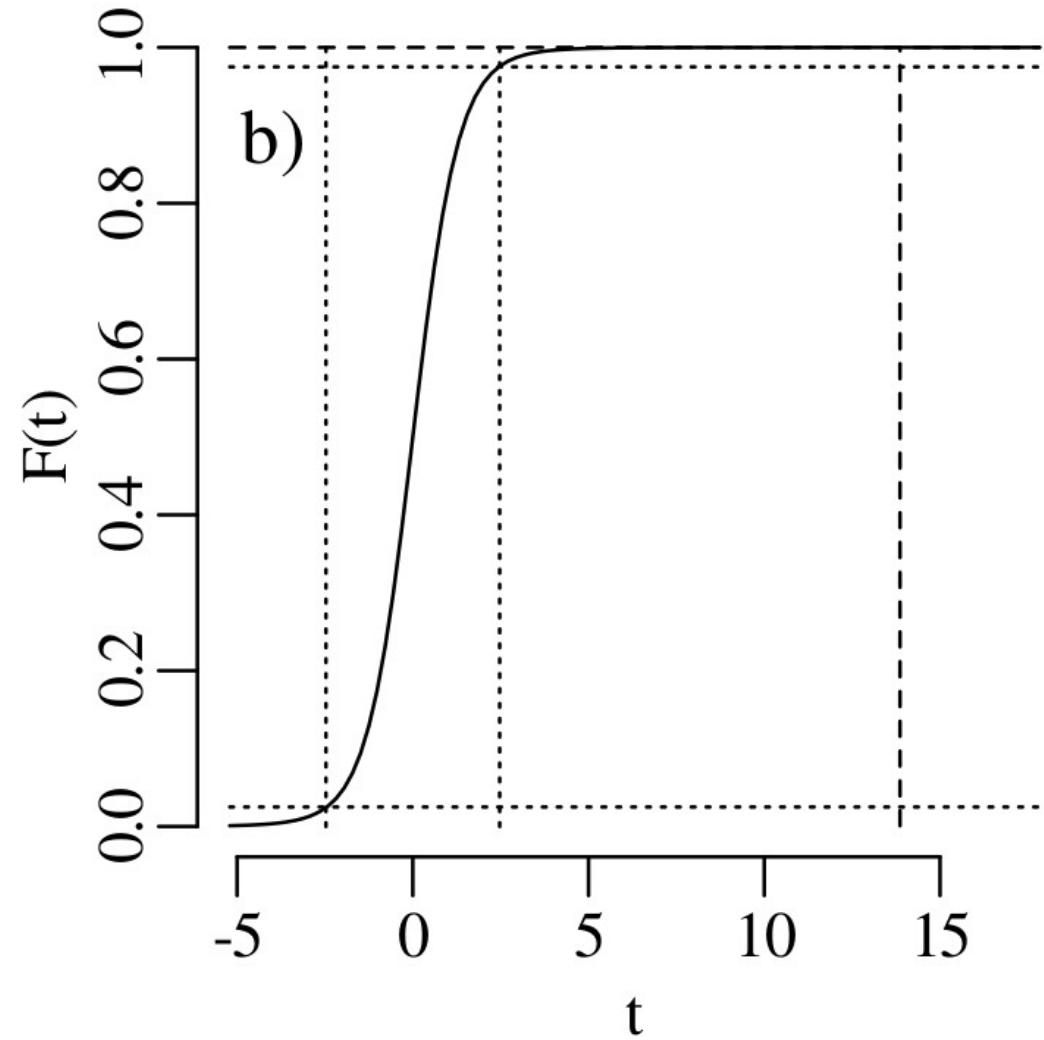
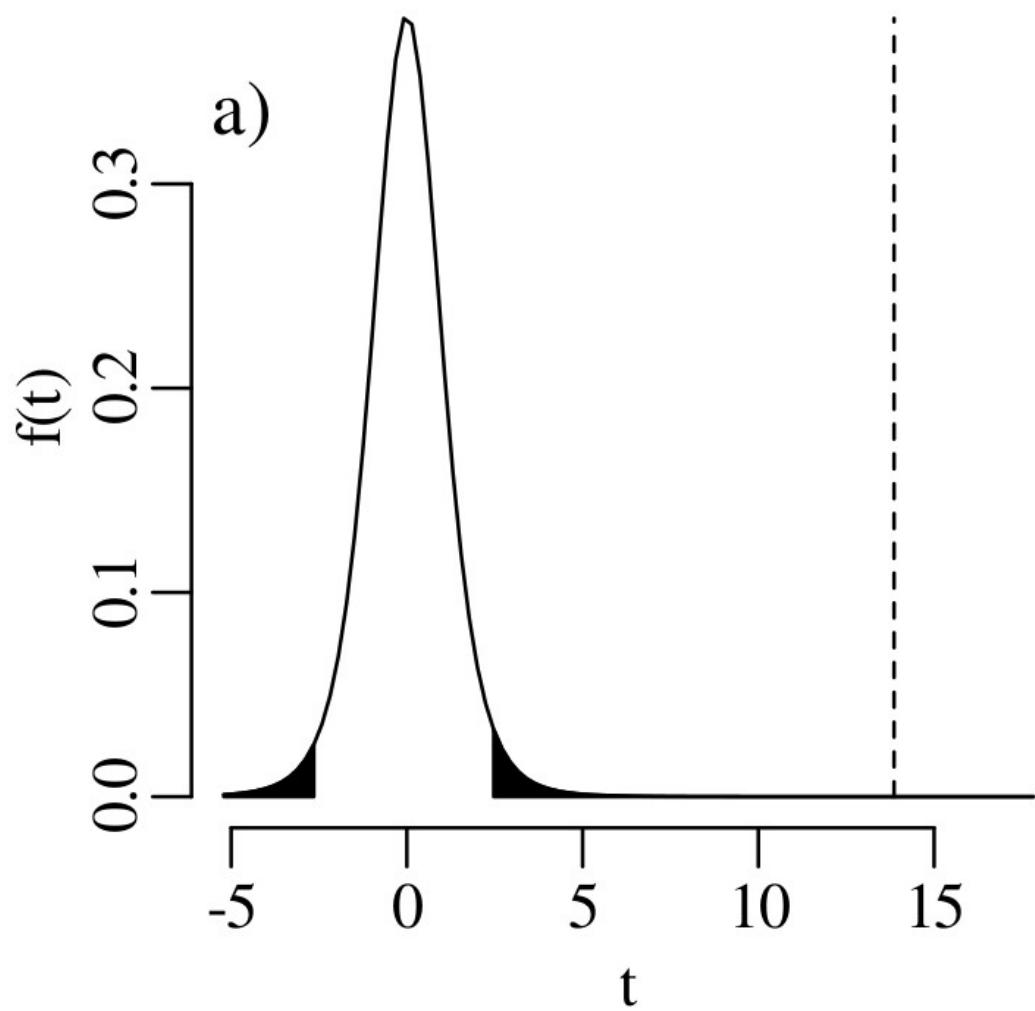
4.  $\alpha = 0.05$

| 5. | $t$           | -3.10 | -2.40 | -1.90 | -1.40 | -0.72 | 0   | 0.72 | 1.40 | 1.90 | 2.40  | 3.10  | 13.98     |
|----|---------------|-------|-------|-------|-------|-------|-----|------|------|------|-------|-------|-----------|
|    | $P(t \leq T)$ | 0.01  | 0.025 | 0.05  | 0.1   | 0.25  | 0.5 | 0.75 | 0.9  | 0.95 | 0.975 | 0.99  | 0.9999958 |
|    | $P(t \geq T)$ | 0.99  | 0.975 | 0.95  | 0.9   | 0.75  | 0.5 | 0.25 | 0.1  | 0.05 | 0.025 | 0.010 | 0.0000042 |

$$R = \{|t| > 2.40\}$$

6.  $t = 13.98 \in R$

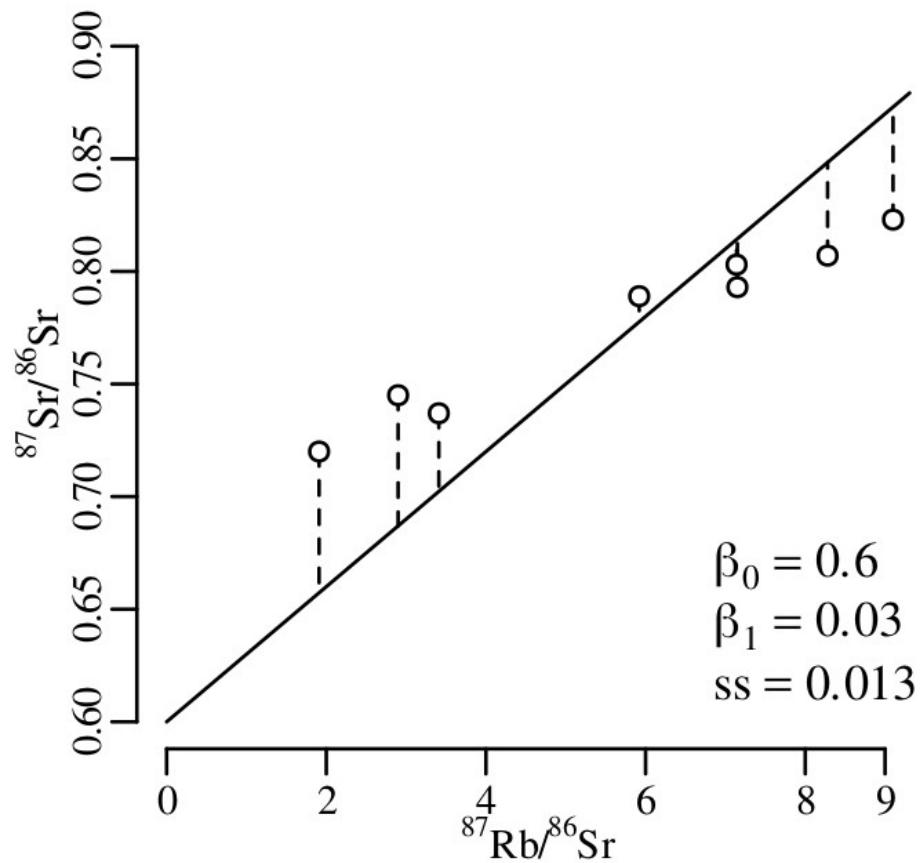
7. p-value =  $2 \times 0.0000042 = 0.0000084 < \alpha$



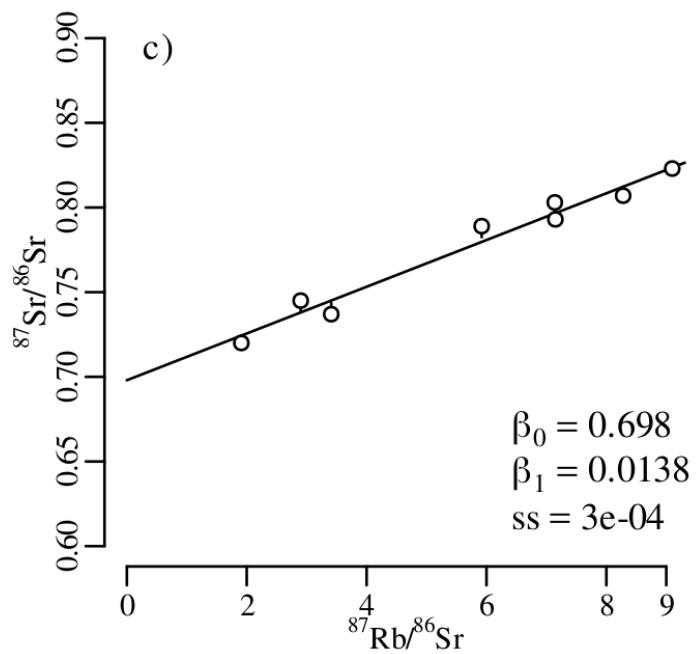
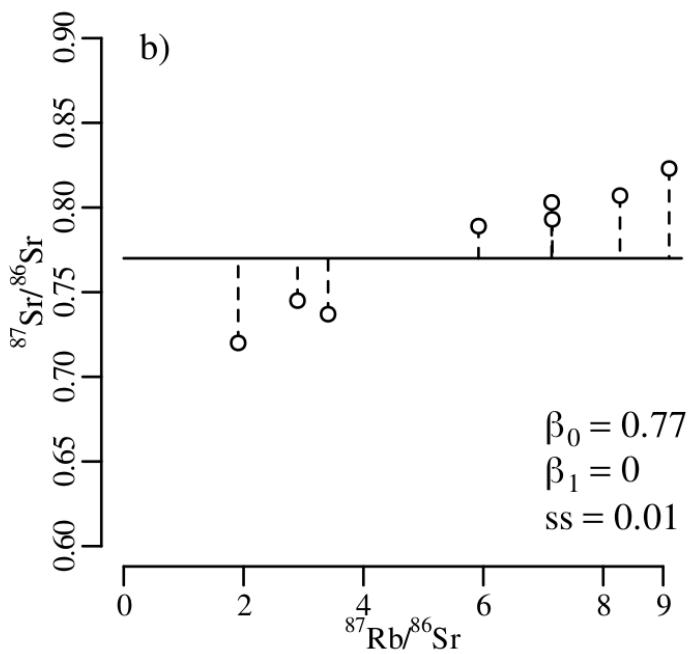
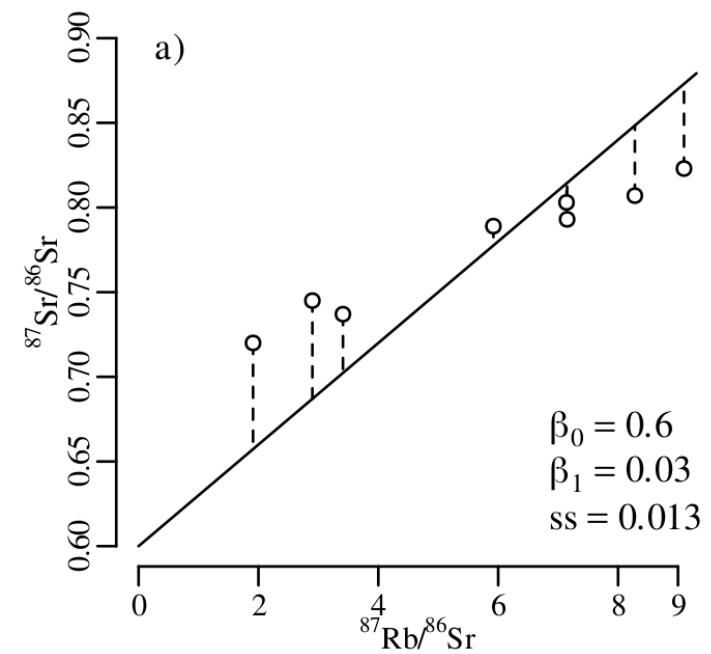
# Least Squares

$$y_i = \beta_0 + \beta_1 x_i + \epsilon_i$$

$$ss \equiv \sum_{i=1}^n \epsilon_i^2 = \sum_{i=1}^n (\beta_0 + \beta_1 x_i - y_i)^2$$



| $i$                     | 1      | 2      | 3     | 4       | 5       | 6      | 7       | 8      |
|-------------------------|--------|--------|-------|---------|---------|--------|---------|--------|
| $x_i$                   | 2.90   | 7.14   | 9.10  | 3.41    | 1.91    | 7.15   | 5.92    | 8.28   |
| $y_i$                   | 0.745  | 0.803  | 0.823 | 0.737   | 0.720   | 0.793  | 0.789   | 0.807  |
| $\beta_0 + \beta_1 x_i$ | 0.687  | 0.8142 | 0.873 | 0.7023  | 0.6573  | 0.8145 | 0.7776  | 0.8484 |
| $\epsilon_i$            | -0.058 | 0.0112 | 0.05  | -0.0347 | -0.0627 | 0.0215 | -0.0114 | 0.0414 |

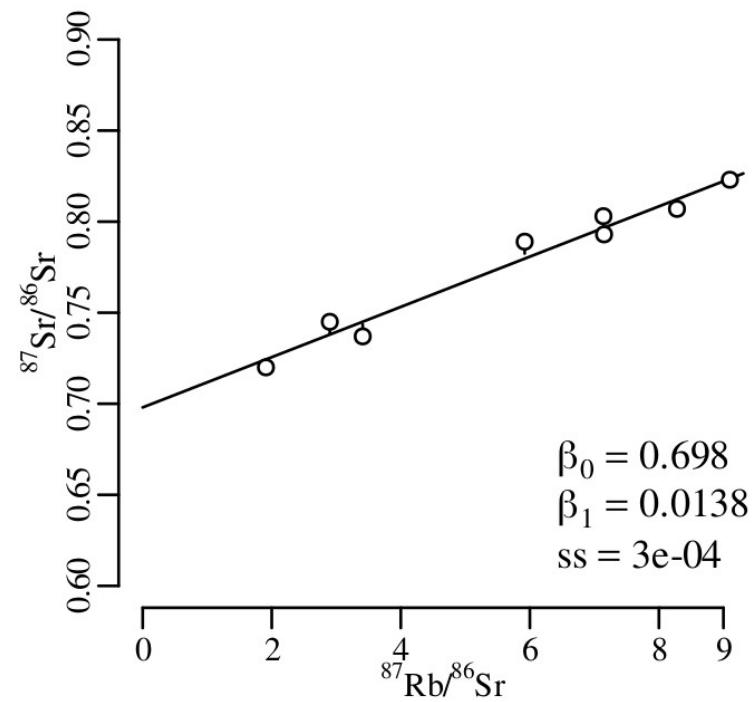


$$\begin{cases} \frac{\partial ss}{\partial \beta_0} = 2 \sum (\beta_0 + \beta_1 x_i - y_i) = 0 \\ \frac{\partial ss}{\partial \beta_1} = 2 \sum (\beta_0 + \beta_1 x_i - y_i) x_i = 0 \end{cases}$$

$$\begin{cases} \hat{\beta}_0 = (\sum x_i^2 \sum y_i - \sum x_i \sum x_i y_i) / \left( n \sum x_i^2 - (\sum x_i)^2 \right) \\ \hat{\beta}_1 = (n \sum x_i y_i - \sum x_i \sum y_i) / \left( n \sum x_i^2 - (\sum x_i)^2 \right) \end{cases}$$

| $i$       | 1     | 2     | 3     | 4     | 5     | 6     | 7     | 8     | $\sum_{i=1}^8$ |
|-----------|-------|-------|-------|-------|-------|-------|-------|-------|----------------|
| $x_i$     | 2.90  | 7.14  | 9.10  | 3.41  | 1.91  | 7.15  | 5.92  | 8.28  | 45.81          |
| $y_i$     | 0.745 | 0.803 | 0.823 | 0.737 | 0.720 | 0.793 | 0.789 | 0.807 | 6.217          |
| $x_i^2$   | 8.41  | 50.98 | 82.81 | 11.63 | 3.648 | 51.12 | 35.05 | 68.56 | 312.2          |
| $x_i y_i$ | 2.160 | 5.733 | 7.489 | 2.513 | 1.375 | 5.670 | 4.671 | 6.682 | 36.29          |

$$\begin{cases} \hat{\beta}_0 = (312.2 \times 6.217 - 45.81 \times 36.29) / (8 \times 312.2 - 45.81^2) = 0.698 \\ \hat{\beta}_1 = (8 \times 36.29 - 45.81 \times 6.217) / (8 \times 312.2 - 45.81^2) = 0.0138 \end{cases}$$



# Maximum Likelihood

$$y_i = \beta_0 + \beta_1 x_i + \epsilon_i$$

$$f(\epsilon_i | \beta_0, \beta_1, \sigma) = \frac{1}{\sigma \sqrt{2\pi}} \exp\left[-\frac{\epsilon_i^2}{2\sigma^2}\right], \text{ where } \epsilon_i = \beta_0 + \beta_1 x_i - y_i$$

$$\mathcal{L}(\beta_0, \beta_1, \sigma | \{(x_1, y_1), \dots, (x_n, y_n)\}) = \prod_{i=1}^n \frac{1}{\sigma \sqrt{2\pi}} \exp\left[-\frac{\epsilon_i^2}{2\sigma^2}\right]$$

$$\begin{aligned}
\max_{\beta_0, \beta_1} \left[ \prod_{i=1}^n \mathcal{L} \right] &= \max_{\beta_0, \beta_1} \left[ \sum_{i=1}^n \ln \mathcal{L} \right] = \max_{\beta_0, \beta_1} \left[ \ln \left( \frac{1}{\sigma \sqrt{2\pi}} \right) - \sum_{i=1}^n \left( \frac{\epsilon_i^2}{2\sigma^2} \right) \right] \\
&= \max_{\beta_0, \beta_1} \left[ - \sum_{i=1}^n \epsilon_i^2 \right] = \min_{\beta_0, \beta_1} \left[ \sum_{i=1}^n \epsilon_i^2 \right] = \min_{\beta_0, \beta_1} (ss)
\end{aligned}$$

$$\Sigma_{\hat{\beta}} = \begin{bmatrix} s[\hat{\beta}_0]^2 & s[\hat{\beta}_0, \hat{\beta}_1] \\ s[\hat{\beta}_0, \hat{\beta}_1] & s[\hat{\beta}_1]^2 \end{bmatrix} = \begin{bmatrix} \frac{\sigma^2 \sum_{i=1}^n x_i^2}{n \sum_{i=1}^n (x_i^2 - \bar{x}^2)} & -\frac{\sigma^2 \bar{x}}{\sum_{i=1}^n (x_i^2 - \bar{x}^2)} \\ -\frac{\sigma^2 \bar{x}}{\sum_{i=1}^n (x_i^2 - \bar{x}^2)} & \frac{\sigma^2}{\sum_{i=1}^n (x_i^2 - \bar{x}^2)} \end{bmatrix}$$

$$\beta_0 \in \left\{ \hat{\beta}_0 \pm t_{df,\alpha/2} \ s[\hat{\beta}_0] \right\}$$

$$\beta_1 \in \left\{ \hat{\beta}_1 \pm t_{df,\alpha/2} \ s[\hat{\beta}_1] \right\}$$

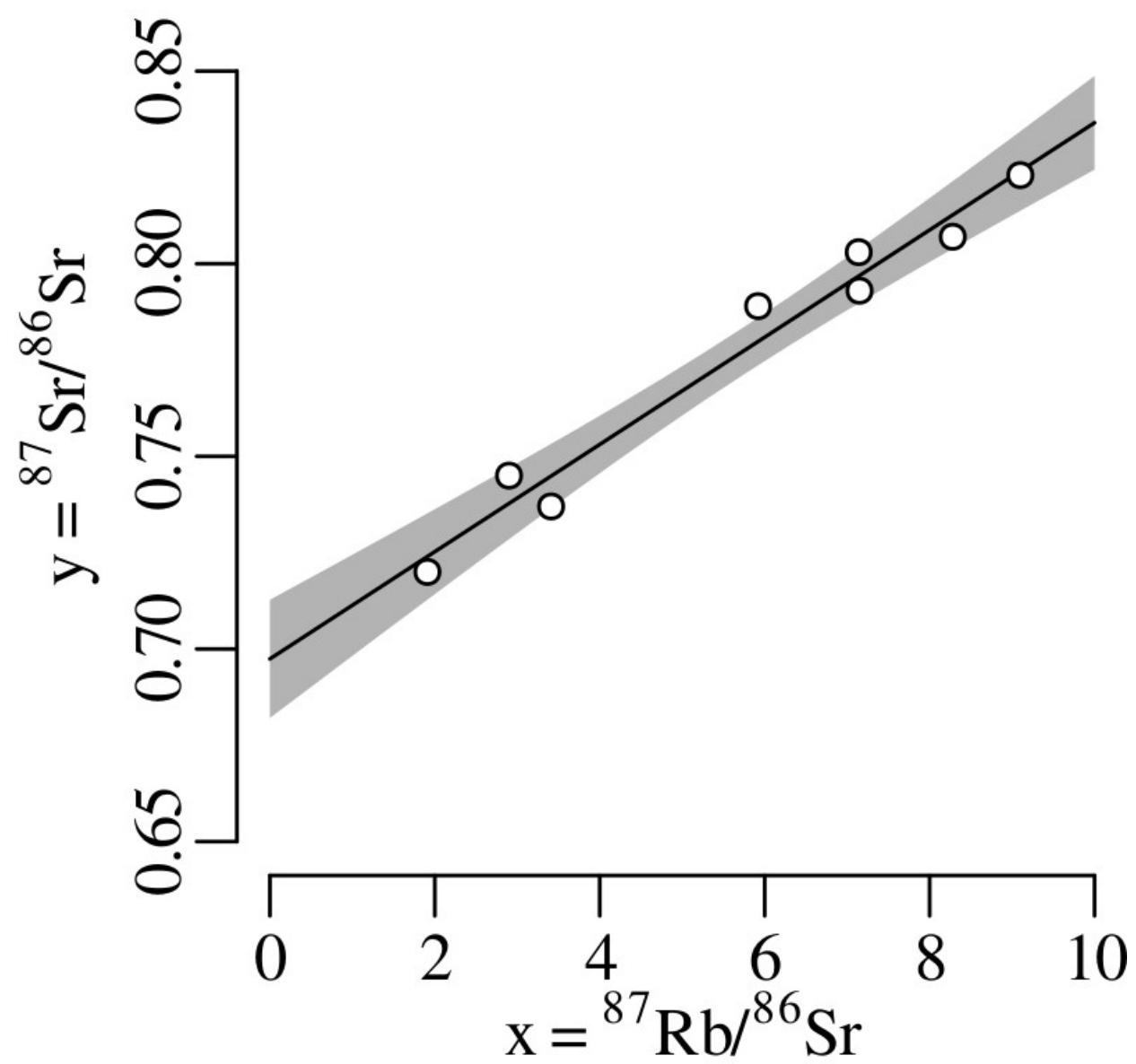
where  $df = n - 2$

$$y = \beta_0 + \beta_1 x$$

$$\Sigma_{\hat{\beta}} = \begin{bmatrix} s[\hat{\beta}_0]^2 & s[\hat{\beta}_0, \hat{\beta}_1] \\ s[\hat{\beta}_0, \hat{\beta}_1] & s[\hat{\beta}_1]^2 \end{bmatrix}$$

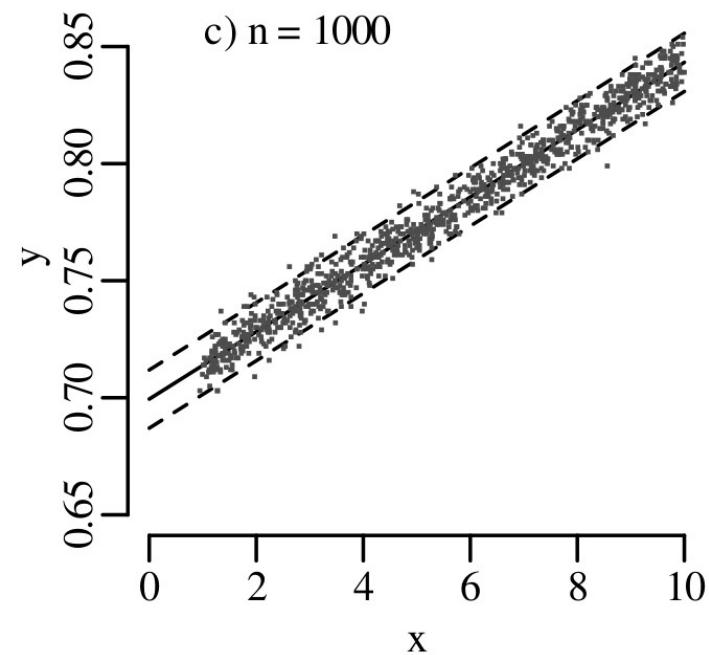
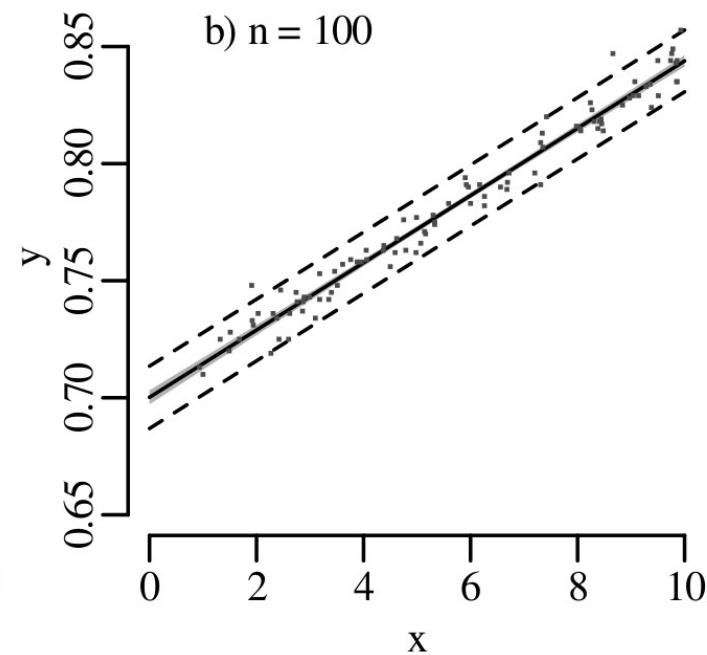
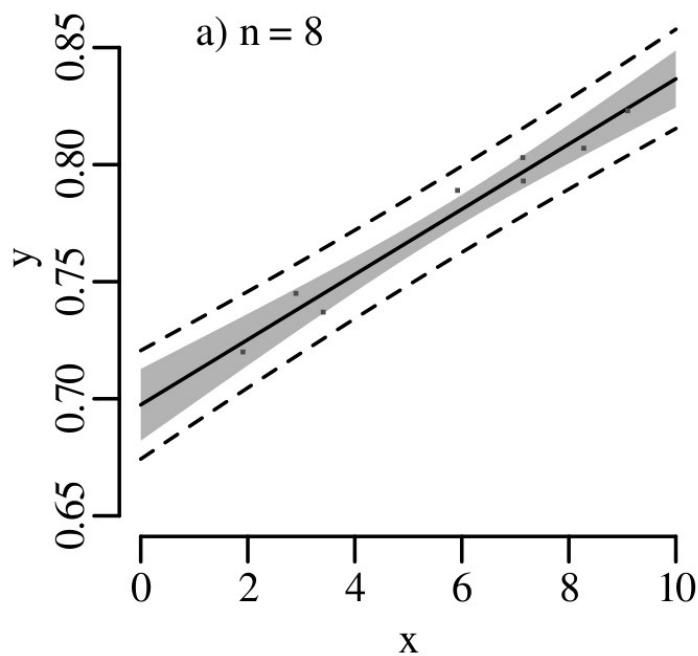


$$y \in \left\{ \hat{\beta}_0 + \hat{\beta}_1 x \pm t_{df, \alpha/2} \sqrt{s[\hat{\beta}_0]^2 + s[\hat{\beta}_1]^2 x^2 + 2s[\hat{\beta}_0, \hat{\beta}_1]x} \right\}$$

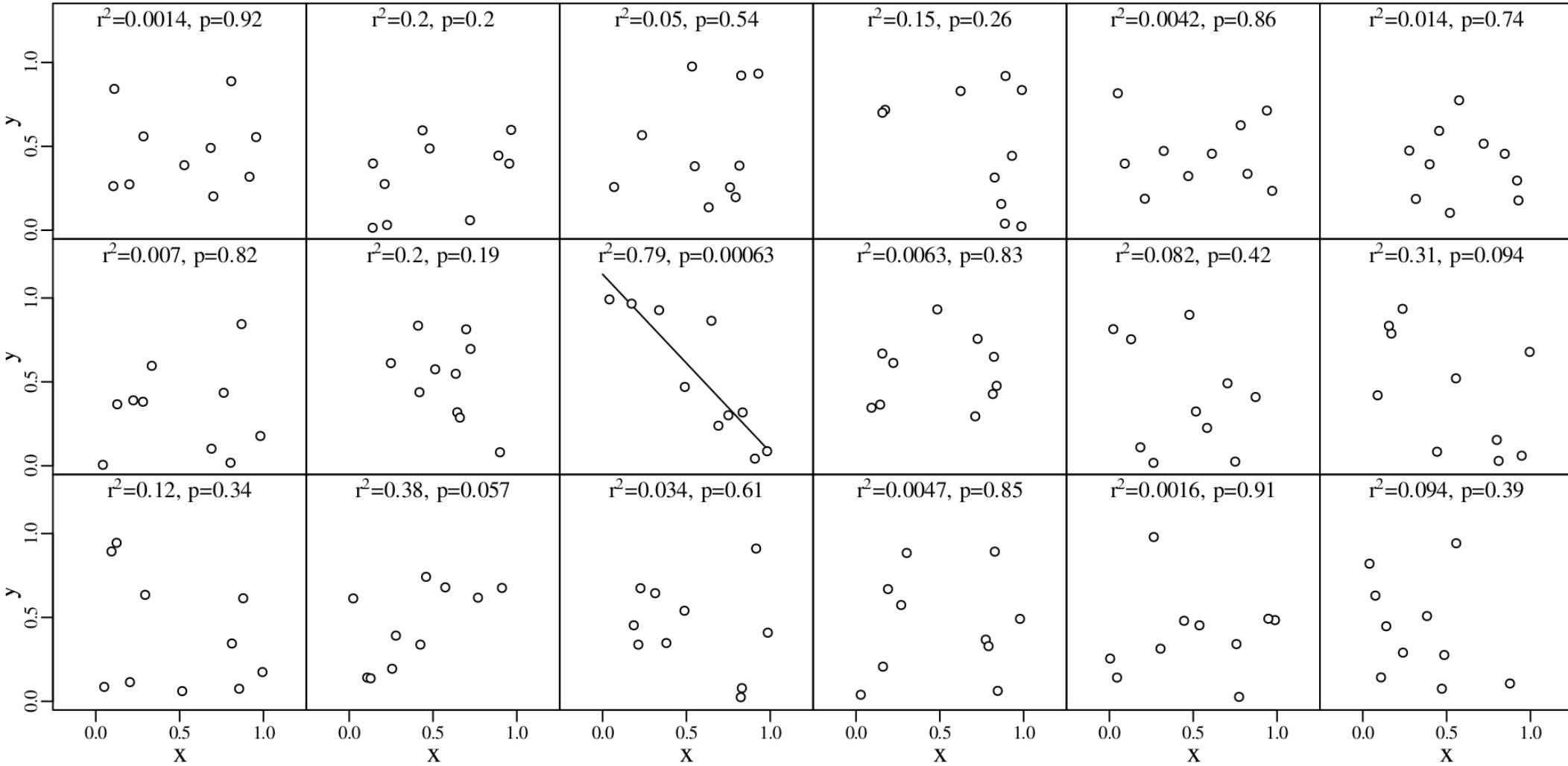


$$y \in \left\{ \hat{\beta}_0 + \hat{\beta}_1 x \pm t_{df,\alpha/2} \sqrt{s[\hat{\beta}_0]^2 + s[\hat{\beta}_1]^2 x^2 + 2s[\hat{\beta}_0, \hat{\beta}_1]x} \right\}$$

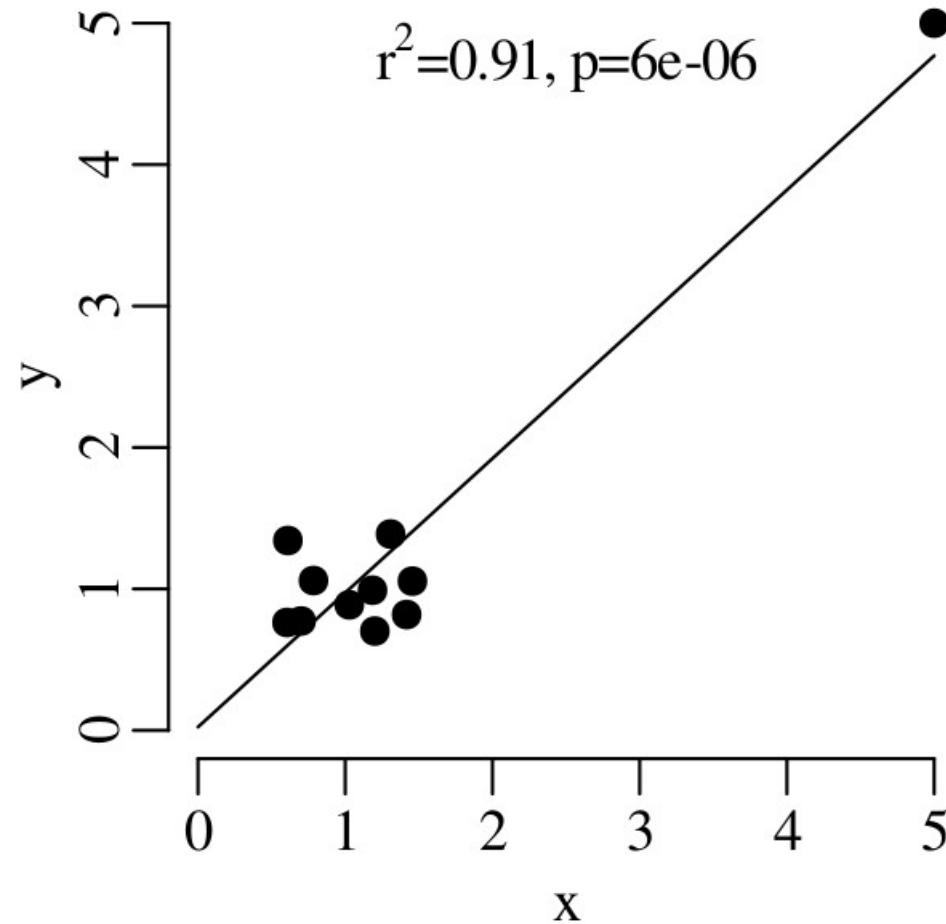
$$y \in \left\{ \hat{\beta}_0 + \hat{\beta}_1 x \pm t_{df,\alpha/2} \sqrt{\hat{\sigma}^2 + s[\hat{\beta}_0]^2 + s[\hat{\beta}_1]^2 x^2 + 2s[\hat{\beta}_0, \hat{\beta}_1]x} \right\}$$



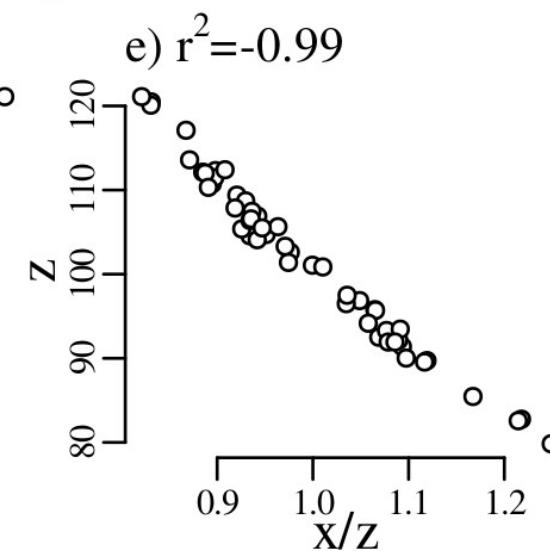
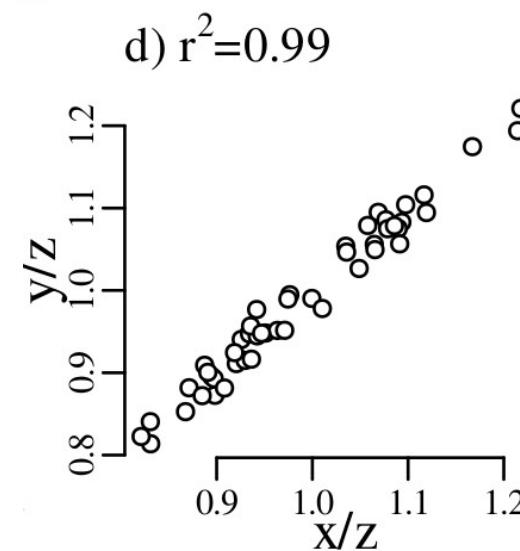
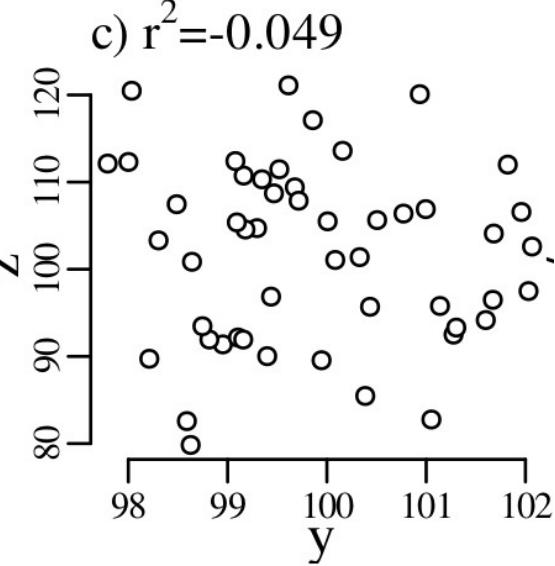
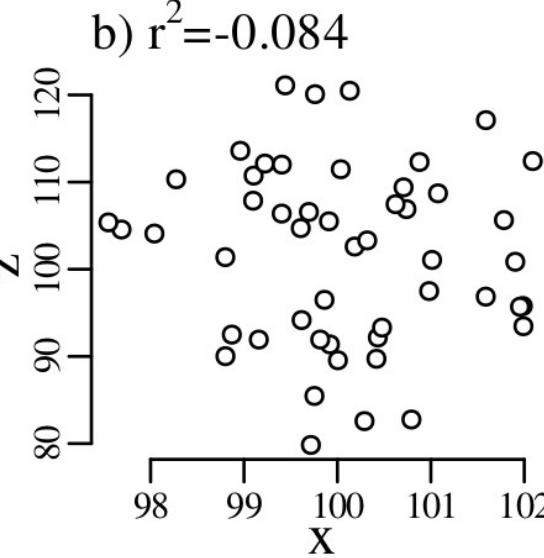
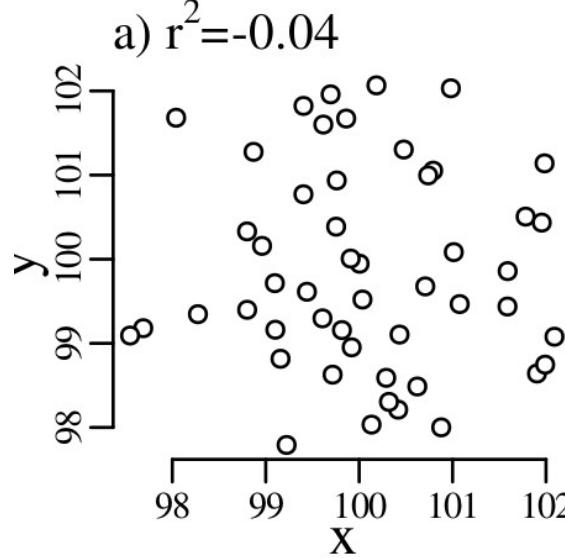
# p-hacking

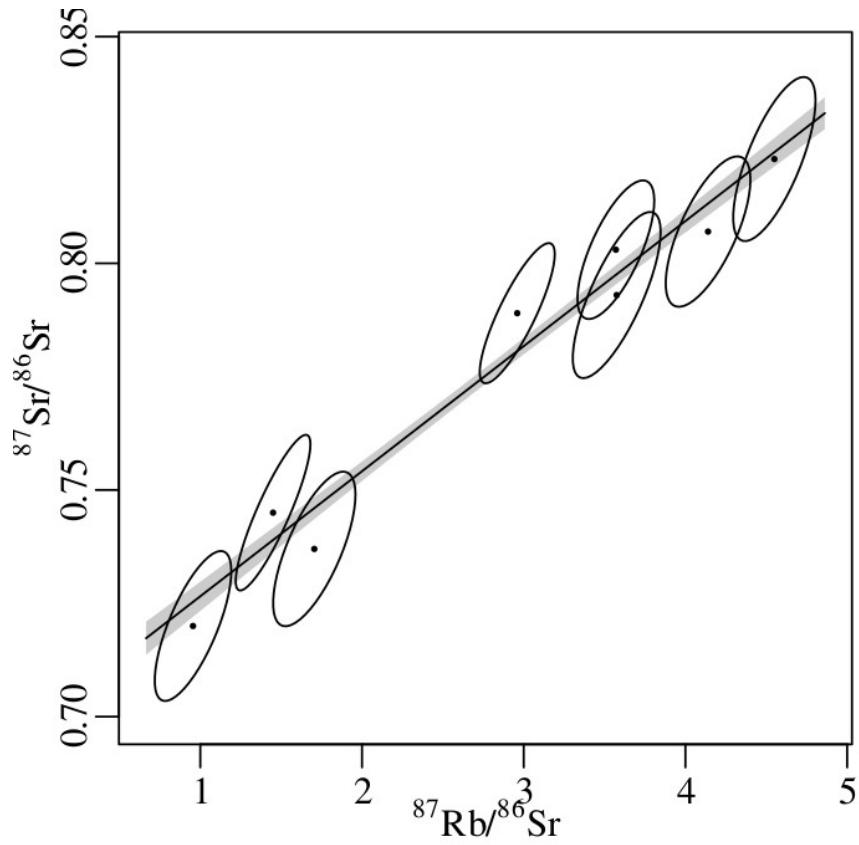
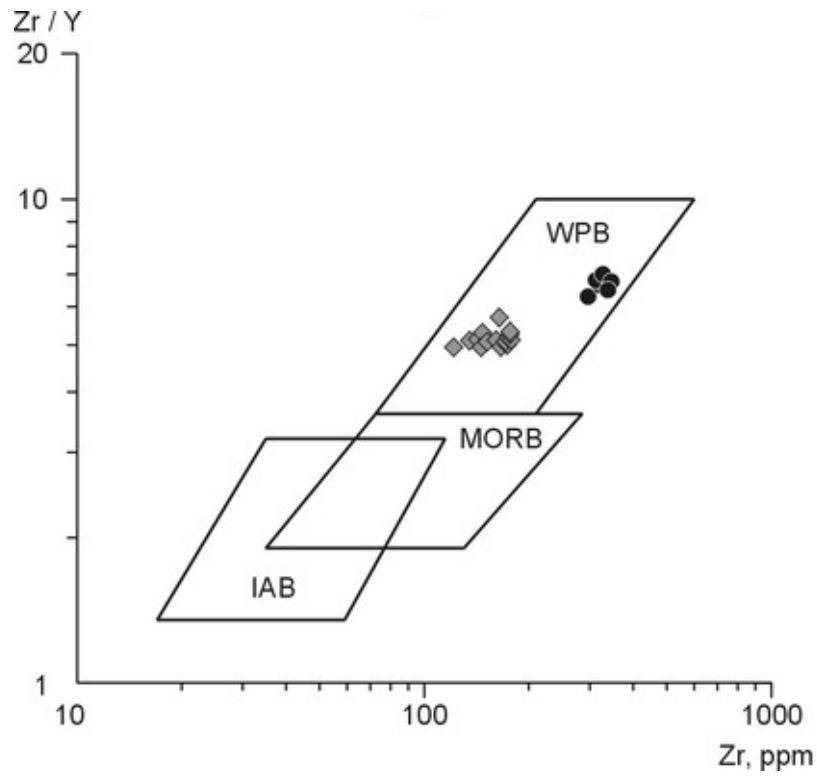


# outliers



# Spurious correlation



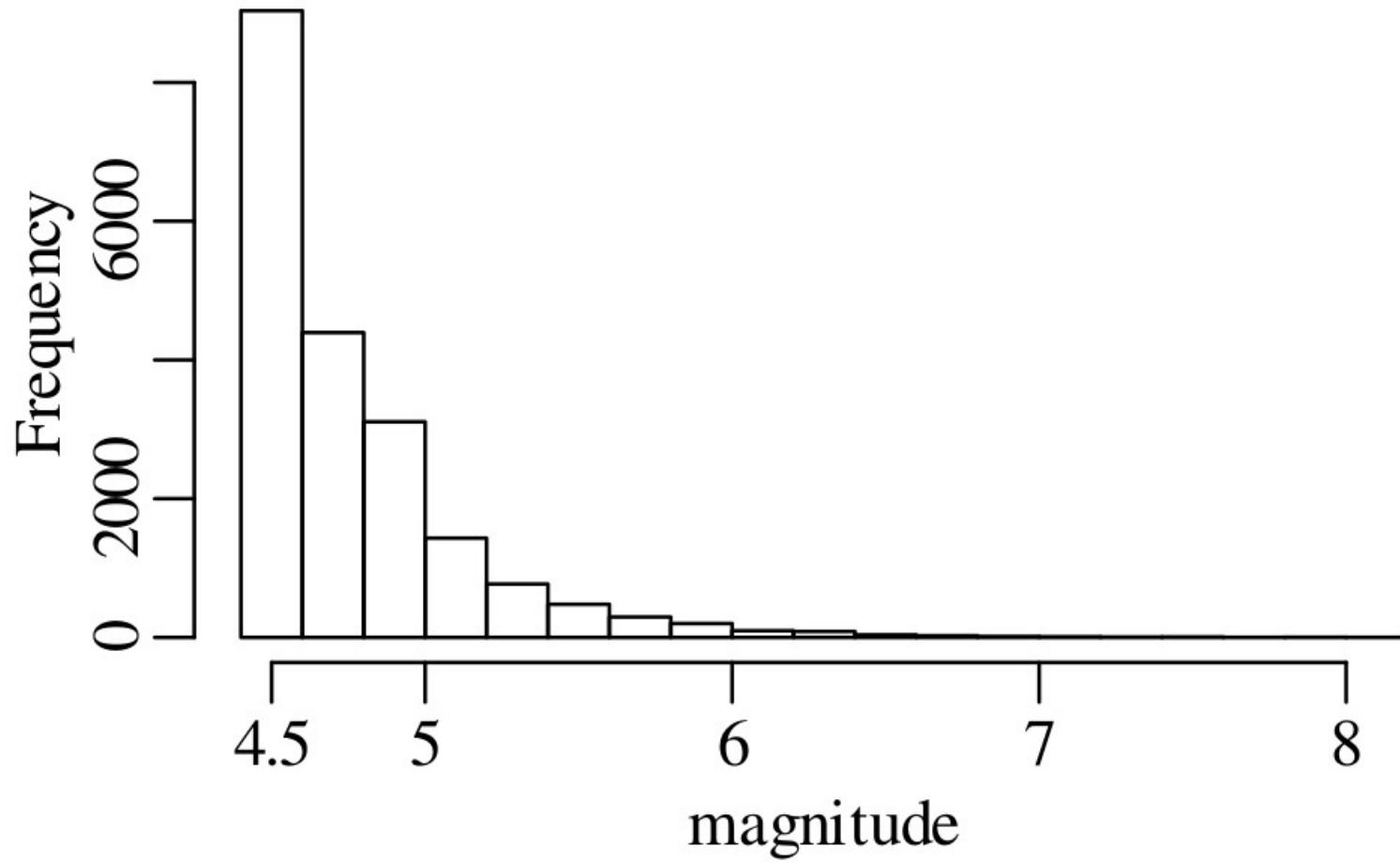


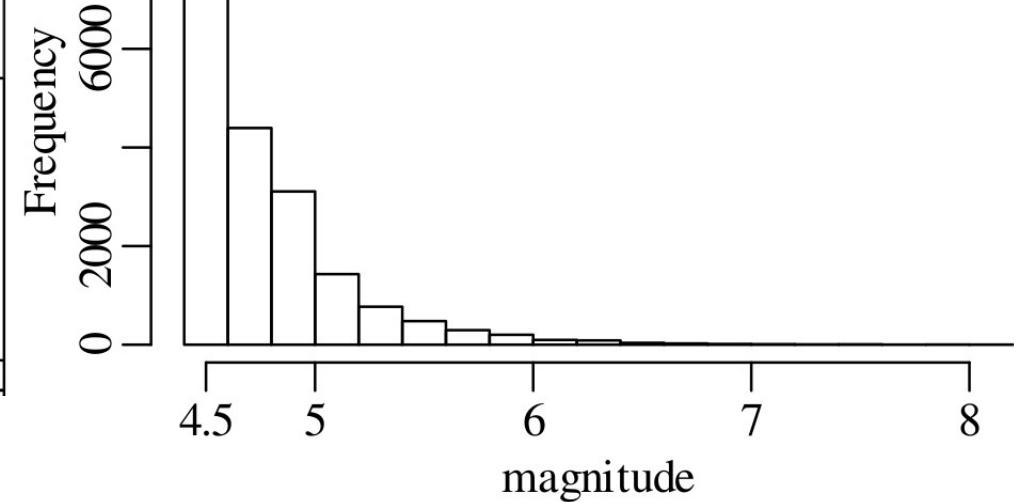
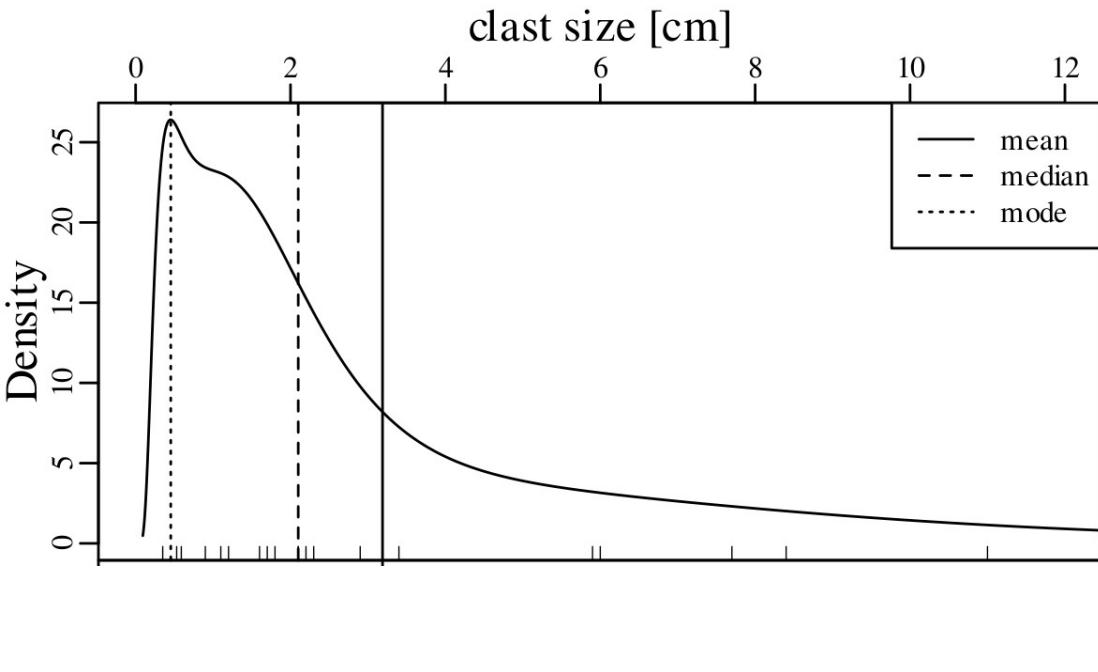
$$\rho_{\frac{w}{y}, \frac{x}{z}} \approx \frac{\rho_{w,x} \left[ \frac{\sigma_w}{\mu_w} \right] \left[ \frac{\sigma_x}{\mu_x} \right] - \rho_{w,z} \left[ \frac{\sigma_w}{\mu_w} \right] \left[ \frac{\sigma_z}{\mu_z} \right] - \rho_{x,y} \left[ \frac{\sigma_x}{\mu_x} \right] \left[ \frac{\sigma_y}{\mu_y} \right] + \rho_{y,z} \left[ \frac{\sigma_y}{\mu_y} \right] \left[ \frac{\sigma_z}{\mu_z} \right]}{\sqrt{\left[ \frac{\sigma_w}{\mu_w} \right]^2 + \left[ \frac{\sigma_y}{\mu_y} \right]^2 - 2\rho_{w,y} \left[ \frac{\sigma_w}{\mu_w} \right] \left[ \frac{\sigma_y}{\mu_y} \right]} \sqrt{\left[ \frac{\sigma_x}{\mu_x} \right]^2 + \left[ \frac{\sigma_z}{\mu_z} \right]^2 - 2\rho_{x,z} \left[ \frac{\sigma_x}{\mu_x} \right] \left[ \frac{\sigma_z}{\mu_z} \right]}}$$

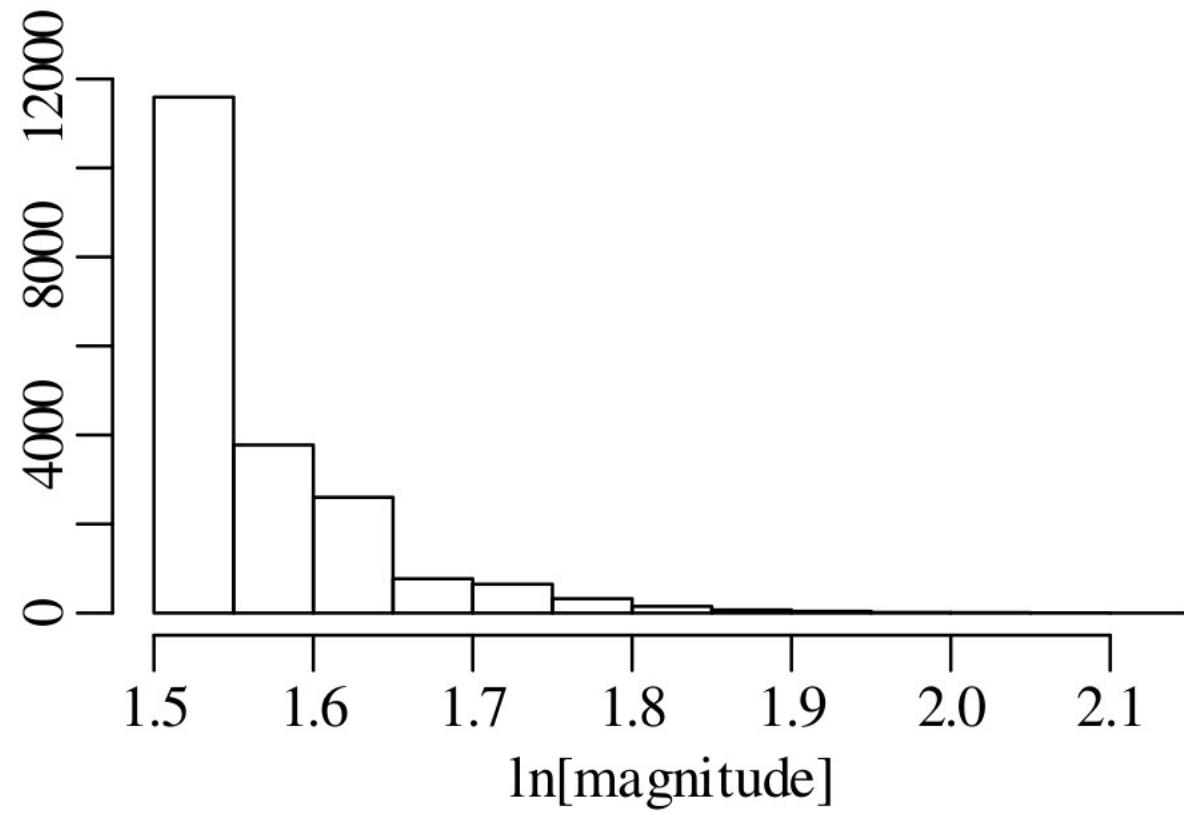
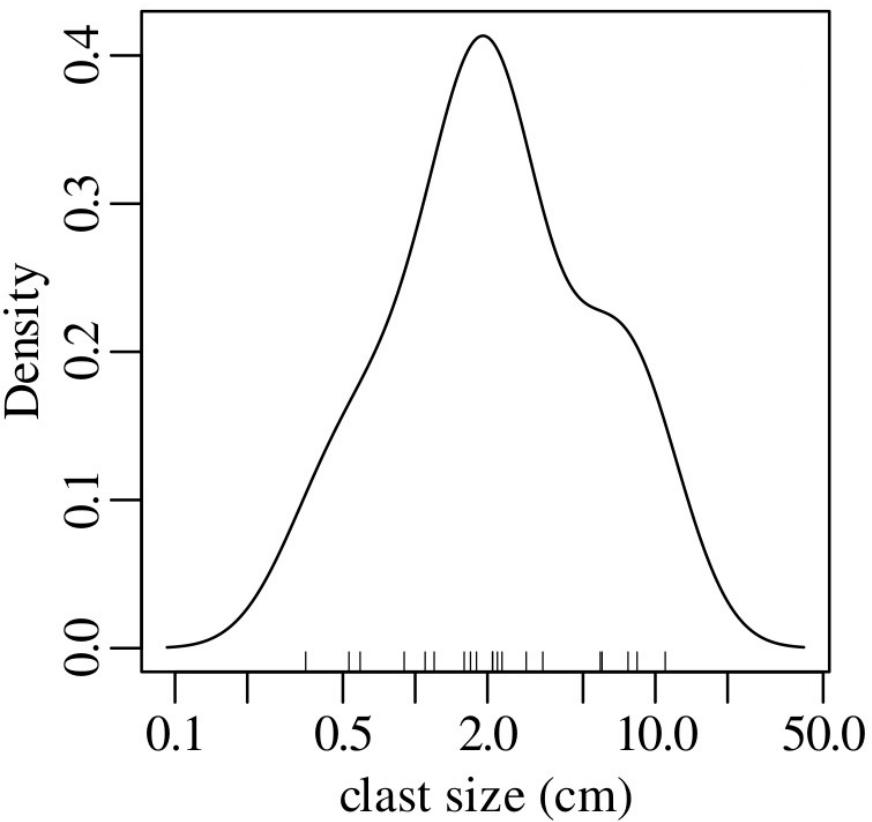
$$\rho_{\frac{y}{z}, \frac{x}{z}} \approx \frac{\left[ \frac{\sigma_z}{\mu_z} \right]^2}{\sqrt{\left[ \frac{\sigma_y}{\mu_y} \right]^2 + \left[ \frac{\sigma_z}{\mu_z} \right]^2} \sqrt{\left[ \frac{\sigma_x}{\mu_x} \right]^2 + \left[ \frac{\sigma_z}{\mu_z} \right]^2}}$$

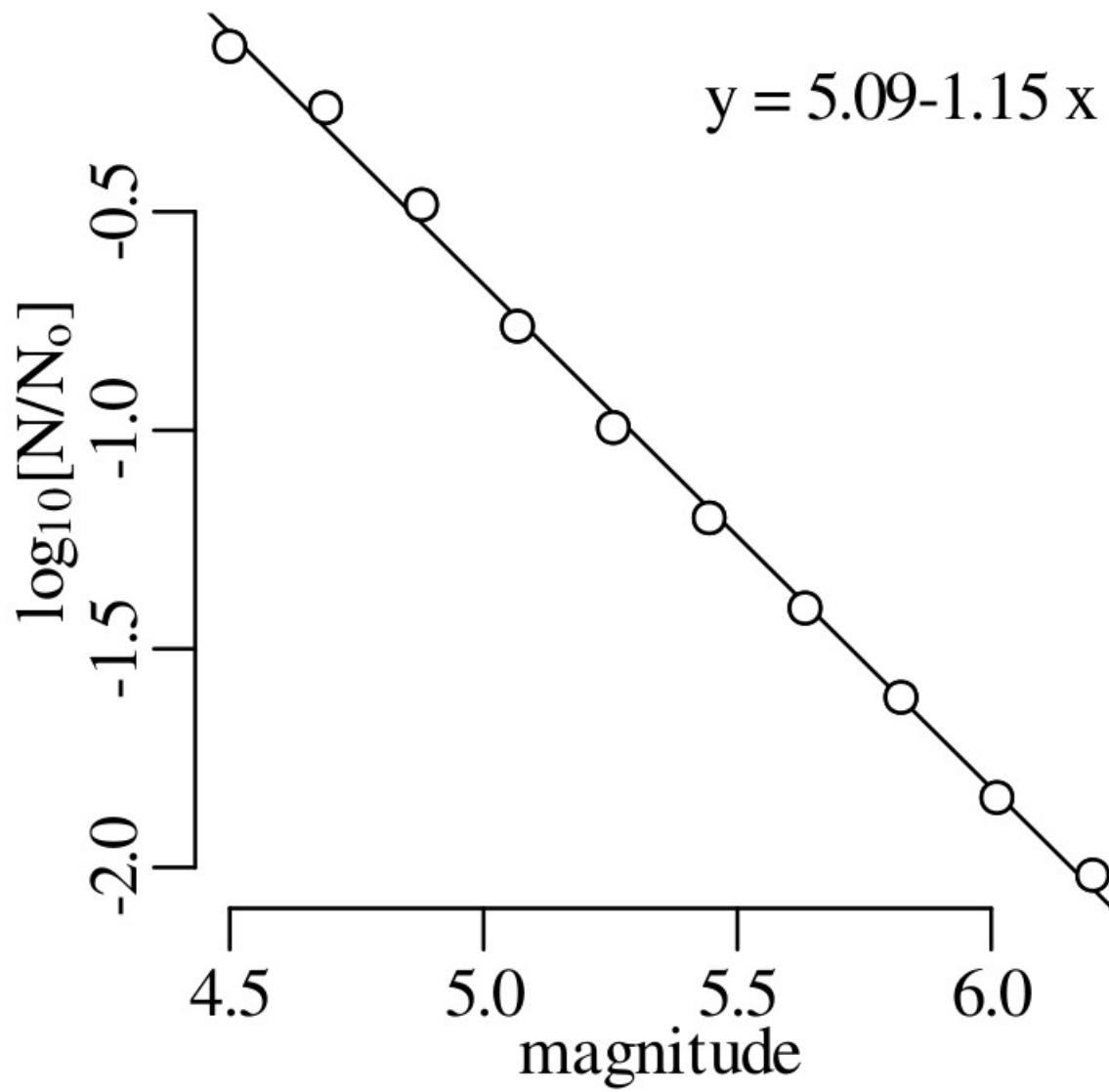
# Statistics for geoscientists

Fractals and chaos

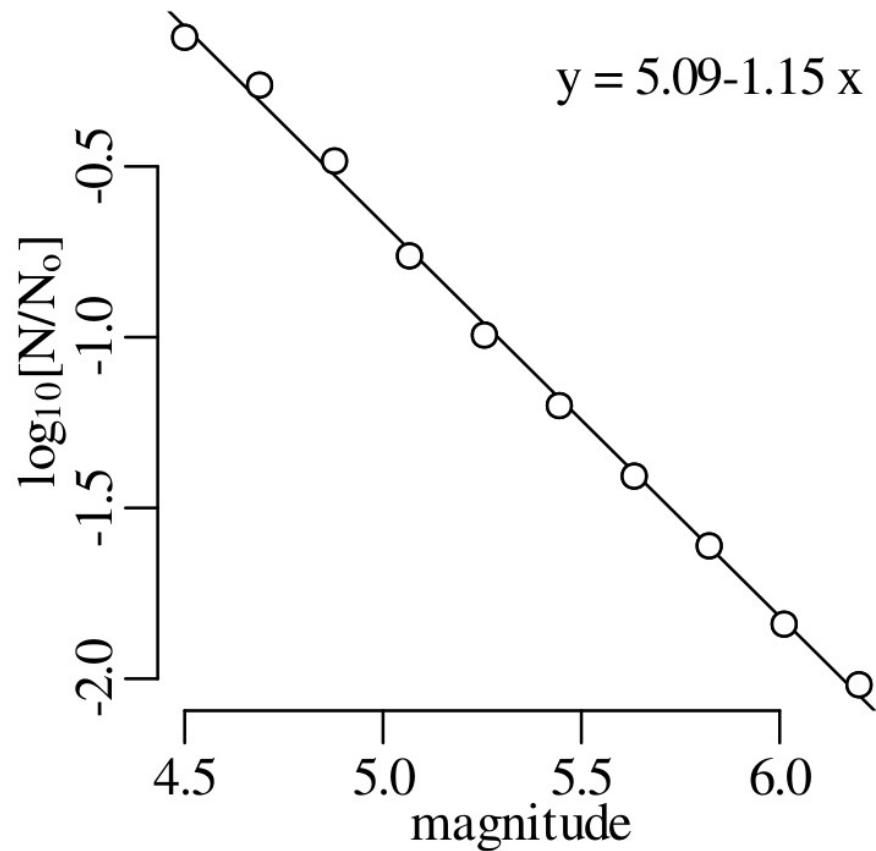








$P(\geq 9.0 \text{ earthquake})?$

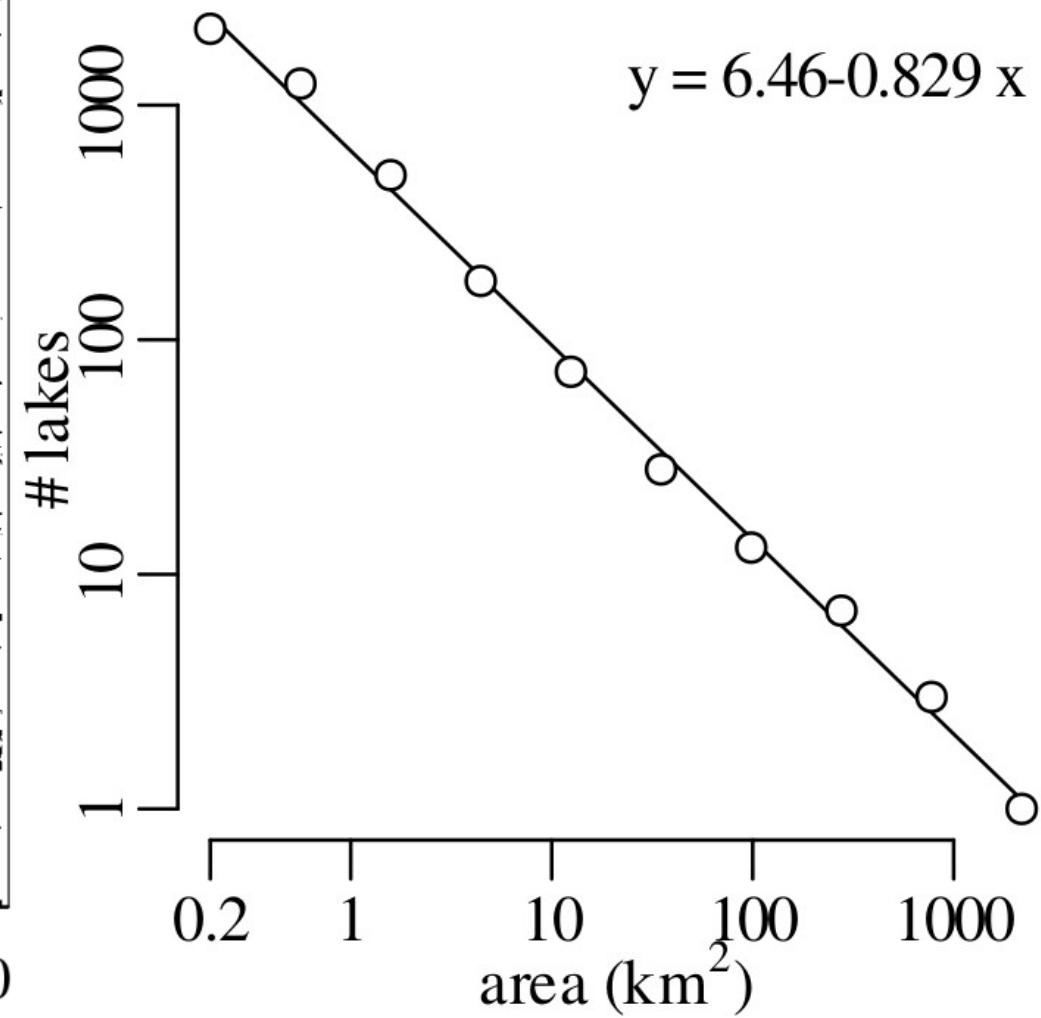
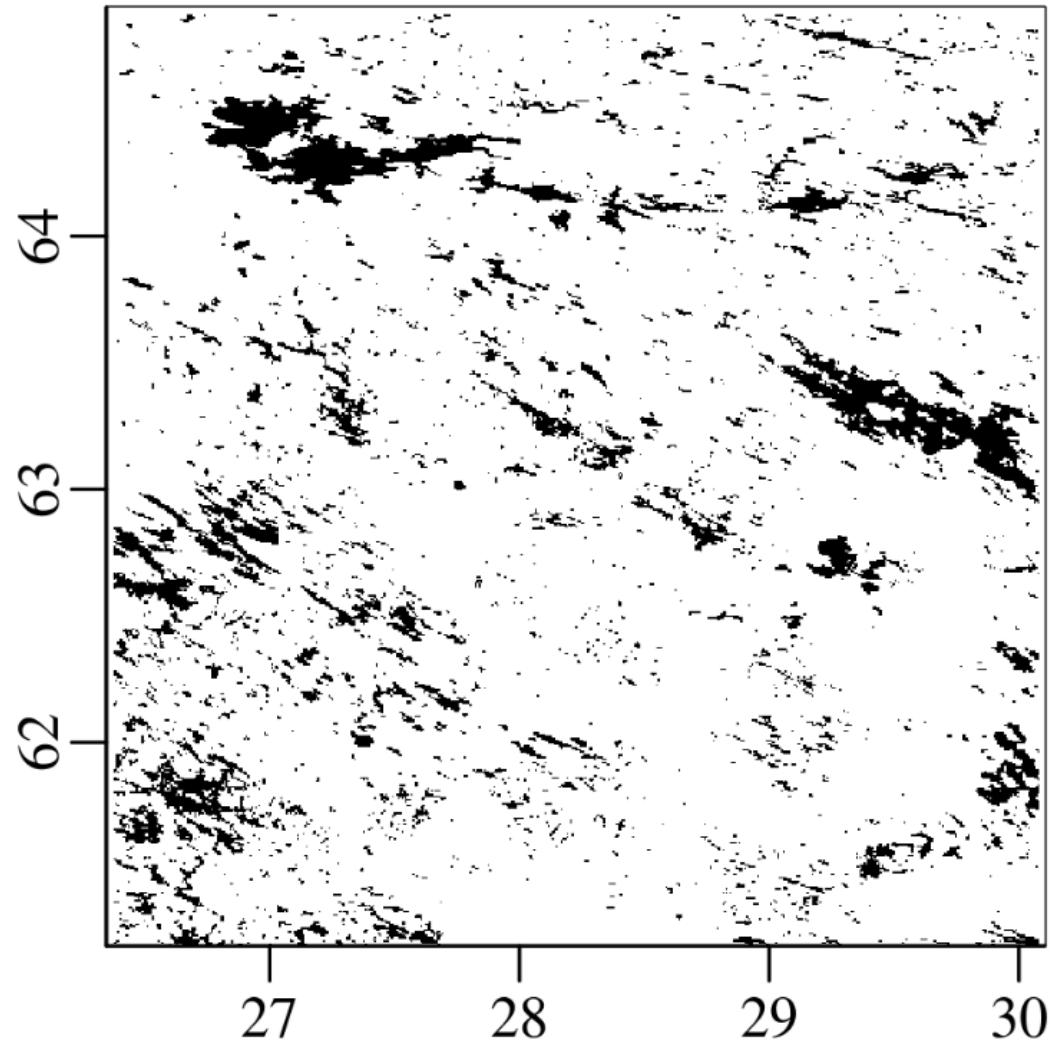


$$\log_{10}[N/N_o] = a + b \text{ magnitude}$$

$$N_o = \frac{20000 \text{ earthquakes}}{1031 \text{ days}} \times 365 \text{ days}$$
$$= 7080 \text{ earthquakes}$$

$$\log_{10}[N/N_o] = 5.09 - 1.15 \times 9.0 = -5.26$$

$$N = 7080 \times 10^{-5.26} = 0.039$$
$$= 3.9\%$$

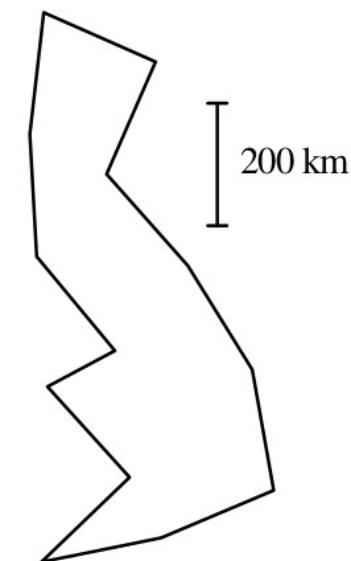


# Reports

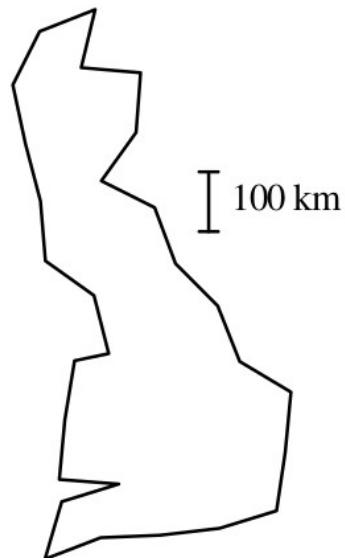
## How Long Is the Coast of Britain? Statistical Self-Similarity and Fractional Dimension

BENOIT MANDELBROT

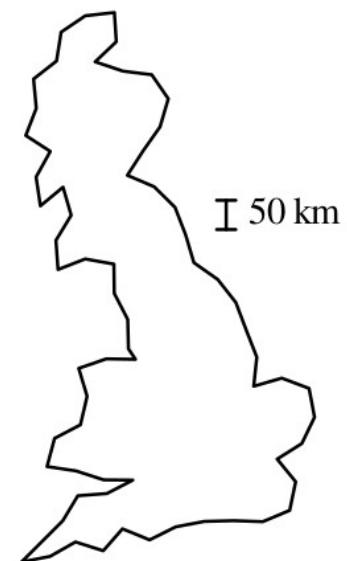
*International Business Machines,  
Thomas J. Watson Research Center,  
Yorktown Heights, New York 10598*



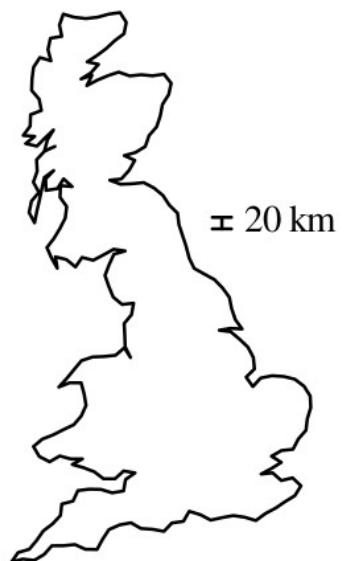
length = 2800km



length = 2900km



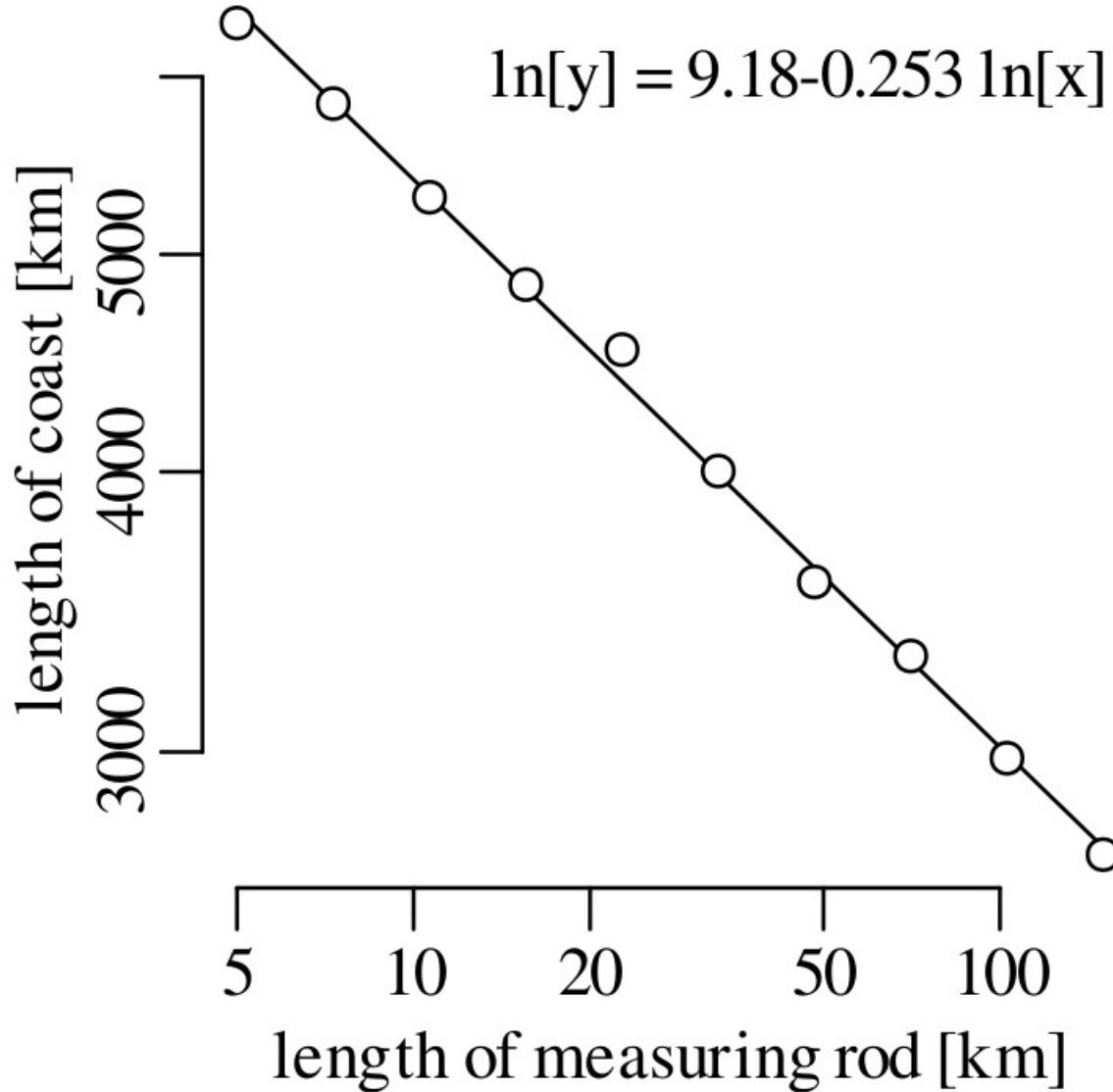
length = 3400km

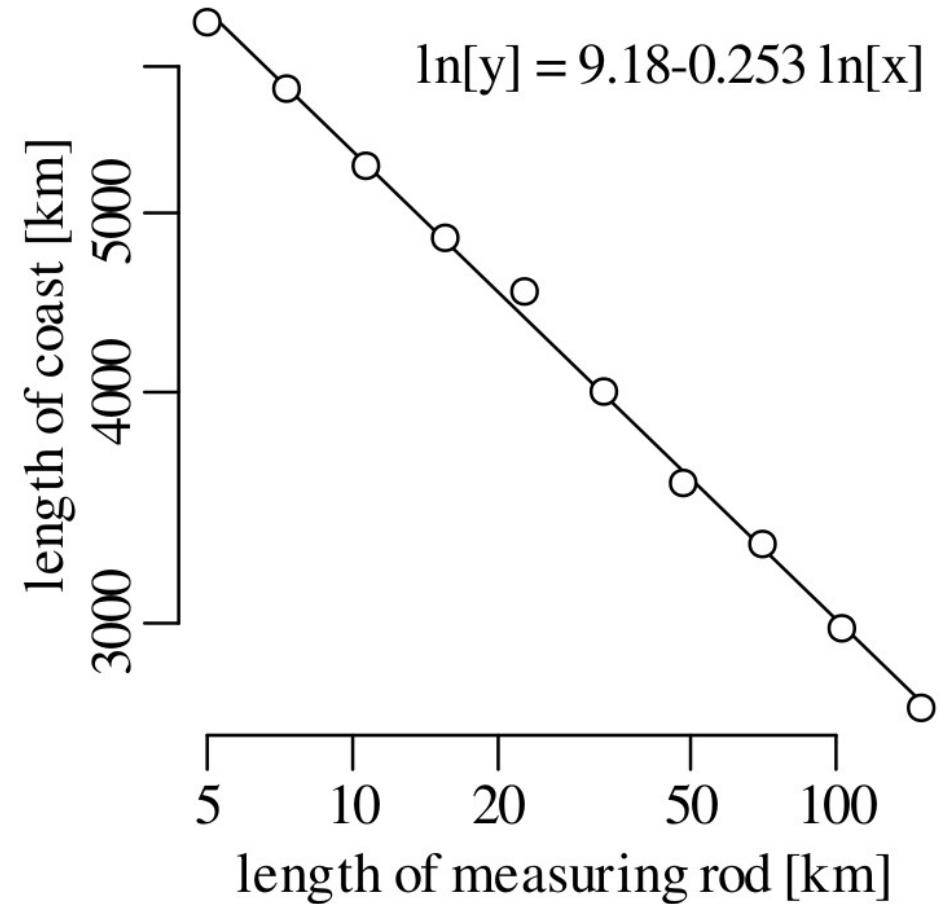


length = 4580km



length = 5490km

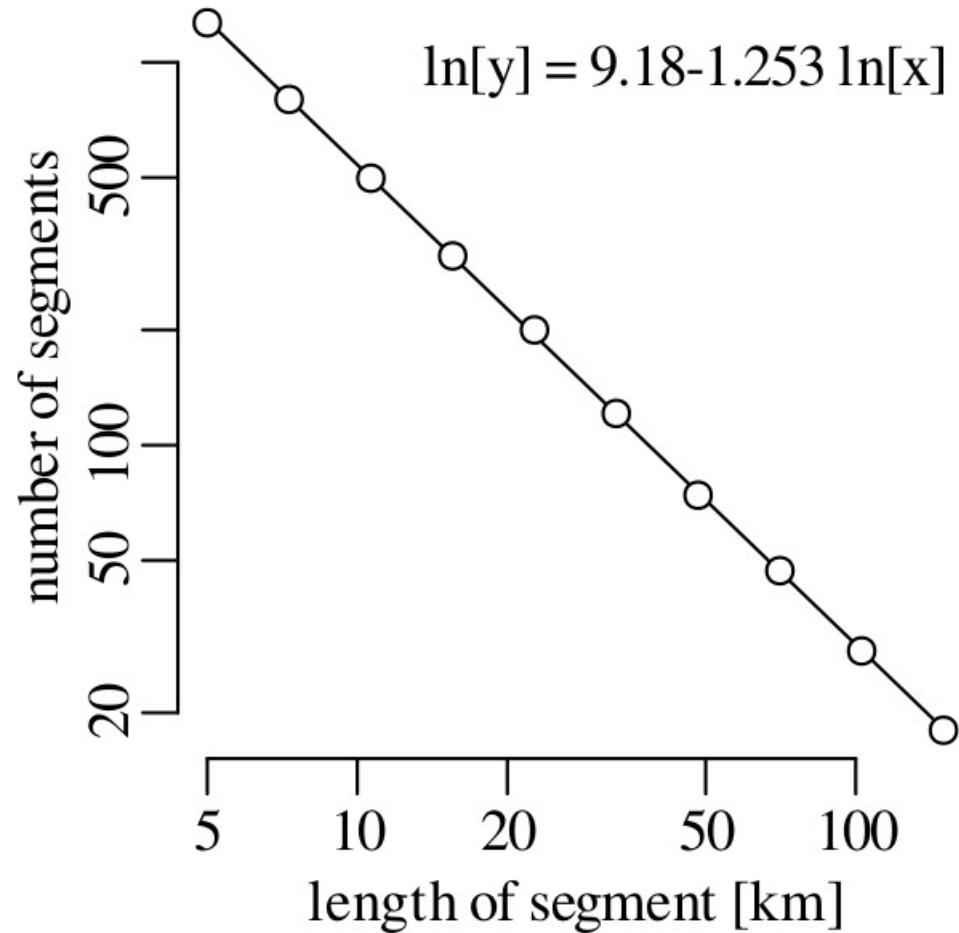
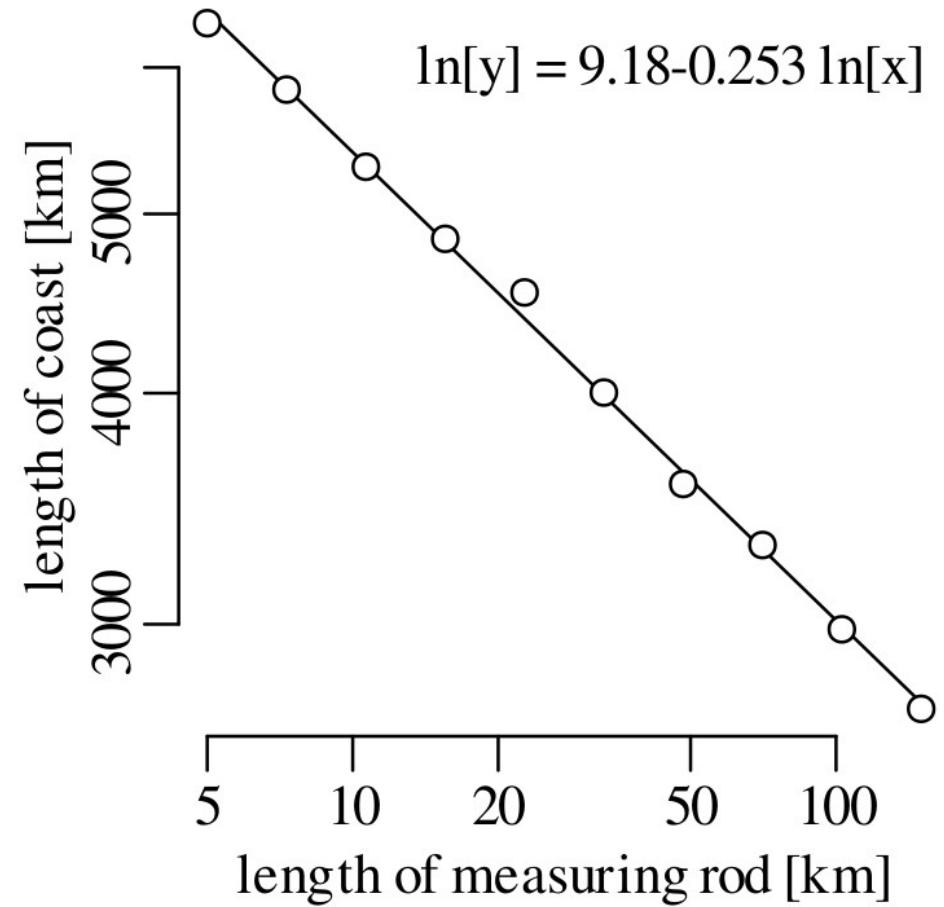


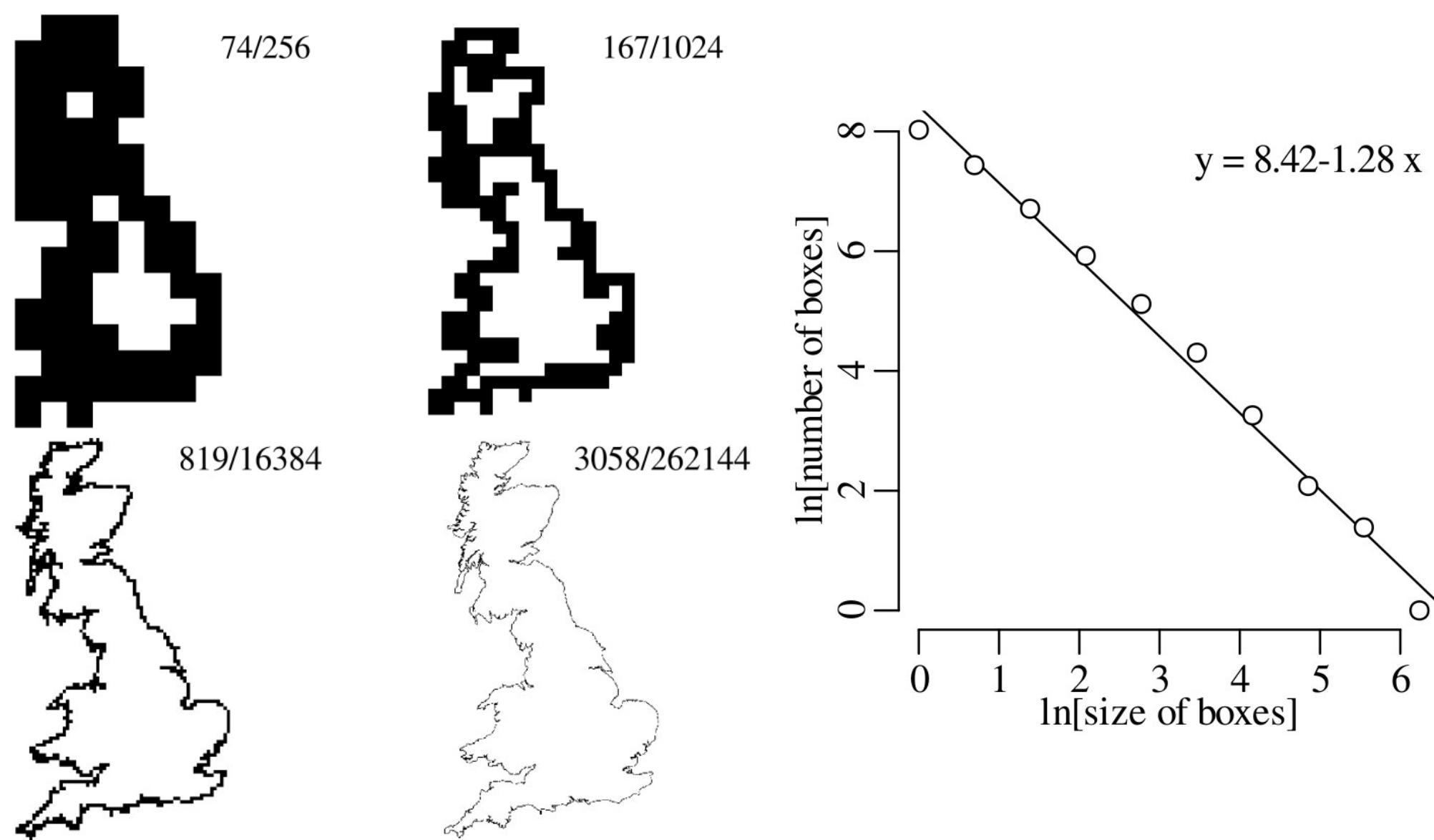


30 cm long ruler



$$\begin{aligned} & \exp(9.18 - 0.253 \ln[3 \times 10^{-4}]) \\ &= 42,060 \text{ km!} \end{aligned}$$

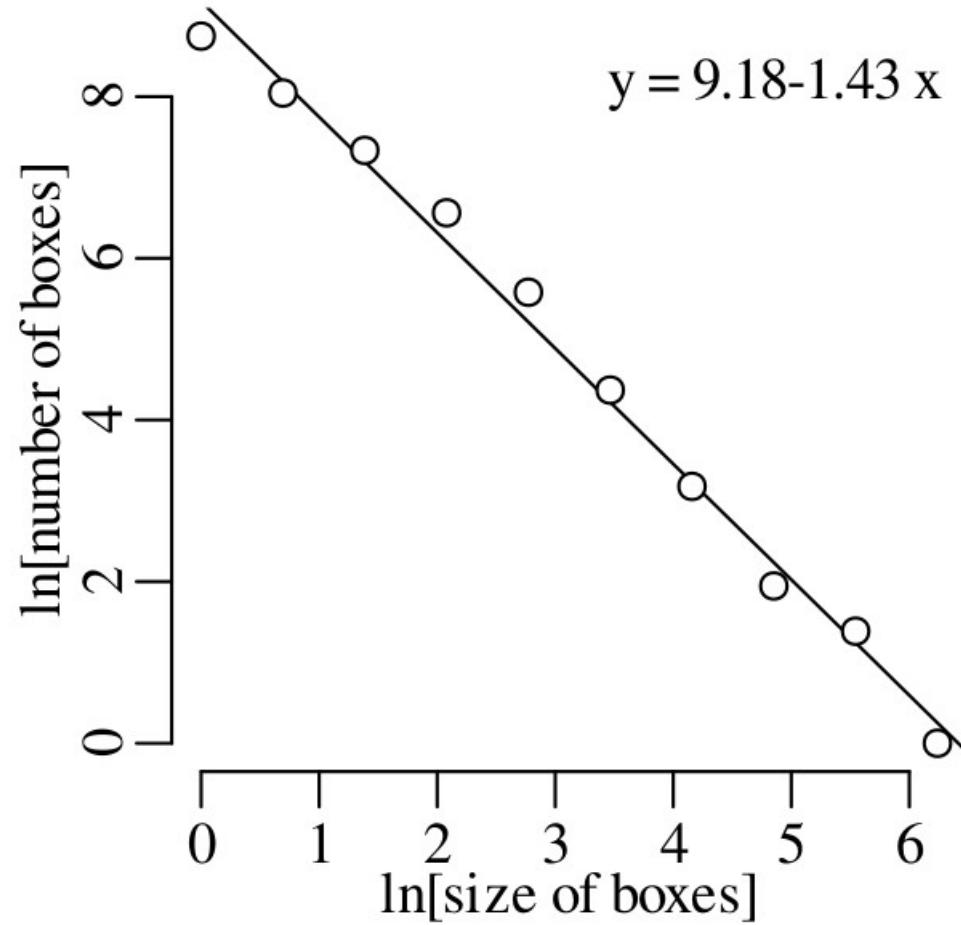




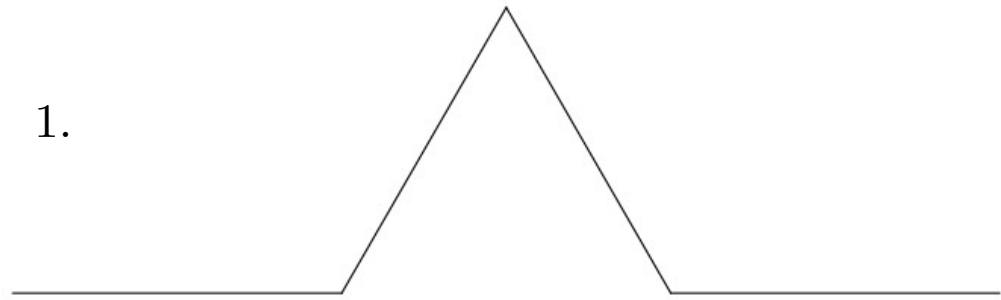
708/4096



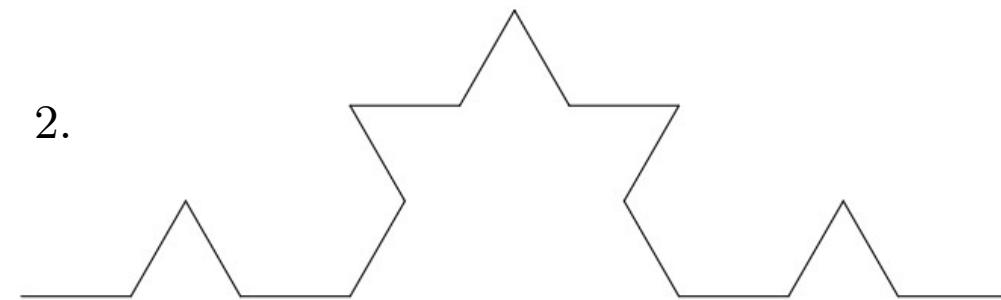
1532/16384



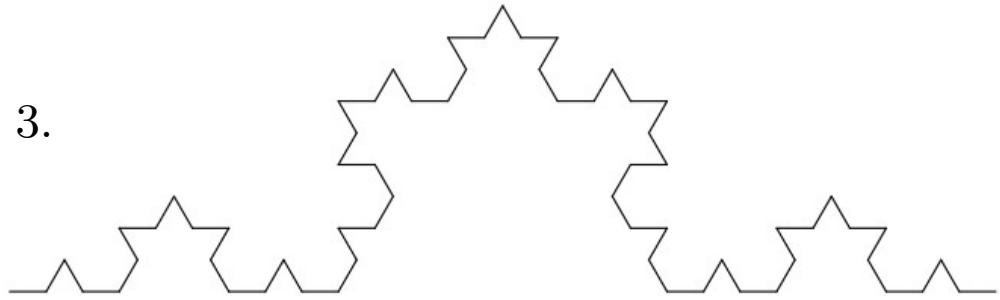
1.



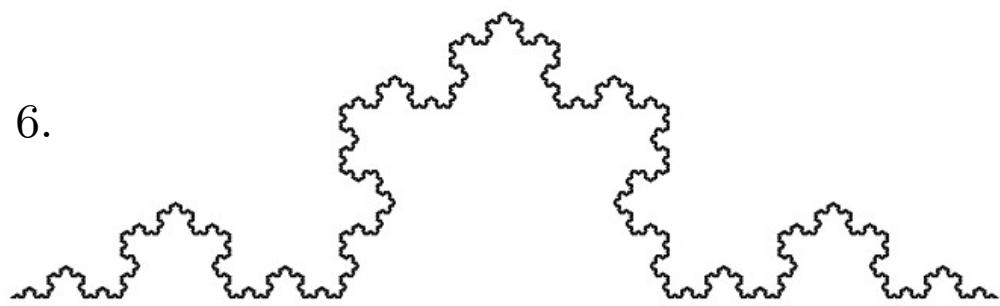
2.

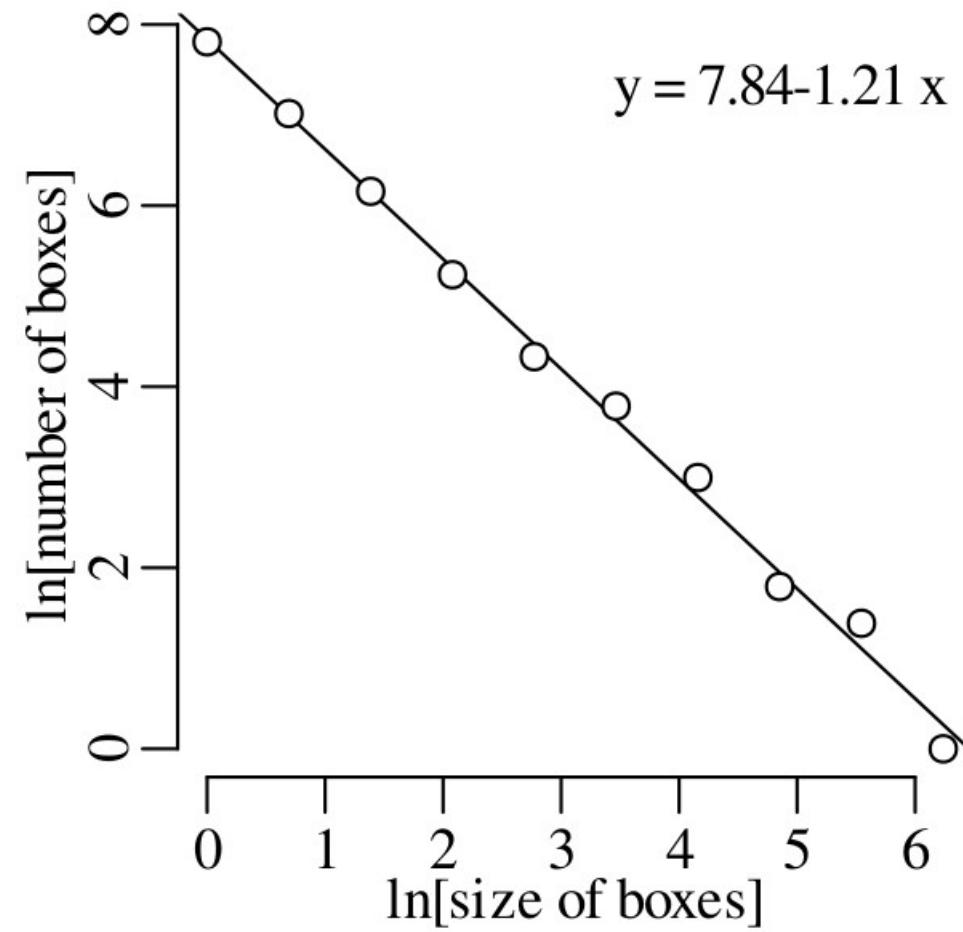
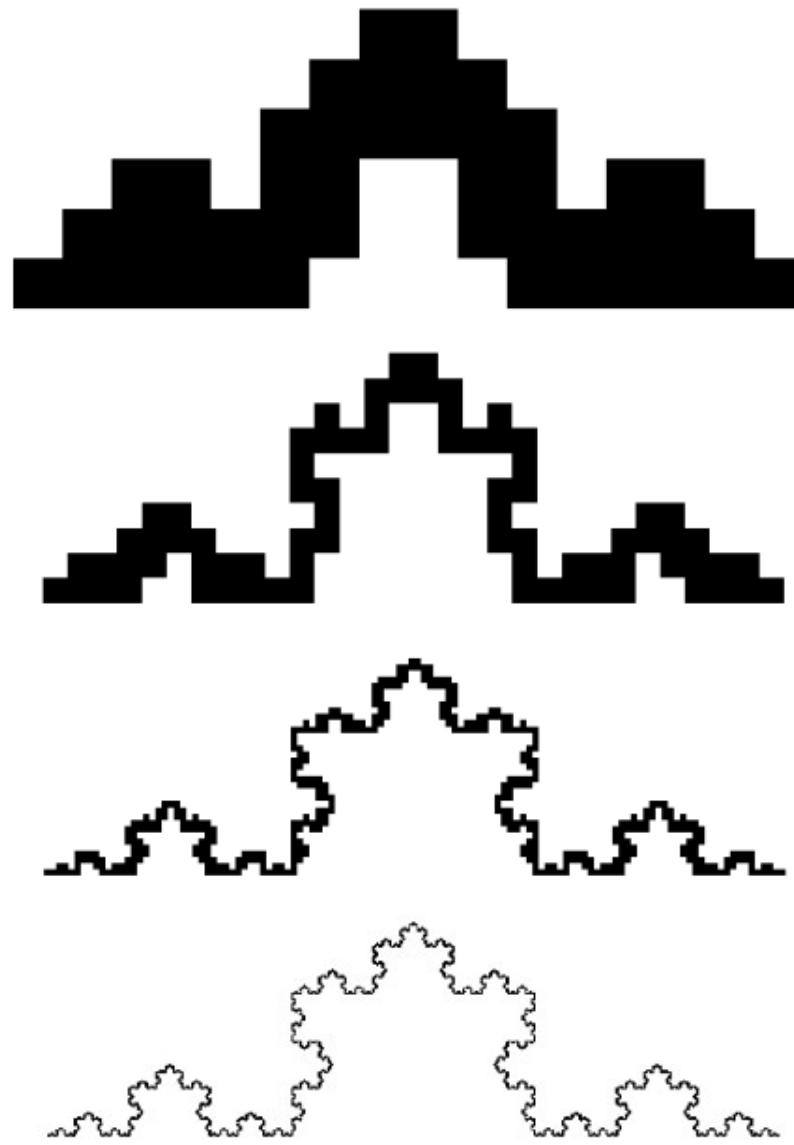


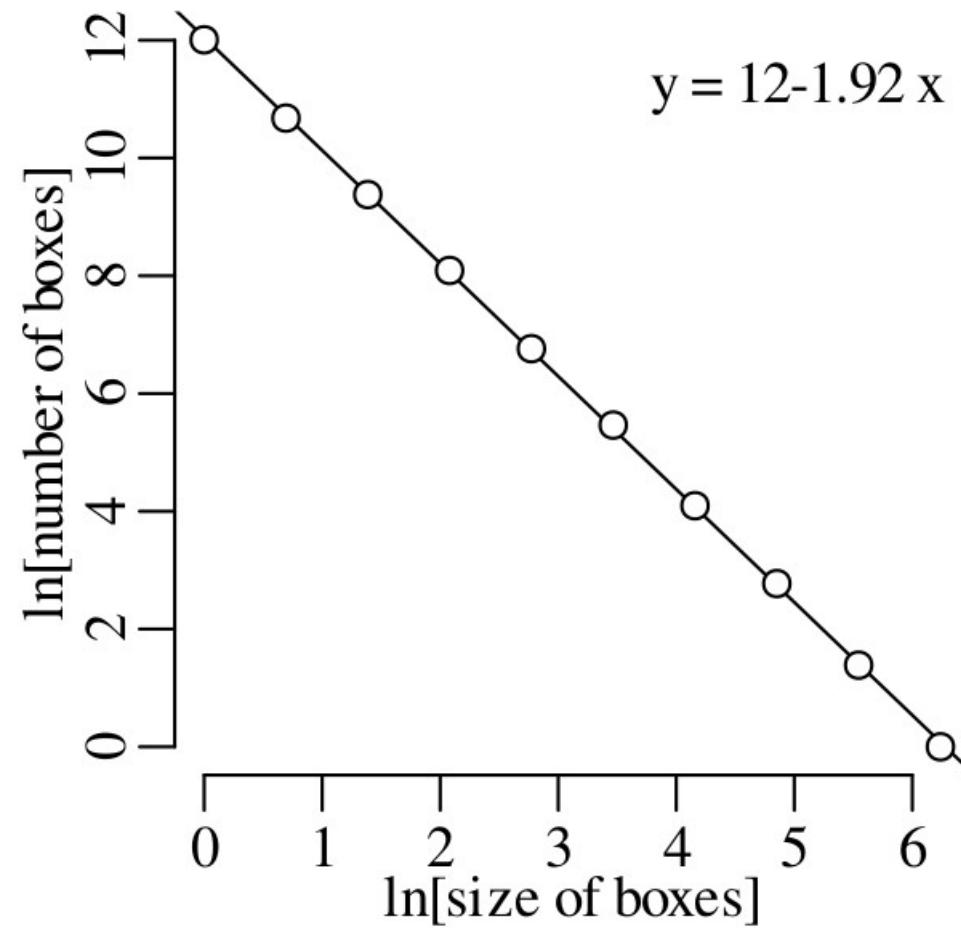
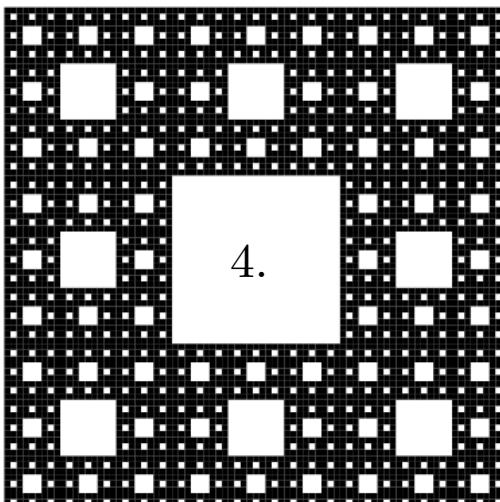
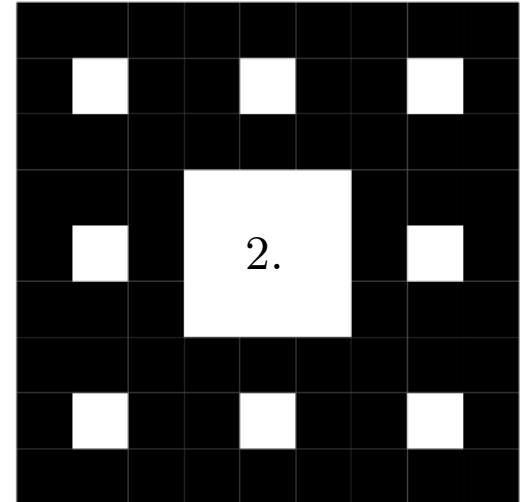
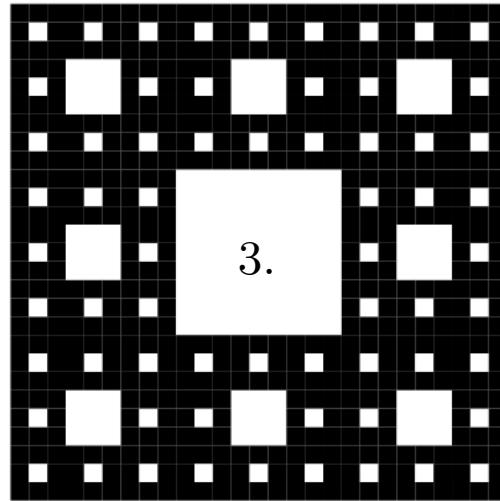
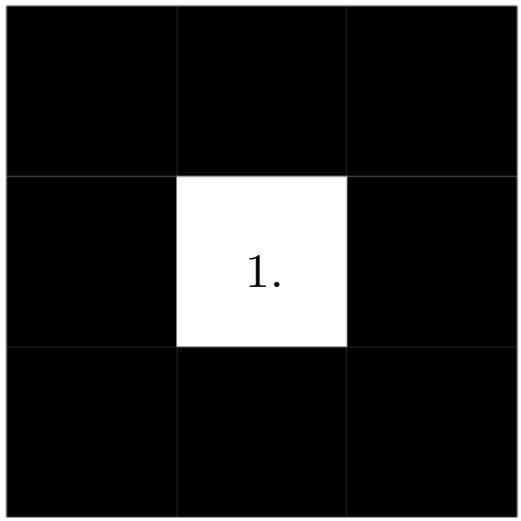
3.



6.



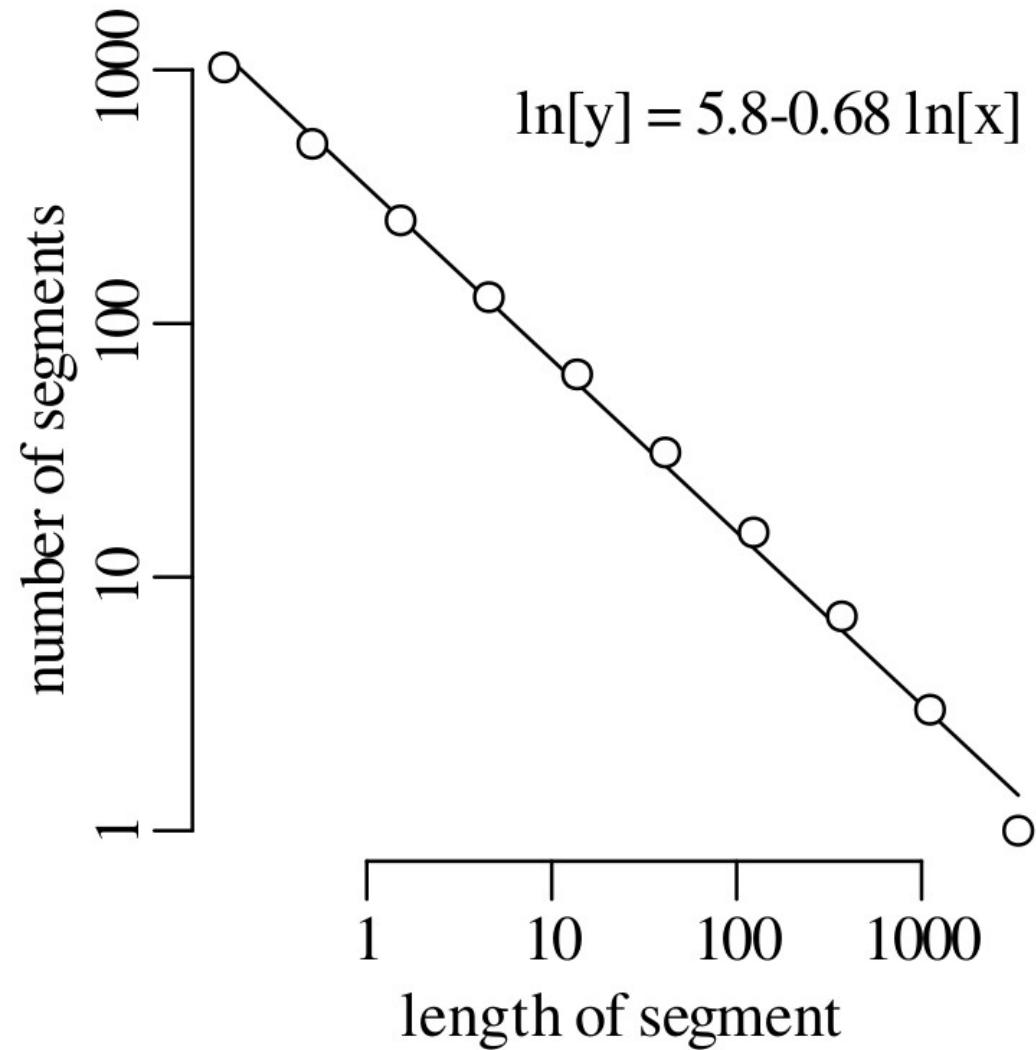




$$\ln[\text{number of boxes}] = C + D \times \ln[\text{size of boxes}]$$



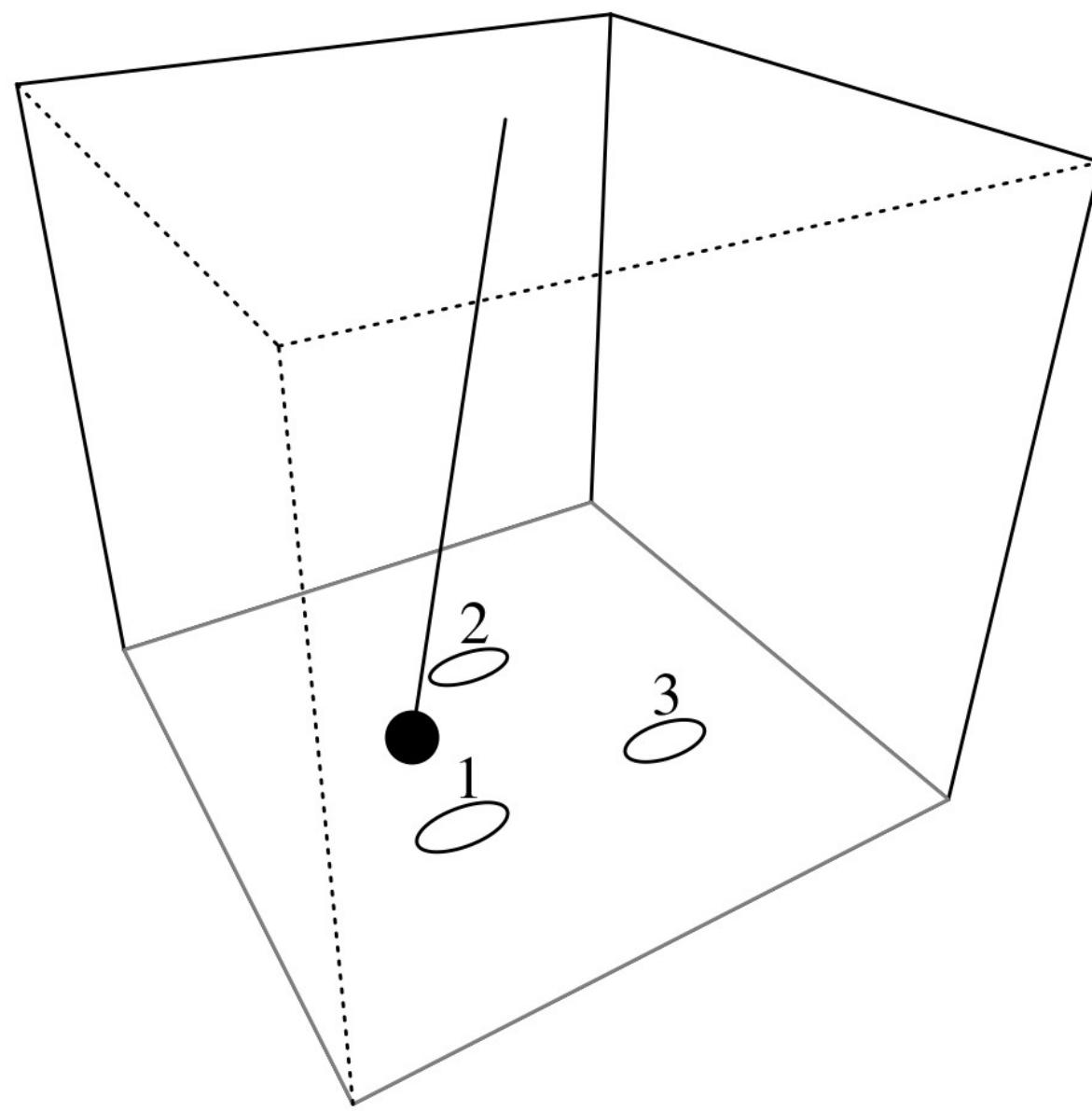
**fractal dimension**

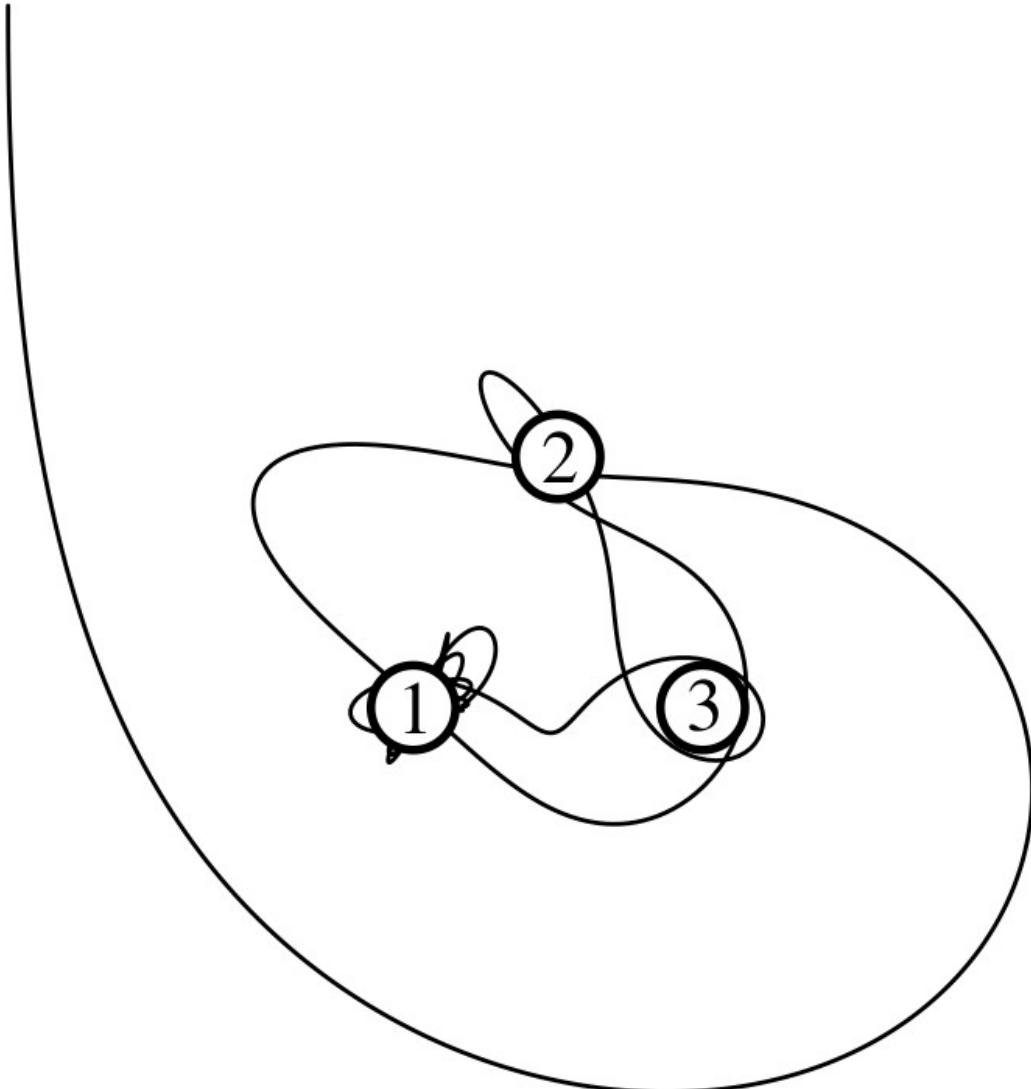
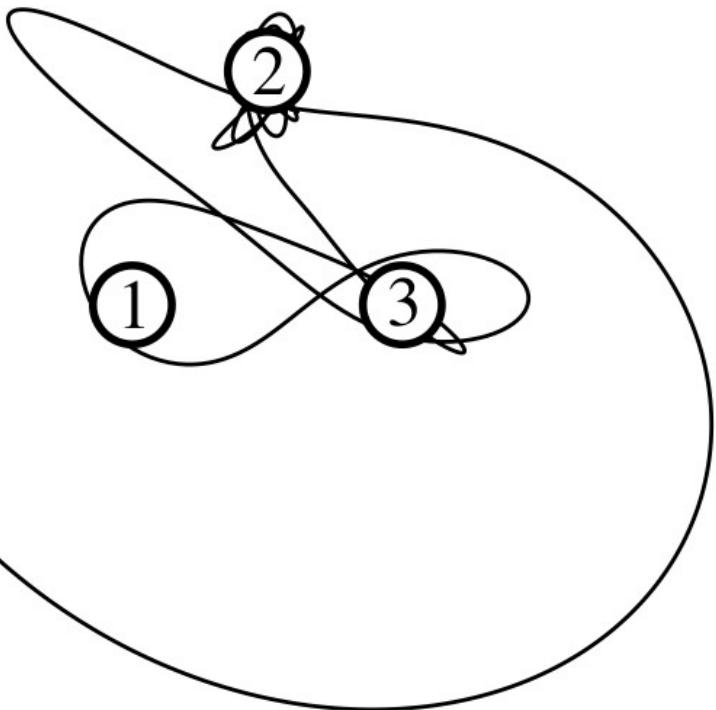


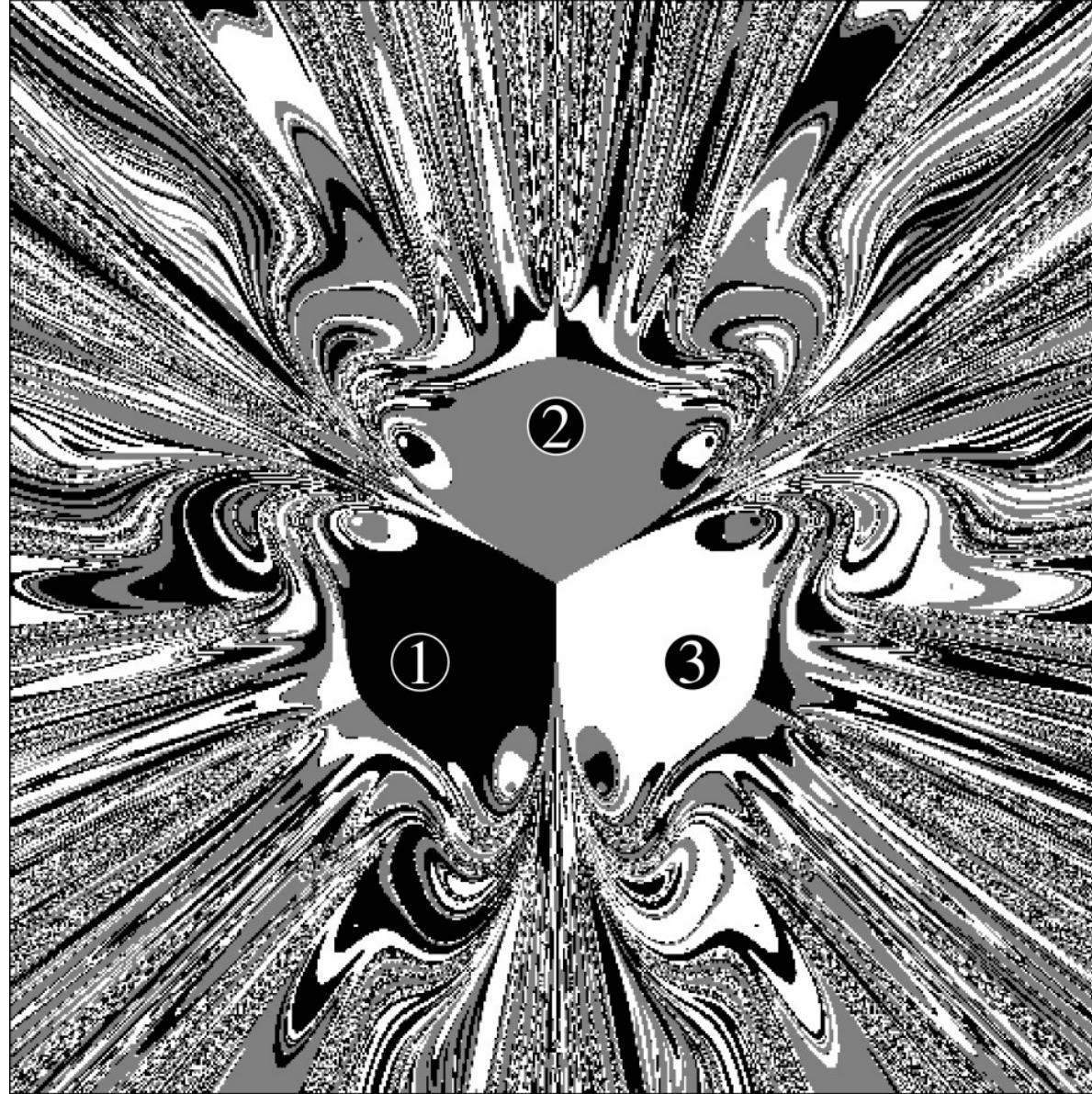
# Chaos

$$F_m(i) \propto 1/|d(i)|^2$$

$$F_f(i) \propto v$$





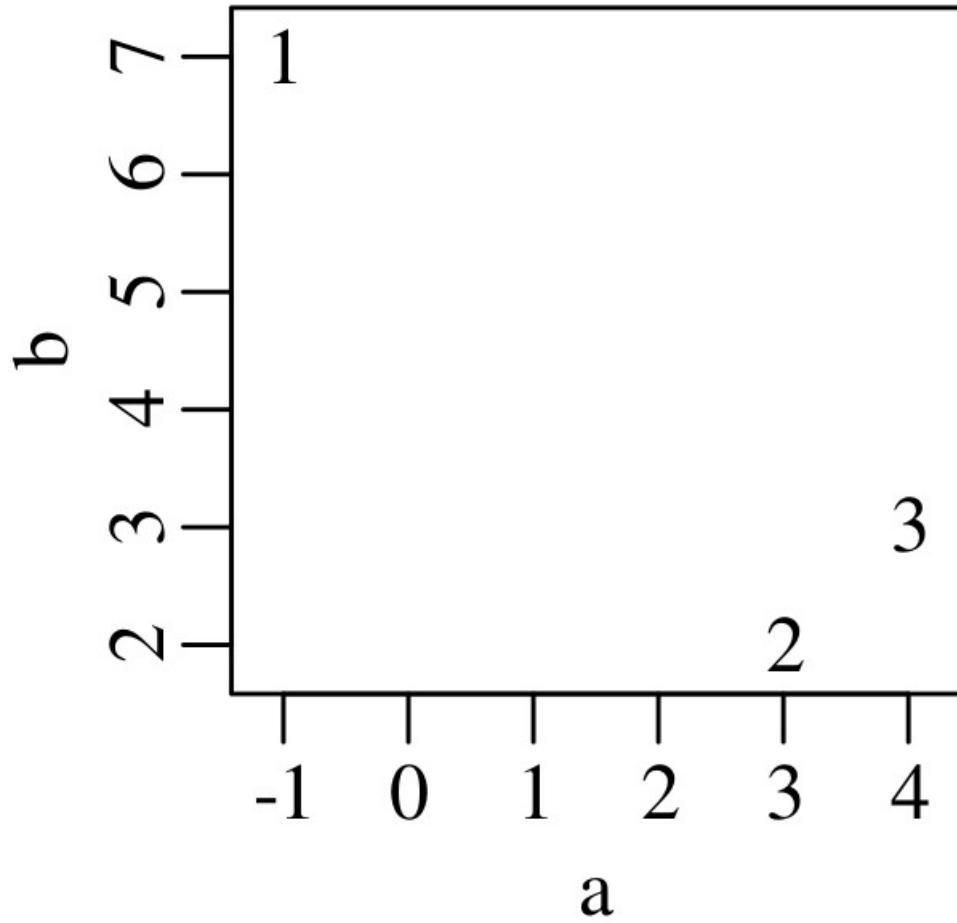


# Statistics for geoscientists

Unsupervised learning

# Principal Component Analysis

$$X = \begin{matrix} & a & b \\ 1 & \begin{bmatrix} -1 & 7 \\ 3 & 2 \end{bmatrix} \\ 2 & \\ 3 & \end{matrix}$$



$$X = \begin{matrix} & \begin{matrix} a & b \end{matrix} \\ \begin{matrix} 1 \\ 2 \\ 3 \end{matrix} & \begin{bmatrix} -1 & 7 \\ 3 & 2 \\ 4 & 3 \end{bmatrix} \end{matrix}$$

$$X = 1_{3,1} \ C + S \ V \ D$$

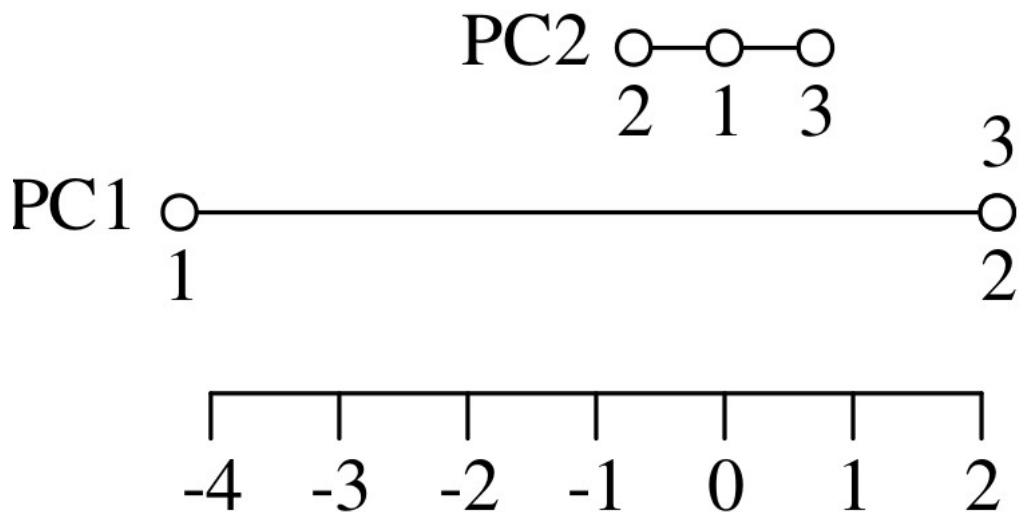
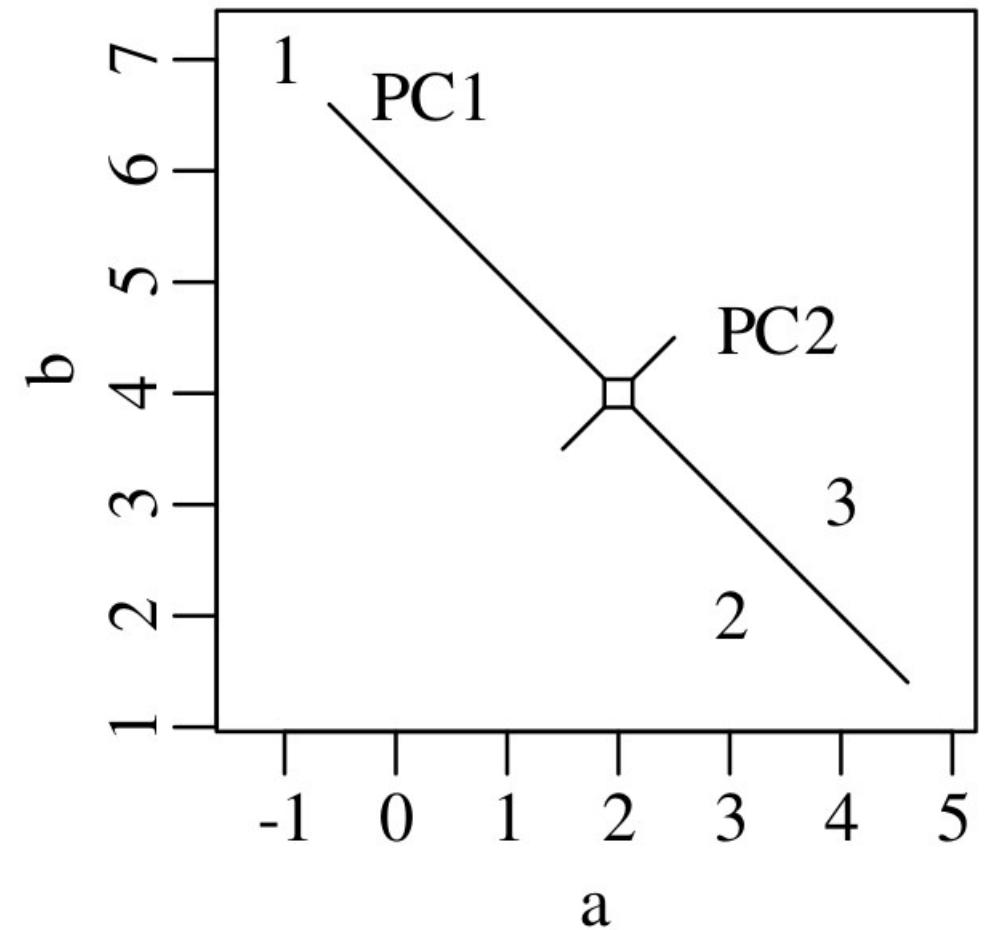
$$= \begin{bmatrix} 1 \\ 1 \\ 1 \end{bmatrix} \begin{bmatrix} 2 & 4 \end{bmatrix} + \begin{bmatrix} -1.15 & 0 \\ 0.58 & -1 \\ 0.58 & 1 \end{bmatrix} \begin{bmatrix} 3.67 & 0 \\ 0 & 0.71 \end{bmatrix} \begin{bmatrix} 0.71 & -0.71 \\ 0.71 & 0.71 \end{bmatrix}$$

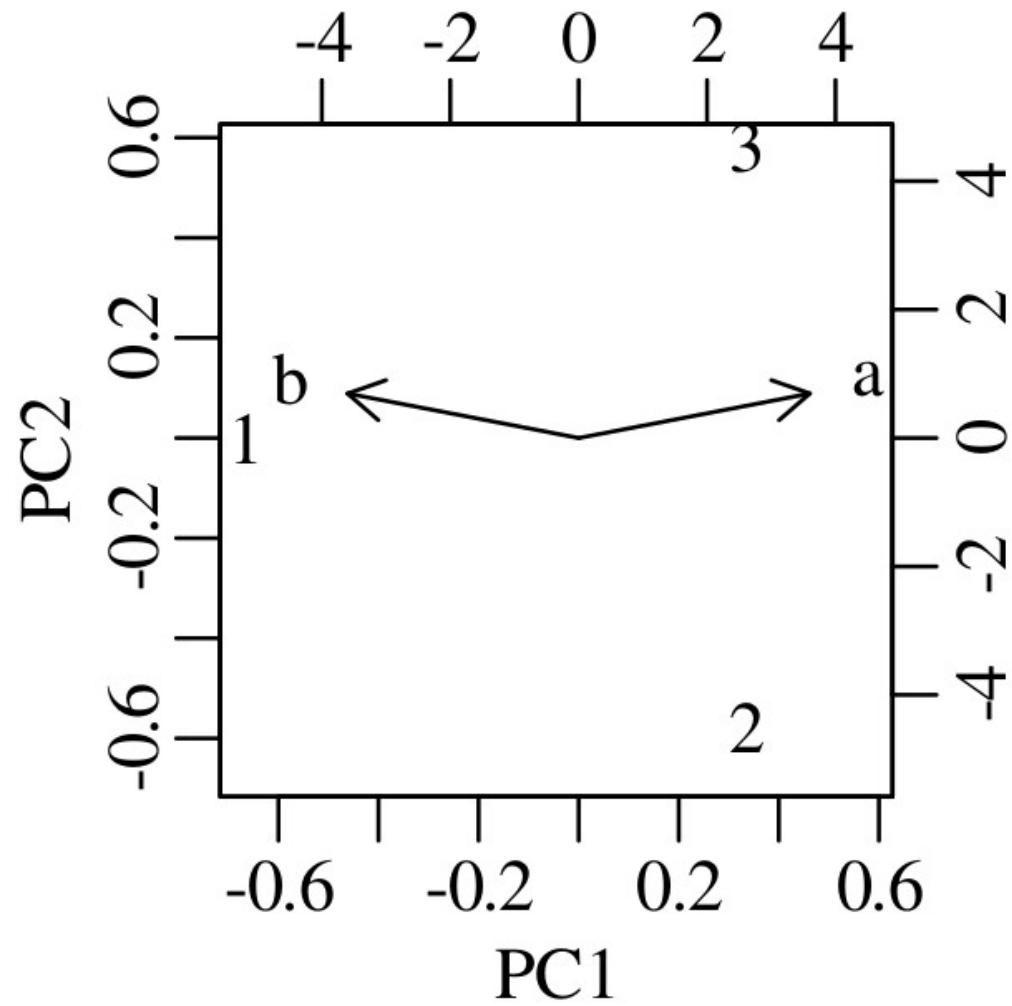
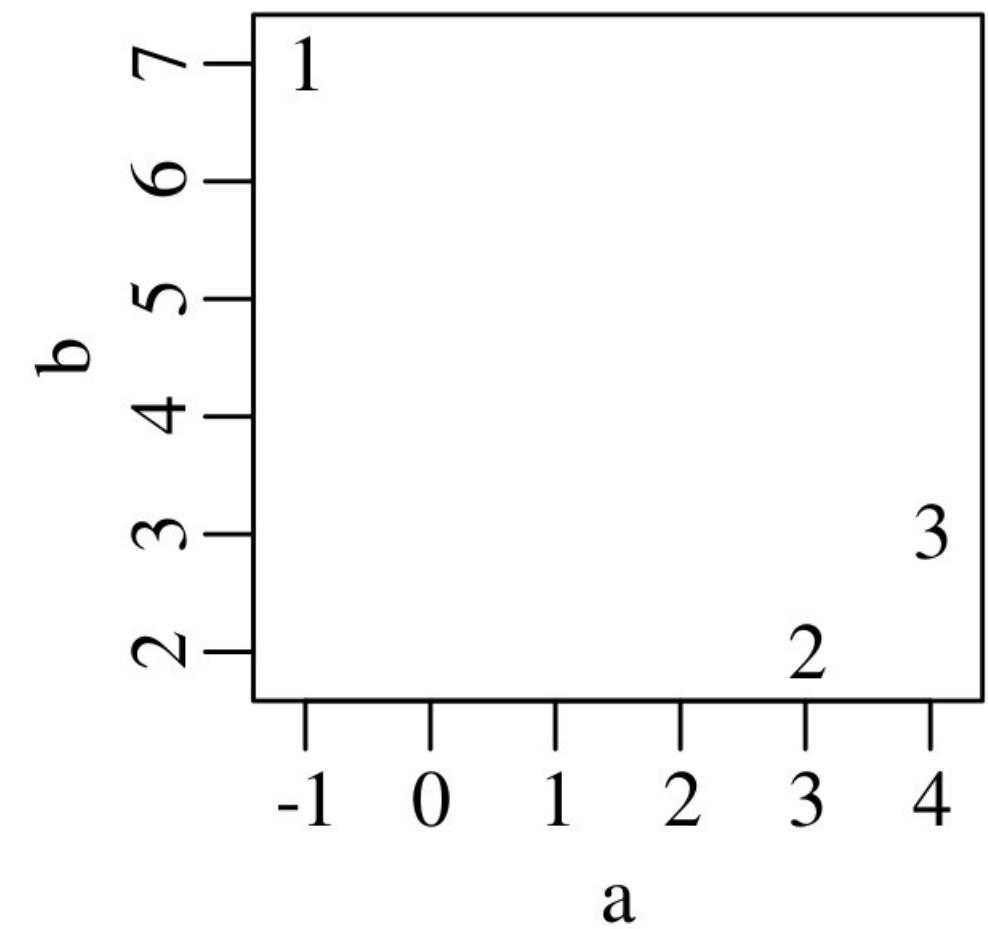
$$X = 1_{3,1} \ C + S \ V \ D$$

$$= \begin{bmatrix} 1 \\ 1 \\ 1 \end{bmatrix} \begin{bmatrix} 2 & 4 \end{bmatrix} + \begin{bmatrix} -1.15 & 0 \\ 0.58 & -1 \\ 0.58 & 1 \end{bmatrix} \begin{bmatrix} 3.67 & 0 \\ 0 & 0.71 \end{bmatrix} \begin{bmatrix} 0.71 & -0.71 \\ 0.71 & 0.71 \end{bmatrix}$$

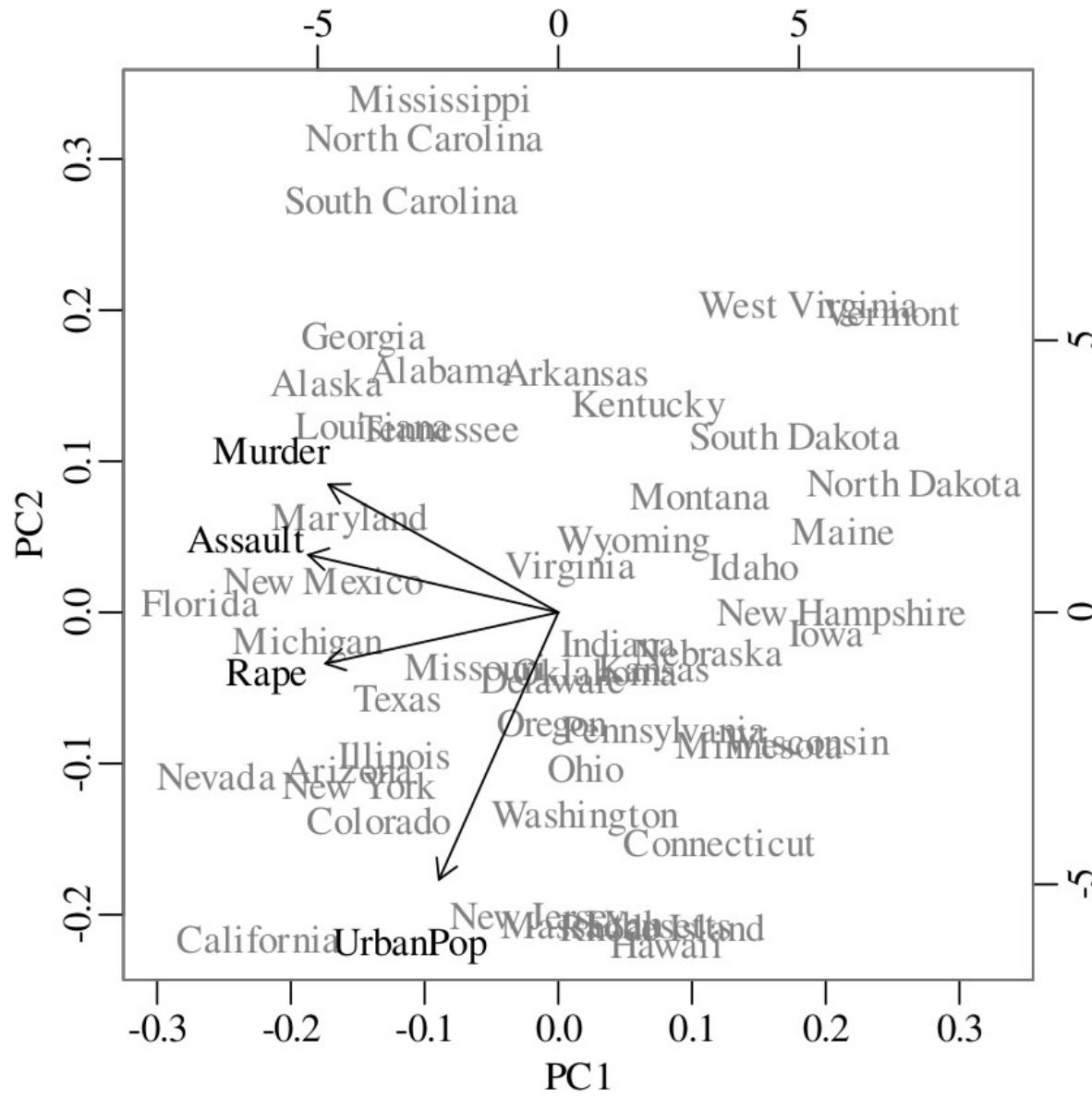
$$P = S \ V = \begin{bmatrix} -4.24 & 0 \\ 2.12 & -0.71 \\ 2.12 & 0.71 \end{bmatrix},$$

$$\text{and } L = V \ D = \begin{bmatrix} 2.6 & -2.6 \\ 0.5 & 0.5 \end{bmatrix}$$



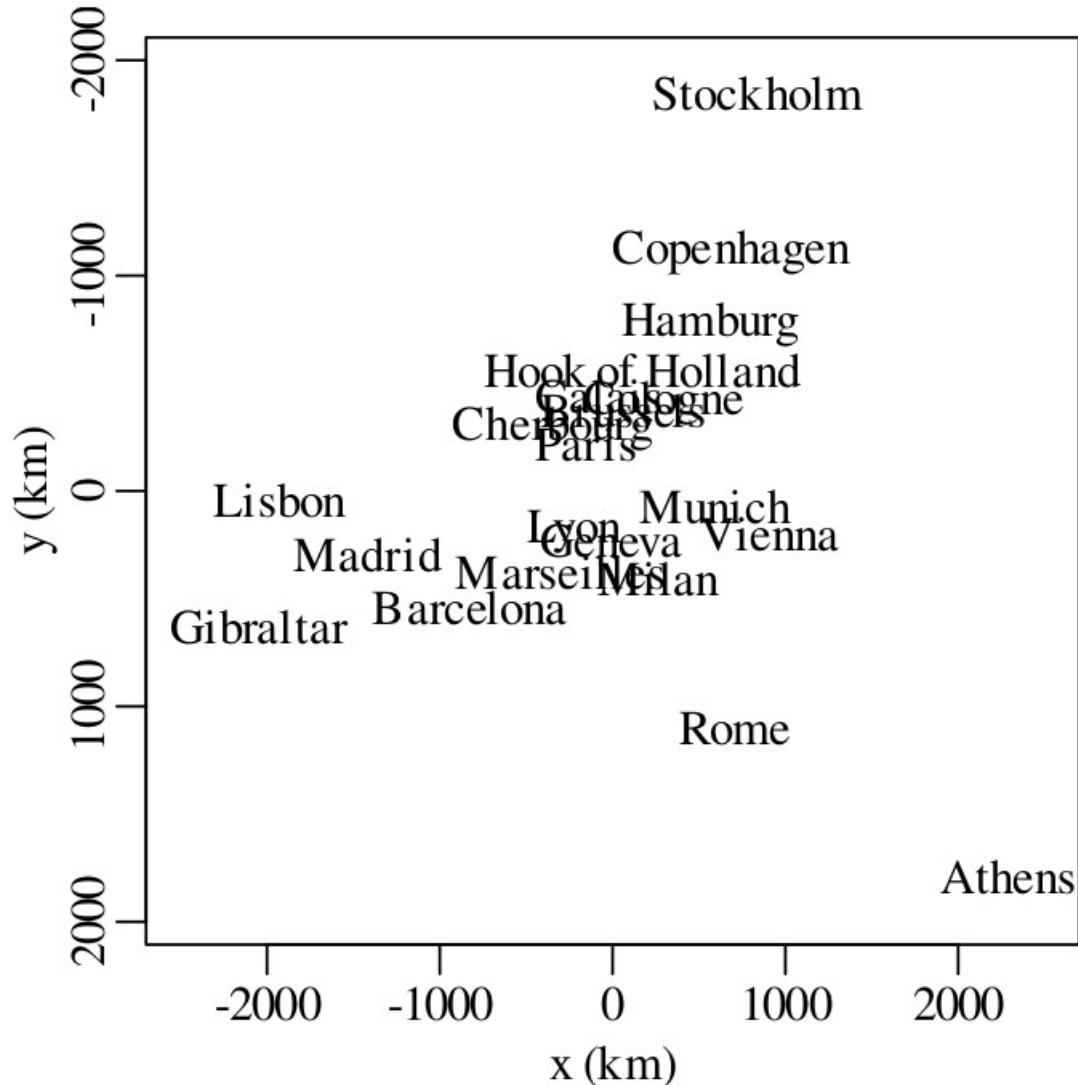


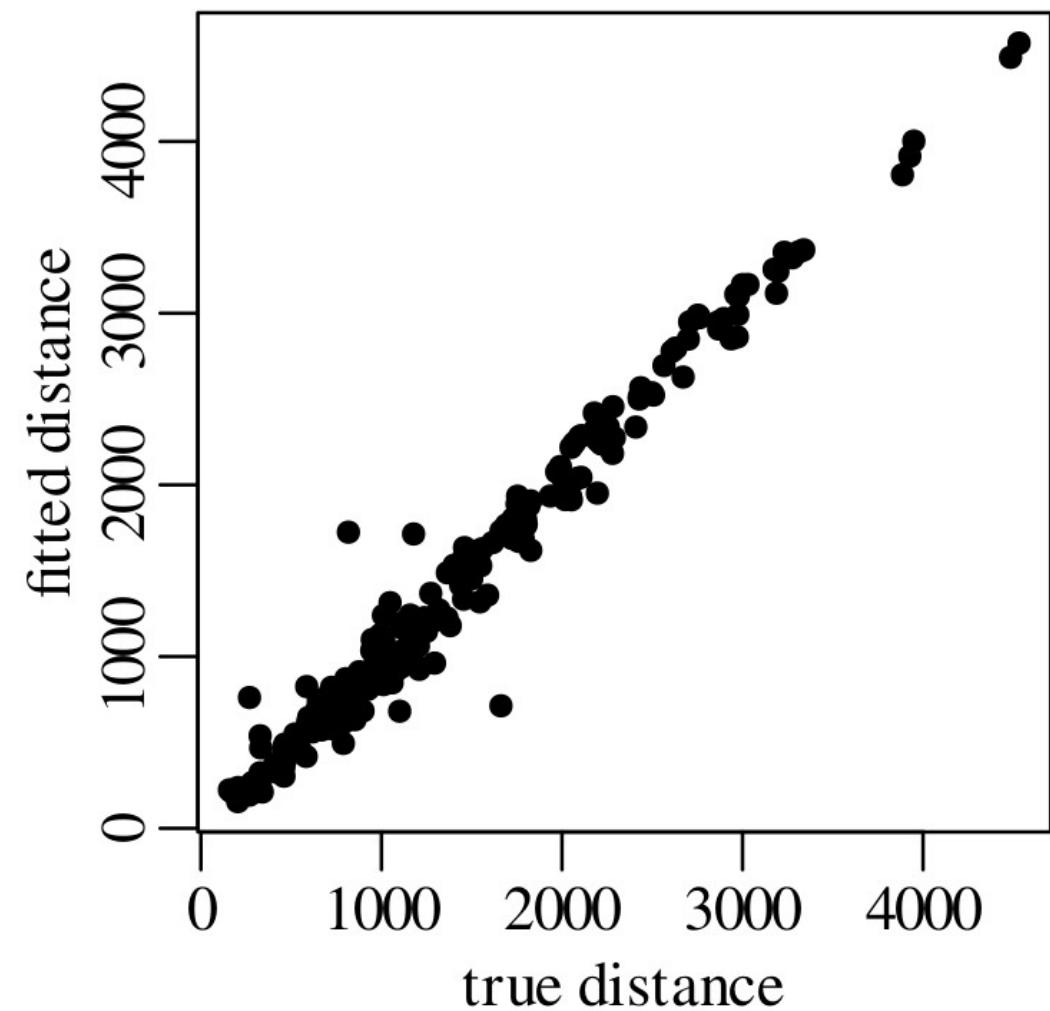
|            | Murder | Assault | Rape | UrbanPop |
|------------|--------|---------|------|----------|
| Alabama    | 13.2   | 236     | 21.2 | 58       |
| Alaska     | 10.0   | 263     | 44.5 | 48       |
| Arizona    | 8.1    | 294     | 31.0 | 80       |
| Arkansas   | 8.8    | 190     | 19.5 | 50       |
| California | 9.0    | 276     | 40.6 | 91       |
| Colorado   | 7.9    | 204     | 38.7 | 78       |
| :          | :      | :       | :    | :        |
| Wisconsin  | 2.6    | 53      | 10.8 | 66       |
| Wyoming    | 6.8    | 161     | 15.6 | 60       |



# Multidimensional Scaling

|           | Athens | Barcelona | Brussels | ... | Rome | Stockholm | Vienna |
|-----------|--------|-----------|----------|-----|------|-----------|--------|
| Athens    | 0      | 3313      | 2963     | ... | 817  | 3927      | 1991   |
| Barcelona | 3313   | 0         | 1326     | ... | 1460 | 2868      | 1802   |
| Brussels  | 2963   | 1318      | 0        | ... | 1511 | 1616      | 1175   |
| :         | :      | :         | :        | ..  | :    | :         | :      |
| Rome      | 817    | 1460      | 1511     | ... | 0    | 2707      | 1209   |
| Stockholm | 3927   | 2868      | 1616     | ... | 2707 | 0         | 2105   |
| Vienna    | 1991   | 1802      | 1175     | ... | 1209 | 2105      | 0      |





$$S = \sqrt{\frac{\sum_{i=1}^n \sum_{j=i+1}^n (f(d[i, j]) - \delta[i, j])^2}{\sum_{i=1}^n \sum_{j=i+1}^n \delta[i, j]^2}}$$

| fit | poor | fair | good | excellent | perfect |
|-----|------|------|------|-----------|---------|
| S   | 0.2  | 0.1  | 0.05 | 0.025     | 0       |

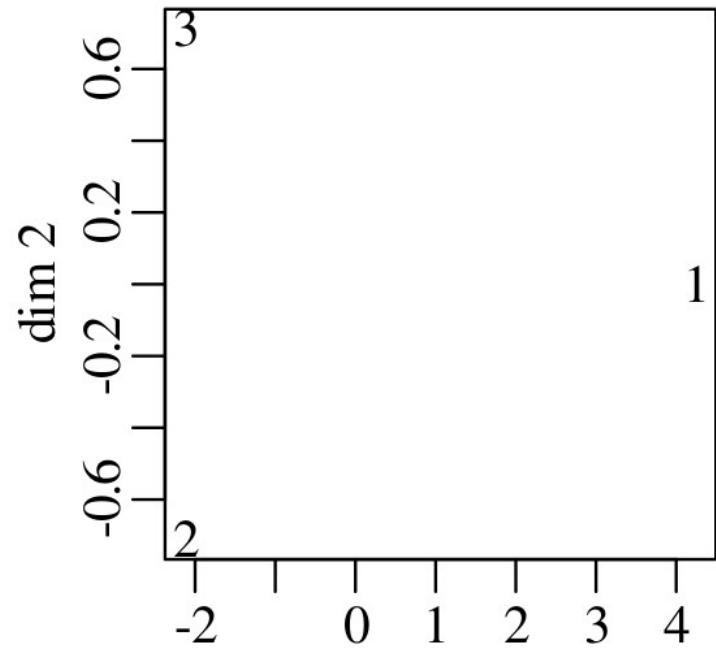
$$X = \begin{matrix} & a & b \\ \begin{matrix} 1 \\ 2 \\ 3 \end{matrix} & \left[ \begin{matrix} -1 & 7 \\ 3 & 2 \\ 4 & 3 \end{matrix} \right] \end{matrix}$$

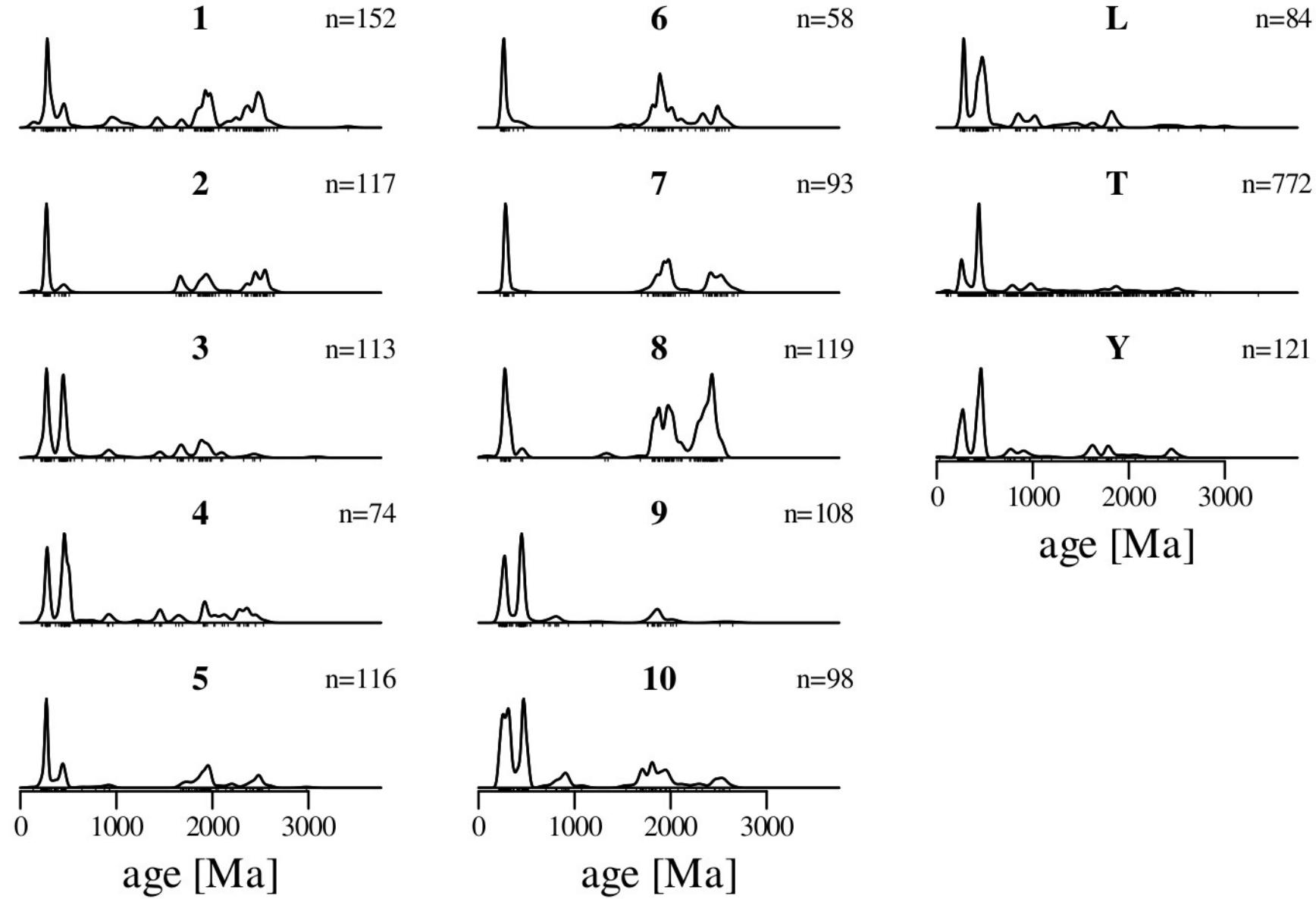
$$d[i, j] = \sqrt{(a[i] - a[j])^2 + (b[i] - b[j])^2}$$



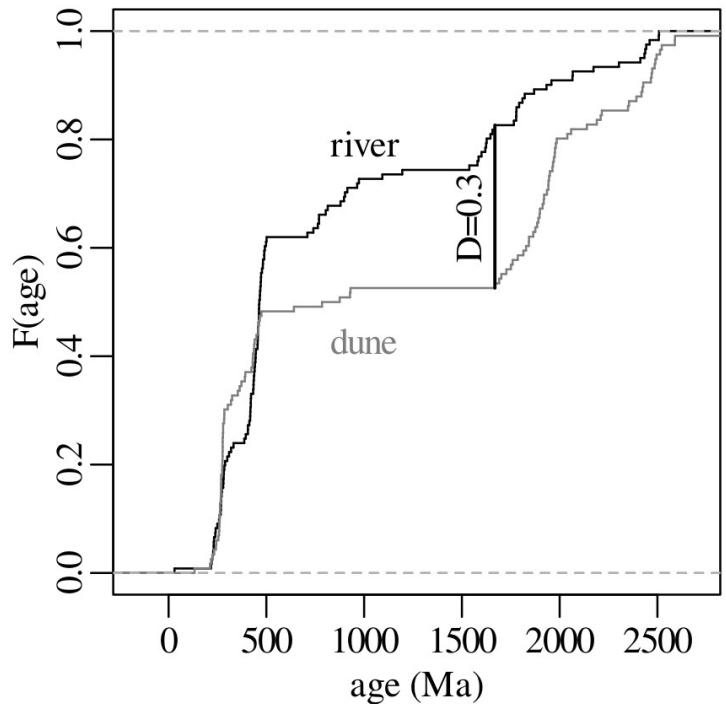
1      2      3

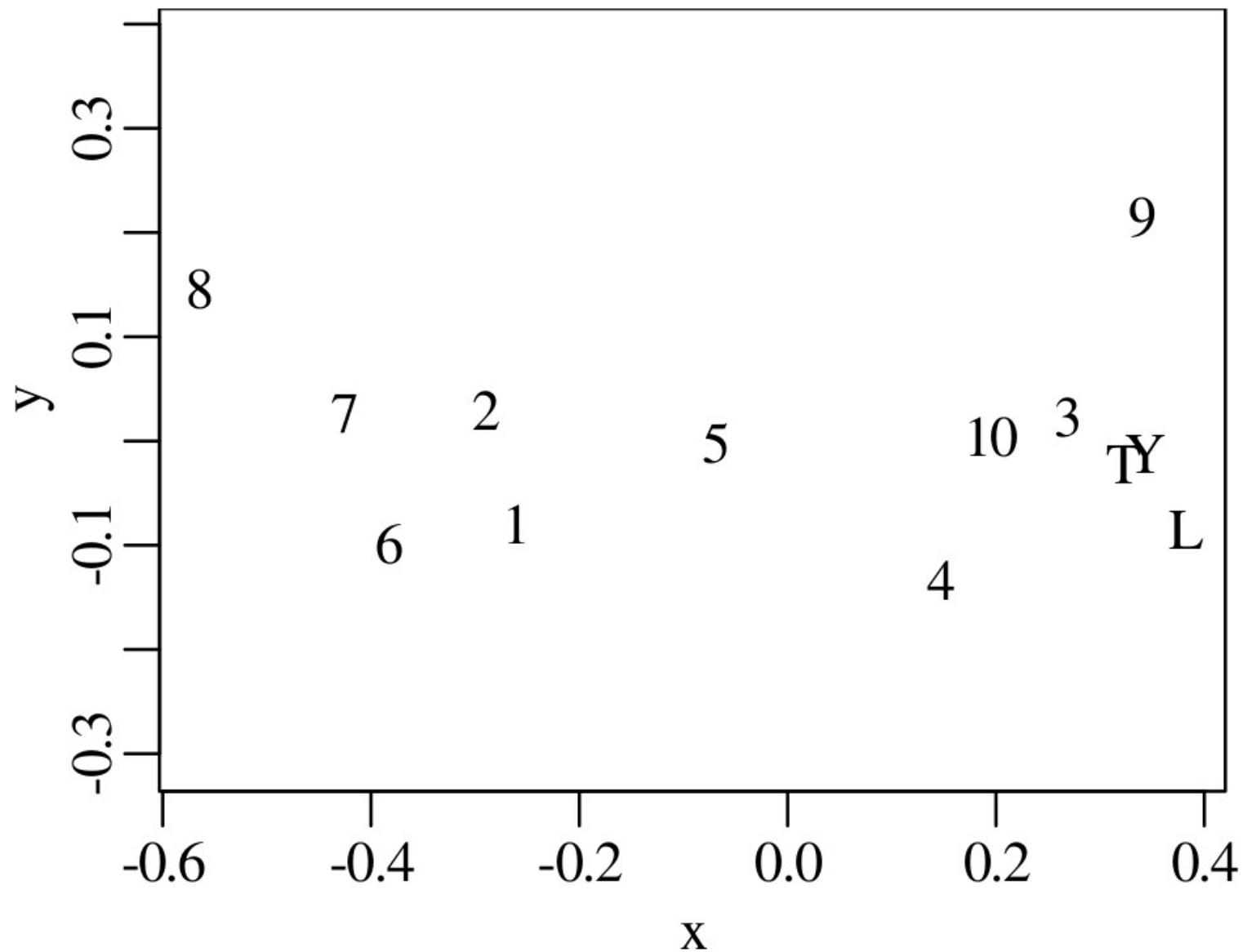
$$d = \begin{matrix} & x & y \\ \begin{matrix} 1 \\ 2 \\ 3 \end{matrix} & \left[ \begin{matrix} 0 & 6.4 & 6.4 \\ 6.4 & 0 & 1.4 \\ 6.4 & 1.4 & 0 \end{matrix} \right] \end{matrix} \longrightarrow m = \begin{matrix} & x & y \\ \begin{matrix} 1 \\ 2 \\ 3 \end{matrix} & \left[ \begin{matrix} 4.24 & 0 \\ -2.12 & -0.71 \\ -2.12 & 0.71 \end{matrix} \right] \end{matrix}$$



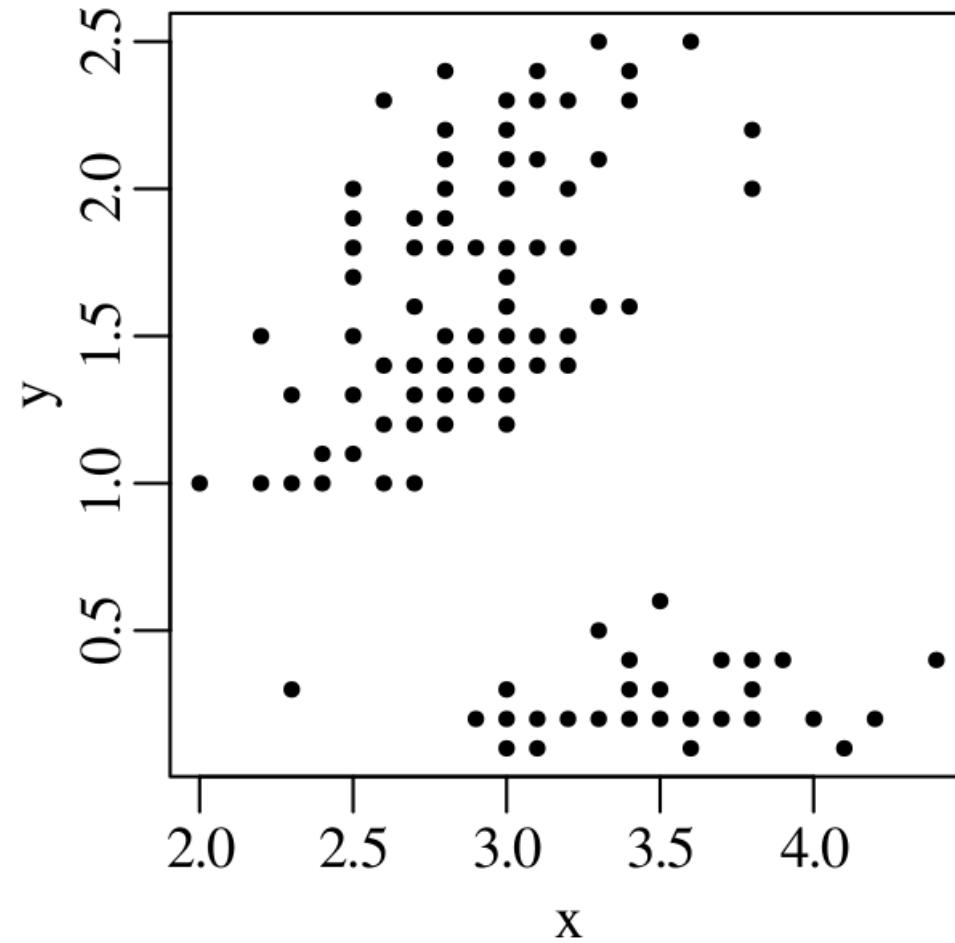


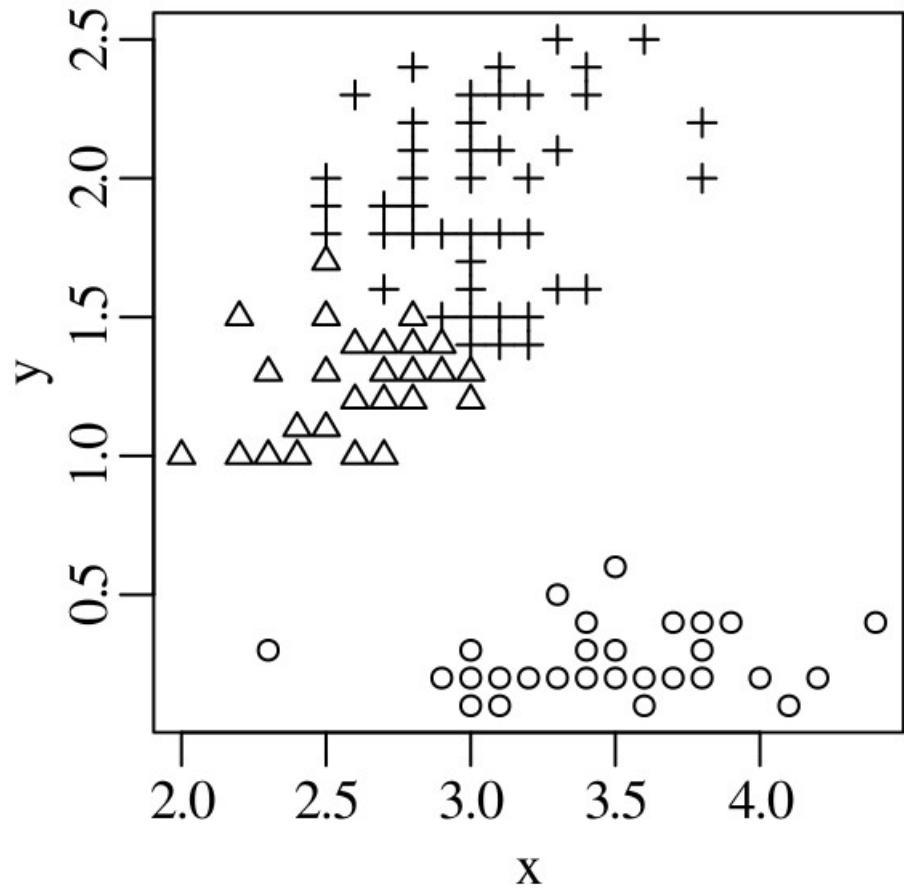
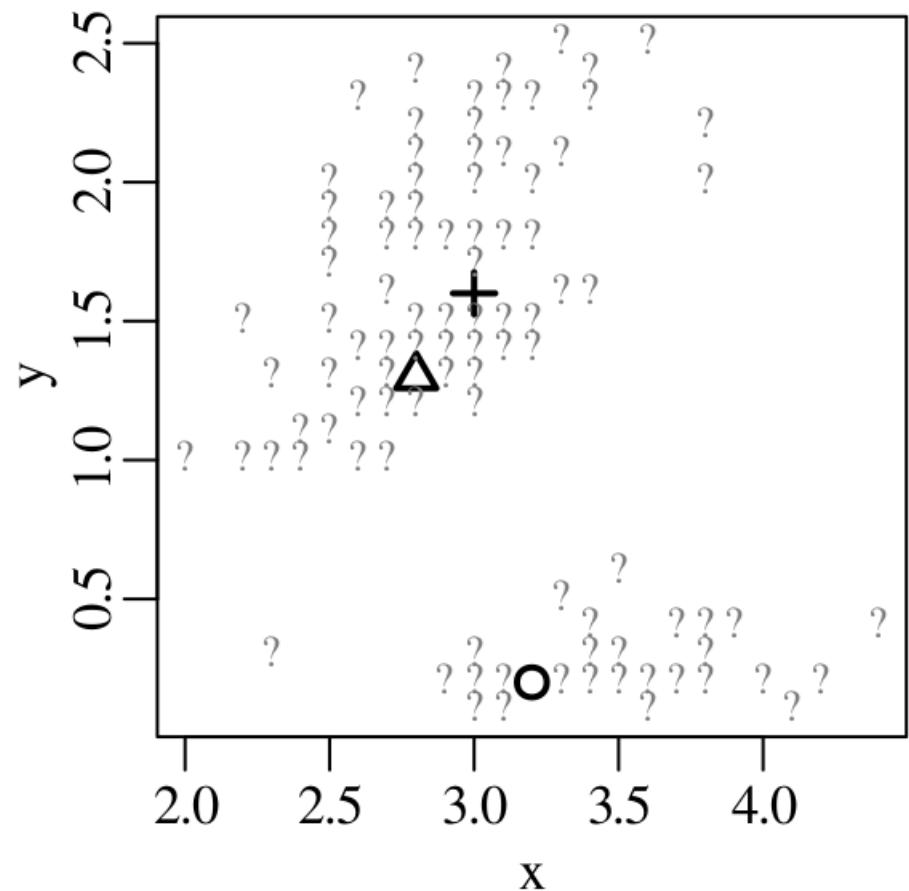
|       | 1  | 2  | 3  | 4  | 5  | 6  | 7  | 8  | 9  | 10 | L  | T  | Y  |
|-------|----|----|----|----|----|----|----|----|----|----|----|----|----|
| 1     | 0  | 14 | 33 | 27 | 18 | 14 | 15 | 22 | 48 | 32 | 42 | 37 | 40 |
| 2     | 14 | 0  | 36 | 33 | 16 | 14 | 15 | 24 | 46 | 32 | 47 | 42 | 43 |
| 3     | 33 | 36 | 0  | 19 | 24 | 44 | 47 | 55 | 17 | 10 | 13 | 12 | 8  |
| 4     | 27 | 33 | 19 | 0  | 20 | 38 | 41 | 48 | 28 | 14 | 21 | 17 | 16 |
| 5     | 18 | 16 | 24 | 20 | 0  | 22 | 24 | 33 | 31 | 20 | 33 | 28 | 30 |
| 6     | 14 | 14 | 44 | 38 | 22 | 0  | 14 | 24 | 52 | 41 | 52 | 48 | 49 |
| d = 7 | 15 | 15 | 47 | 41 | 24 | 14 | 0  | 16 | 51 | 43 | 54 | 49 | 52 |
| 8     | 22 | 24 | 55 | 48 | 33 | 24 | 16 | 0  | 61 | 53 | 63 | 59 | 62 |
| 9     | 48 | 46 | 17 | 28 | 31 | 52 | 51 | 61 | 0  | 20 | 22 | 18 | 16 |
| 10    | 32 | 32 | 10 | 14 | 20 | 41 | 43 | 53 | 20 | 0  | 17 | 15 | 13 |
| L     | 42 | 47 | 13 | 21 | 33 | 52 | 54 | 63 | 22 | 17 | 0  | 10 | 11 |
| T     | 37 | 42 | 12 | 17 | 28 | 48 | 49 | 59 | 18 | 15 | 10 | 0  | 7  |
| Y     | 40 | 43 | 8  | 16 | 30 | 49 | 52 | 62 | 16 | 13 | 11 | 7  | 0  |

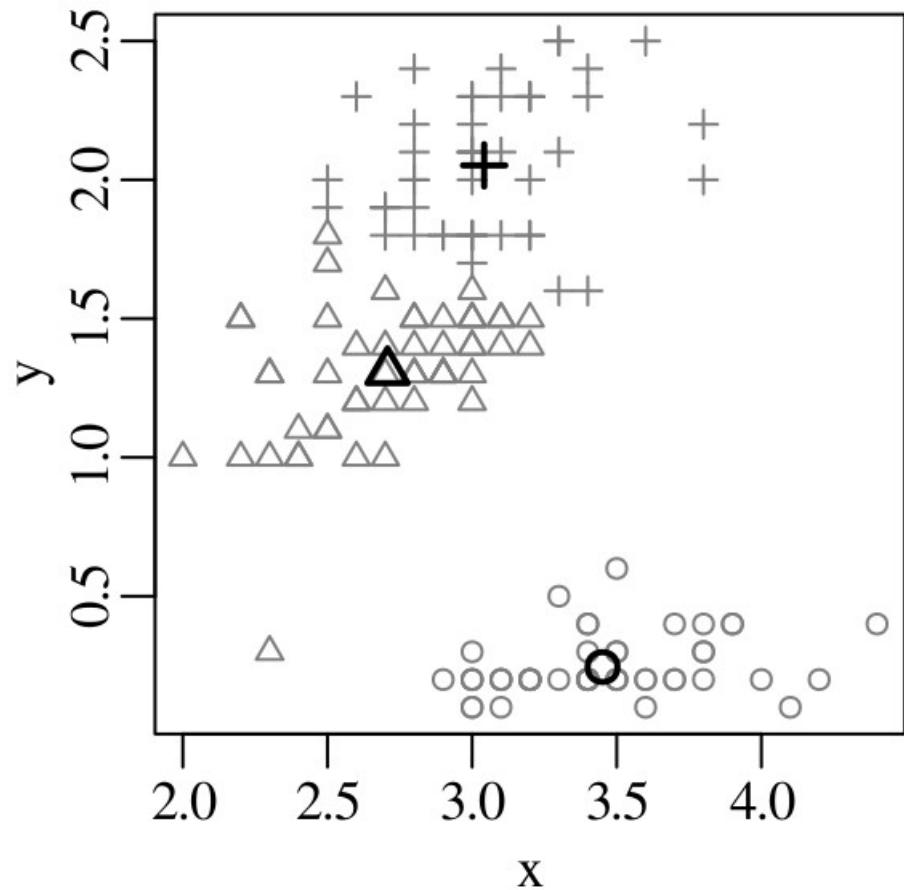
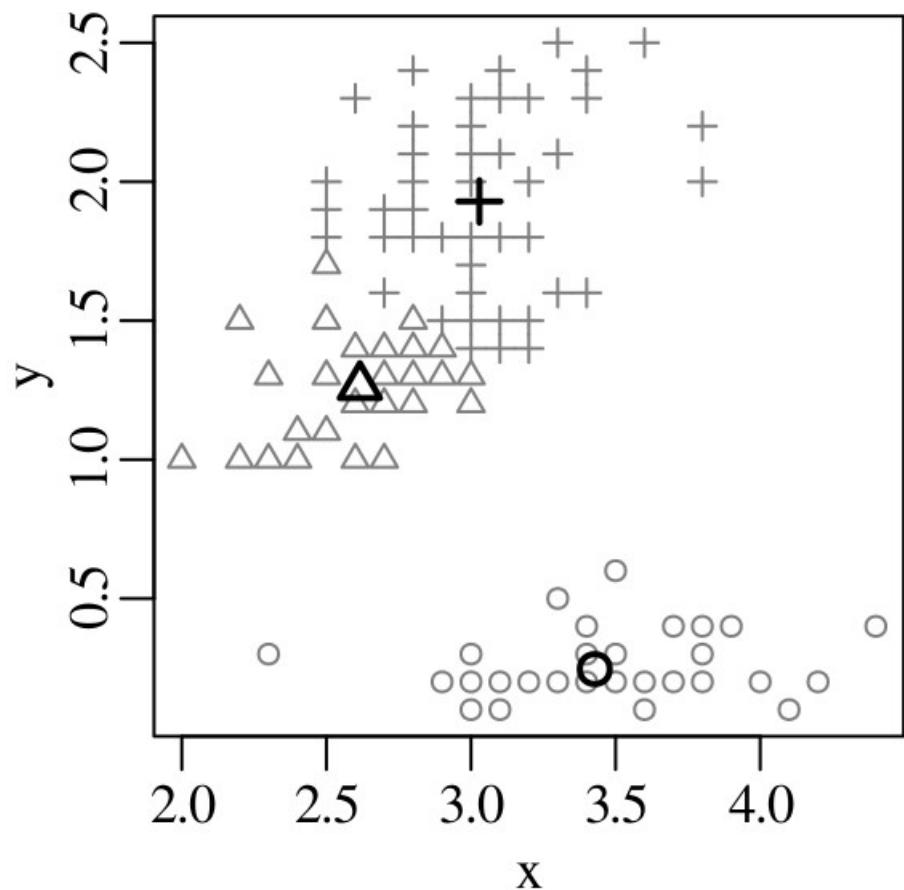


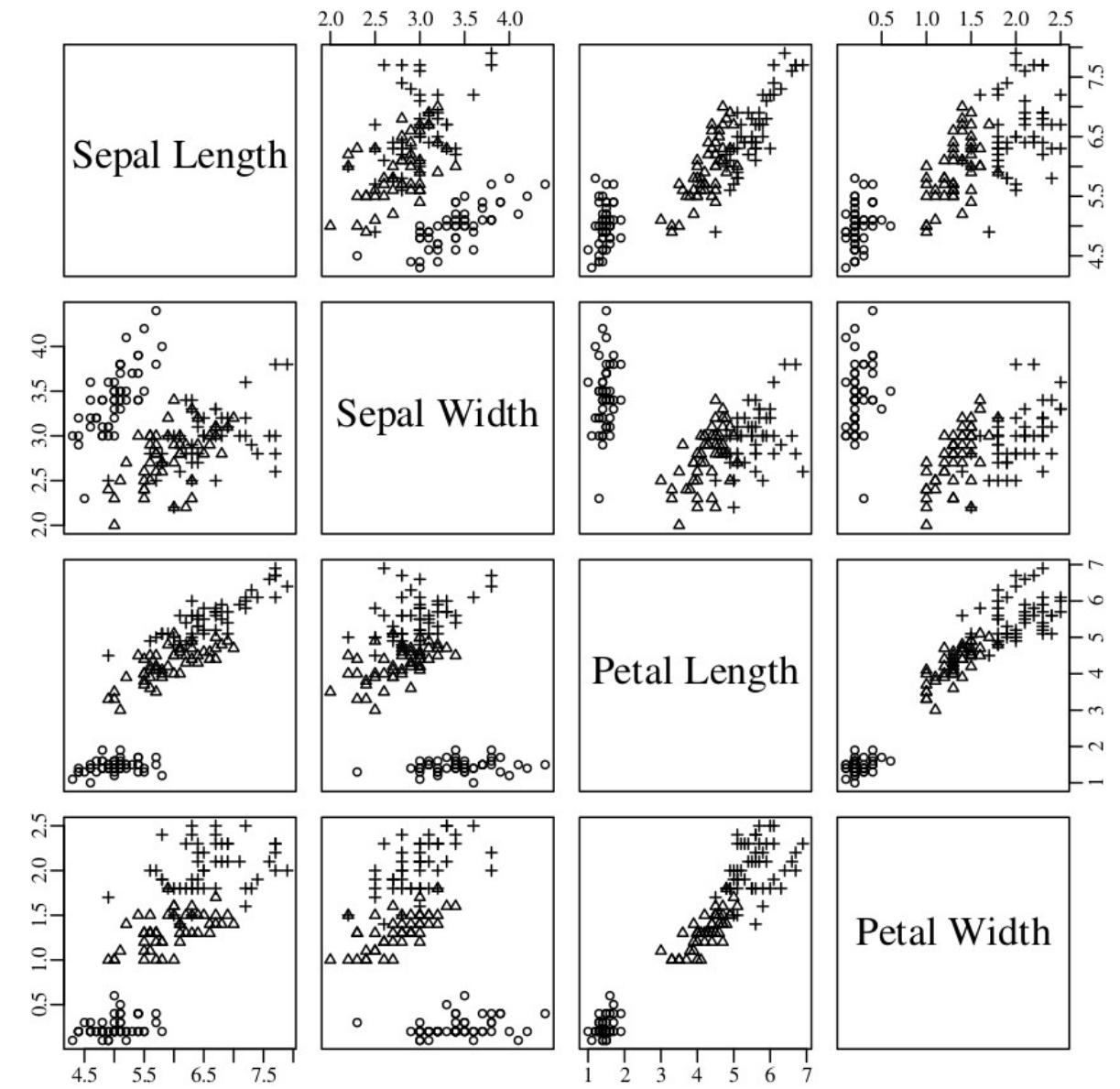


# K-means clustering





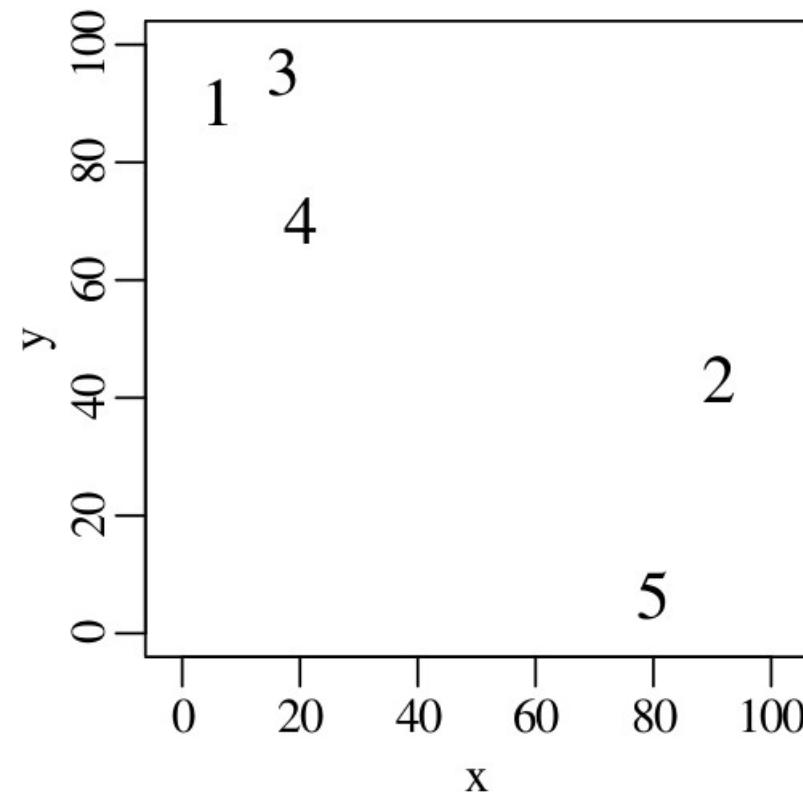




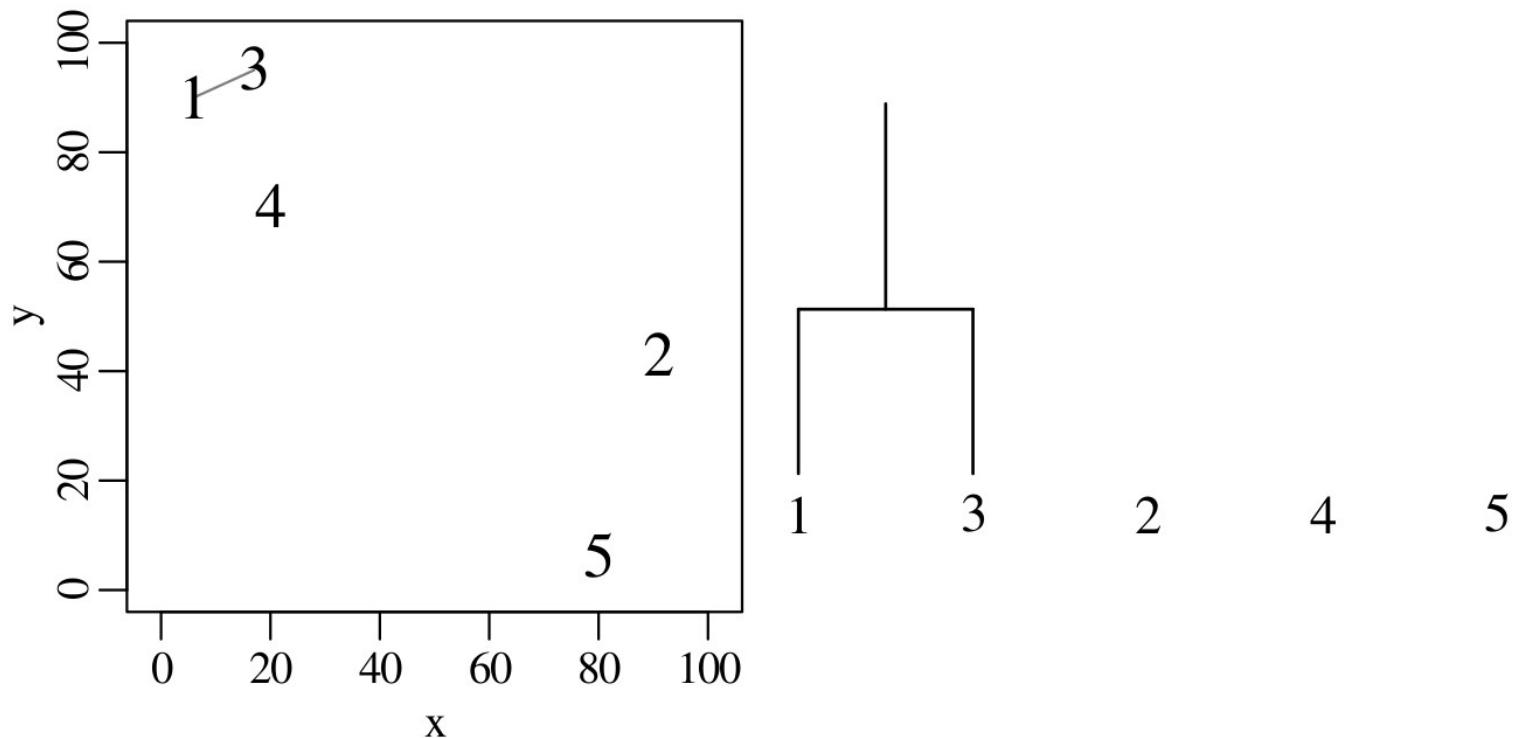
| cluster | setosa | versicolor | virginica |
|---------|--------|------------|-----------|
| 1       | 50     | 0          | 0         |
| 2       | 0      | 48         | 14        |
| 3       | 0      | 2          | 36        |

# Hierarchical clustering

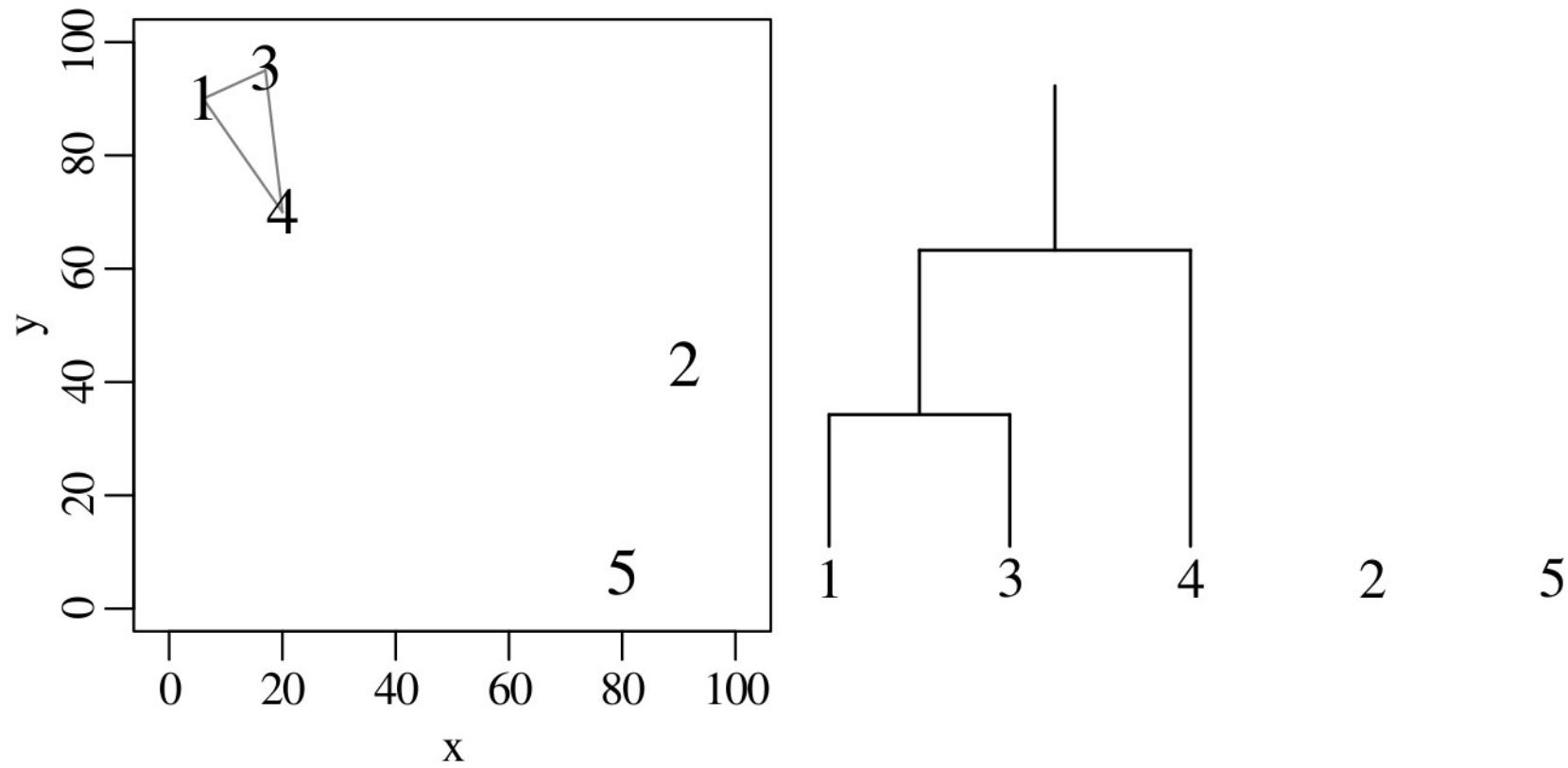
$$X = \begin{matrix} & x & y \\ \begin{matrix} 1 \\ 2 \\ 3 \\ 4 \\ 5 \end{matrix} & \left[ \begin{matrix} 6 & 90 \\ 91 & 43 \\ 17 & 95 \\ 20 & 70 \\ 80 & 6 \end{matrix} \right] \end{matrix}$$



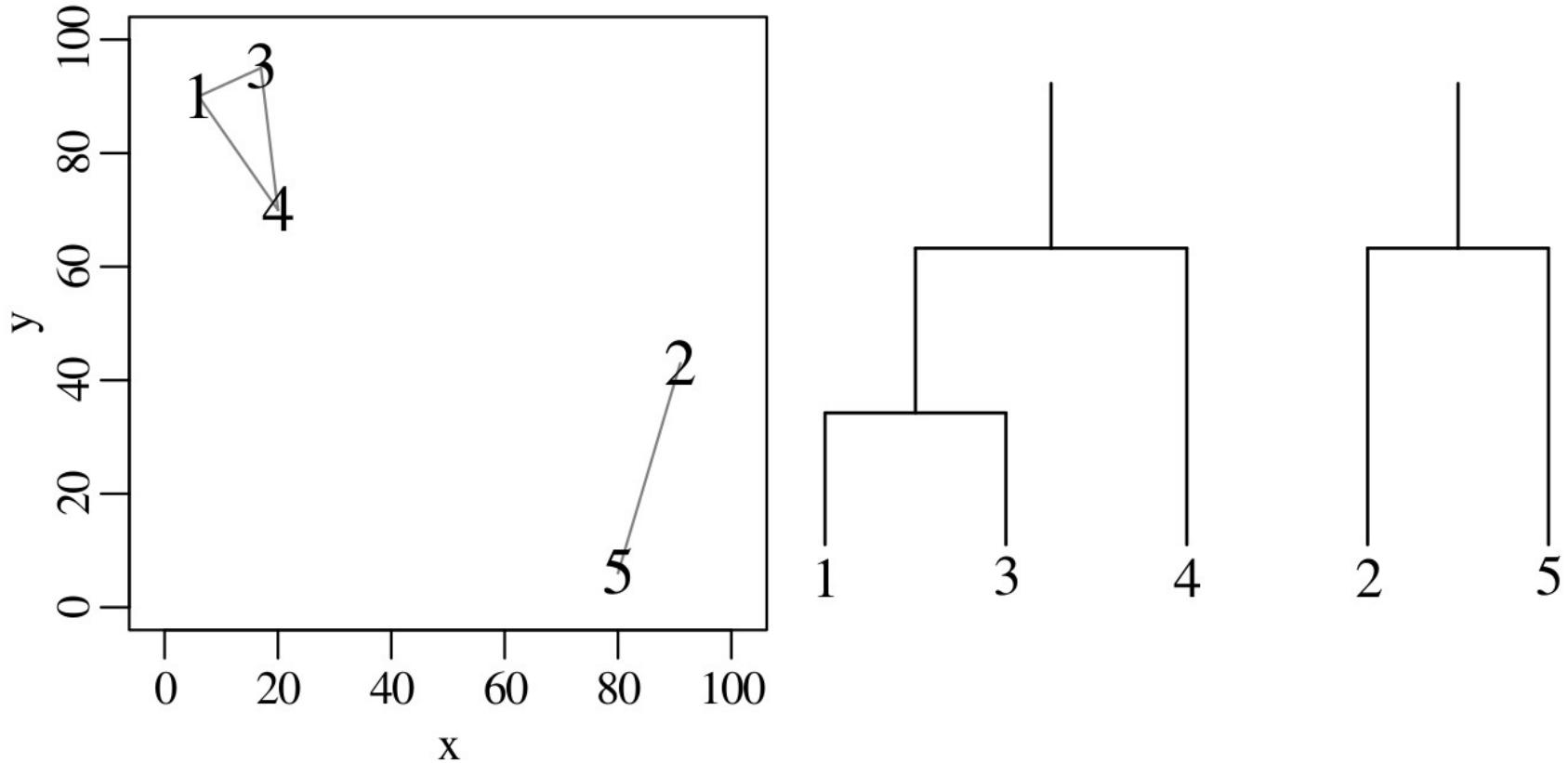
|   | 1    | 2    | 3    | 4    | 5    |
|---|------|------|------|------|------|
| 1 | 0    | 97.1 | 12.1 | 24.4 | 112  |
| 2 | 97.1 | 0    | 90.4 | 76.0 | 38.6 |
| 3 | 12.1 | 90.4 | 0    | 25.2 | 109  |
| 4 | 24.4 | 76.0 | 25.2 | 0    | 87.7 |
| 5 | 112  | 38.6 | 109  | 87.7 | 0    |

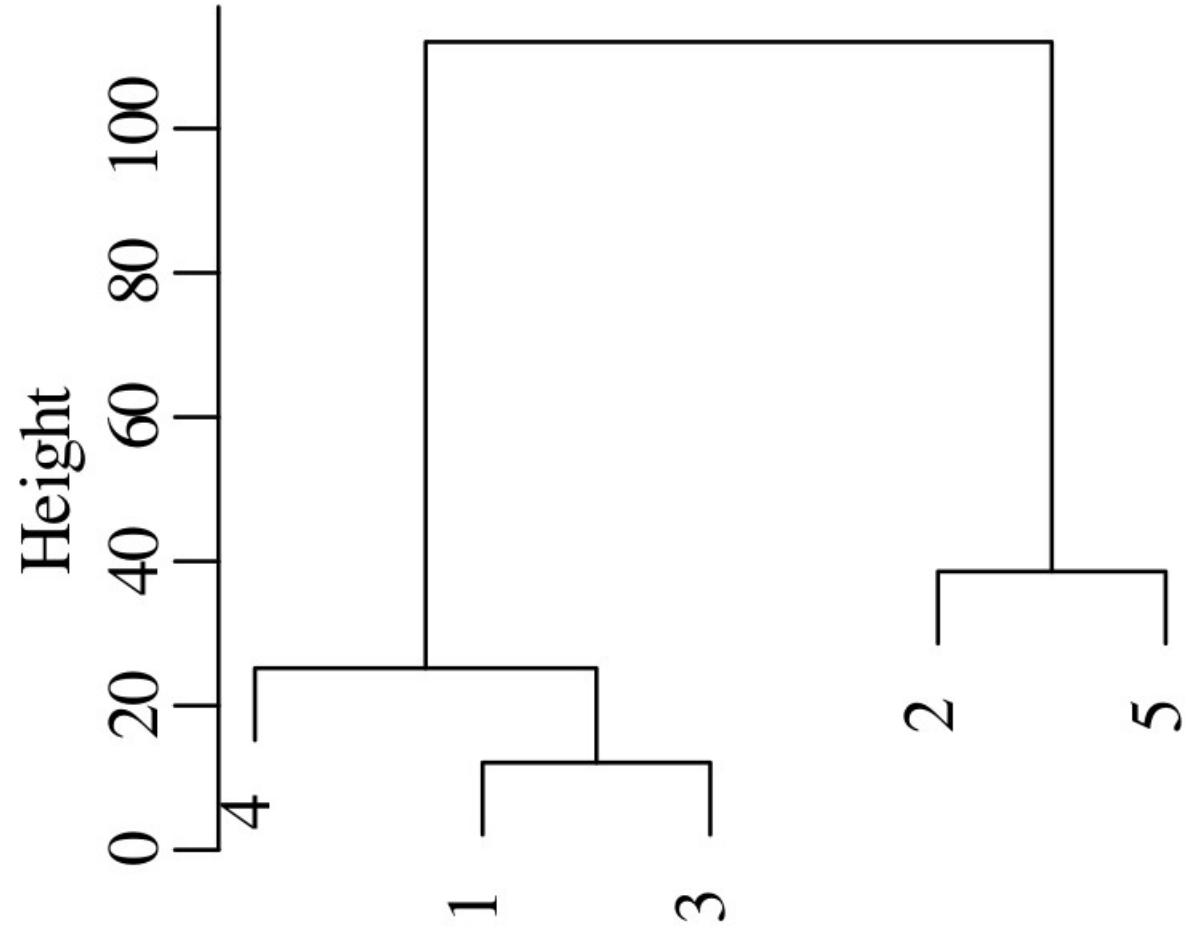


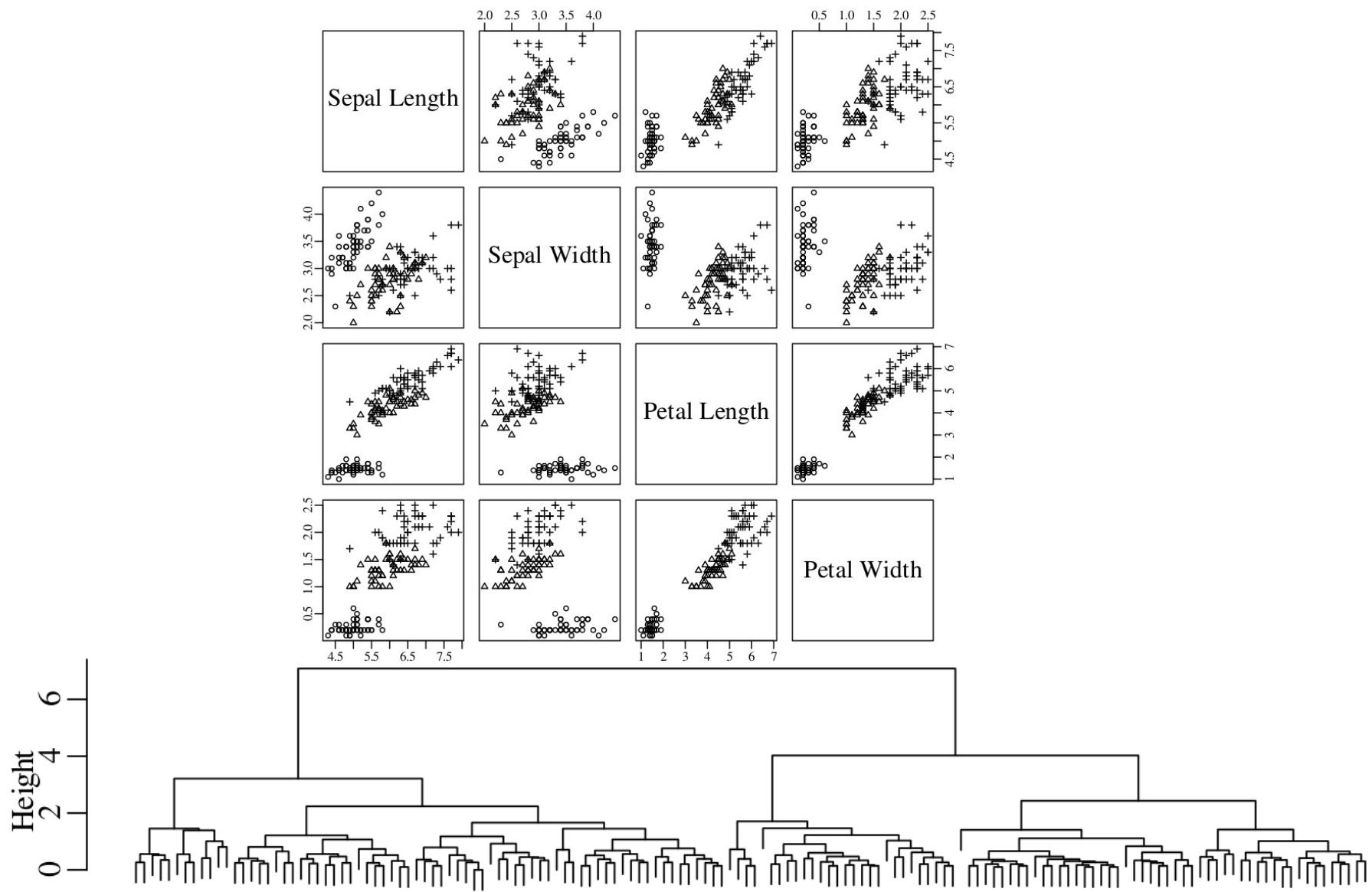
$$d = \begin{bmatrix} 13 & 2 & 4 & 5 \\ 13 & 0 & 97.1 & \boxed{25.2} & 112 \\ 2 & 97.1 & 0 & 76.0 & 38.6 \\ 4 & \boxed{25.2} & 76.0 & 0 & 87.7 \\ 5 & 112 & 38.6 & 87.7 & 0 \end{bmatrix} \text{ where } d[13, i] = \max(d[1, i], d[3, i]) \text{ for } i \in \{2, 4, 5\}$$

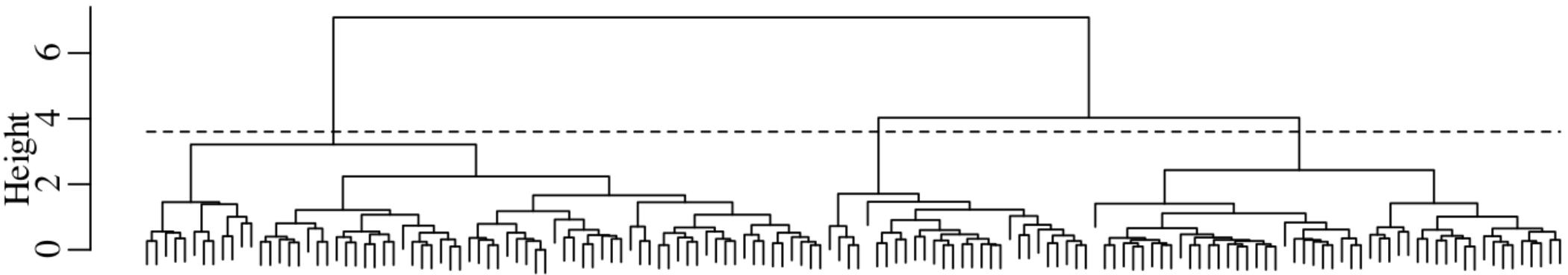


$$d = \begin{matrix} & 134 & 2 & 5 \\ & 134 & \left[ \begin{matrix} 0 & 97.1 & 112 \\ 97.1 & 0 & 76.0 \\ 112 & 76.0 & 0 \end{matrix} \right] \\ 2 & & & \\ 5 & & & \end{matrix}$$







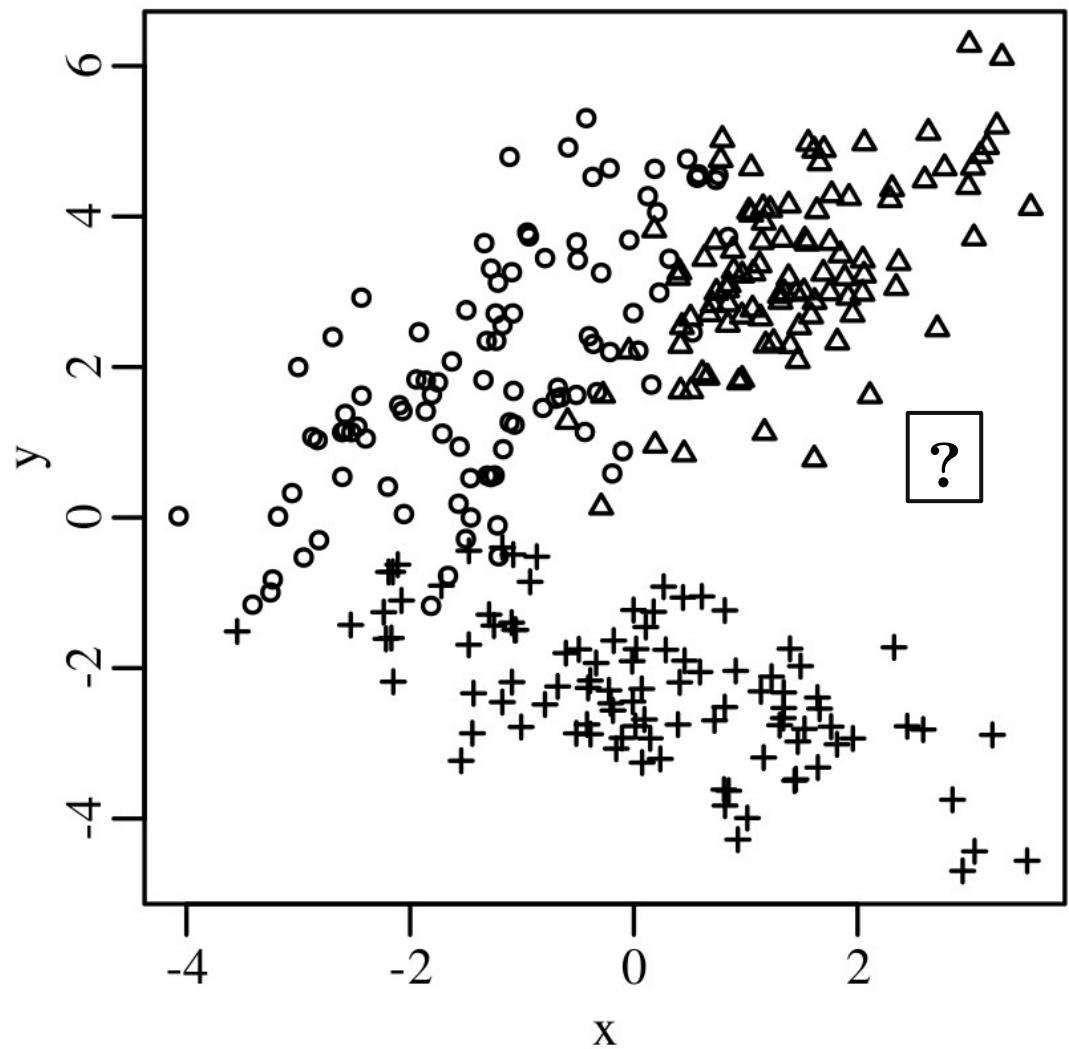


| cluster | setosa | versicolor | virginica |
|---------|--------|------------|-----------|
| 1       | 50     | 0          | 0         |
| 2       | 0      | 23         | 49        |
| 3       | 0      | 27         | 1         |

# Statistics for geoscientists

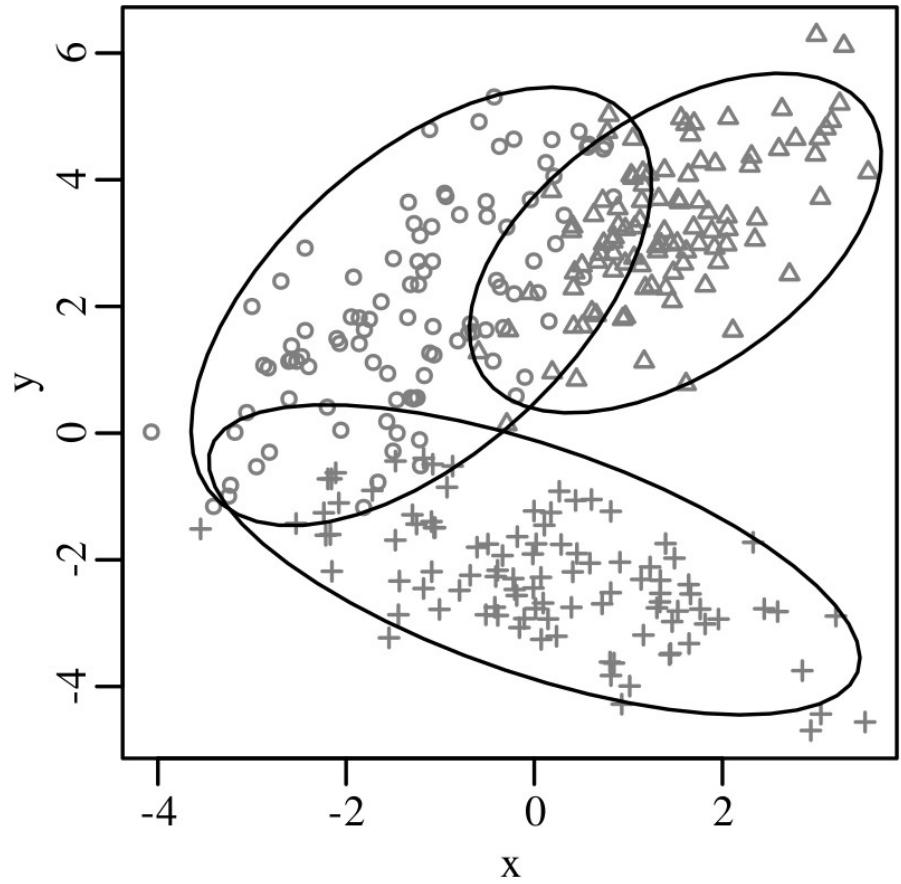
Supervised learning

# Discriminant Analysis



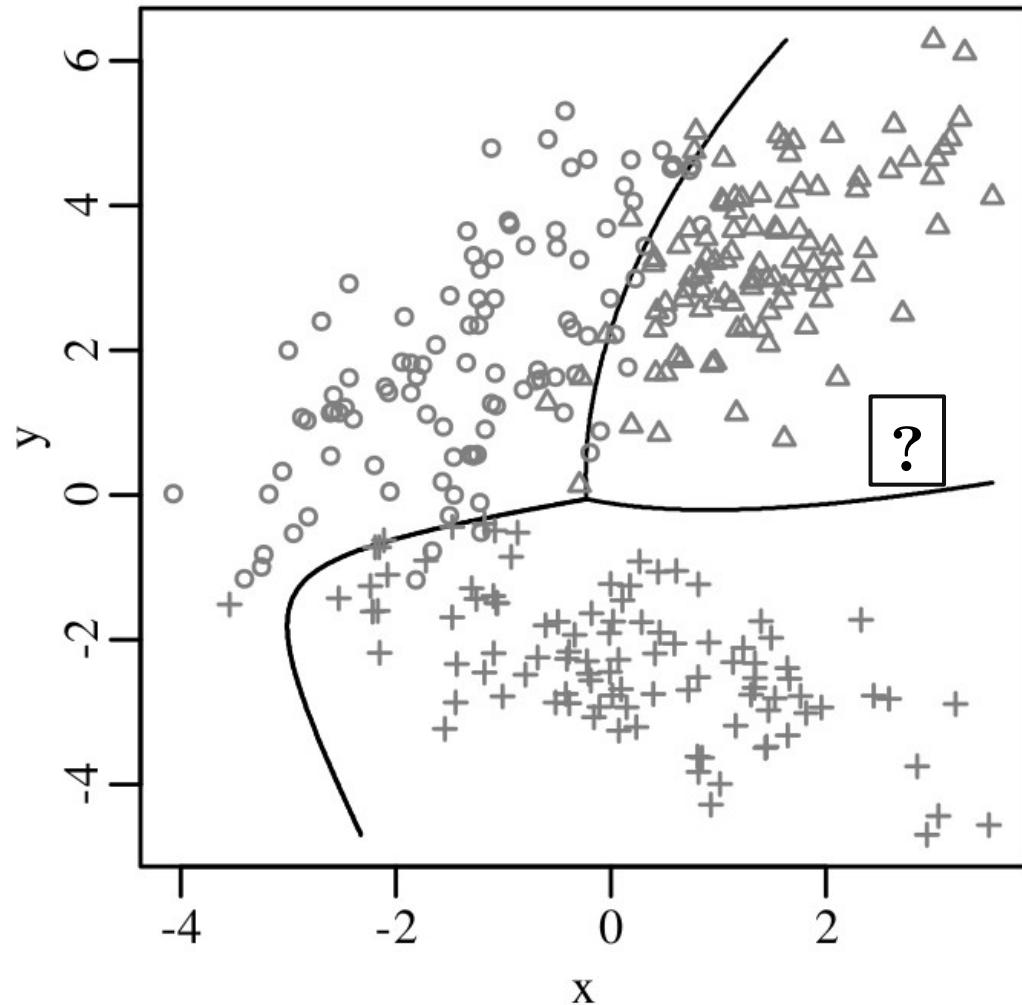
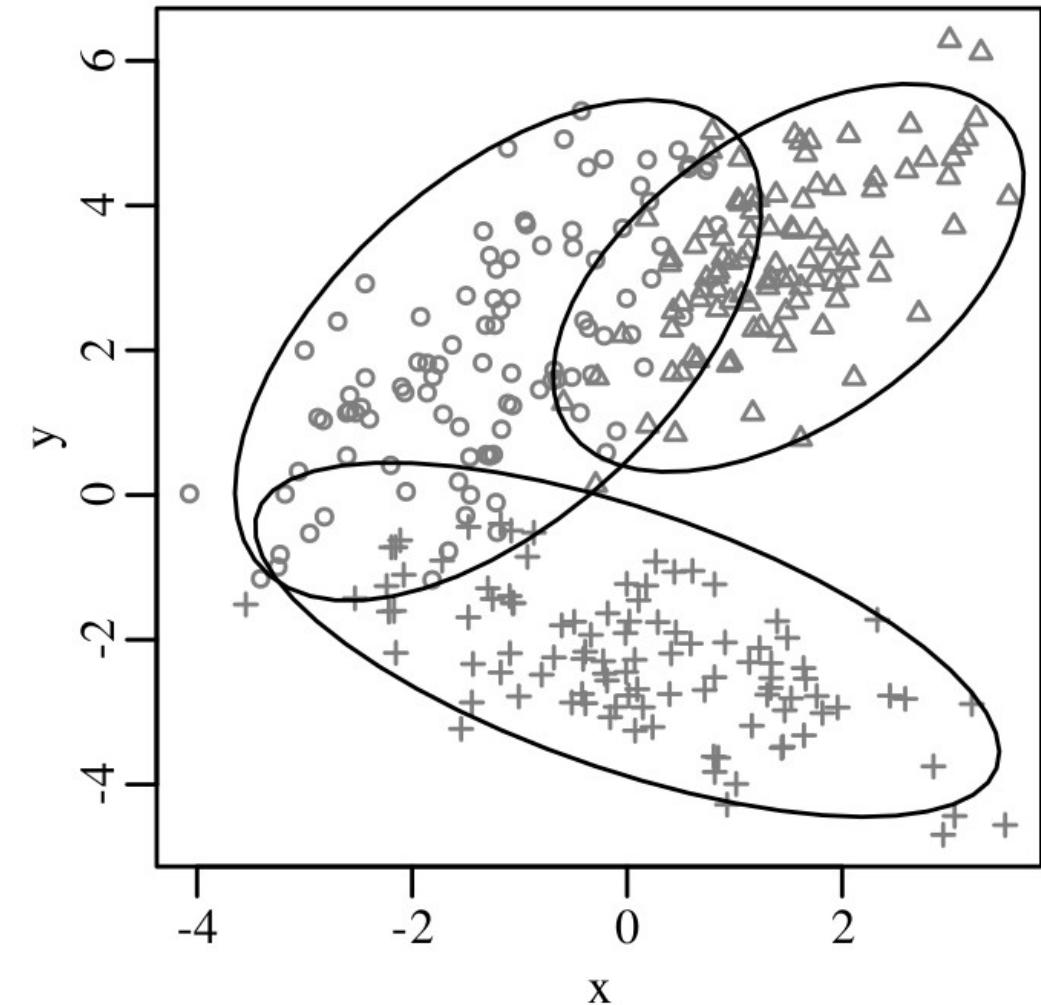
$$d = \max_{k=1,\dots,K} P(G = k | X = x)$$

$$P(G|X) \propto P(X|G)P(G)$$



$$P(X = x | G = k) = \frac{\exp\left(-\frac{1}{2}(x - \mu_k)^T \Sigma_k^{-1} (x - \mu_k)\right)}{\sqrt{(2\pi)^N |\Sigma_k|}}$$

$$d = \max_{k=1,\dots,K} \left[ -\frac{1}{2} \ln |\Sigma_k| - \frac{1}{2}(x - \mu_k)^T \Sigma_k^{-1} (x - \mu_k) \right]$$

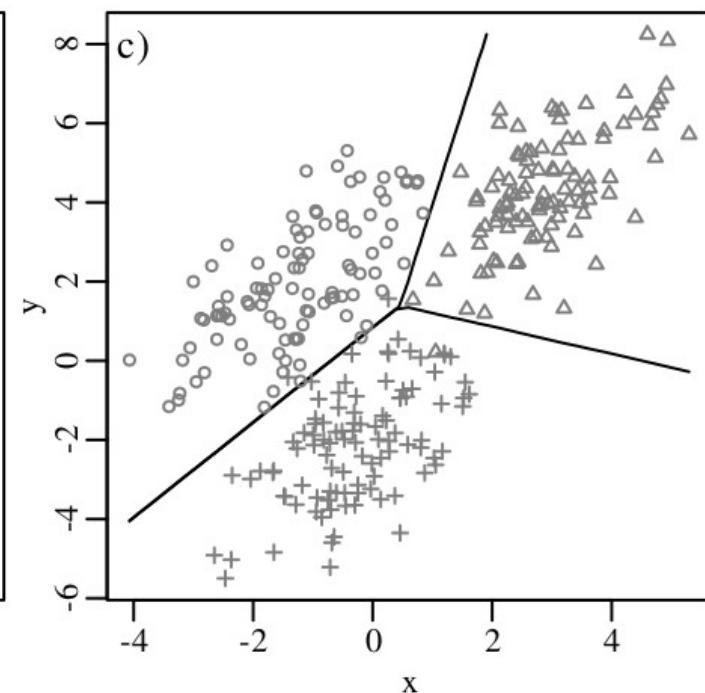
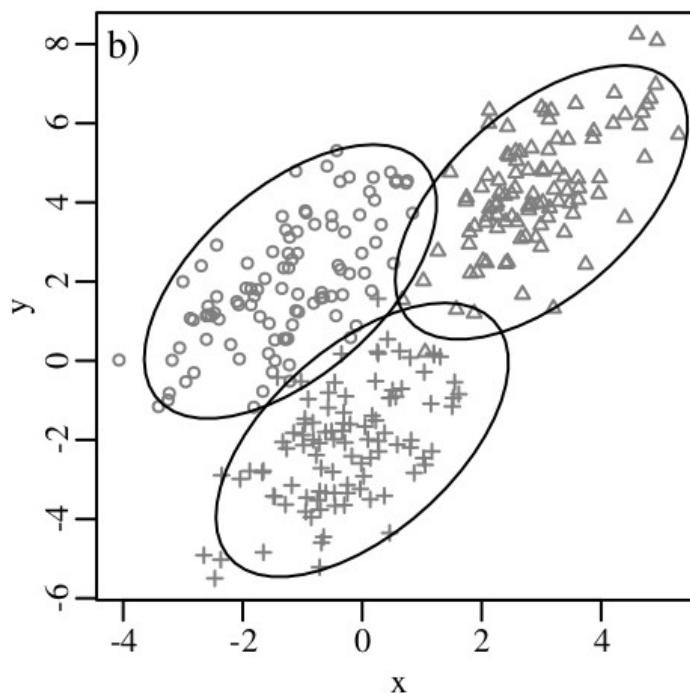
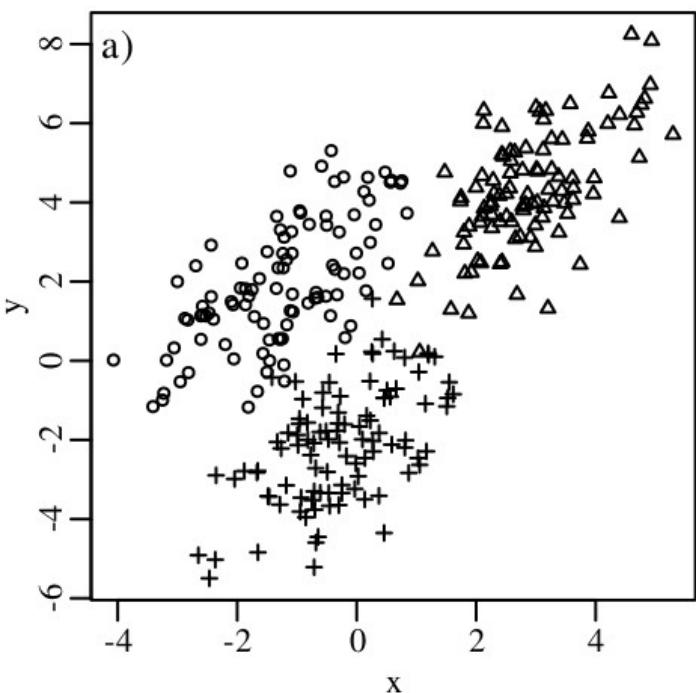


QDA

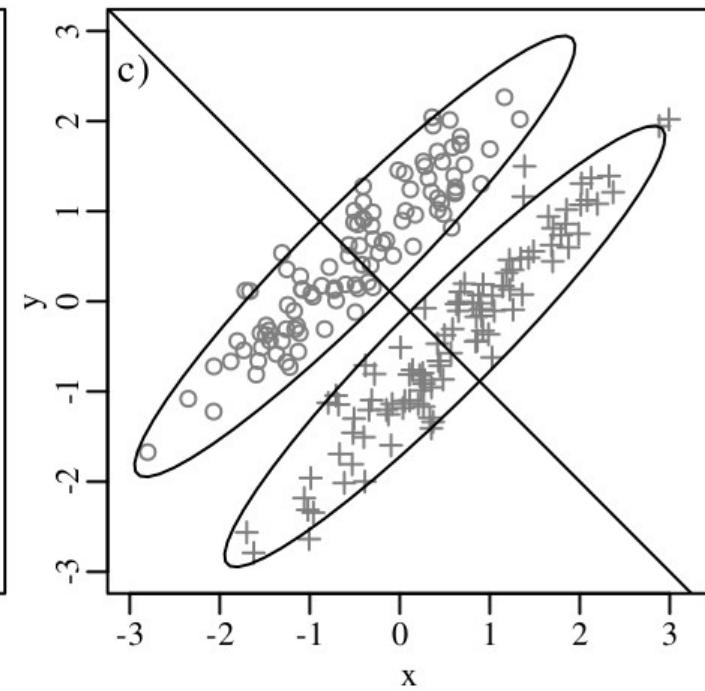
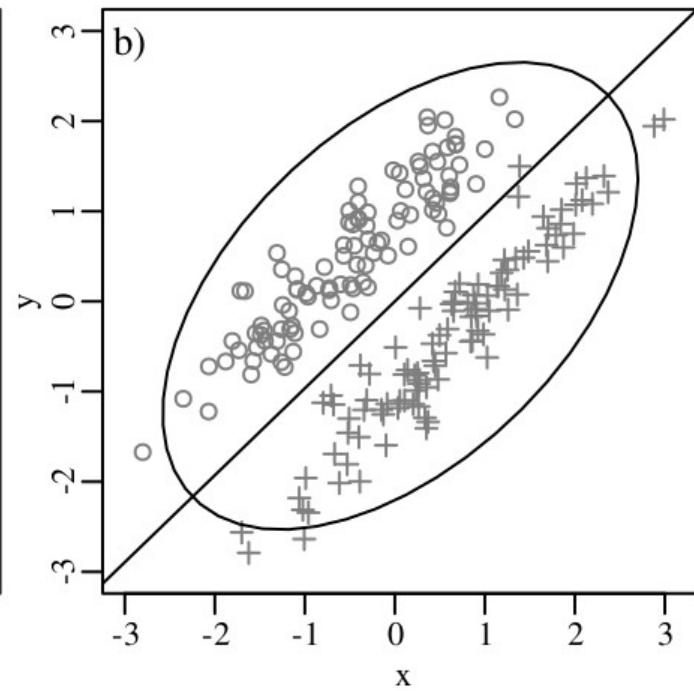
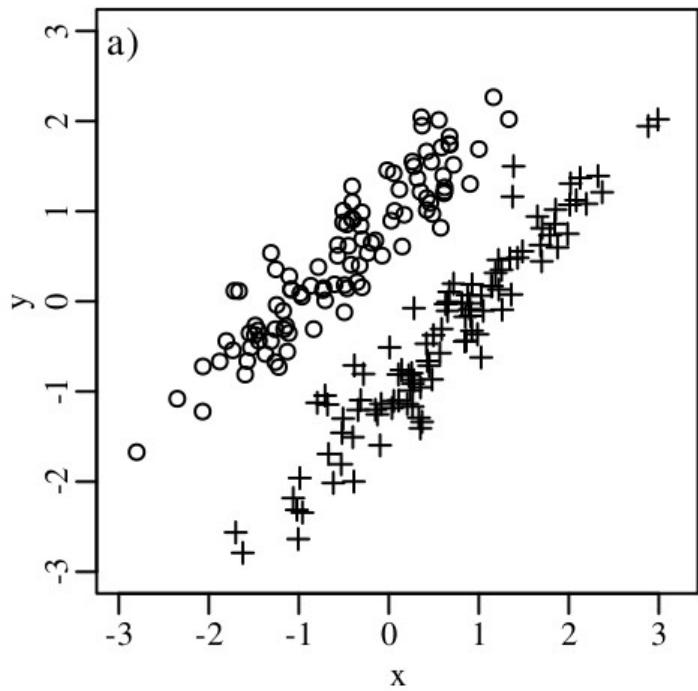
$$d = \max_{k=1,\dots,K} \left[ -\frac{1}{2} \ln |\Sigma_k| - \frac{1}{2} (x - \mu_k)^T \Sigma_k^{-1} (x - \mu_k) \right]$$

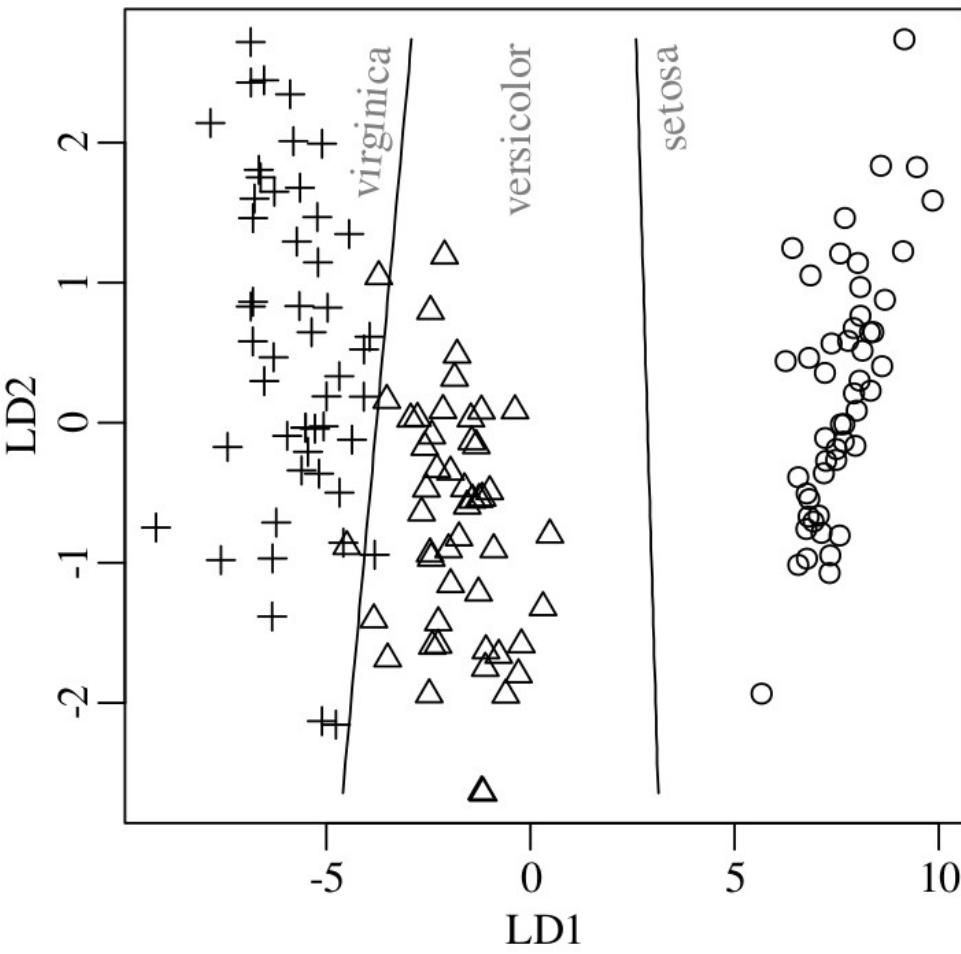
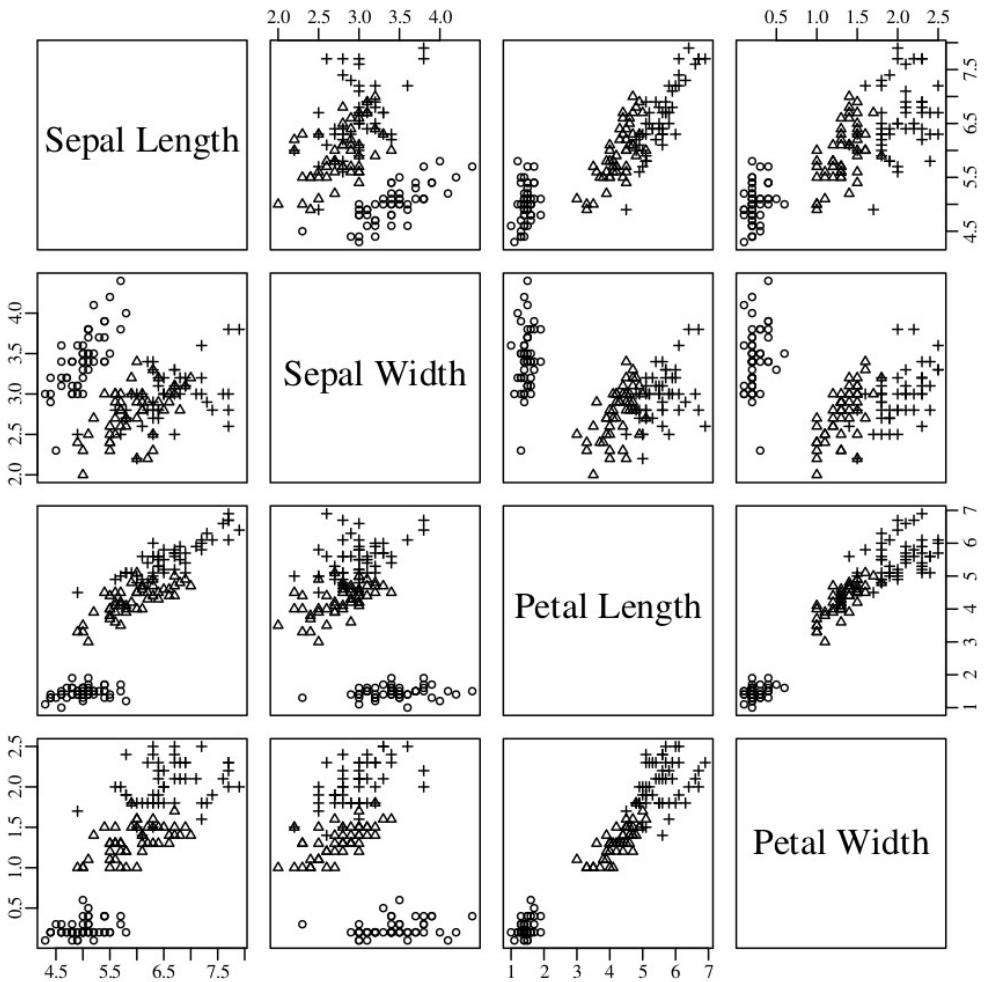
LDA

$$d = \max_{k=1,\dots,K} \left[ x^T \Sigma^{-1} \mu_k - \frac{1}{2} \mu_k^T \Sigma^{-1} \mu_k \right]$$



# PCA vs. LDA

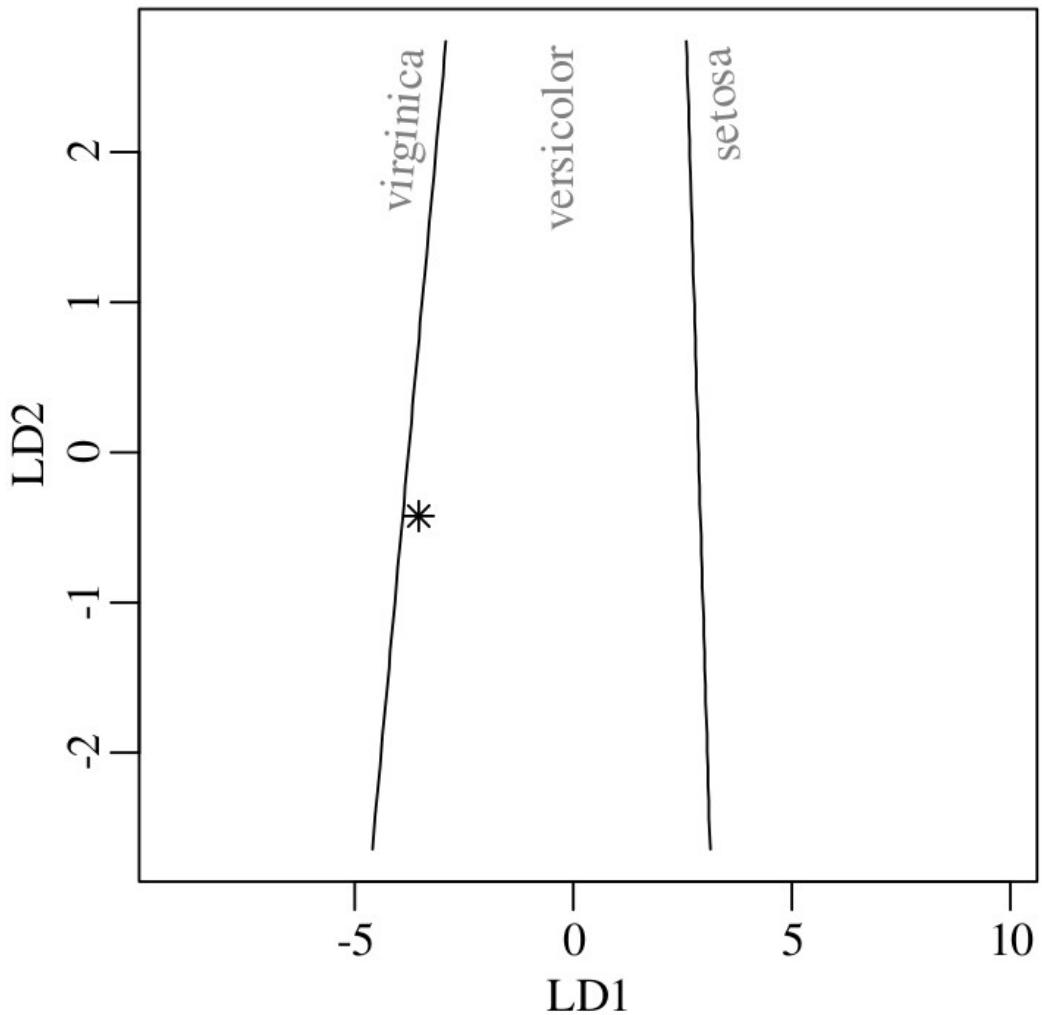




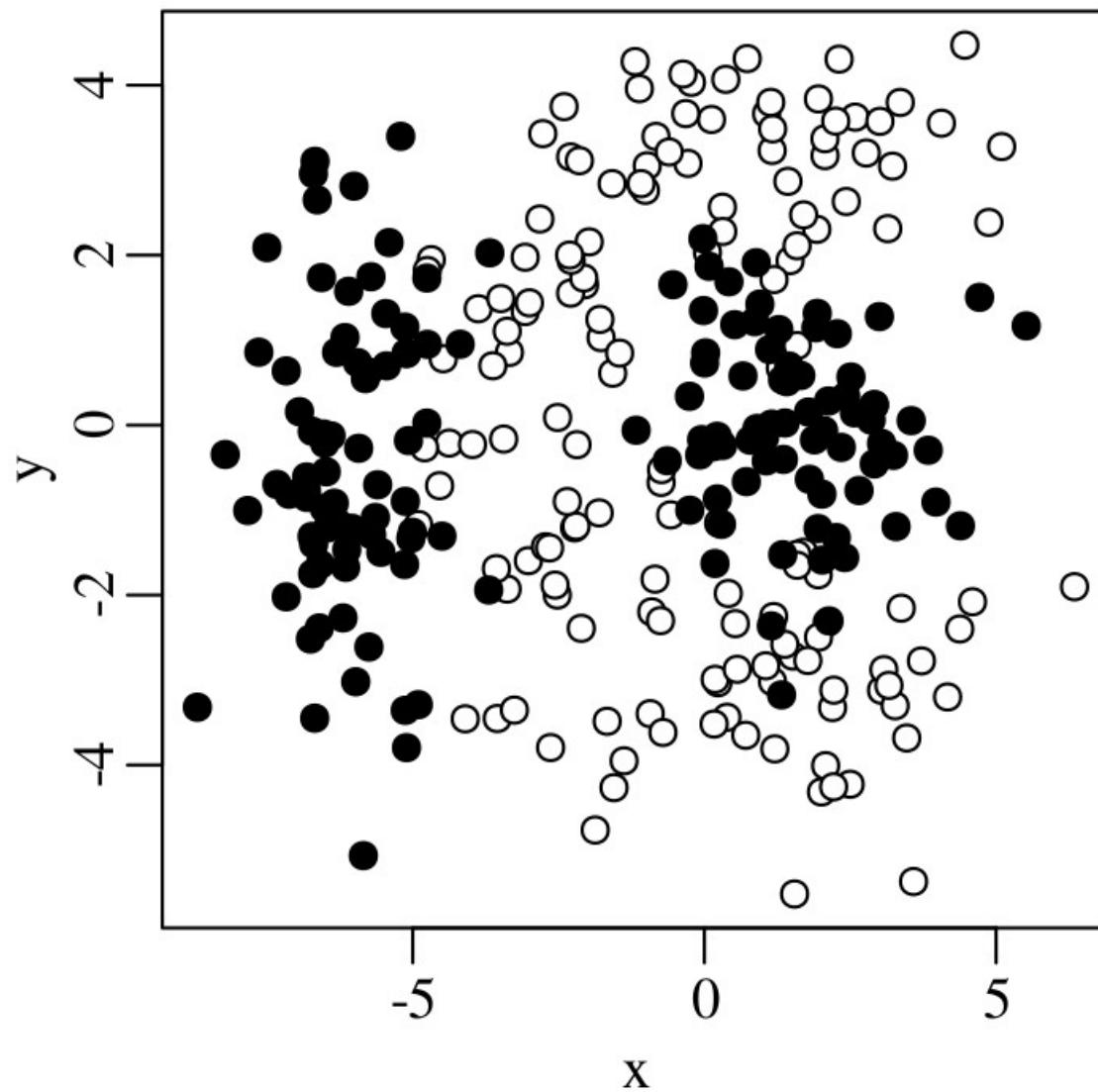
$$\begin{aligned} \text{LD1} = & 0.83 \times (\text{sepal length} - 5.84) \\ & + 1.53 \times (\text{sepal width} - 3.06) \\ & - 2.20 \times (\text{petal length} - 3.76) \\ & - 2.81 \times (\text{petal width} - 1.20) \end{aligned}$$

$$\begin{aligned} \text{LD2} = & 0.024 \times (\text{sepal length} - 5.85) \\ & + 2.16 \times (\text{sepal width} - 3.06) \\ & - 0.93 \times (\text{petal length} - 3.76) \\ & + 2.84 \times (\text{petal width} - 1.20) \end{aligned}$$

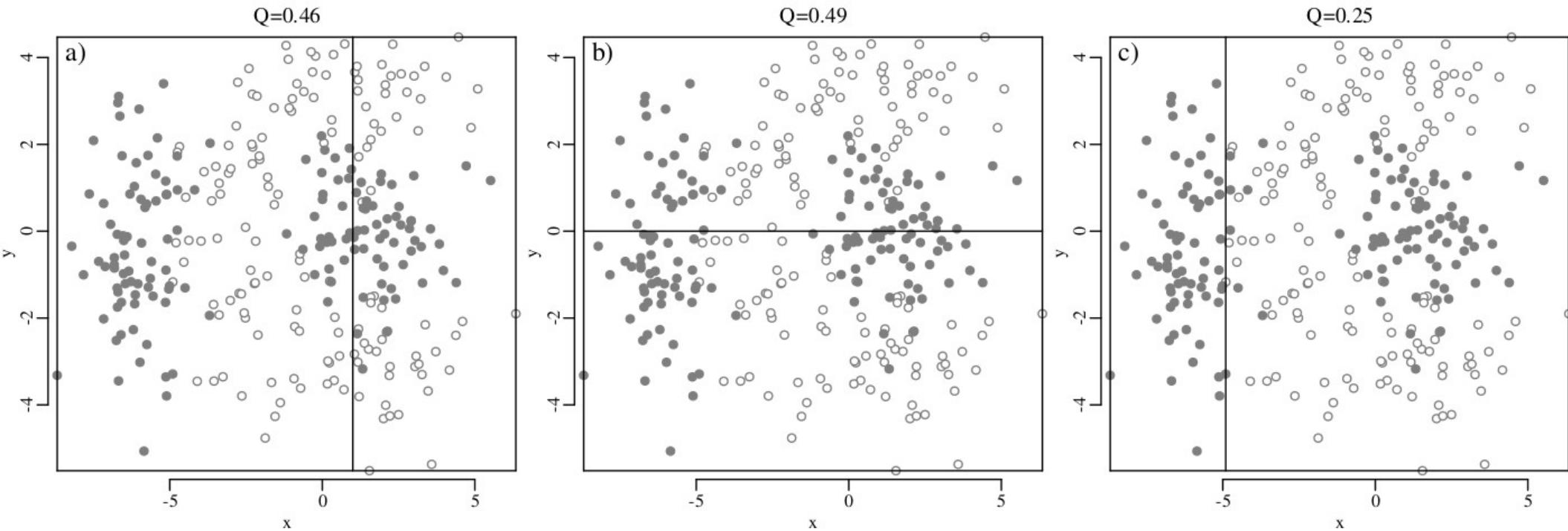
|          | <i>virginica</i> | <i>versicolor</i> | <i>setosa</i>         |
|----------|------------------|-------------------|-----------------------|
| $P(G X)$ | 0.19             | 0.81              | $3.2 \times 10^{-27}$ |

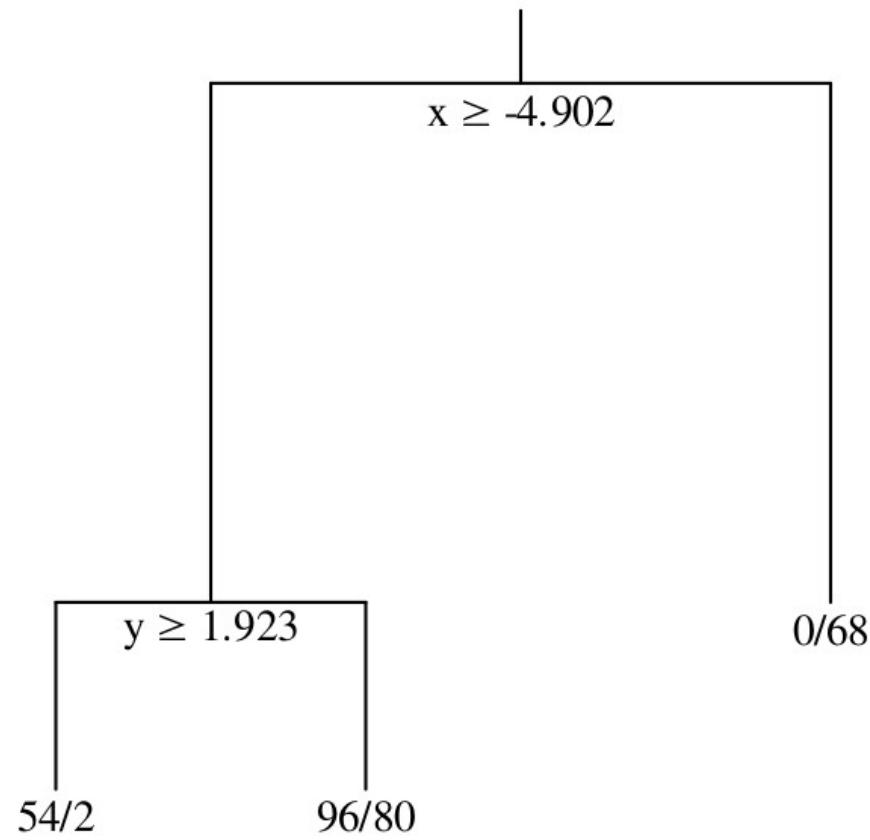
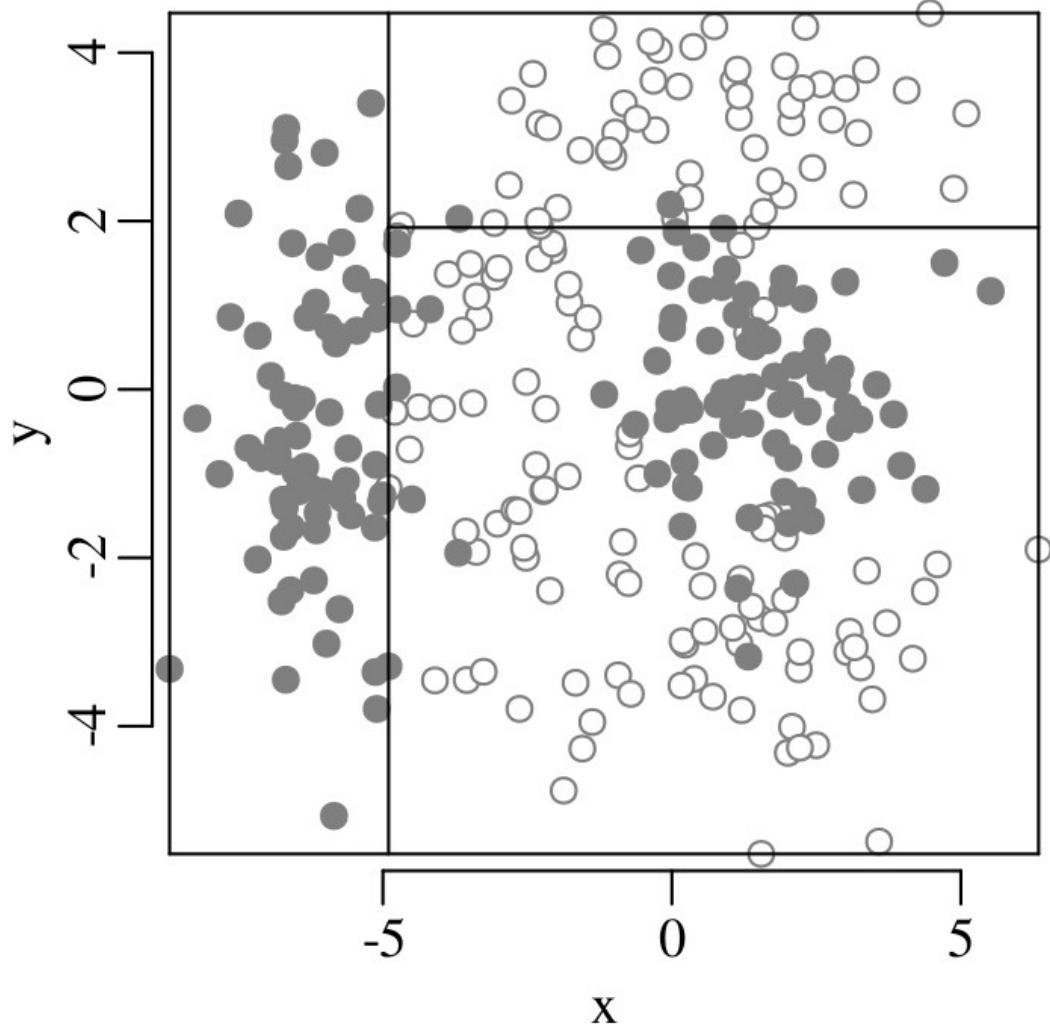


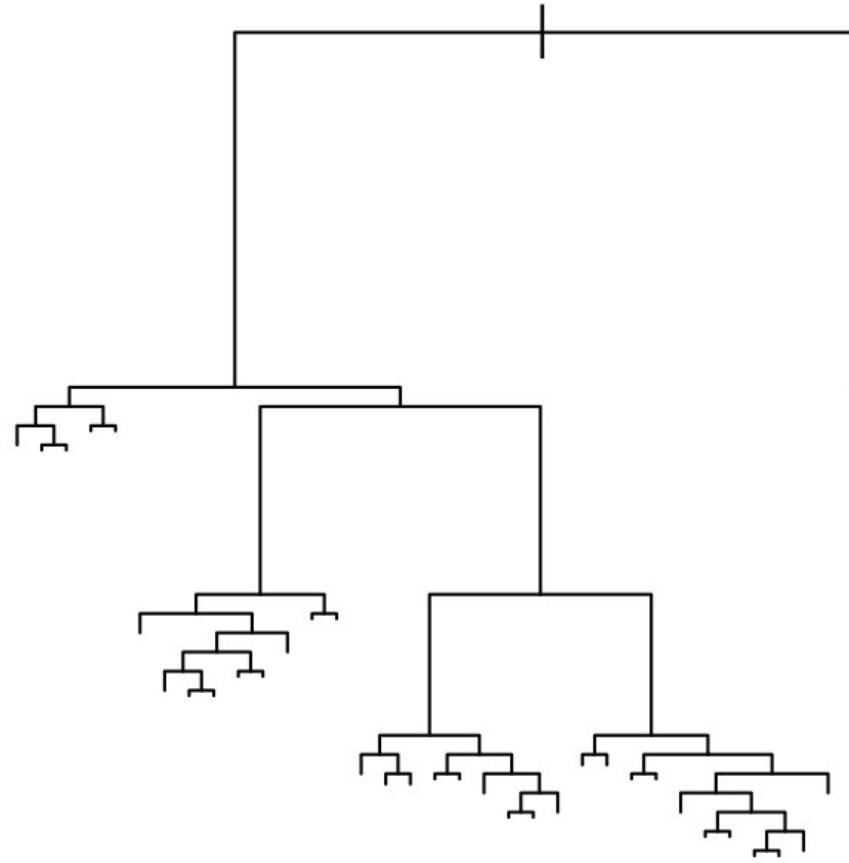
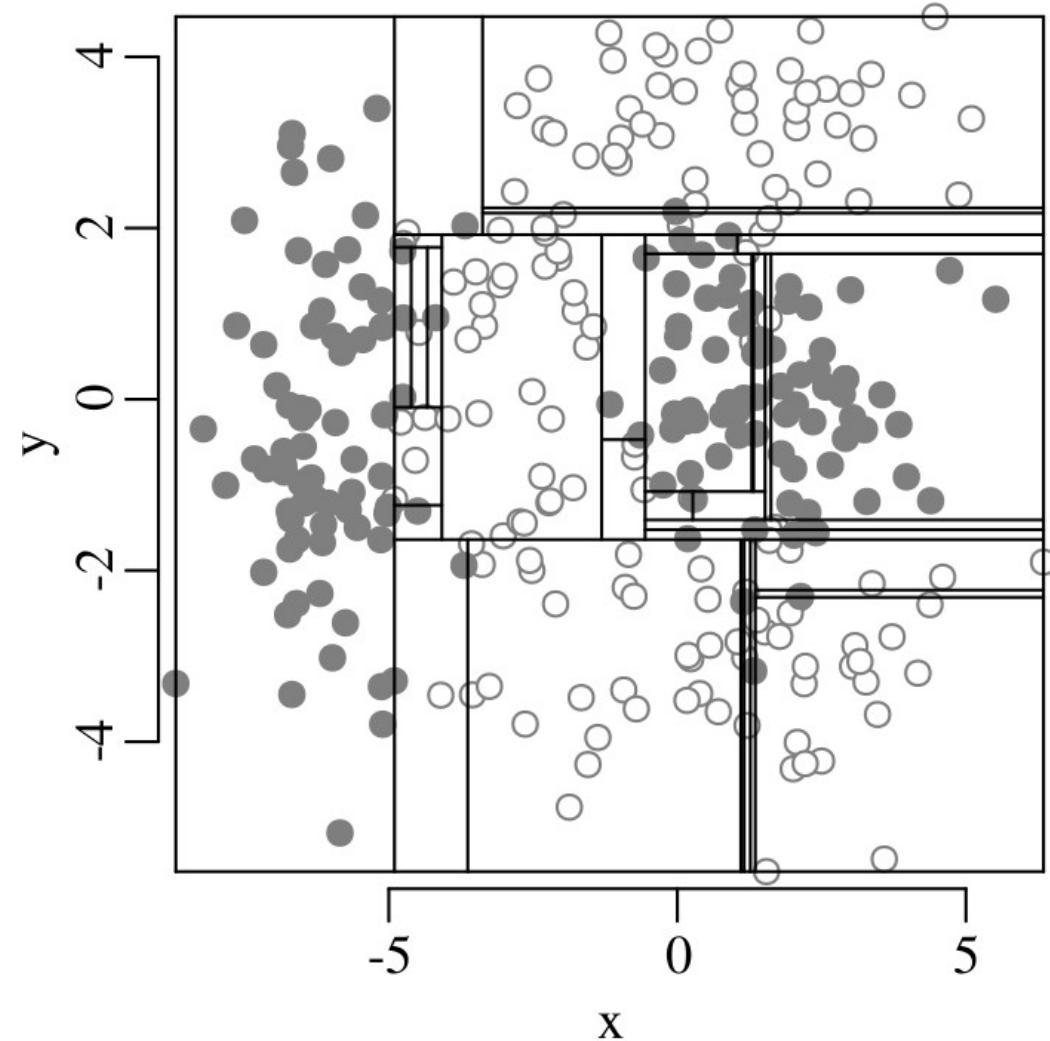
## Decision trees

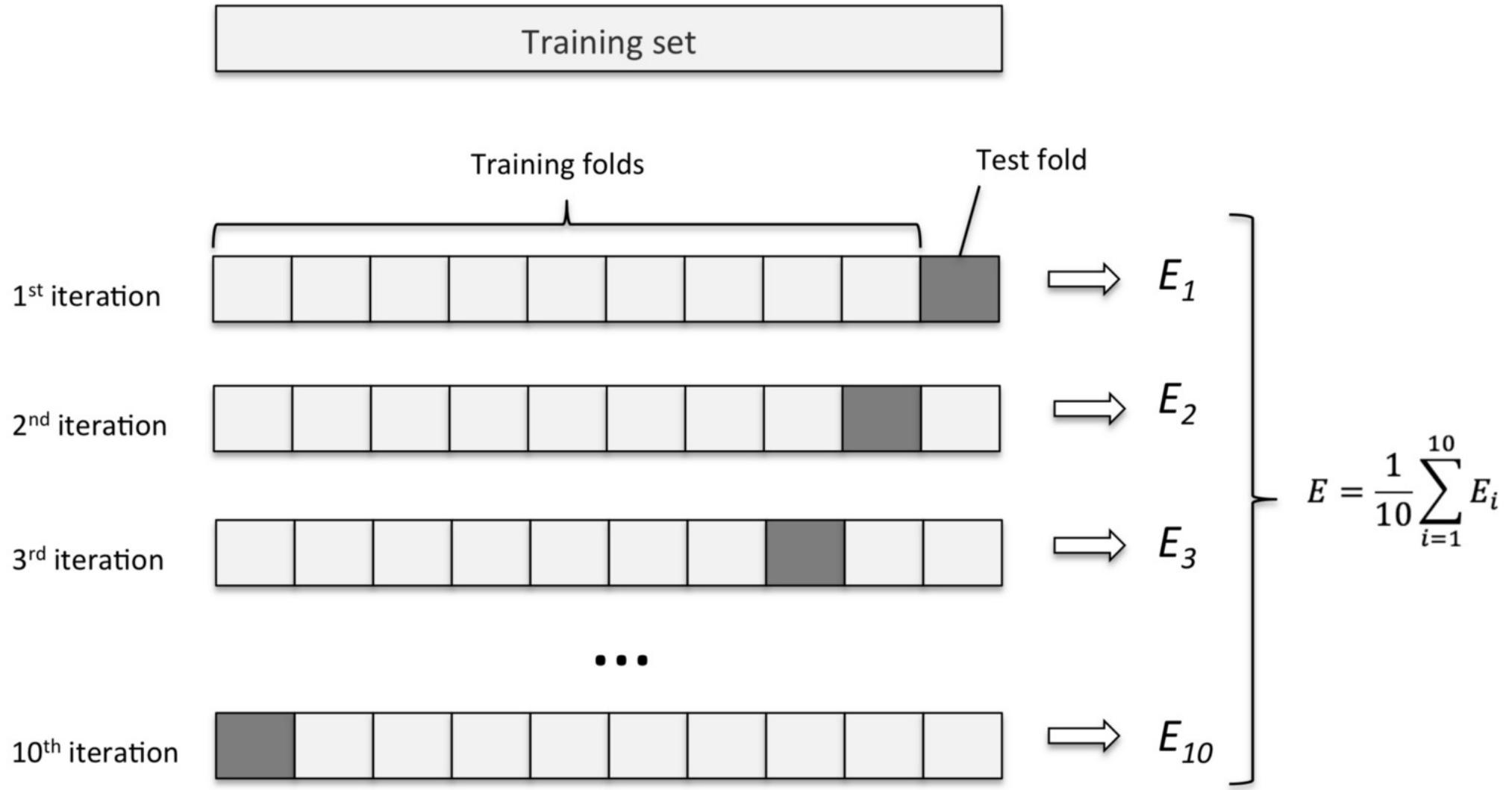


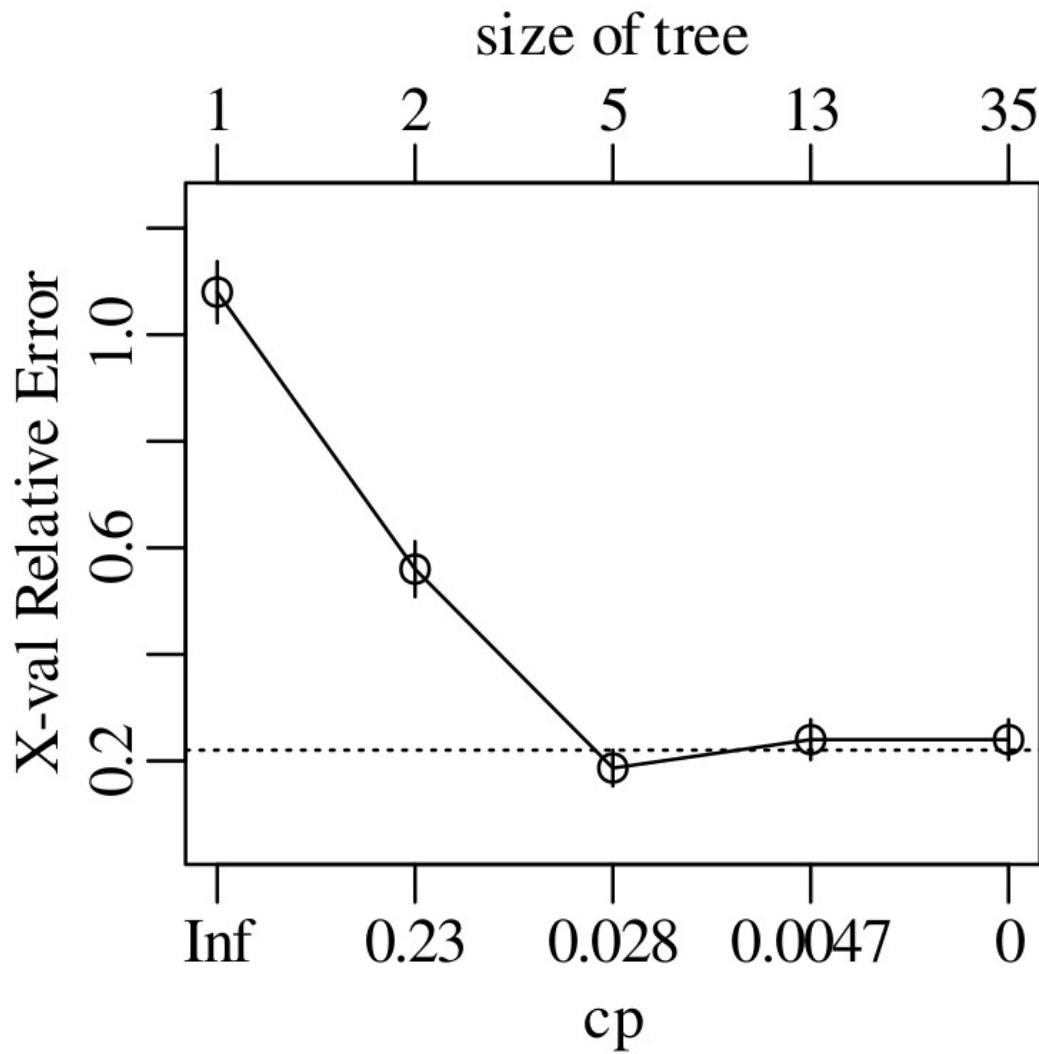
$$Q = p_{\circ}(1 - p_{\circ}) + p_{\bullet}(1 - p_{\bullet})$$



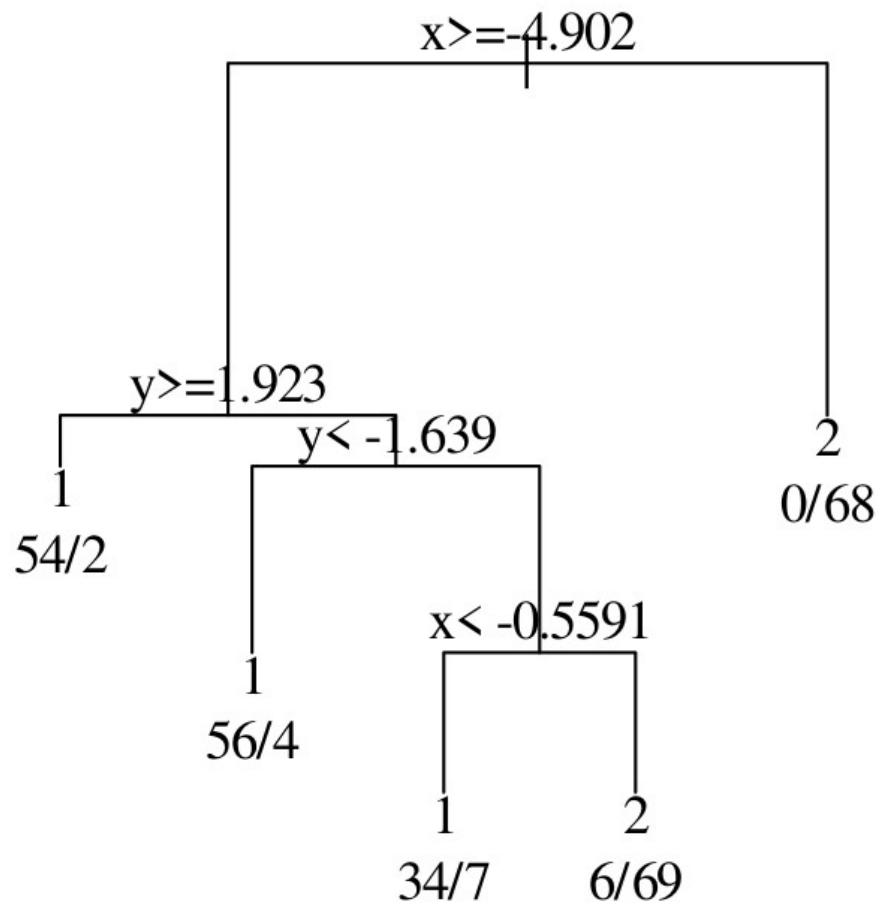
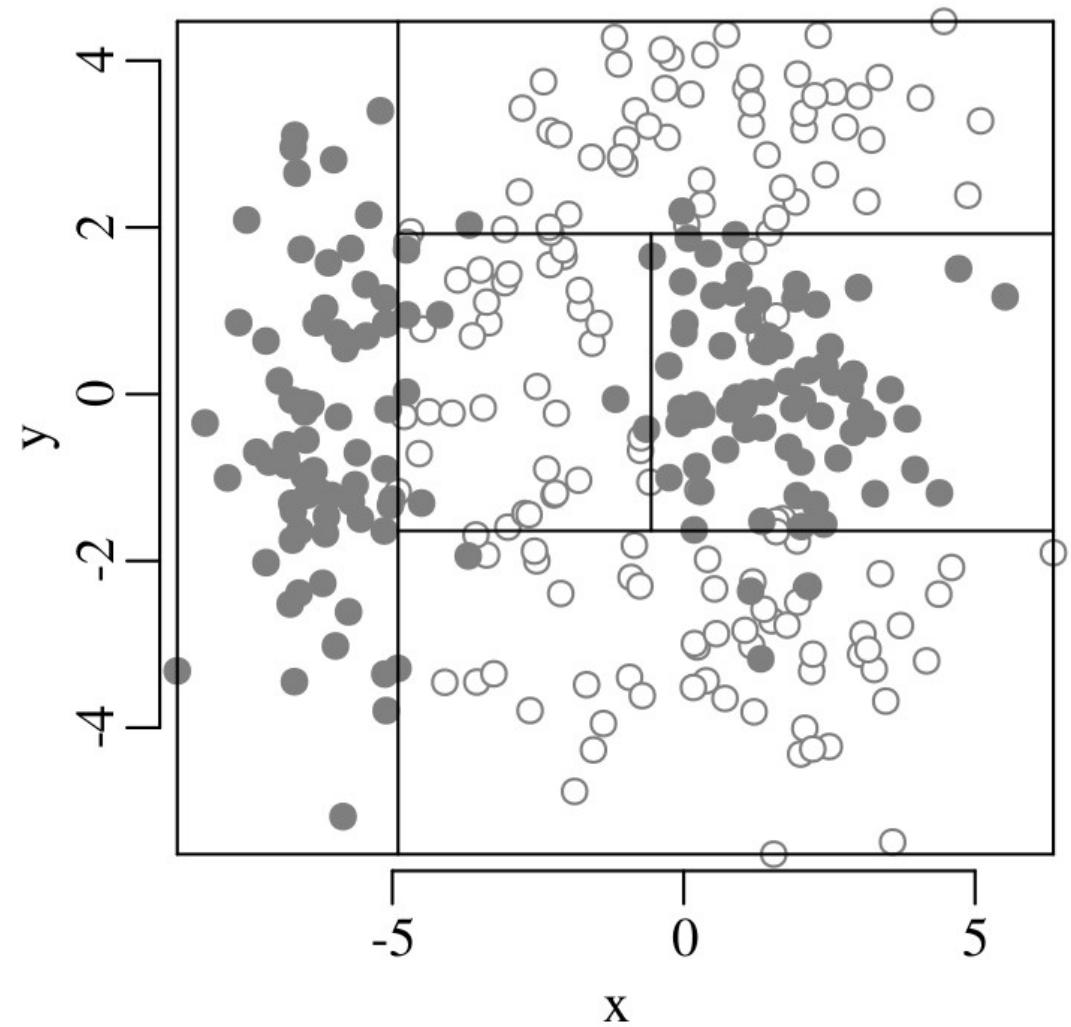


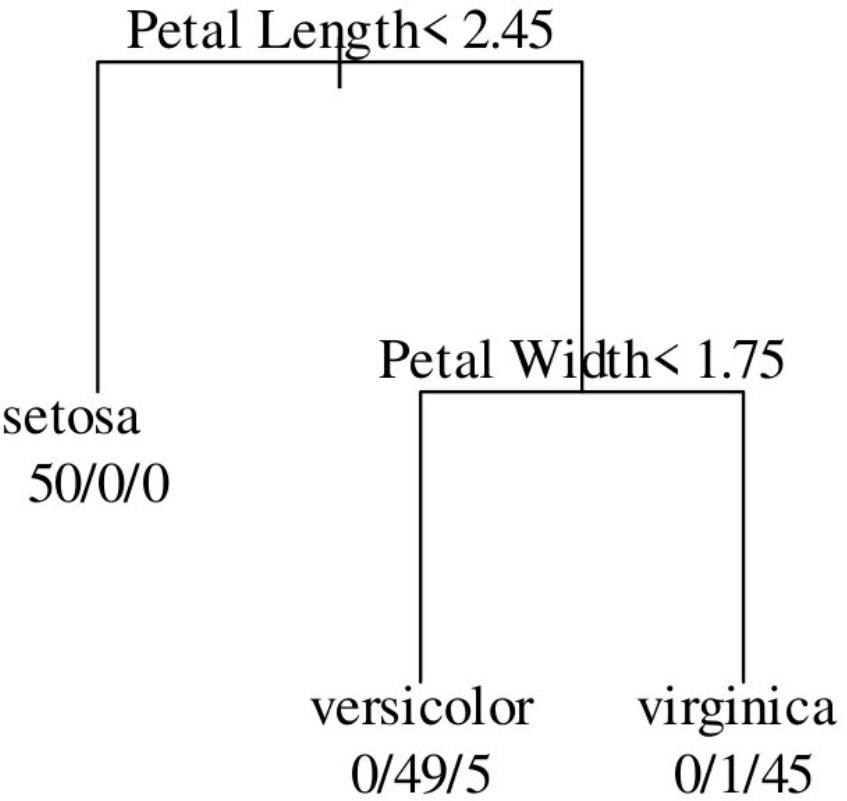
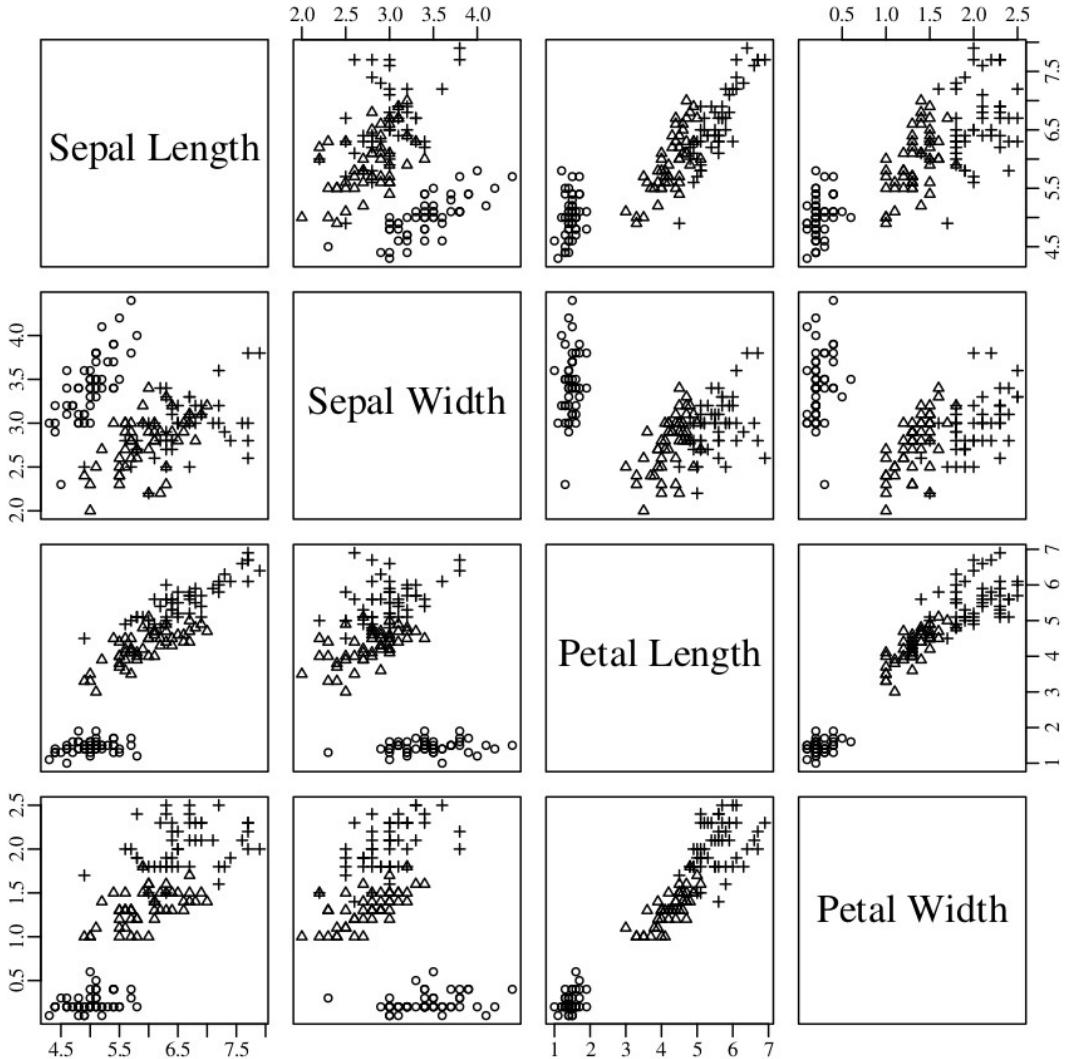






$$cp_\alpha(T) = \sum_{m=1}^{|T|} N_m Q_m + \alpha |T|$$
$$T_\alpha = \min_{T \subset T_0} cp_\alpha(T)$$





# Statistics for geoscientists

Compositional data

## Ratio data

|   | 1    | 2    | 3    | 4    | 5    | 6    | 7    | 8    | 9    | 10    |
|---|------|------|------|------|------|------|------|------|------|-------|
| A | 0.27 | 0.37 | 0.57 | 0.91 | 0.20 | 0.90 | 0.94 | 0.66 | 0.63 | 0.062 |
| B | 0.21 | 0.18 | 0.69 | 0.38 | 0.77 | 0.50 | 0.72 | 0.99 | 0.38 | 0.780 |

|     | 1    | 2    | 3    | 4    | 5    | 6    | 7    | 8    | 9    | 10    |
|-----|------|------|------|------|------|------|------|------|------|-------|
| $A$ | 0.27 | 0.37 | 0.57 | 0.91 | 0.20 | 0.90 | 0.94 | 0.66 | 0.63 | 0.062 |
| $B$ | 0.21 | 0.18 | 0.69 | 0.38 | 0.77 | 0.50 | 0.72 | 0.99 | 0.38 | 0.780 |

|       | 1    | 2    | 3    | 4    | 5    | 6    | 7    | 8    | 9    | 10    |
|-------|------|------|------|------|------|------|------|------|------|-------|
| $A/B$ | 1.30 | 2.10 | 0.83 | 2.40 | 0.26 | 1.80 | 1.30 | 0.67 | 1.70 | 0.079 |
| $B/A$ | 0.78 | 0.47 | 1.20 | 0.42 | 3.80 | 0.55 | 0.76 | 1.50 | 0.60 | 13.0  |

|     | 1    | 2    | 3    | 4    | 5    | 6    | 7    | 8    | 9    | 10    |
|-----|------|------|------|------|------|------|------|------|------|-------|
| $A$ | 0.27 | 0.37 | 0.57 | 0.91 | 0.20 | 0.90 | 0.94 | 0.66 | 0.63 | 0.062 |
| $B$ | 0.21 | 0.18 | 0.69 | 0.38 | 0.77 | 0.50 | 0.72 | 0.99 | 0.38 | 0.780 |

|           | 1    | 2    | 3    | 4    | 5    | 6    | 7    | 8    | 9    | 10    |
|-----------|------|------|------|------|------|------|------|------|------|-------|
| $A/B$     | 1.30 | 2.10 | 0.83 | 2.40 | 0.26 | 1.80 | 1.30 | 0.67 | 1.70 | 0.079 |
| $B/A$     | 0.78 | 0.47 | 1.20 | 0.42 | 3.80 | 0.55 | 0.76 | 1.50 | 0.60 | 13.0  |
| $1/(A/B)$ | 0.78 | 0.47 | 1.20 | 0.42 | 3.80 | 0.55 | 0.76 | 1.50 | 0.60 | 13.0  |

|   | 1    | 2    | 3    | 4    | 5    | 6    | 7    | 8    | 9    | 10    |
|---|------|------|------|------|------|------|------|------|------|-------|
| A | 0.27 | 0.37 | 0.57 | 0.91 | 0.20 | 0.90 | 0.94 | 0.66 | 0.63 | 0.062 |
| B | 0.21 | 0.18 | 0.69 | 0.38 | 0.77 | 0.50 | 0.72 | 0.99 | 0.38 | 0.780 |

|           | 1    | 2    | 3    | 4    | 5    | 6    | 7    | 8    | 9    | 10    | mean |
|-----------|------|------|------|------|------|------|------|------|------|-------|------|
| $A/B$     | 1.30 | 2.10 | 0.83 | 2.40 | 0.26 | 1.80 | 1.30 | 0.67 | 1.70 | 0.079 | 1.20 |
| $B/A$     | 0.78 | 0.47 | 1.20 | 0.42 | 3.80 | 0.55 | 0.76 | 1.50 | 0.60 | 13.0  | 2.30 |
| $1/(A/B)$ | 0.78 | 0.47 | 1.20 | 0.42 | 3.80 | 0.55 | 0.76 | 1.50 | 0.60 | 13.0  |      |

$$\frac{1}{\overline{A/B}} = \frac{1}{1.20} = 0.81 \neq 2.30 = \overline{B/A}$$

and  $\frac{1}{\overline{B/A}} = \frac{1}{2.30} = 0.44 \neq 1.20 = \overline{A/B}$

|   | 1    | 2    | 3    | 4    | 5    | 6    | 7    | 8    | 9    | 10    |
|---|------|------|------|------|------|------|------|------|------|-------|
| A | 0.27 | 0.37 | 0.57 | 0.91 | 0.20 | 0.90 | 0.94 | 0.66 | 0.63 | 0.062 |
| B | 0.21 | 0.18 | 0.69 | 0.38 | 0.77 | 0.50 | 0.72 | 0.99 | 0.38 | 0.780 |

|            | 1     | 2     | 3     | 4     | 5     | 6     | 7     | 8     | 9     | 10    |
|------------|-------|-------|-------|-------|-------|-------|-------|-------|-------|-------|
| $\ln[A/B]$ | 0.25  | 0.75  | -0.18 | 0.86  | -1.30 | 0.59  | 0.27  | -0.41 | 0.50  | -2.50 |
| $\ln[B/A]$ | -0.25 | -0.75 | 0.18  | -0.86 | 1.30  | -0.59 | -0.27 | 0.41  | -0.50 | 2.50  |

|            | mean  | $\exp[\text{mean}]$ |
|------------|-------|---------------------|
| $\ln[A/B]$ | -0.12 | 0.88                |
| $\ln[B/A]$ | 0.12  | 1.13                |

$$\frac{1}{g(A/B)} = \frac{1}{0.88} = 1.13 = g(B/A)$$

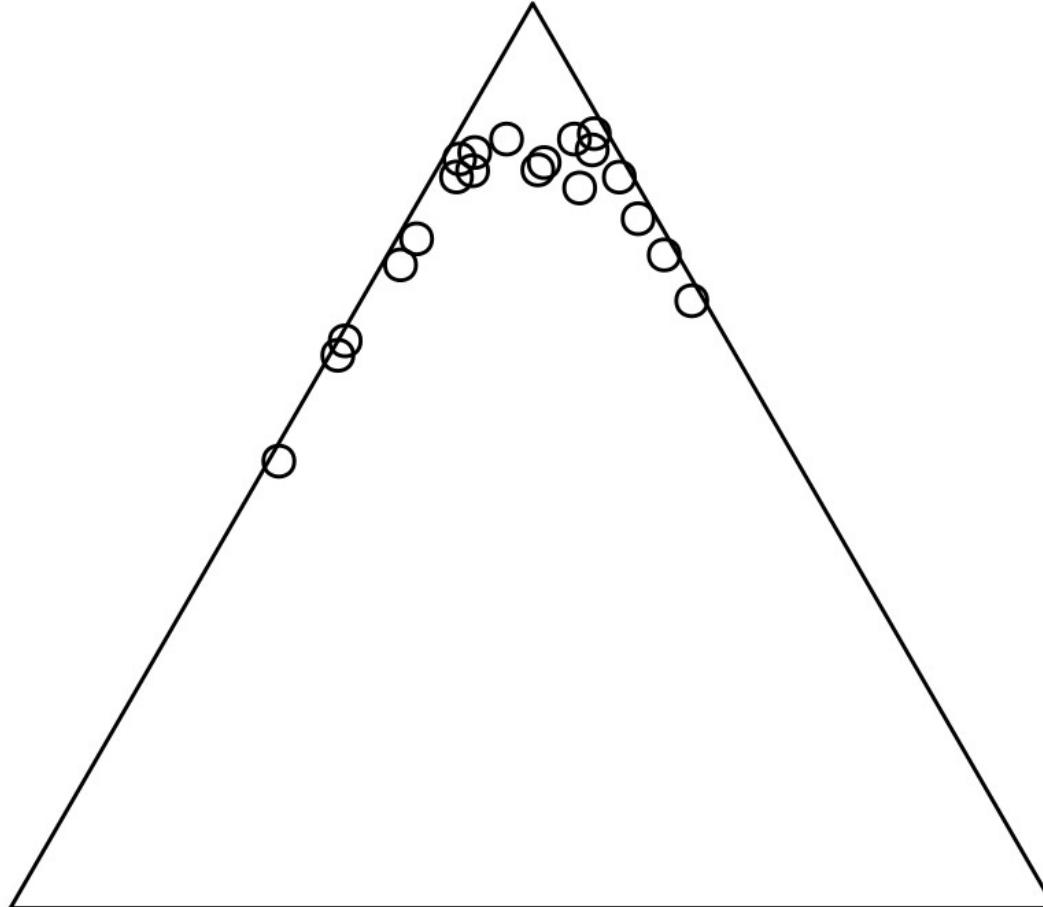
and  $\frac{1}{g(B/A)} = \frac{1}{1.13} = 0.88 = g(A/B)$

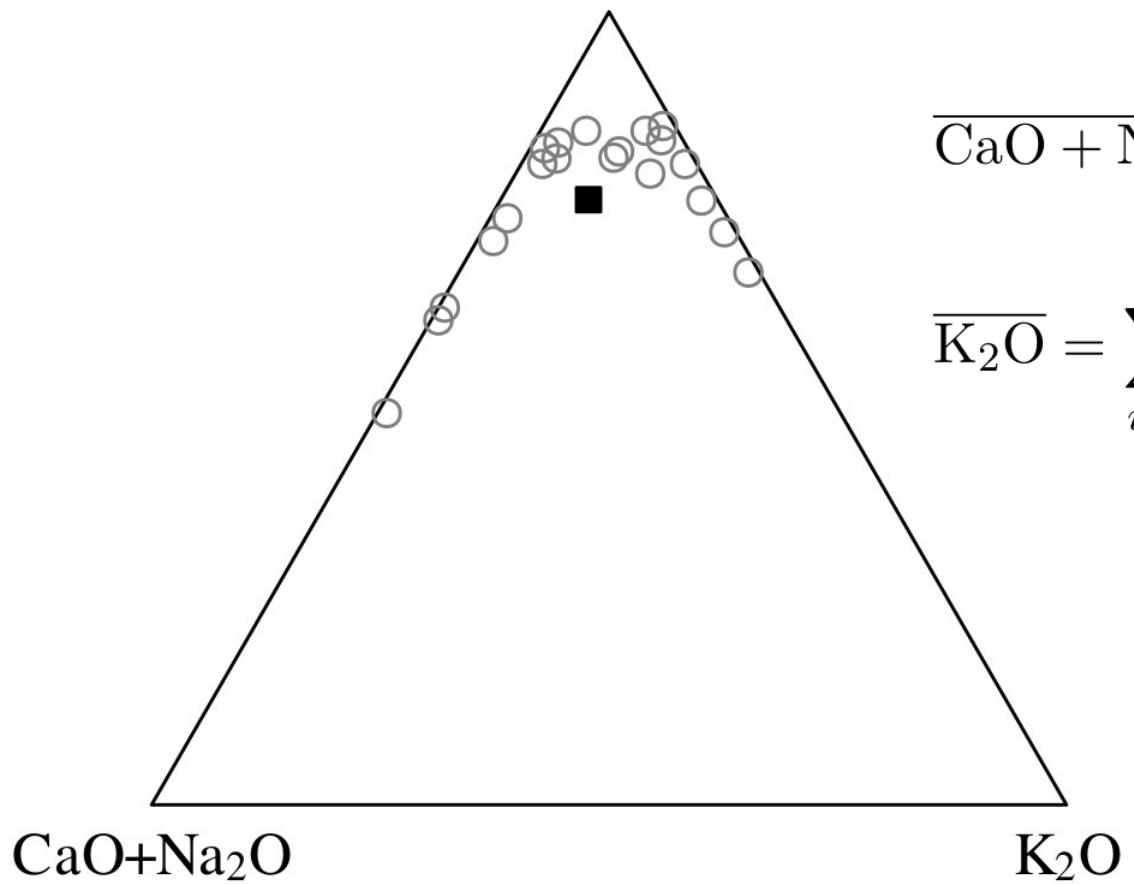
$$\sum_{i=1}^n C_i = 1$$

Al<sub>2</sub>O<sub>3</sub>

CaO+Na<sub>2</sub>O

K<sub>2</sub>O



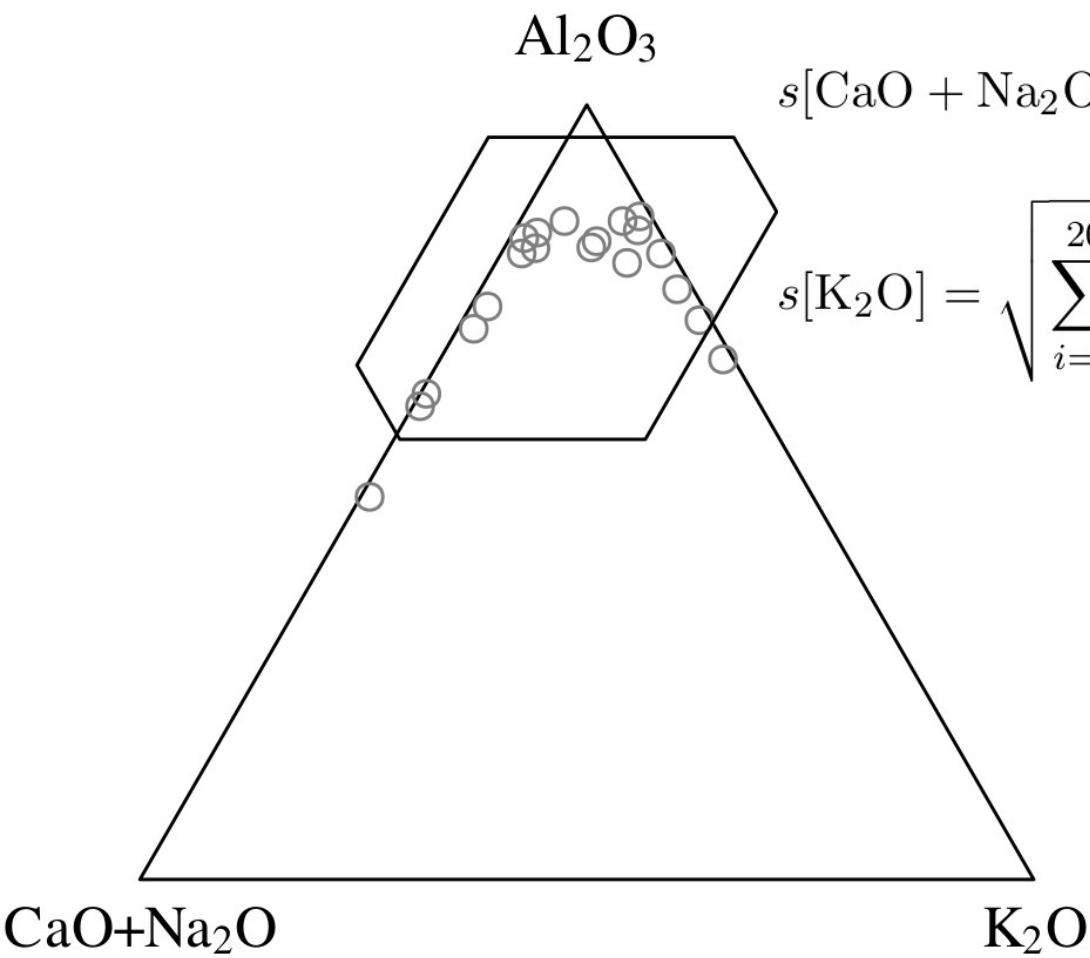


$$\overline{\text{Al}_2\text{O}_3} = \sum_{i=1}^{20} (\text{Al}_2\text{O}_3)_i / 20 = 0.763$$

$$\overline{\text{CaO} + \text{Na}_2\text{O}} = \sum_{i=1}^{20} (\text{CaO}_2 + \text{Na}_2\text{O})_i / 20 = 0.141$$

$$\overline{\text{K}_2\text{O}} = \sum_{i=1}^{20} (\text{K}_2\text{O})_i / 20 = 0.096$$

$$s[\text{Al}_2\text{O}_3] = \sqrt{\sum_{i=1}^{20} \frac{((\text{Al}_2\text{O}_3)_i - 0.763)^2}{19}} = 0.0975$$



$$s[\text{CaO} + \text{Na}_2\text{O}] = \sqrt{\sum_{i=1}^{20} \frac{((\text{CaO}_2 + \text{Na}_2\text{O})_i - 0.141)^2}{19}} = 0.142$$

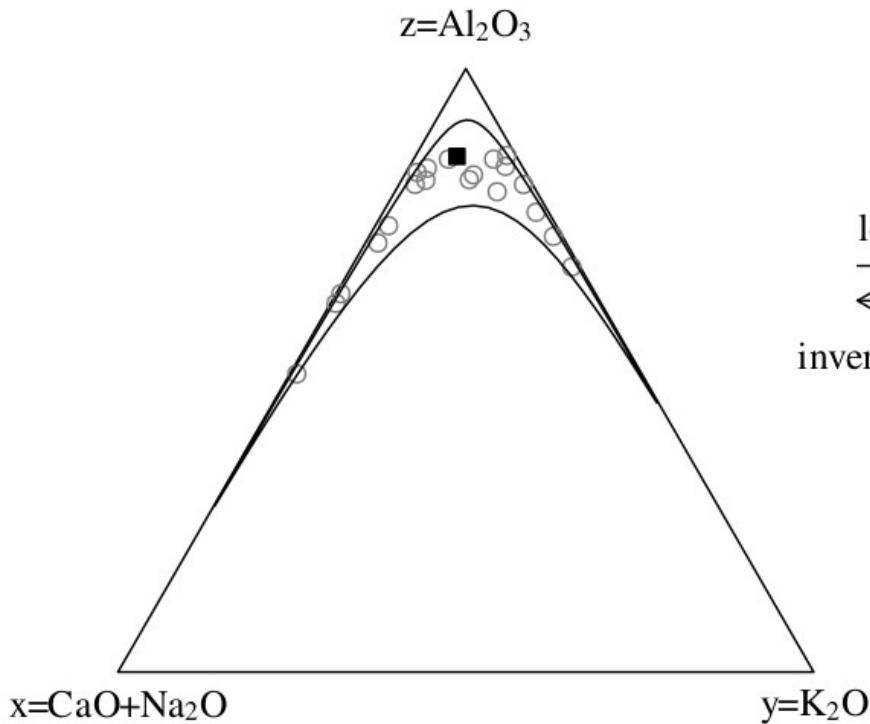
$$s[\text{K}_2\text{O}] = \sqrt{\sum_{i=1}^{20} \frac{((\text{K}_2\text{O})_i - 0.096)^2}{19}} = 0.0926$$

$\text{Al}_2\text{O}_3 : 0.763 \pm 0.195$

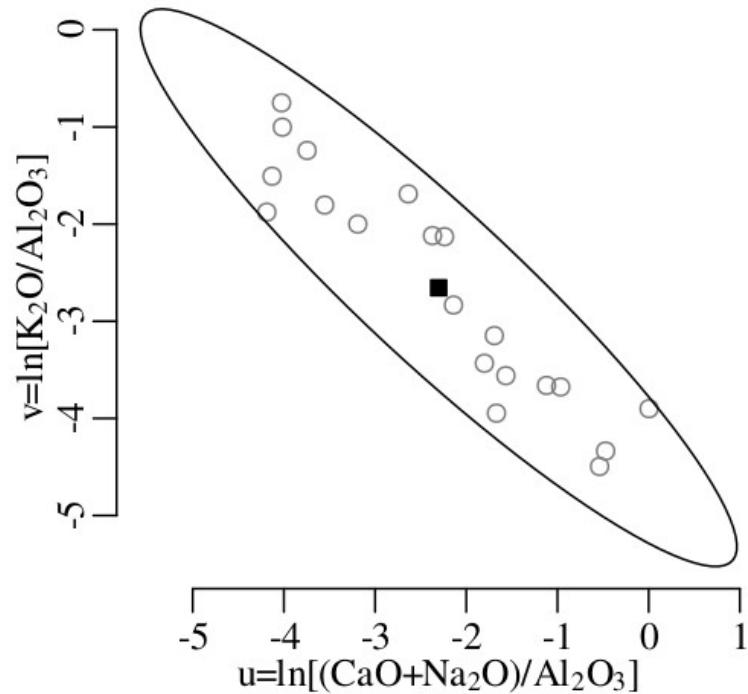
$\text{CaO} + \text{Na}_2\text{O} : 0.141 \pm 0.284$

$\text{K}_2\text{O} : 0.096 \pm 0.185$

$$v = \ln \left[ \frac{x}{z} \right], w = \ln \left[ \frac{y}{z} \right]$$



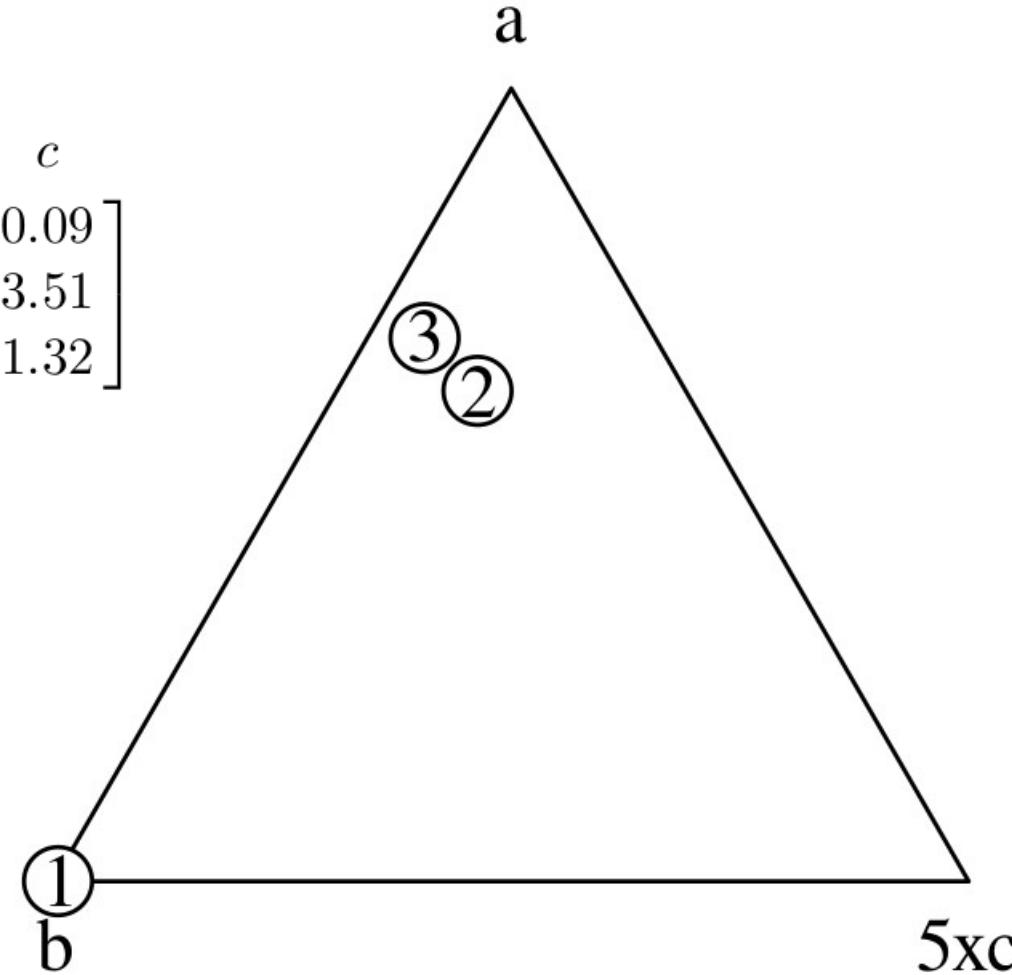
logratio transformation  
 ← →  
 inverse logratio transformation



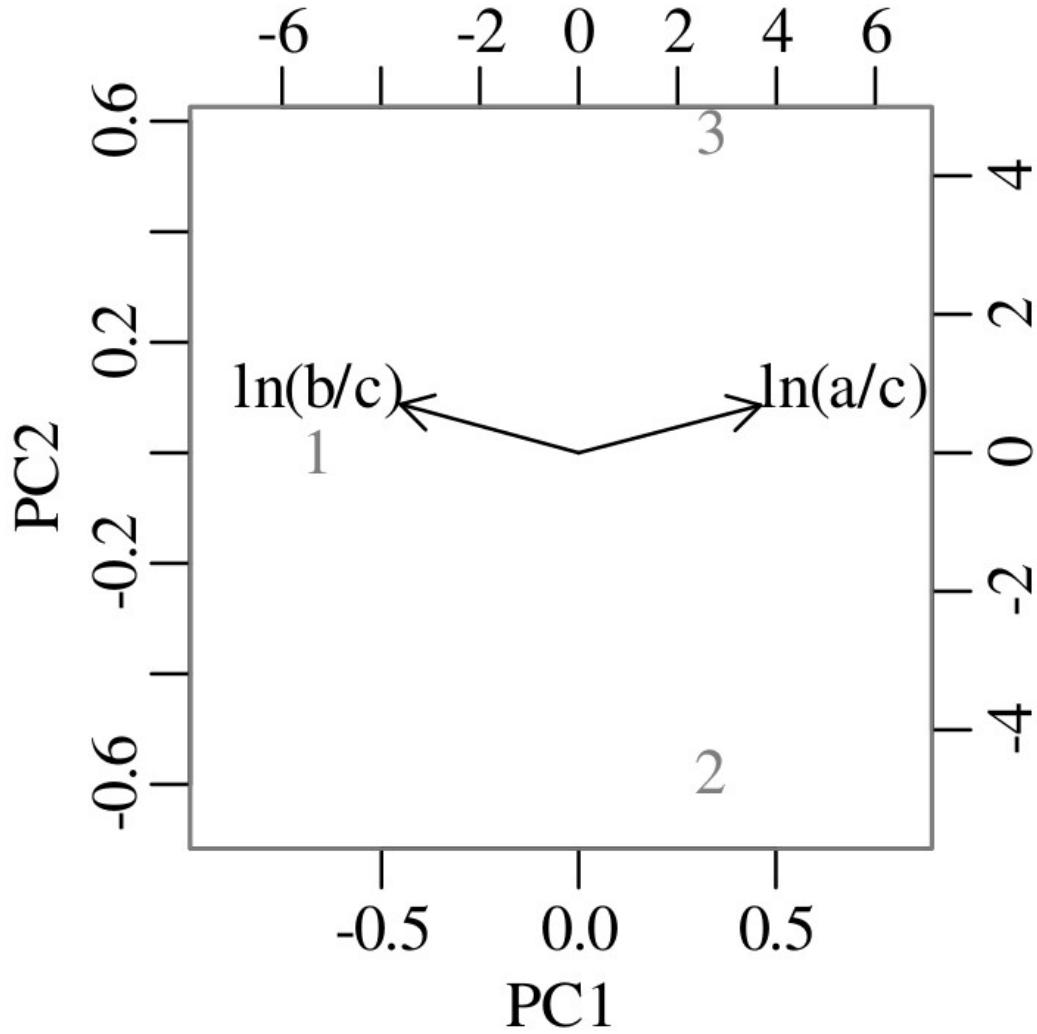
$$x = \frac{\exp[v]}{\exp[v] + \exp[w] + 1}, y = \frac{\exp[w]}{\exp[v] + \exp[w] + 1}, z = \frac{1}{\exp[v] + \exp[w] + 1}$$

# PCA of compositional data

$$X = \begin{matrix} & a & b & c \\ 1 & [0.03 & 99.88 & 0.09] \\ 2 & [70.54 & 25.95 & 3.51] \\ 3 & [72.14 & 26.54 & 1.32] \end{matrix}$$



$$X_a = \begin{bmatrix} \ln(a/c) & \ln(b/c) \\ 1 & -1 \\ 2 & 3 \\ 3 & 4 \end{bmatrix}$$



## centred logratio transformation

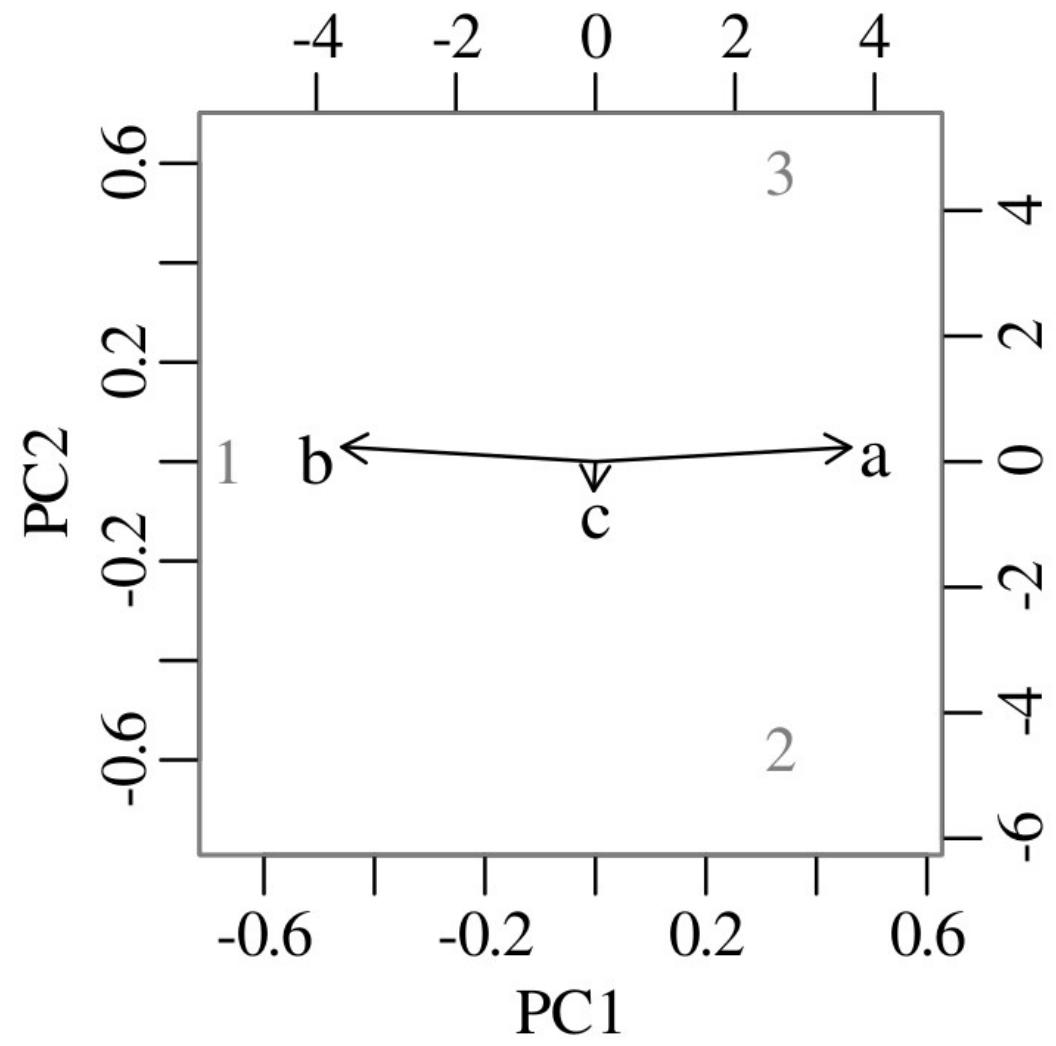
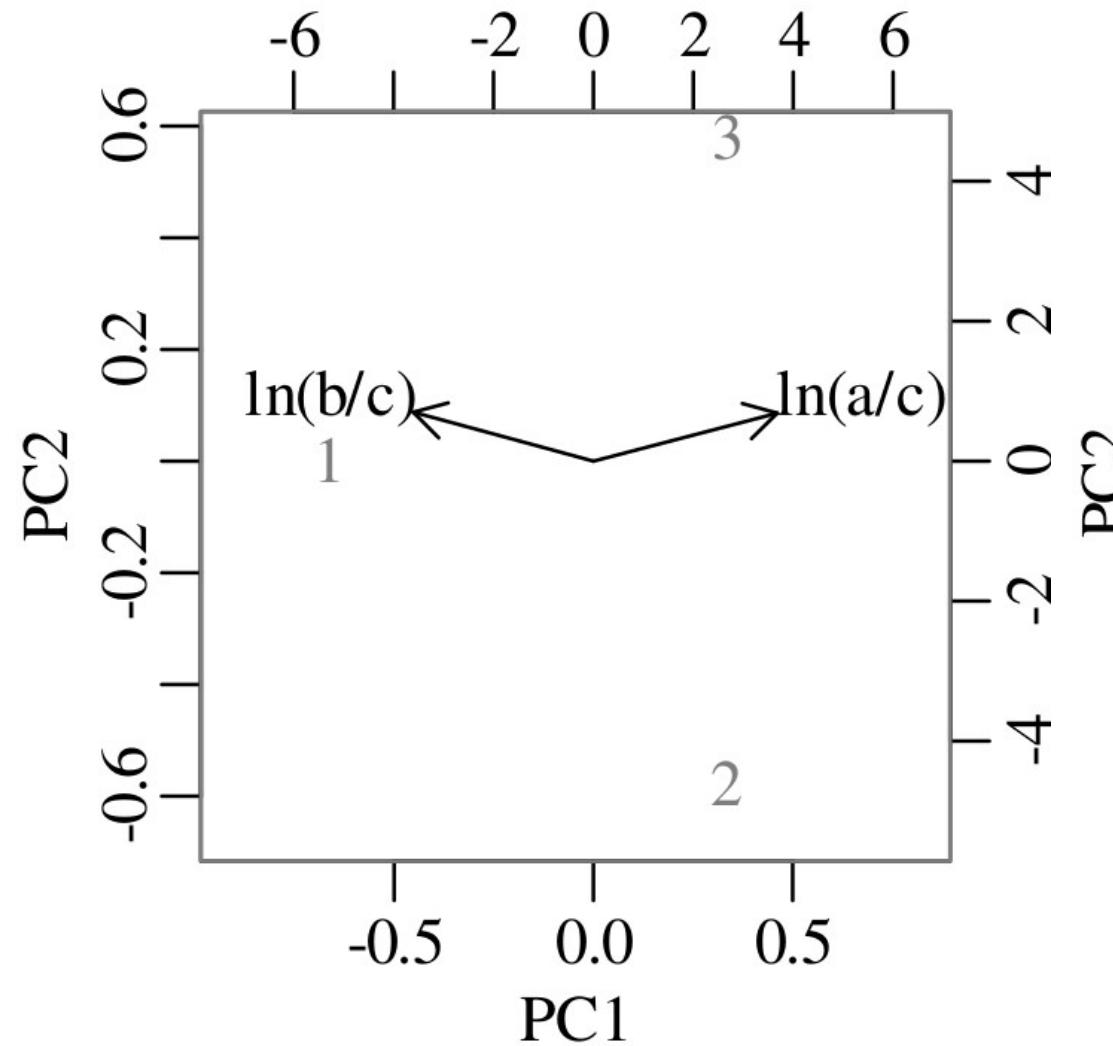
$$u_i = \ln \left[ \frac{x_i}{g_i} \right], v_i = \ln \left[ \frac{y_i}{g_i} \right], \text{ and } w_i = \ln \left[ \frac{z_i}{g_i} \right]$$

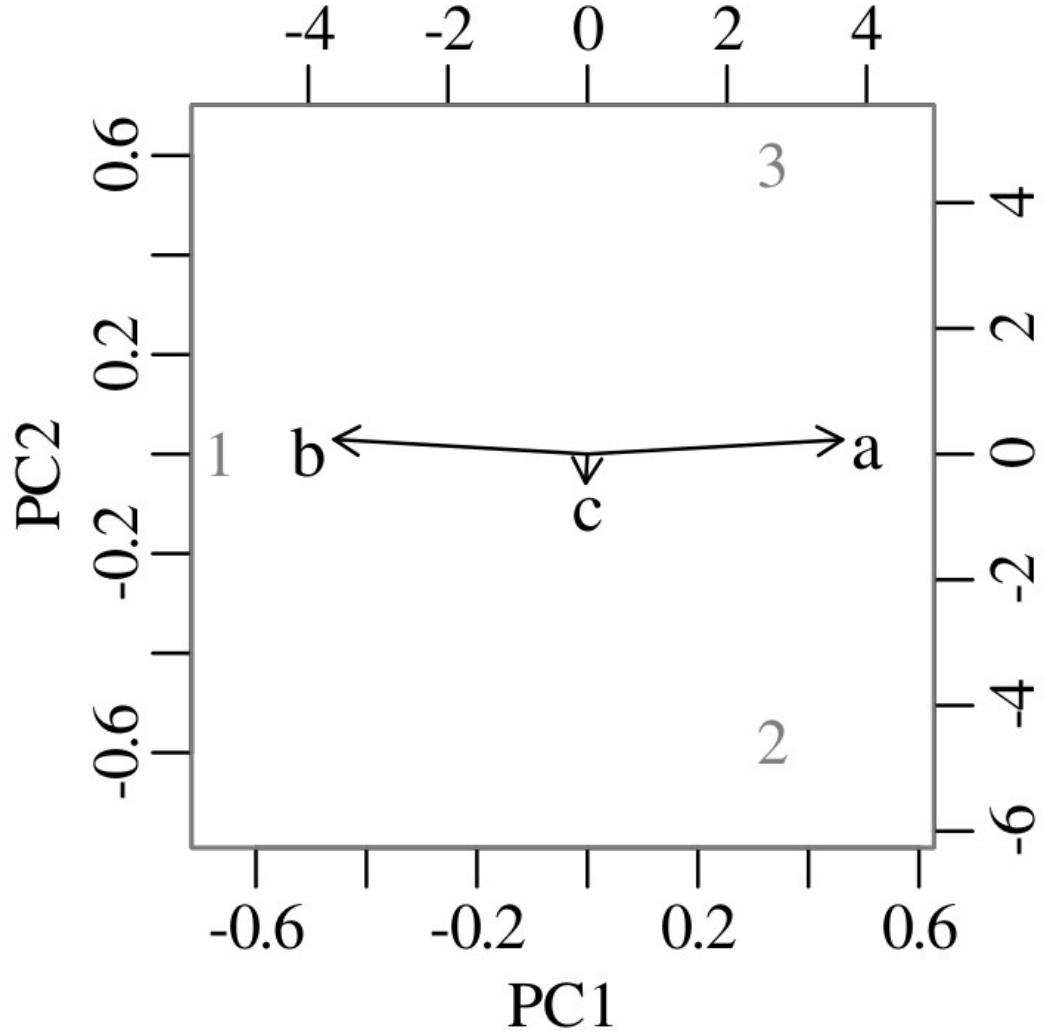
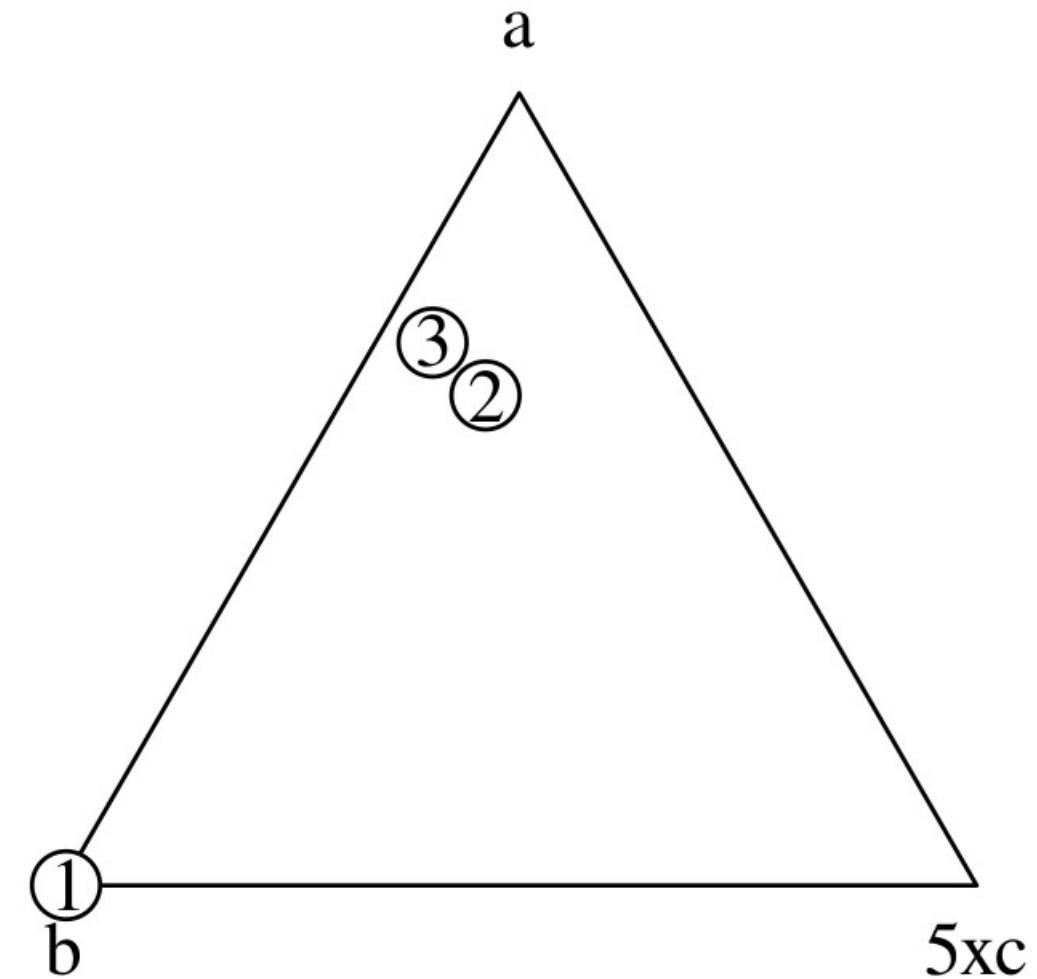
$$\text{where } g_i = \exp \left[ \frac{\ln[x_i] + \ln[y_i] + \ln[z_i]}{3} \right]$$

$$X_c = \begin{matrix} & \ln(a/g) & \ln(b/g) & \ln(c/g) \\ \begin{matrix} 1 \\ 2 \\ 3 \end{matrix} & \begin{bmatrix} -3 & 5 & -2 \\ 1.33 & 0.33 & -1.67 \\ 1.67 & 0.67 & -2.33 \end{bmatrix} \end{matrix}$$

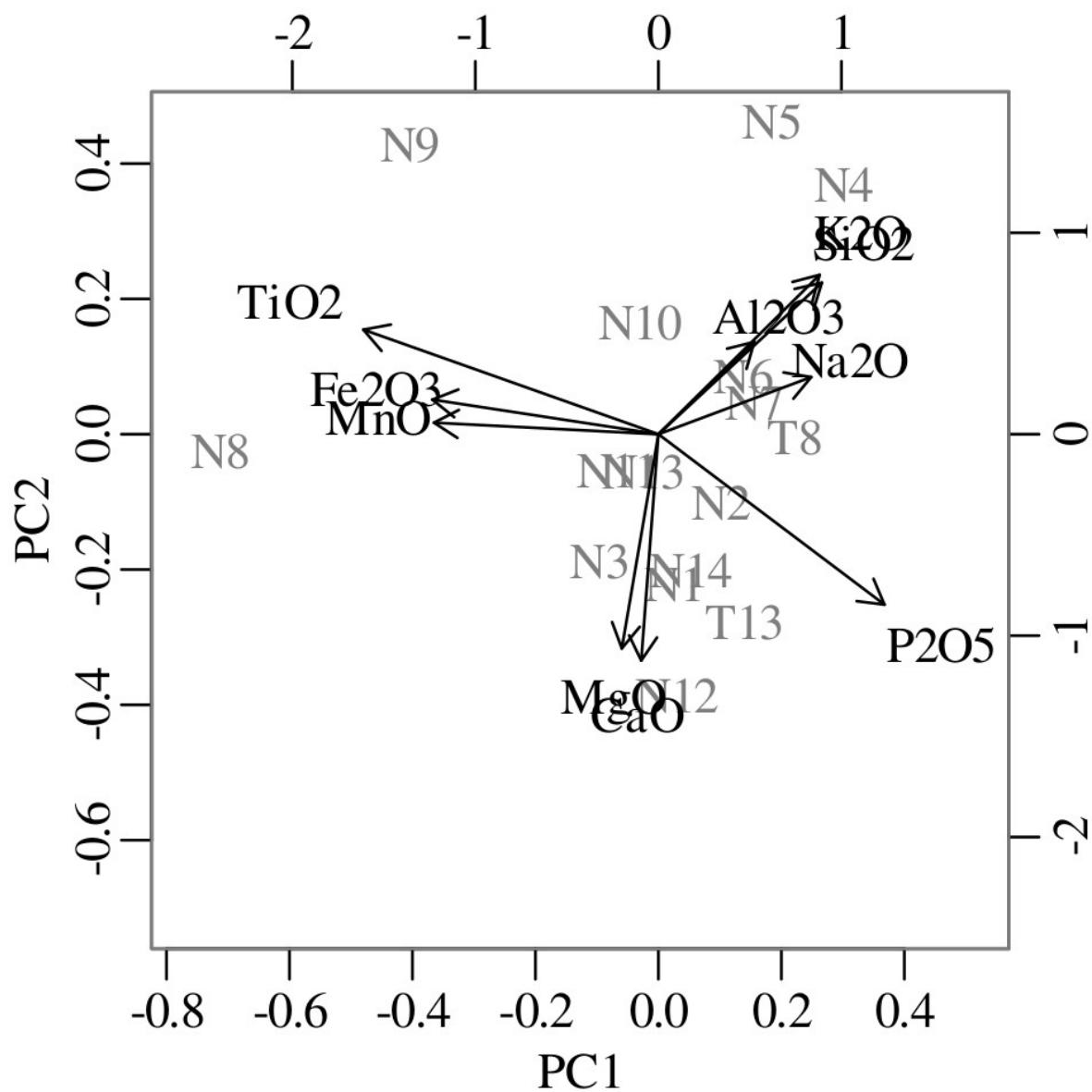
$$X_c = \begin{matrix} & \ln(a/g) & \ln(b/g) & \ln(c/g) \\ \begin{matrix} 1 \\ 2 \\ 3 \end{matrix} & \begin{bmatrix} -3 & 5 & -2 \\ 1.33 & 0.33 & -1.67 \\ 1.67 & 0.67 & -2.33 \end{bmatrix} \end{matrix}$$

$$X_c = 1_{3,1} \ C + S \ V \ D = \begin{bmatrix} 1 \\ 1 \\ 1 \end{bmatrix} \begin{bmatrix} 0 & 2 & -2 \end{bmatrix} + \begin{bmatrix} -1.15 & 0 & 0.82 \\ 0.58 & -1 & 0.82 \\ 0.58 & 1 & 0.82 \end{bmatrix} \begin{bmatrix} 3.67 & 0 & 0 \\ 0 & 0.41 & 0 \\ 0 & 0 & 0 \end{bmatrix} \begin{bmatrix} 0.71 & -0.71 & 0 \\ 0.41 & 0.41 & -0.82 \\ 0.58 & 0.58 & 0.58 \end{bmatrix}$$

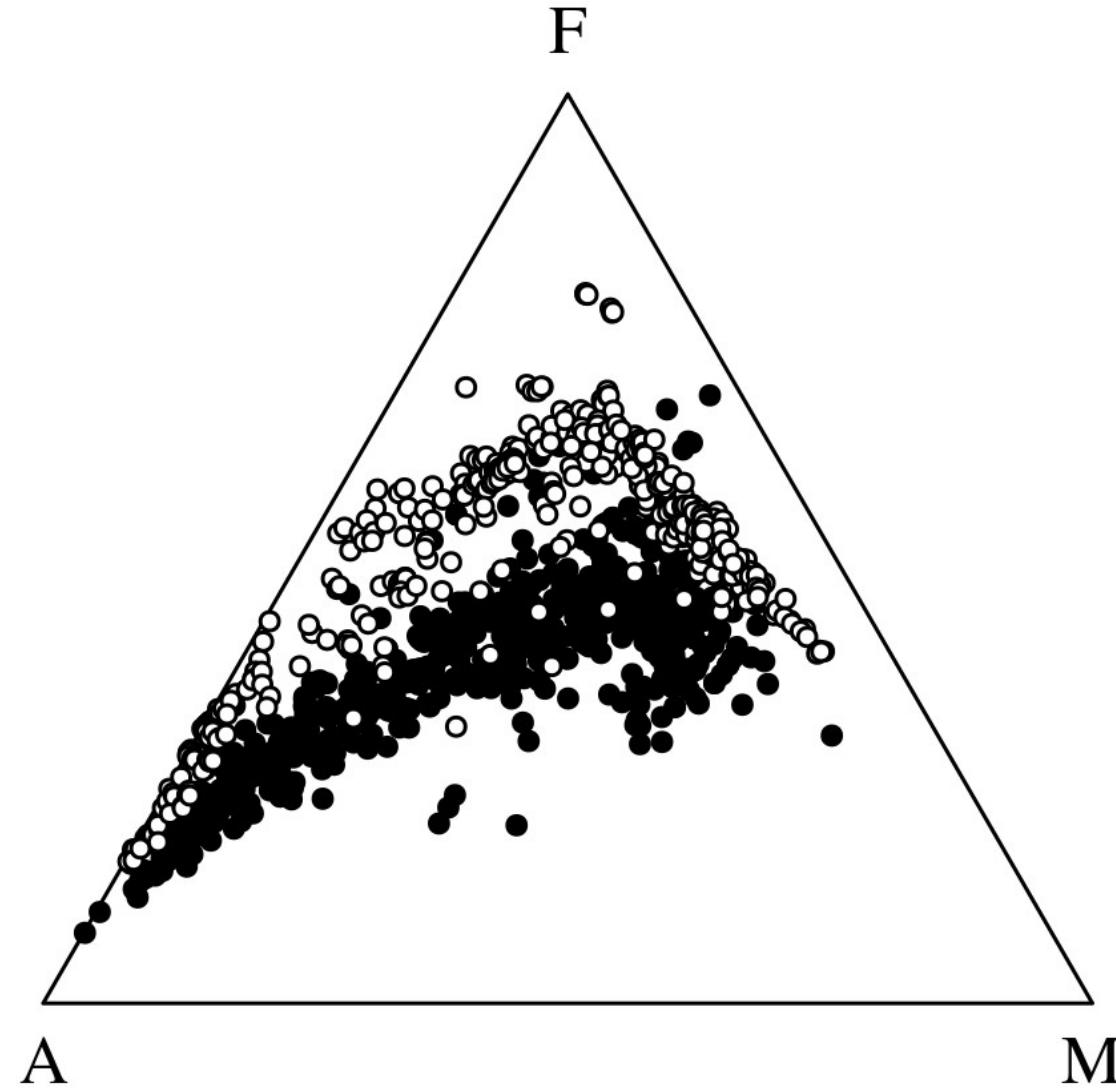


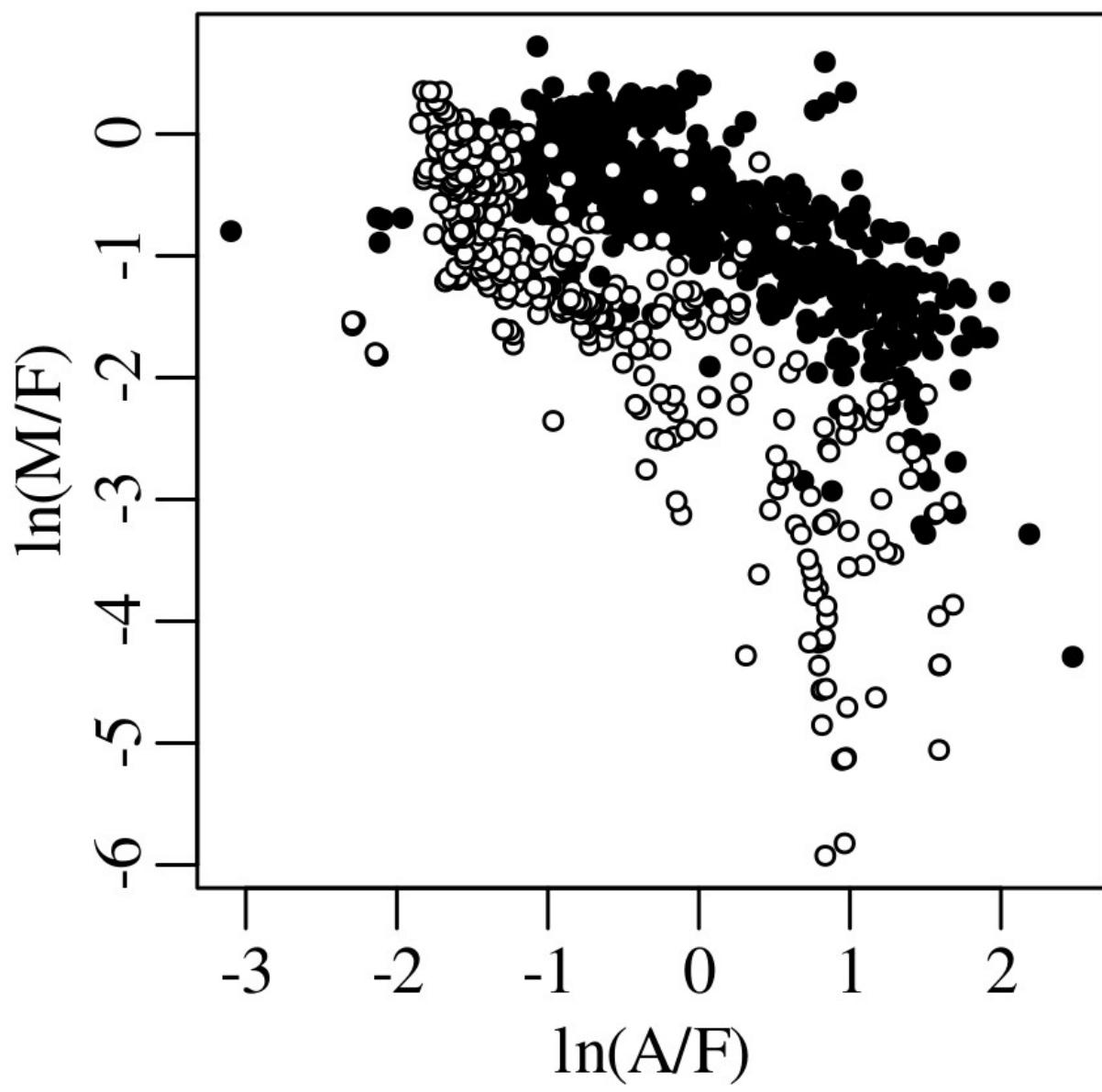


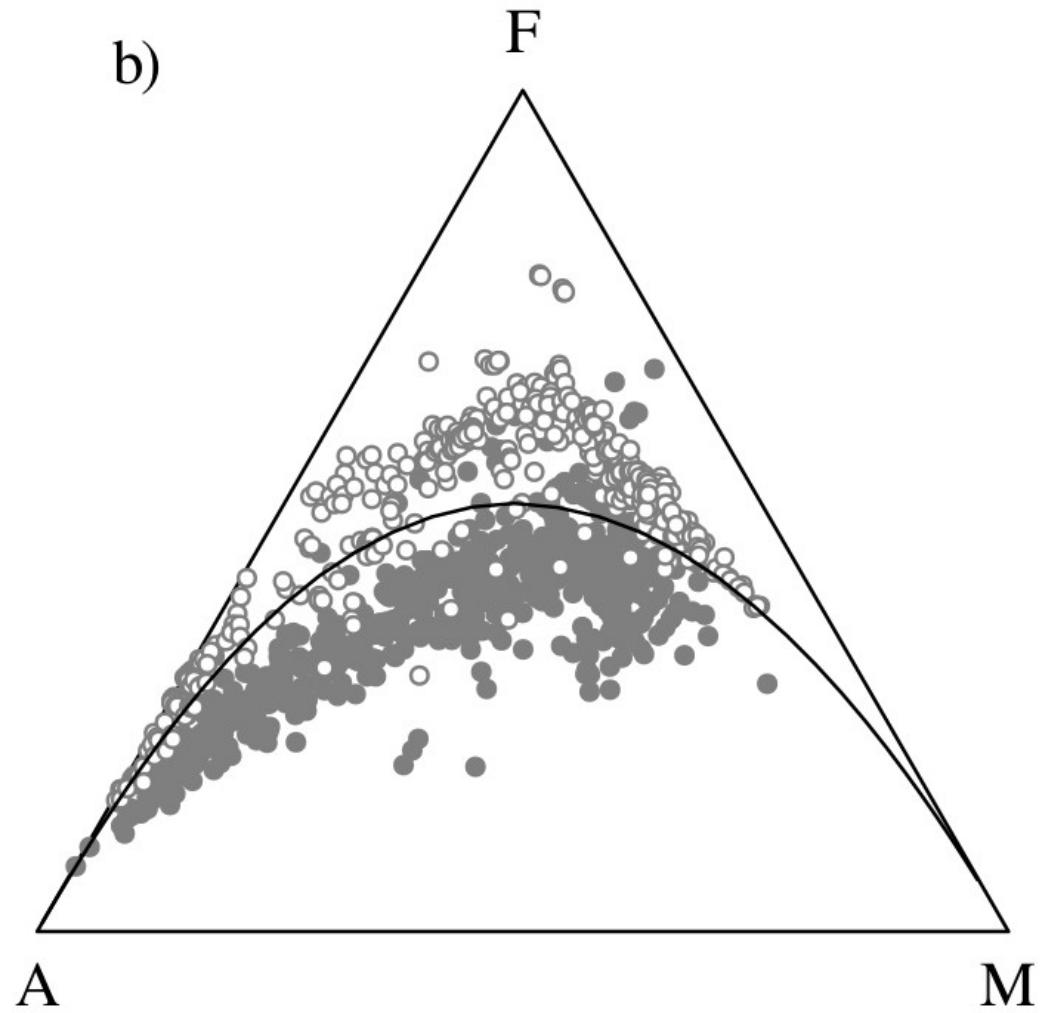
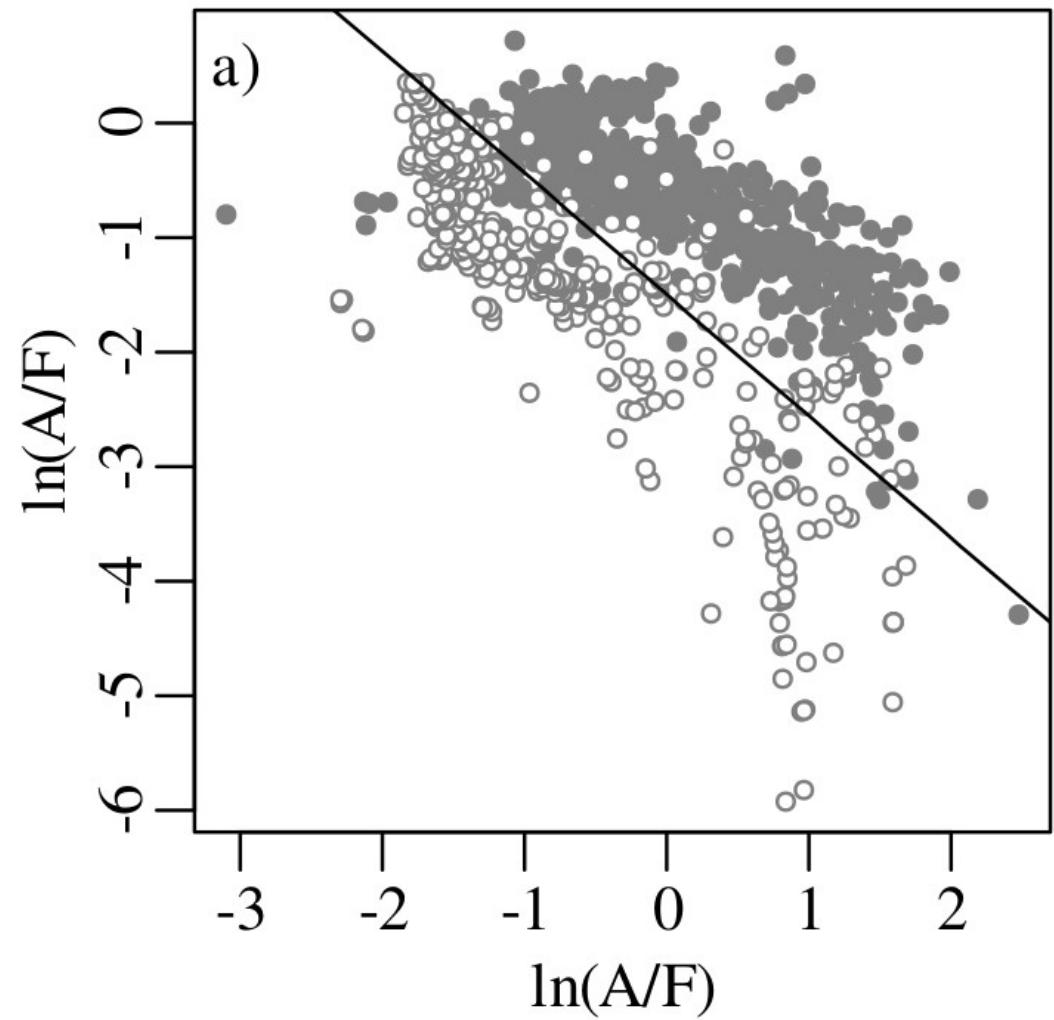
|     | SiO <sub>2</sub> | Al <sub>2</sub> O <sub>3</sub> | Fe <sub>2</sub> O <sub>3</sub> | MgO  | CaO  | Na <sub>2</sub> O | K <sub>2</sub> O | TiO <sub>2</sub> | P <sub>2</sub> O <sub>5</sub> | MnO  |
|-----|------------------|--------------------------------|--------------------------------|------|------|-------------------|------------------|------------------|-------------------------------|------|
| N1  | 82.54            | 6.14                           | 3.18                           | 1.65 | 2.24 | 1.16              | 1.35             | 0.44             | 0.11                          | 0.06 |
| N2  | 83.60            | 6.42                           | 2.55                           | 1.25 | 1.83 | 1.21              | 1.48             | 0.36             | 0.08                          | 0.04 |
| :   | :                | :                              | :                              | :    | :    | :                 | :                | :                | :                             | :    |
| N13 | 79.96            | 6.41                           | 3.19                           | 1.31 | 2.10 | 1.09              | 1.62             | 0.51             | 0.06                          | 0.05 |
| N14 | 73.62            | 9.96                           | 4.07                           | 2.01 | 3.45 | 1.86              | 2.29             | 0.44             | 0.10                          | 0.07 |
| T8  | 85.70            | 5.89                           | 1.82                           | 0.81 | 1.44 | 1.12              | 1.67             | 0.23             | 0.07                          | 0.03 |
| T13 | 82.54            | 6.02                           | 2.30                           | 1.42 | 3.07 | 1.19              | 1.46             | 0.34             | 0.11                          | 0.04 |



# LDA of compositional data



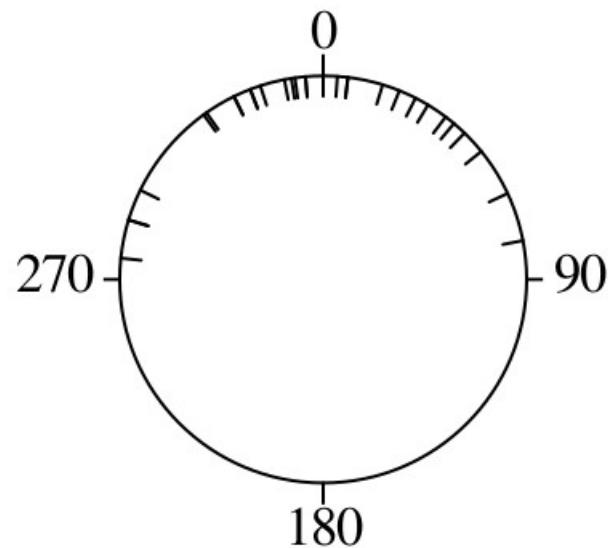




# Statistics for geoscientists

Directional data

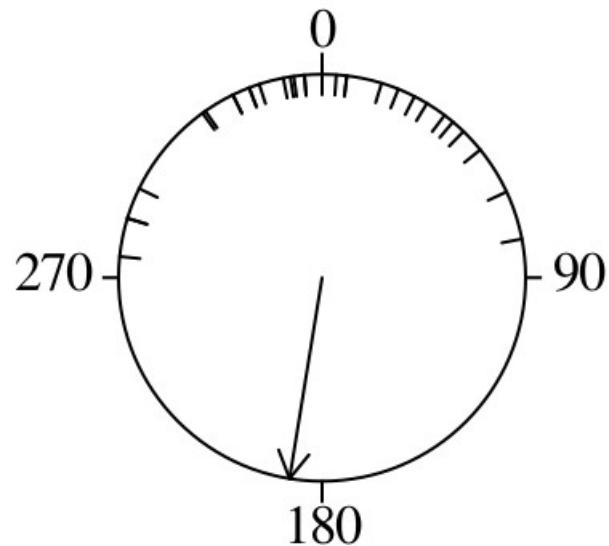
|     |     |     |     |     |    |    |     |     |     |     |    |     |     |     |
|-----|-----|-----|-----|-----|----|----|-----|-----|-----|-----|----|-----|-----|-----|
| 44  | 51  | 79  | 65  | 27  | 31 | 4  | 355 | 22  | 352 | 287 | 7  | 287 | 339 | 0   |
| 276 | 342 | 355 | 334 | 296 | 7  | 17 | 351 | 349 | 37  | 339 | 40 | 324 | 325 | 334 |



$$\begin{array}{r} 44 + 51 + 79 + 65 + 27 + 31 + 4 + 355 + 22 + 352 + 287 + 7 + 287 + 339 + 0 \\ 276 + 342 + 355 + 334 + 296 + 7 + 17 + 351 + 349 + 37 + 339 + 40 + 324 + 325 + 334 \end{array}$$

30

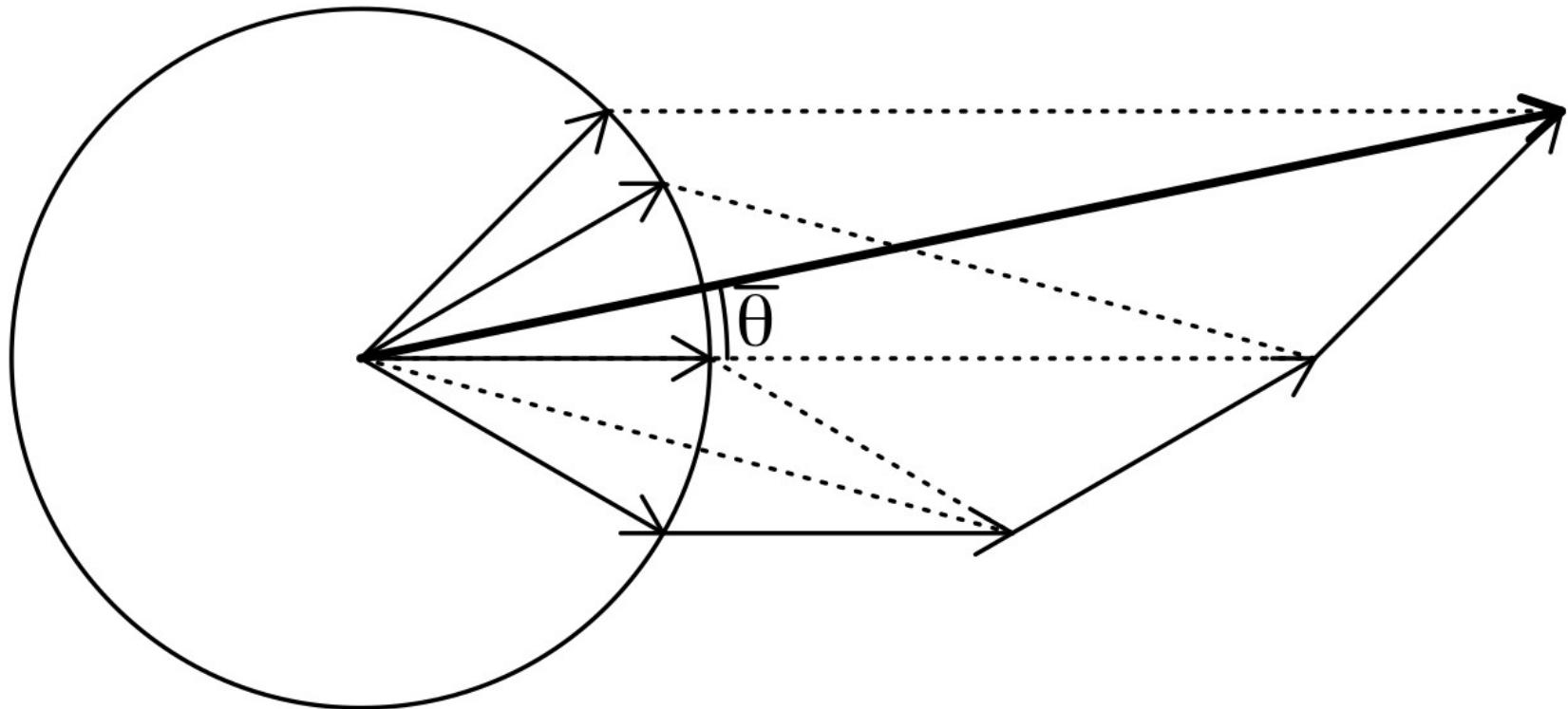
$$= 189.2^\circ$$



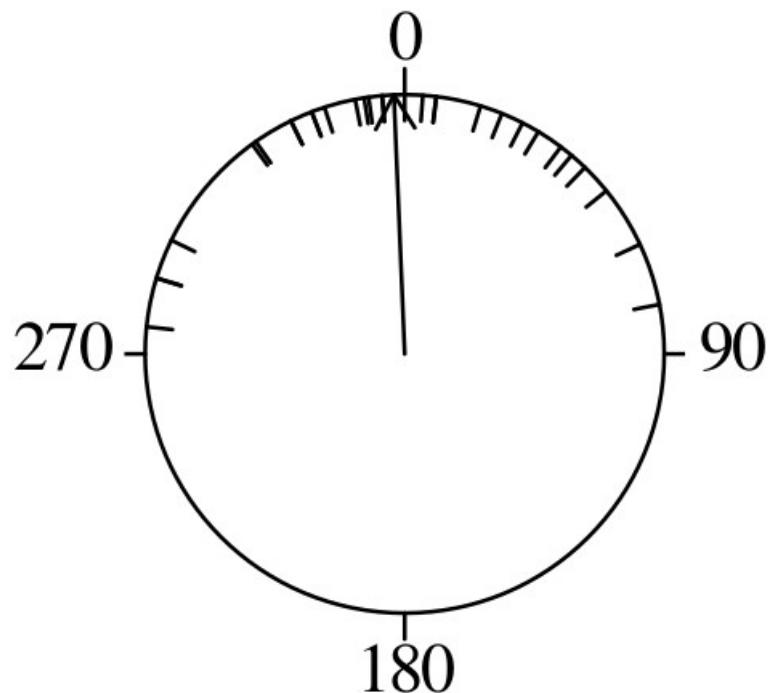
$$\sin[\theta \pm \phi] = \sin[\theta] \cos[\phi] \pm \cos[\theta] \sin[\phi]$$

$$\theta - \phi = \arcsin (\sin[\theta] \cos[\phi] - \cos[\theta] \sin[\phi])$$

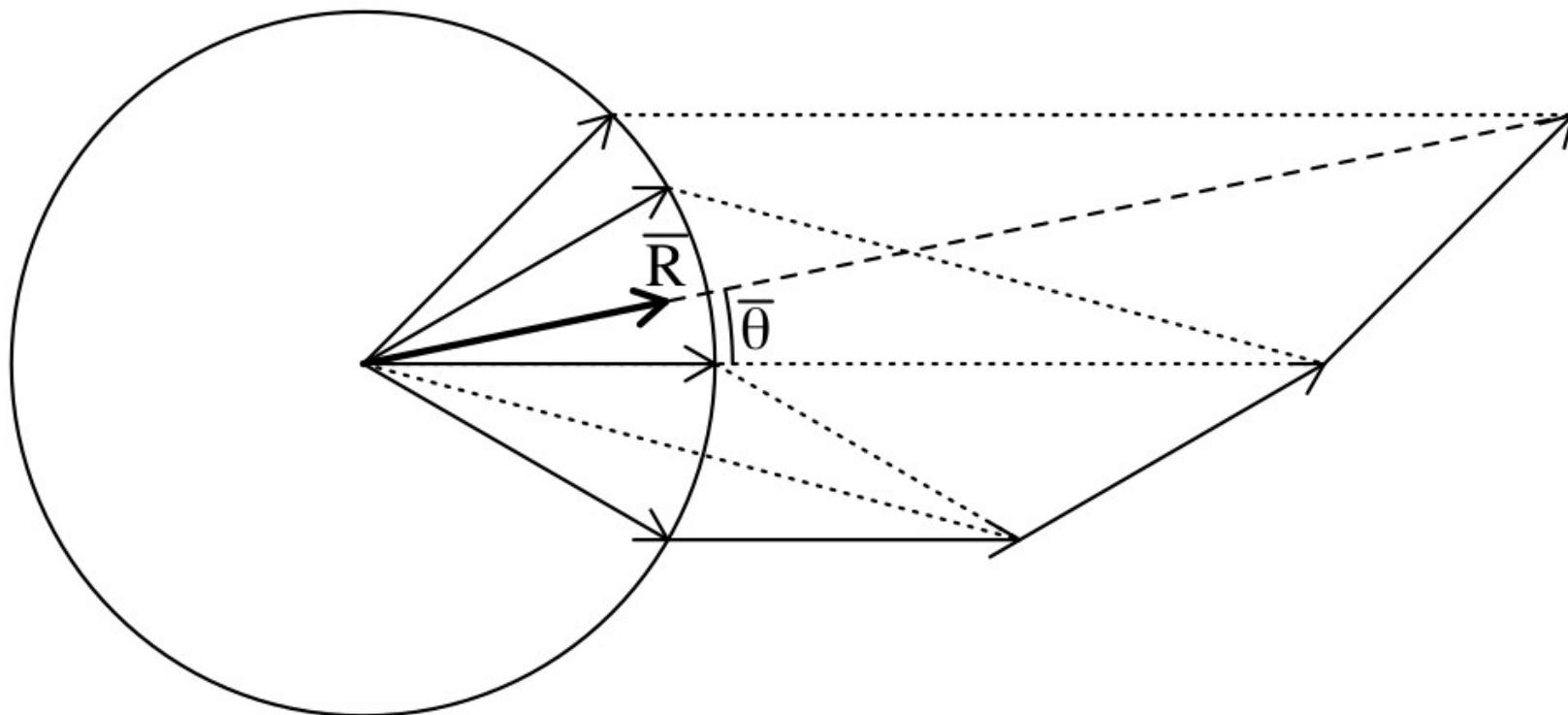
$$\bar{\theta} = \arctan \left( \frac{\sum_{i=1}^n \sin[\theta_i]}{\sum_{i=1}^n \cos[\theta_i]} \right)$$



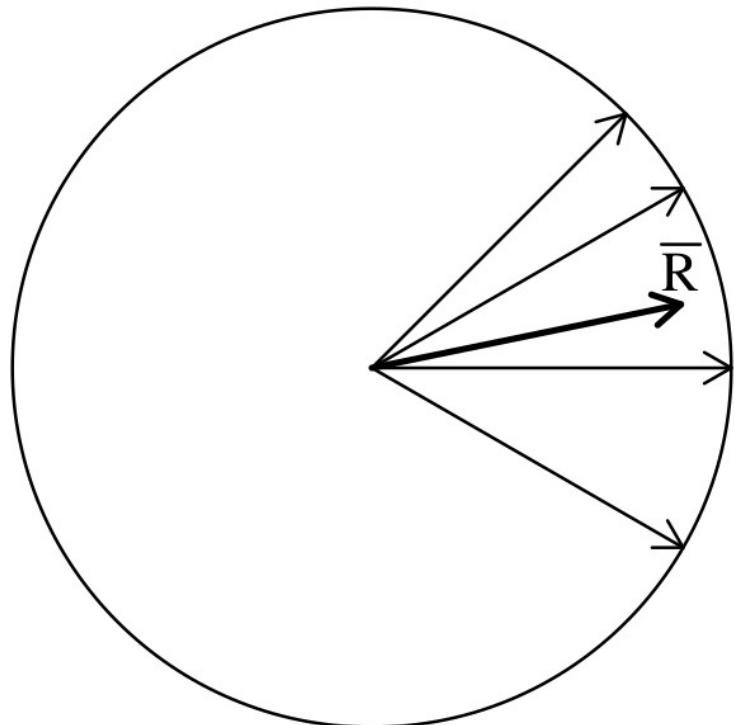
|     |     |     |     |     |    |    |     |     |     |     |    |     |     |     |
|-----|-----|-----|-----|-----|----|----|-----|-----|-----|-----|----|-----|-----|-----|
| 44  | 51  | 79  | 65  | 27  | 31 | 4  | 355 | 22  | 352 | 287 | 7  | 287 | 339 | 0   |
| 276 | 342 | 355 | 334 | 296 | 7  | 17 | 351 | 349 | 37  | 339 | 40 | 324 | 325 | 334 |



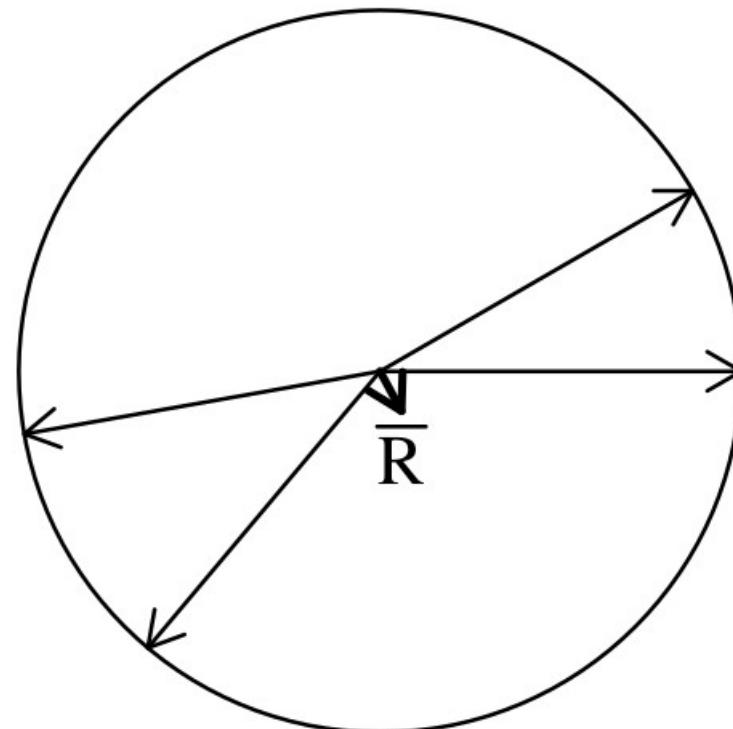
$$\bar{R} = \sqrt{\left( \sum_{i=1}^n \sin[\theta_i]/n \right)^2 + \left( \sum_{i=1}^n \cos[\theta_i]/n \right)^2}$$



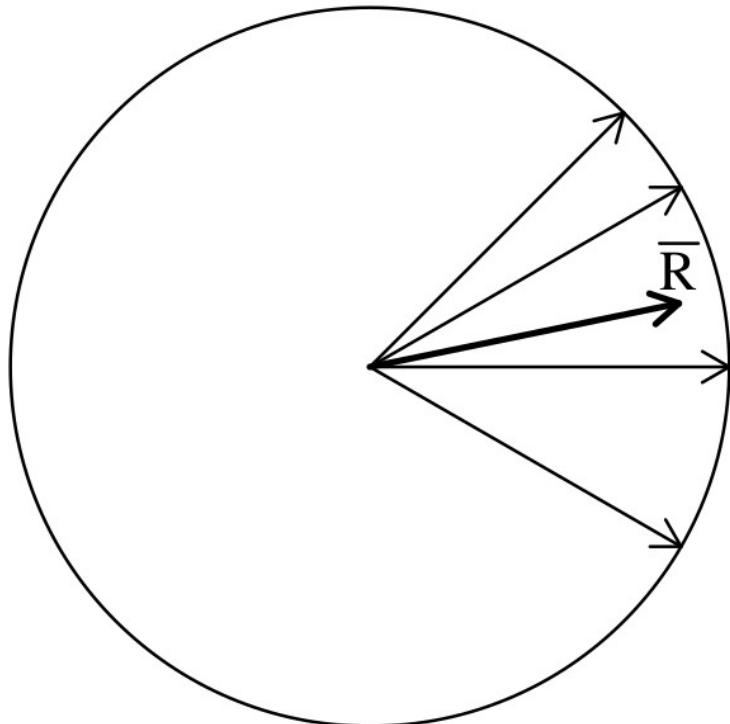
$$\bar{R} = 0.88$$



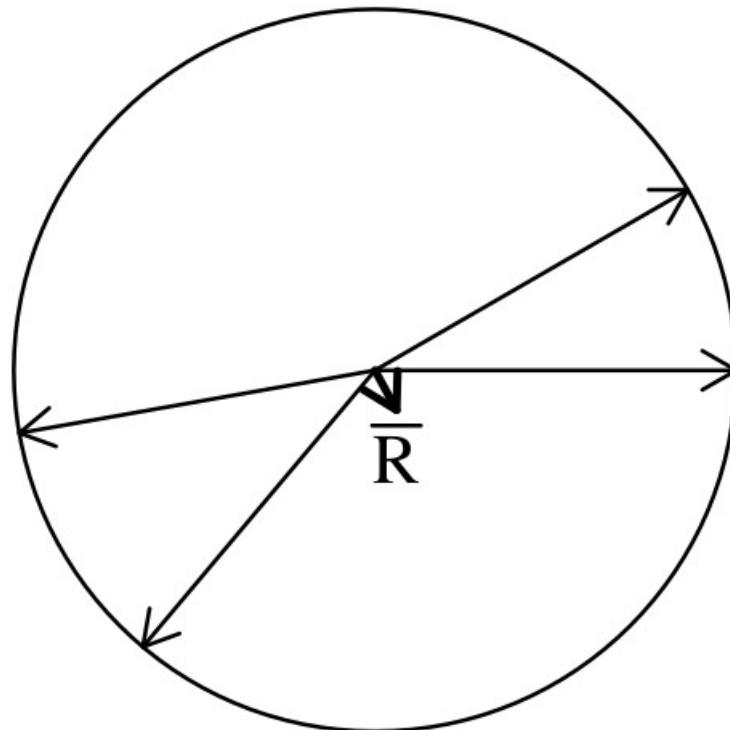
$$\bar{R} = 0.125$$



$$s_c = \sqrt{\ln(1/\bar{R}^2)}$$



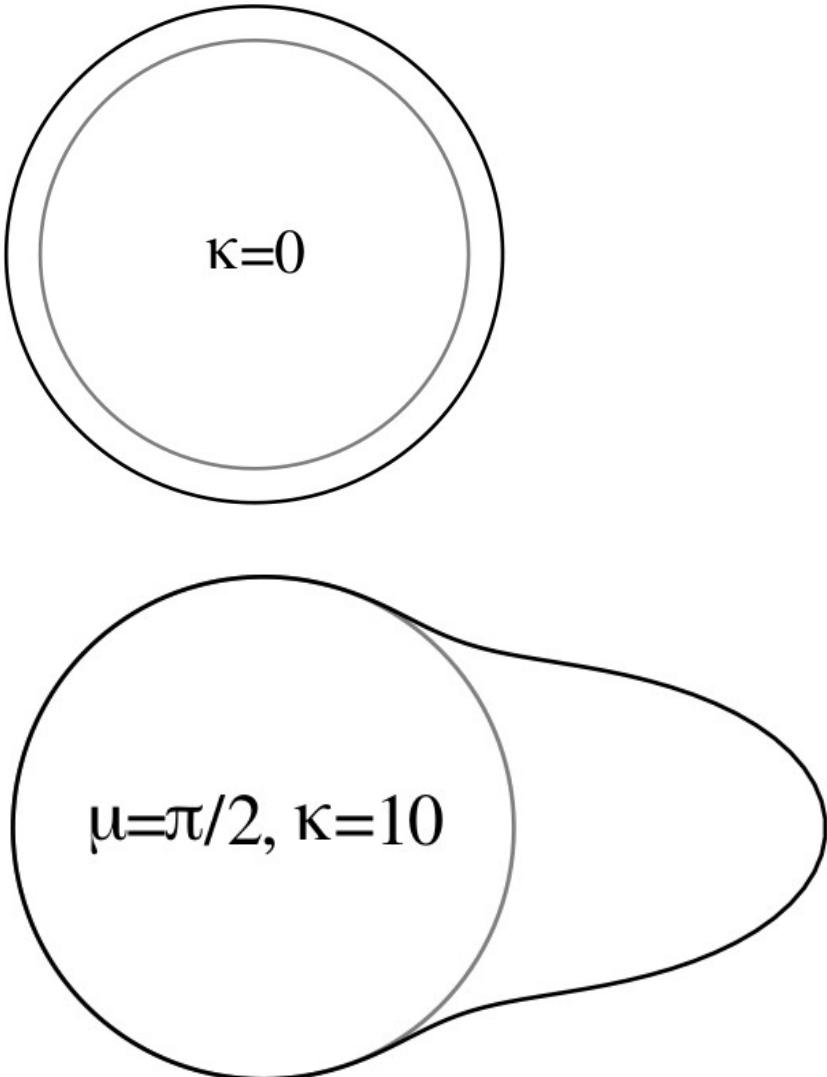
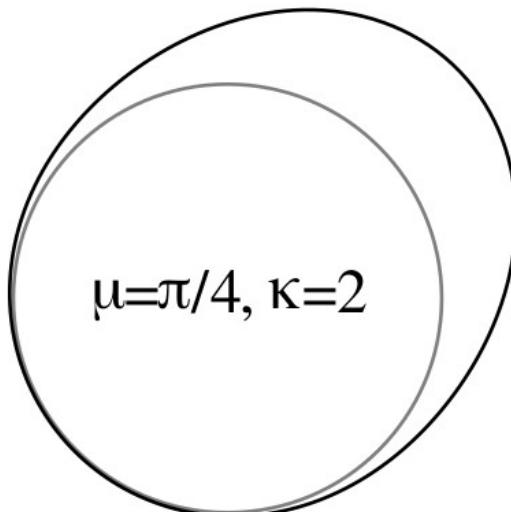
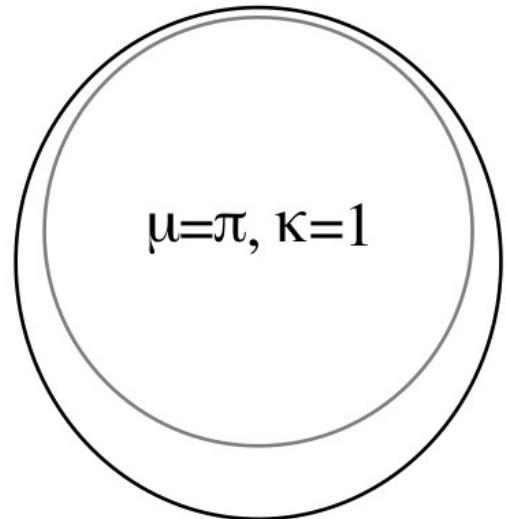
$$s_c = 0.51$$



$$s_c = 2.04$$

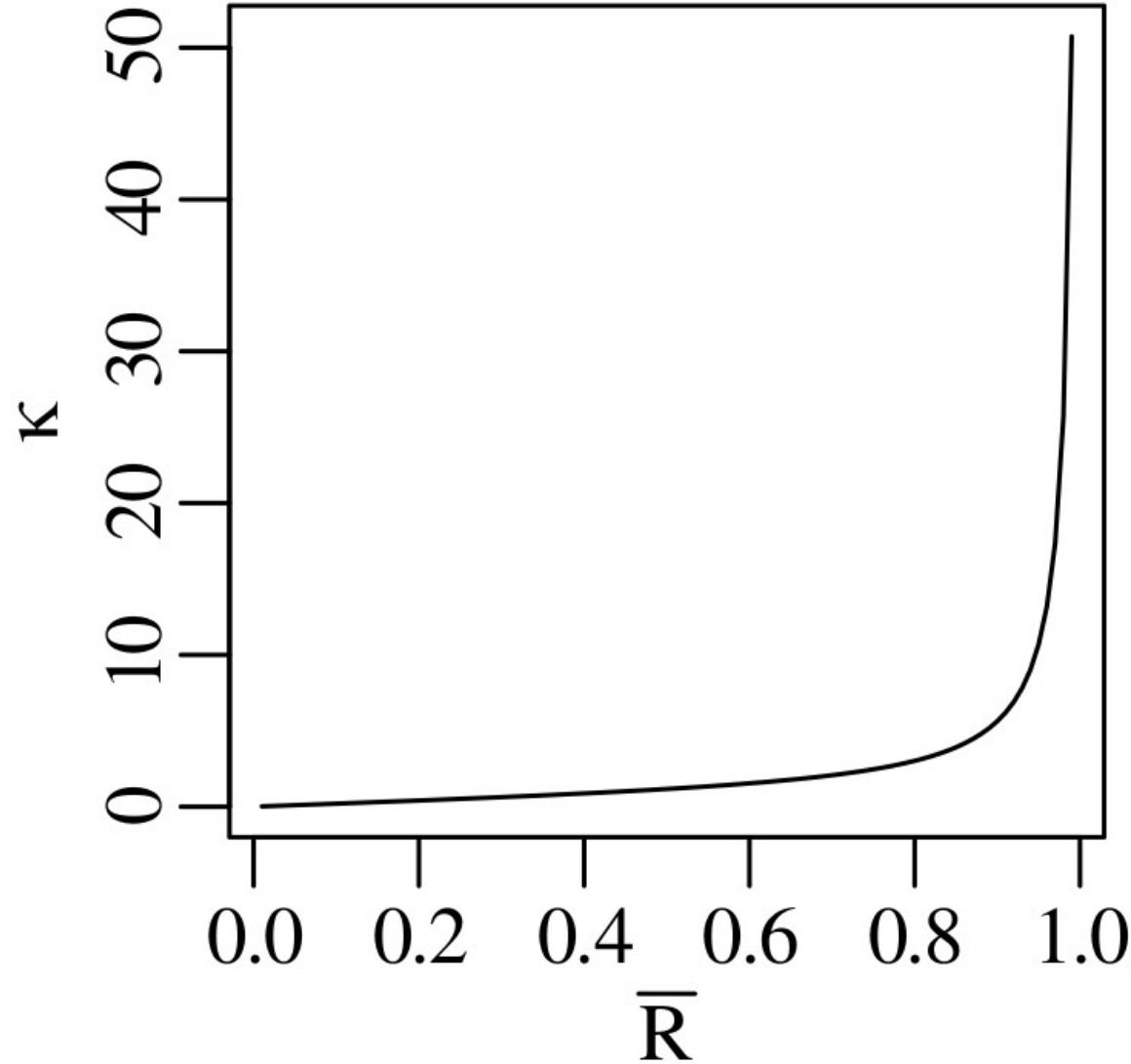
# The von Mises distribution

$$f(\theta|\mu, \kappa) \propto \exp[\kappa \cos(\theta - \mu)]$$

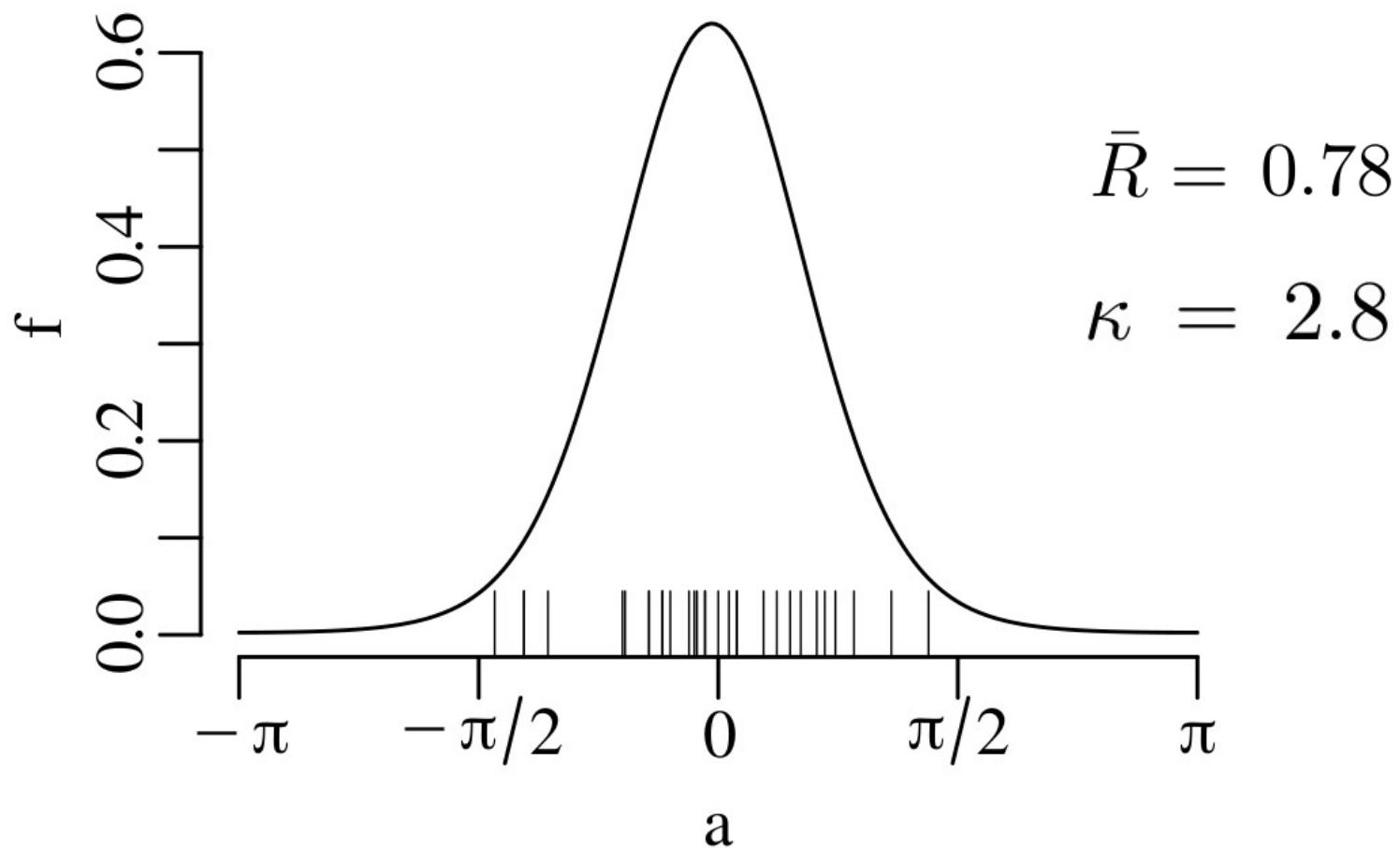


$\kappa=0$

$\mu=\pi/2, \kappa=10$



|     |     |     |     |     |    |    |     |     |     |     |    |     |     |     |
|-----|-----|-----|-----|-----|----|----|-----|-----|-----|-----|----|-----|-----|-----|
| 44  | 51  | 79  | 65  | 27  | 31 | 4  | 355 | 22  | 352 | 287 | 7  | 287 | 339 | 0   |
| 276 | 342 | 355 | 334 | 296 | 7  | 17 | 351 | 349 | 37  | 339 | 40 | 324 | 325 | 334 |



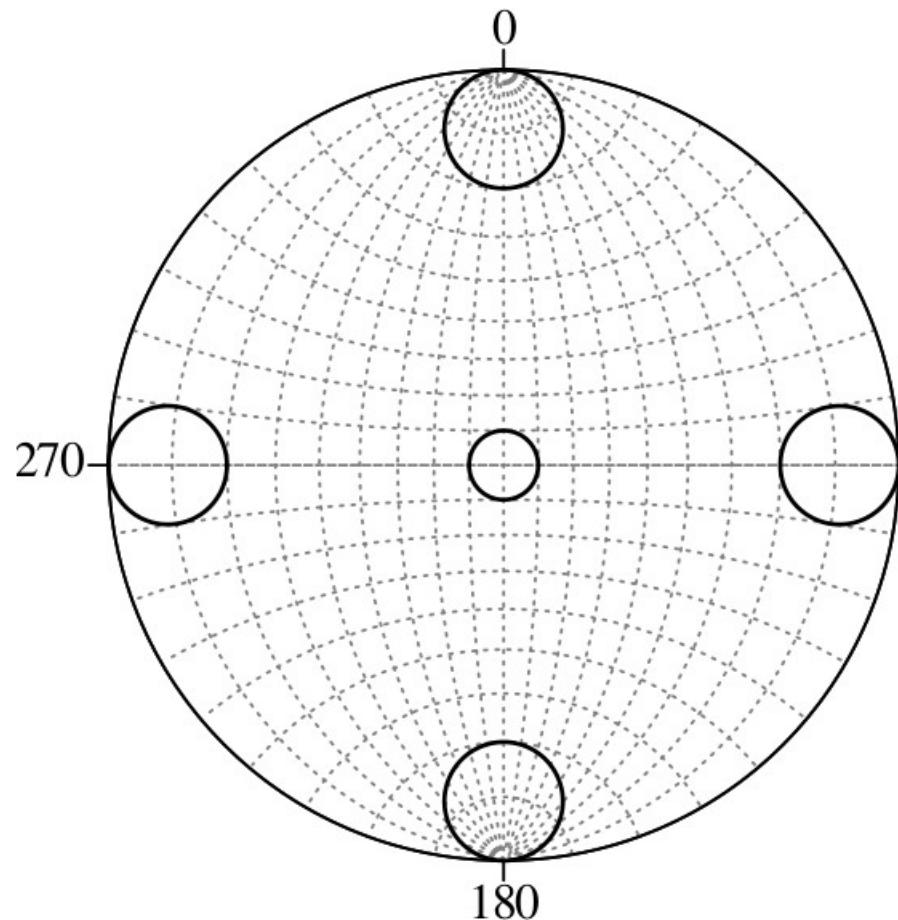
# Spherical data

$$\text{latitude } (l) \text{ and longitude } (L) \quad \left\{ \begin{array}{l} x = \cos[l] \sin[L] \\ y = \sin[l] \\ z = -\cos[l] \cos[L] \end{array} \right.$$

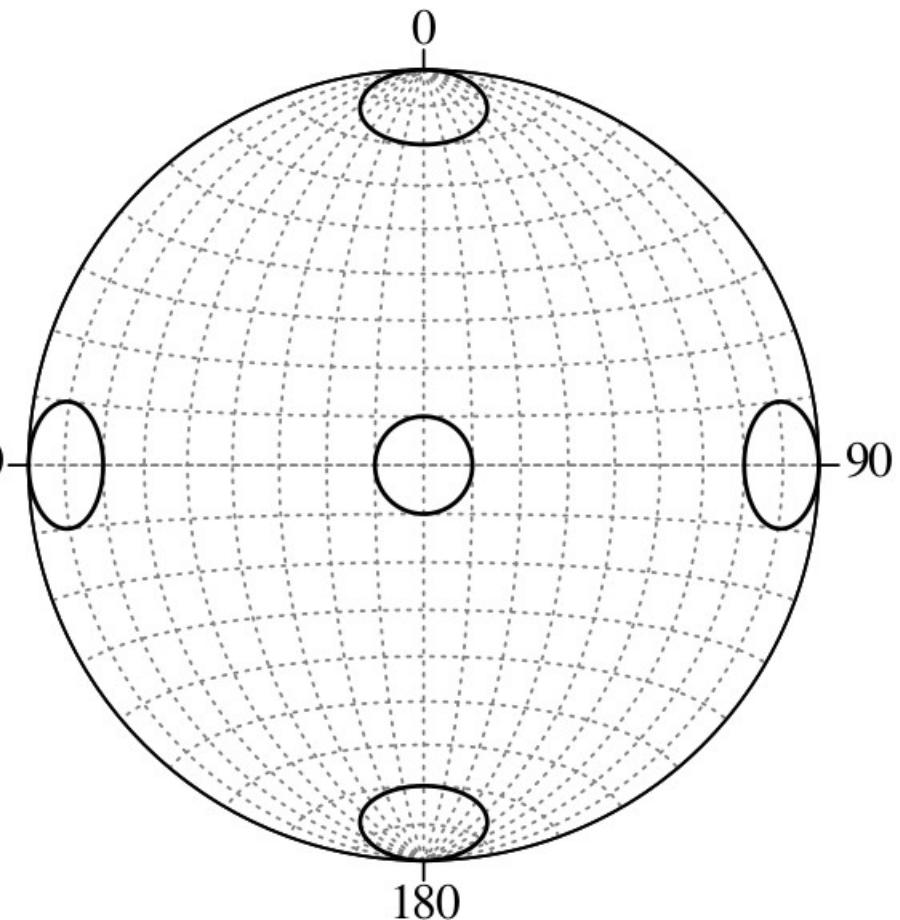
$$\text{strike } (S) \text{ and dip } (D) \quad \left\{ \begin{array}{l} x = -\cos[D] \sin[S] \\ y = \cos[D] \cos[S] \\ z = \sin[D] \end{array} \right.$$

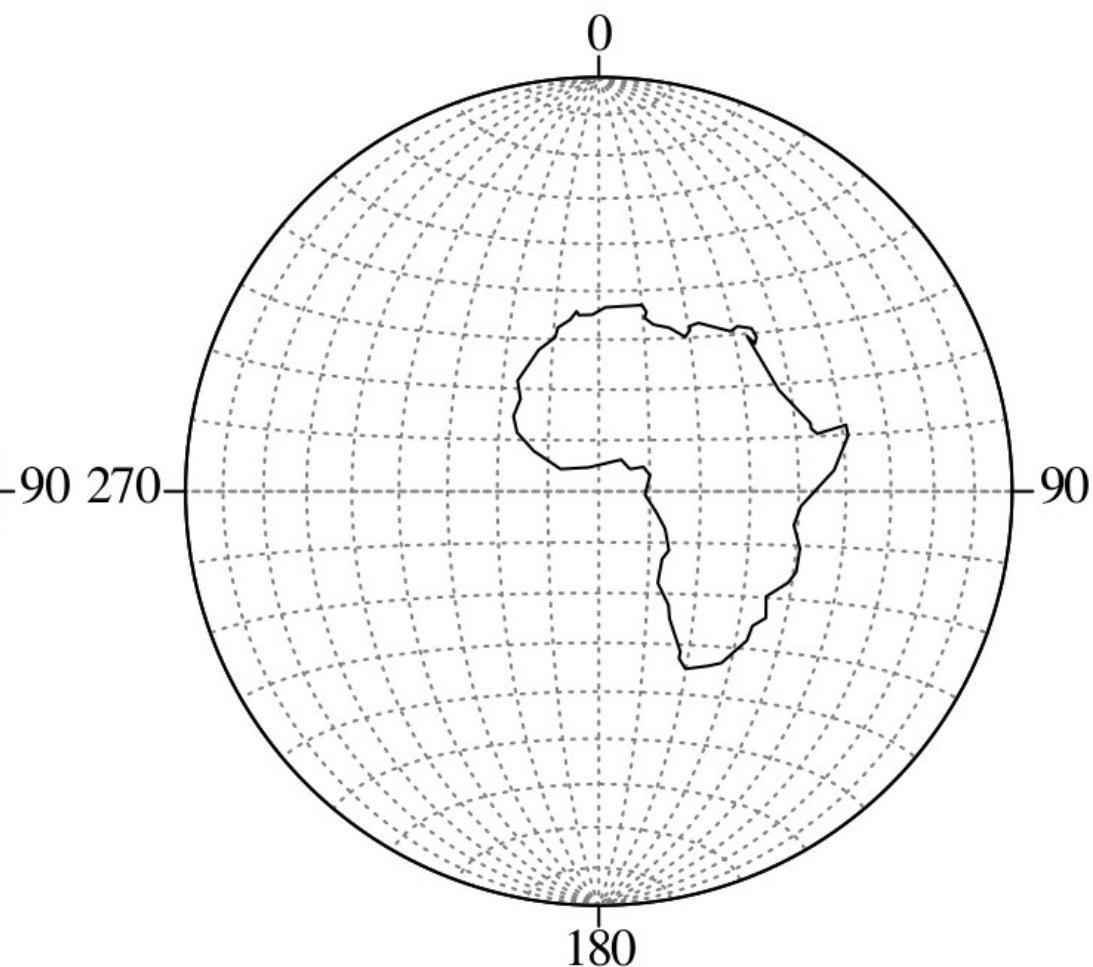
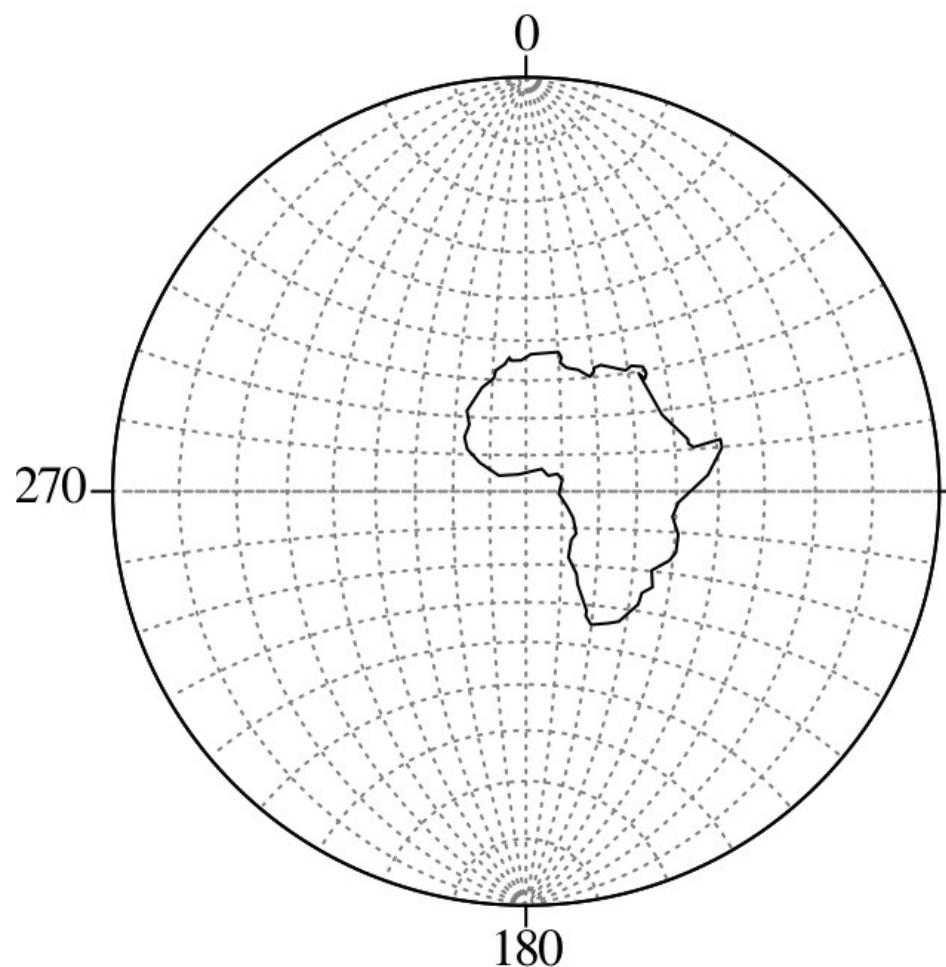
$$\text{dip } (D) \text{ and azimuth } (A) \quad \left\{ \begin{array}{l} x = \cos[D] \cos[A] \\ y = \cos[D] \sin[A] \\ z = \sin[D] \end{array} \right.$$

Wulff

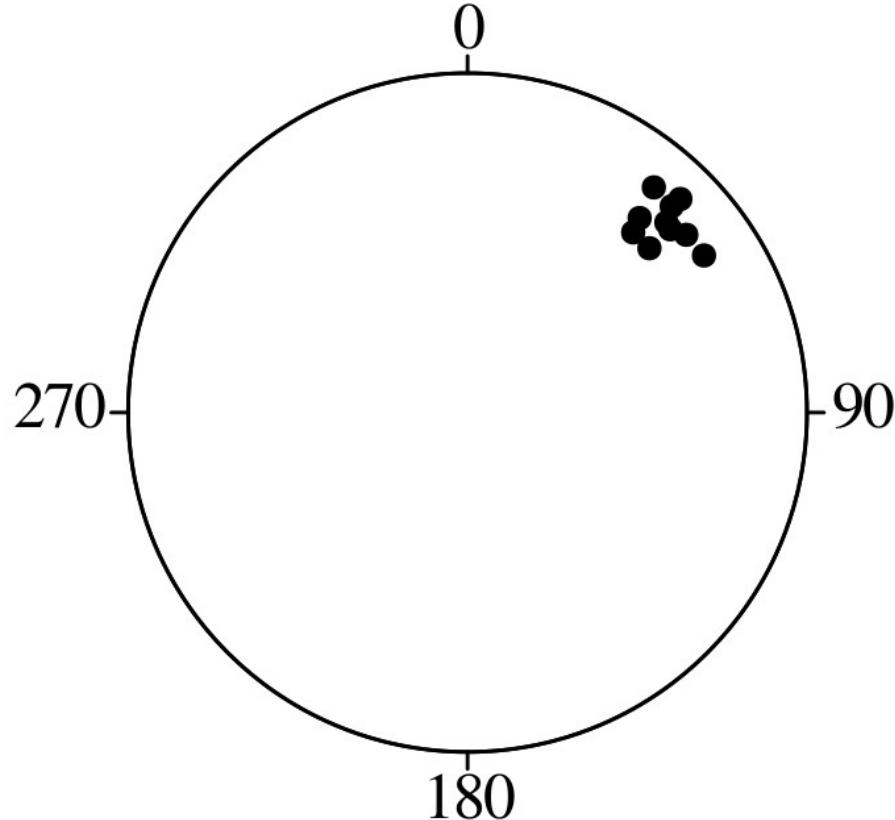
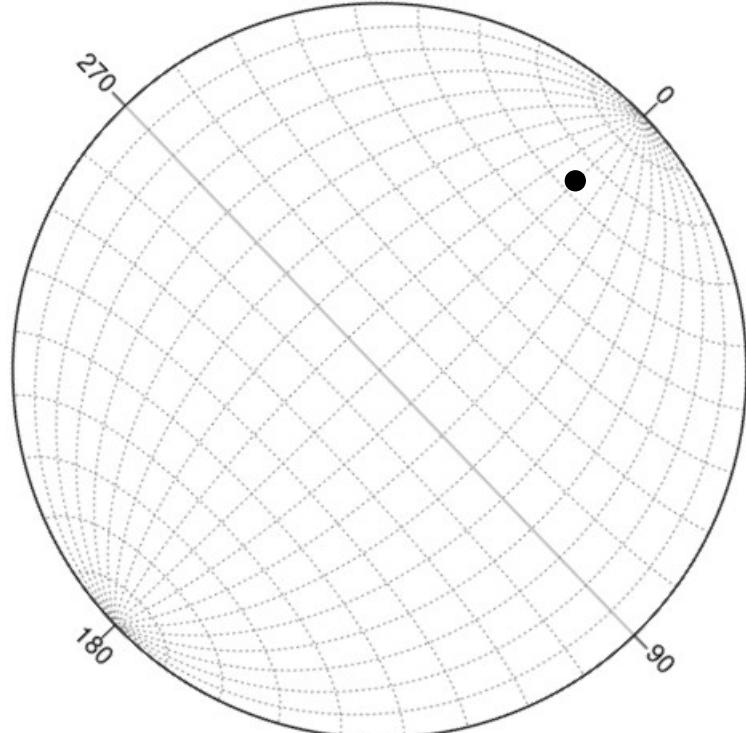


Schmidt

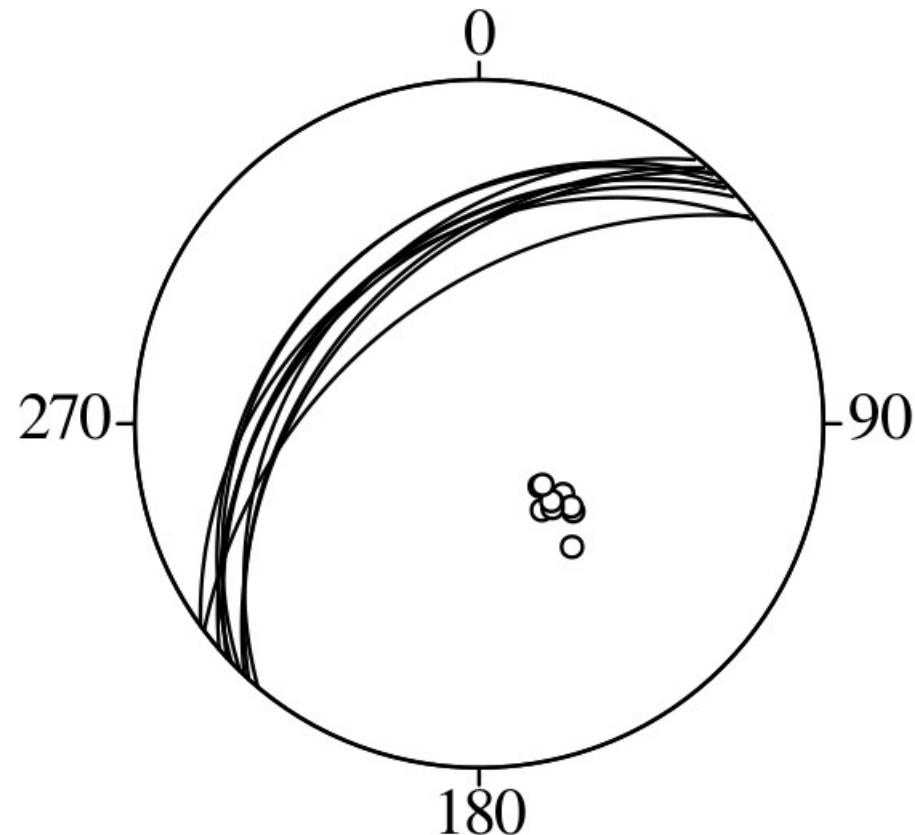
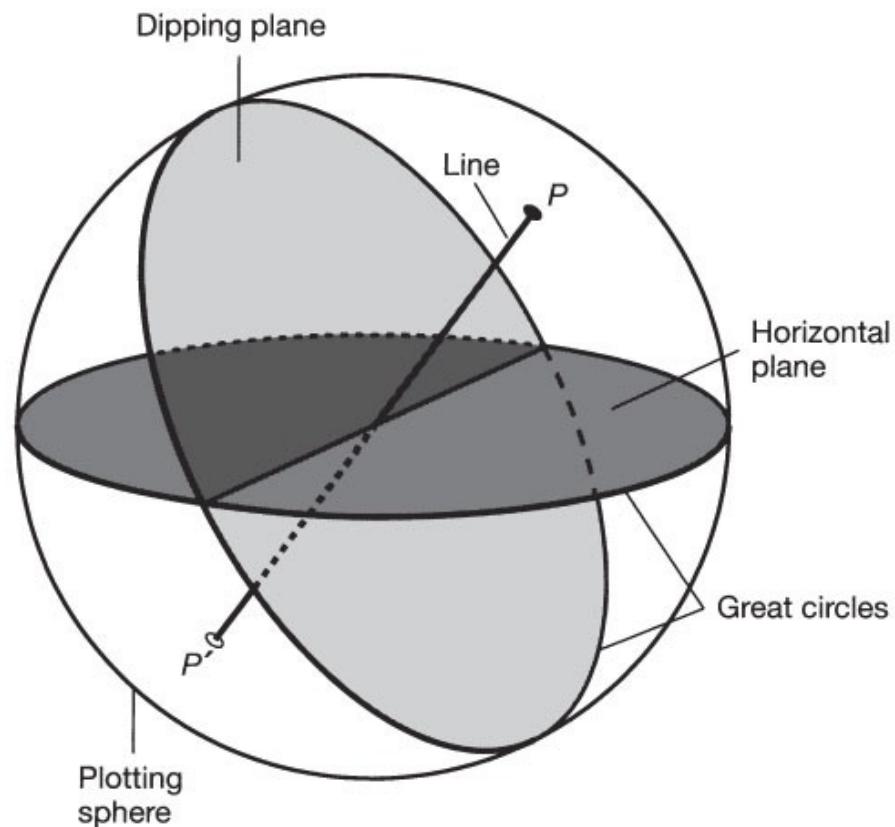




|             |      |      |      |      |      |      |      |      |      |      |
|-------------|------|------|------|------|------|------|------|------|------|------|
| declination | 47.9 | 46.3 | 44.7 | 50.9 | 56.4 | 42.6 | 44.9 | 41.5 | 47.9 | 39.6 |
| inclination | 28.6 | 20.1 | 15.6 | 18.1 | 17.5 | 28.7 | 12.2 | 24.5 | 20.6 | 15.0 |



|        |      |      |      |      |      |      |      |      |      |      |
|--------|------|------|------|------|------|------|------|------|------|------|
| strike | 226  | 220  | 223  | 222  | 233  | 227  | 234  | 229  | 227  | 224  |
| dip    | 28.4 | 35.3 | 41.0 | 39.6 | 48.3 | 34.7 | 34.5 | 36.0 | 34.2 | 28.7 |



$$\begin{cases} x = \cos[l] \sin[L] \\ y = \sin[l] \\ z = -\cos[l] \cos[L] \end{cases}$$

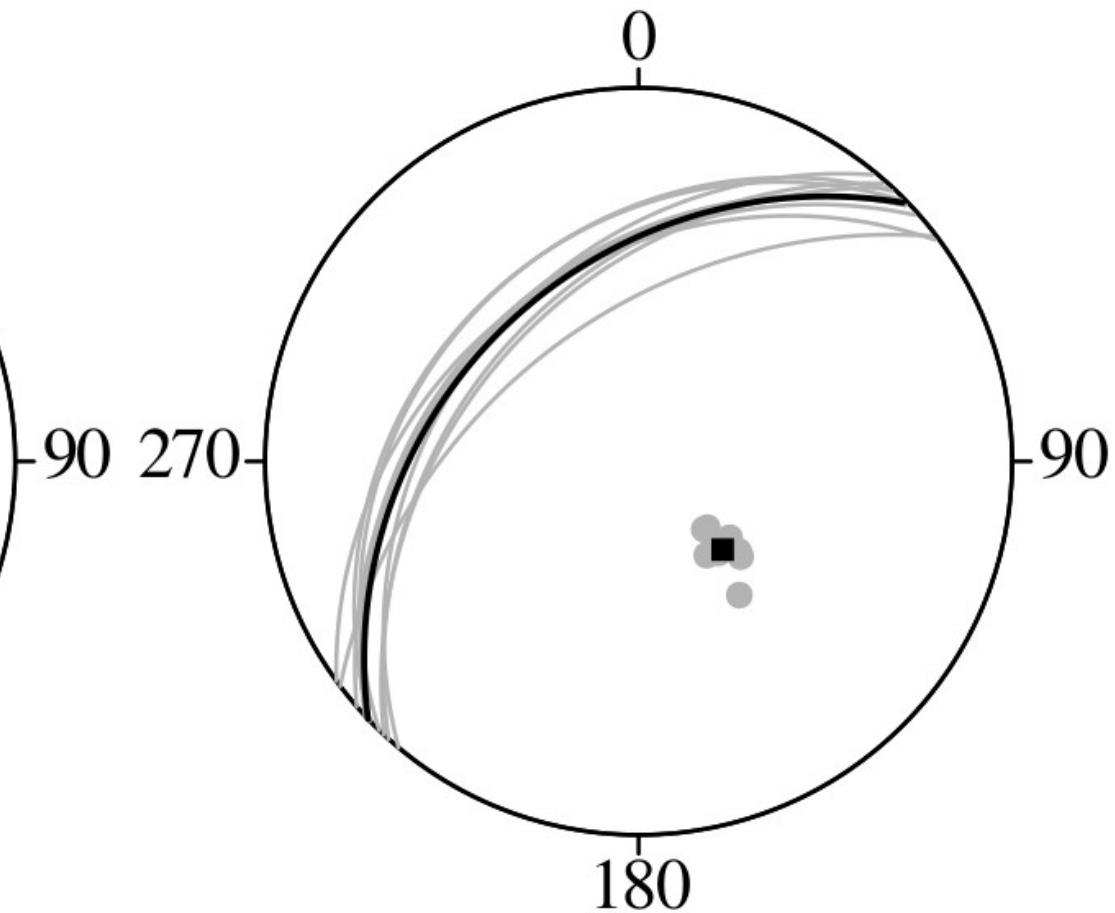
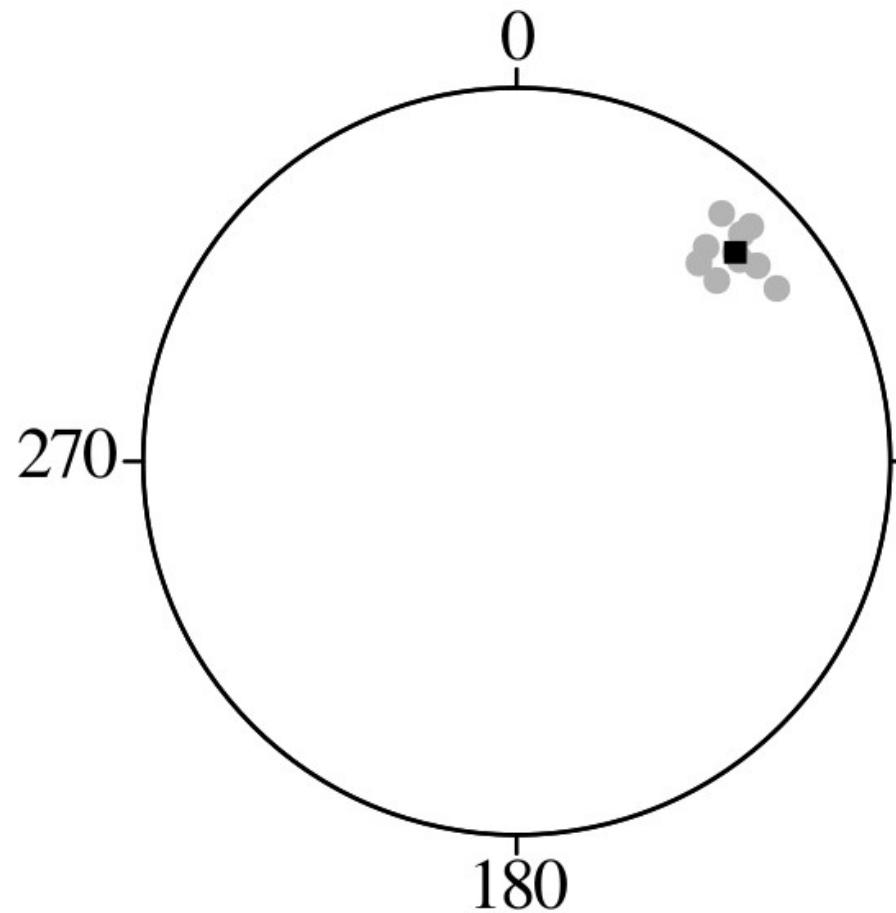
$$\begin{cases} \bar{L} = \arctan(-\bar{x}/\bar{z}) \\ \bar{l} = \arcsin(\bar{y}) \end{cases}$$

$$\begin{cases} x = -\cos[D] \sin[S] \\ y = \cos[D] \cos[S] \\ z = \sin[D] \end{cases}$$

$$\begin{cases} \bar{S} = \arctan(-\bar{x}/\bar{y}) \\ \bar{D} = \arcsin(\bar{z}) \end{cases}$$

$$\begin{cases} x = \cos[D] \cos[A] \\ y = \cos[D] \sin[A] \\ z = \sin[D] \end{cases}$$

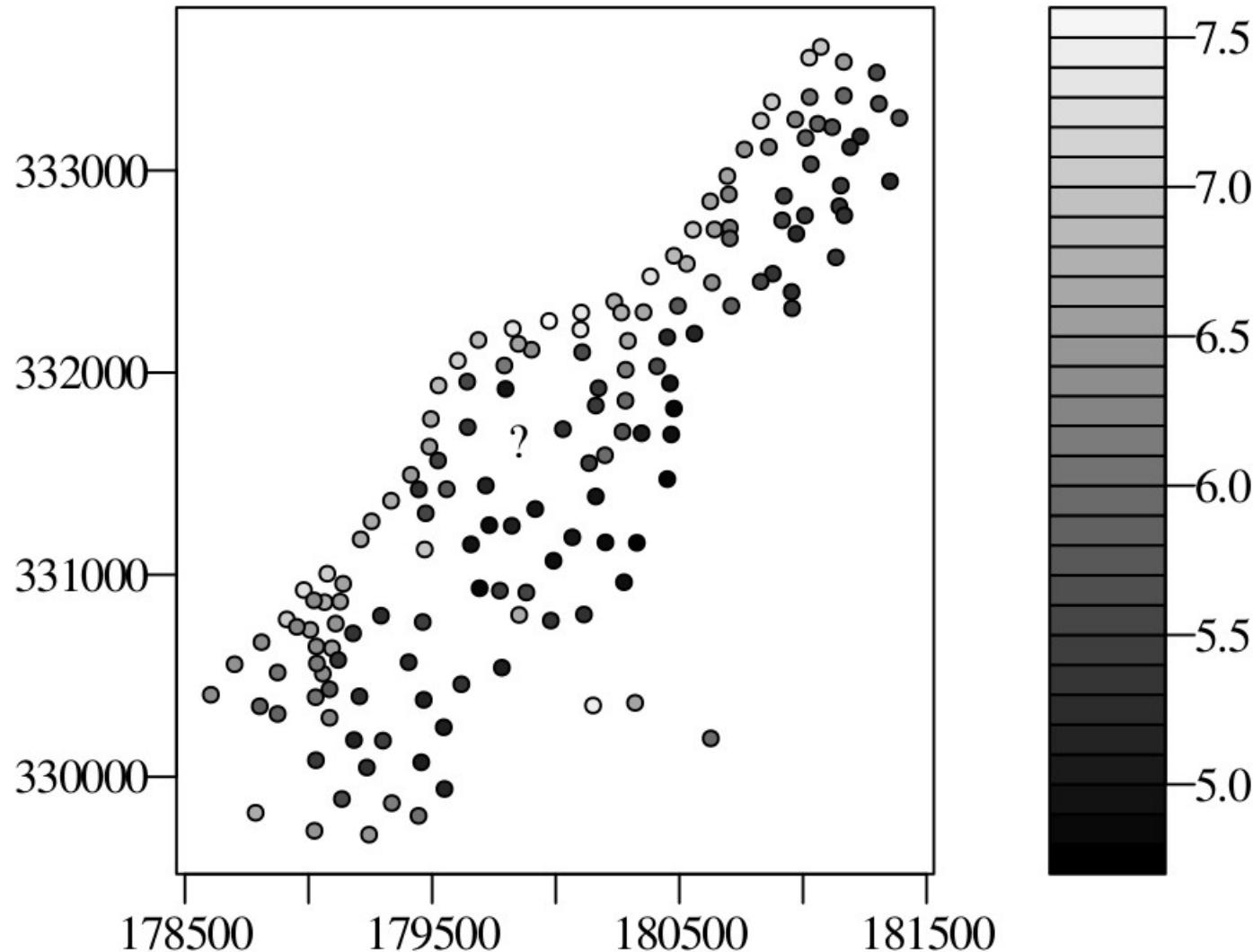
$$\begin{cases} \bar{D} = \arctan(\bar{y}/\bar{x}) \\ \bar{A} = \arcsin(\bar{z}) \end{cases}$$



# Statistics for geoscientists

Spatial data

| #        | $x$      | $y$      | Zn       | $z = \ln[\text{Zn}]$ |
|----------|----------|----------|----------|----------------------|
| 1        | 181072   | 333611   | 1022     | 6.93                 |
| 2        | 181025   | 333558   | 1141     | 7.04                 |
| 3        | 181165   | 333537   | 640      | 6.46                 |
| $\vdots$ | $\vdots$ | $\vdots$ | $\vdots$ |                      |
| 50       | 180199   | 331591   | 375      | 5.93                 |
| $\vdots$ | $\vdots$ | $\vdots$ | $\vdots$ |                      |
| 155      | 180627   | 330190   | 375      | 5.93                 |

$\ln[\text{Zn}]$ 

“everything is related to everything else, but near things are more related than distant things.”

– Waldo R. Tobler (1969)



|     | 1    | 2    | 3    | 4     | 5     | 6    | 7      | 8    | 9    | 10  | ... | 155 |
|-----|------|------|------|-------|-------|------|--------|------|------|-----|-----|-----|
| 1   | 0    |      |      |       |       |      |        |      |      |     |     |     |
| 2   | 0.16 | 0    |      |       |       |      |        |      |      |     |     |     |
| 3   | 0.66 | 0.82 | 0    |       |       |      |        |      |      |     |     |     |
| 4   | 2.0  | 2.1  | 1.3  | 0     |       |      |        |      |      |     |     |     |
| 5   | 1.9  | 2.0  | 1.2  | 0.065 | 0     |      |        |      |      |     |     |     |
| 6   | 1.8  | 2.0  | 1.2  | 0.13  | 0.062 | 0    |        |      |      |     |     |     |
| 7   | 1.5  | 1.7  | 0.87 | 0.42  | 0.36  | 0.29 | 0      |      |      |     |     |     |
| 8   | 1.3  | 1.5  | 0.64 | 0.65  | 0.58  | 0.52 | 0.23   | 0    |      |     |     |     |
| 9   | 1.5  | 1.7  | 0.87 | 0.42  | 0.36  | 0.30 | 0.0041 | 0.22 | 0    |     |     |     |
| 10  | 2.4  | 2.6  | 1.8  | 0.48  | 0.54  | 0.61 | 0.90   | 1.1  | 0.90 | 0   |     |     |
| ⋮   | ⋮    | ⋮    | ⋮    | ⋮     | ⋮     | ⋮    | ⋮      | ⋮    | ⋮    | ⋮   | ⋮   | ⋮   |
| 155 | 1.4  | 1.6  | 0.76 | 0.53  | 0.47  | 0.41 | 0.11   | 0.11 | 0.11 | 1.0 | ... | 0   |

|     | 1    | 2    | 3    | 4     | 5     | 6    | 7      | 8    | 9    | 10  | ... | 155  |
|-----|------|------|------|-------|-------|------|--------|------|------|-----|-----|------|
| 1   | 0    | 71   | 120  | 260   | 370   | 470  | 260    | 250  | 380  | 470 | ... | 3400 |
| 2   | 0.16 | 0    | 140  | 280   | 360   | 470  | 230    | 200  | 330  | 440 | ... | 3400 |
| 3   | 0.66 | 0.82 | 0    | 140   | 250   | 360  | 170    | 220  | 320  | 380 | ... | 3400 |
| 4   | 2.0  | 2.1  | 1.3  | 0     | 150   | 240  | 180    | 300  | 350  | 320 | ... | 3400 |
| 5   | 1.9  | 2.0  | 1.2  | 0.065 | 0     | 110  | 150    | 280  | 270  | 180 | ... | 3200 |
| 6   | 1.8  | 2.0  | 1.2  | 0.13  | 0.062 | 0    | 250    | 380  | 330  | 180 | ... | 3200 |
| 7   | 1.5  | 1.7  | 0.87 | 0.42  | 0.36  | 0.29 | 0      | 140  | 170  | 210 | ... | 3200 |
| 8   | 1.3  | 1.5  | 0.64 | 0.65  | 0.58  | 0.52 | 0.23   | 0    | 140  | 280 | ... | 3200 |
| 9   | 1.5  | 1.7  | 0.87 | 0.42  | 0.36  | 0.30 | 0.0041 | 0.22 | 0    | 180 | ... | 3100 |
| 10  | 2.4  | 2.6  | 1.8  | 0.48  | 0.54  | 0.61 | 0.90   | 1.1  | 0.90 | 0   | ... | 3000 |
| :   | :    | :    | :    | :     | :     | :    | :      | :    | :    | :   | ..  | :    |
| 155 | 1.4  | 1.6  | 0.76 | 0.53  | 0.47  | 0.41 | 0.11   | 0.11 | 0.11 | 1.0 | ... | 0    |

# semivariance

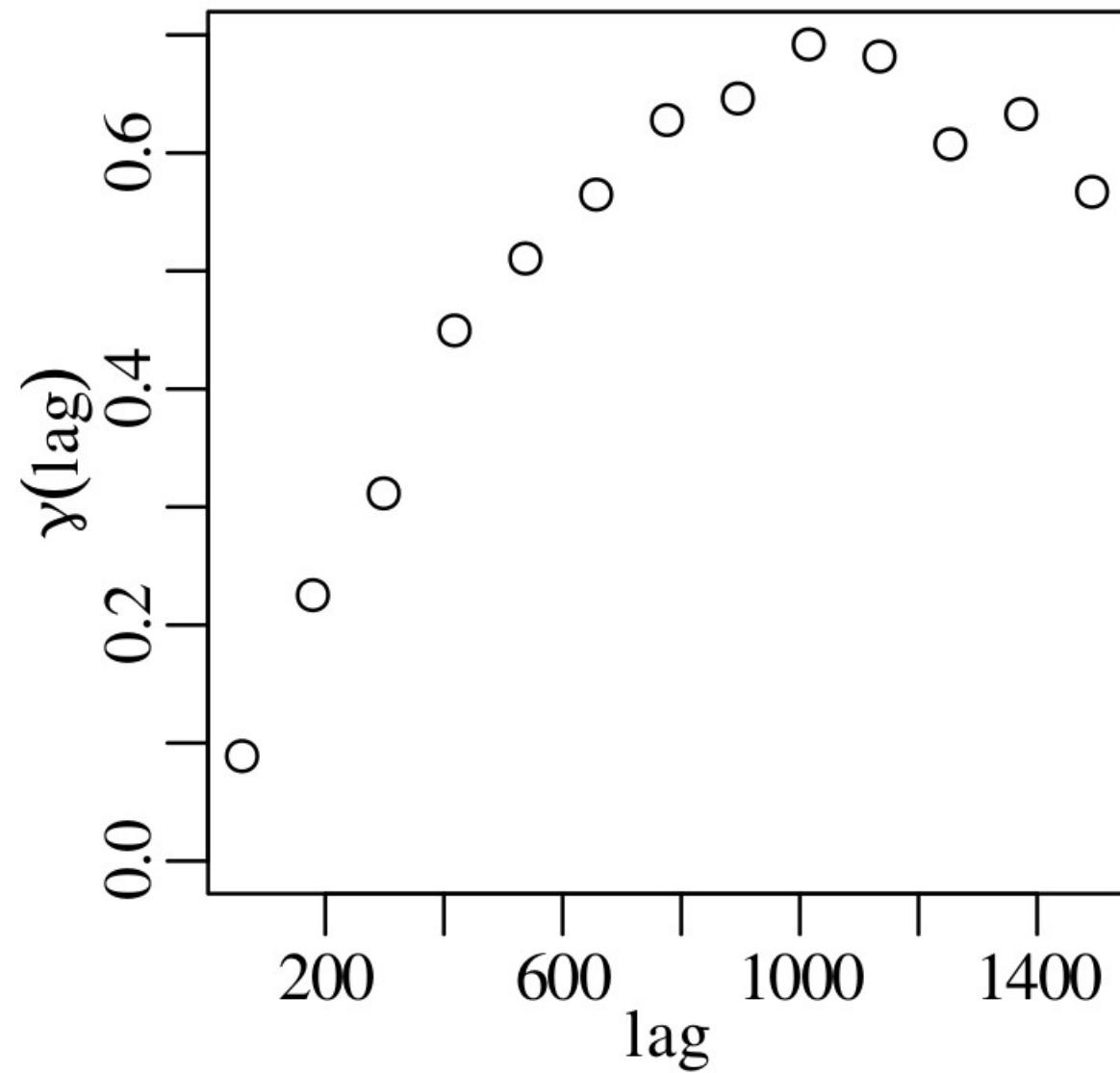
$$\gamma[h] = \sum_k^{n(h)} \frac{(z_{i[k]} - z_{j[k]})^2}{2n(h)}$$

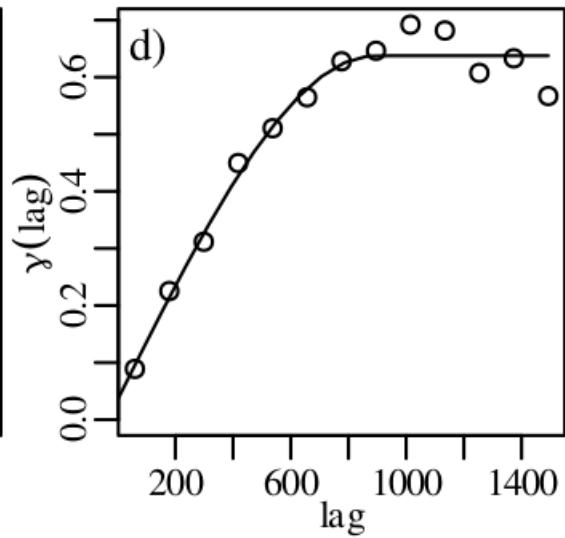
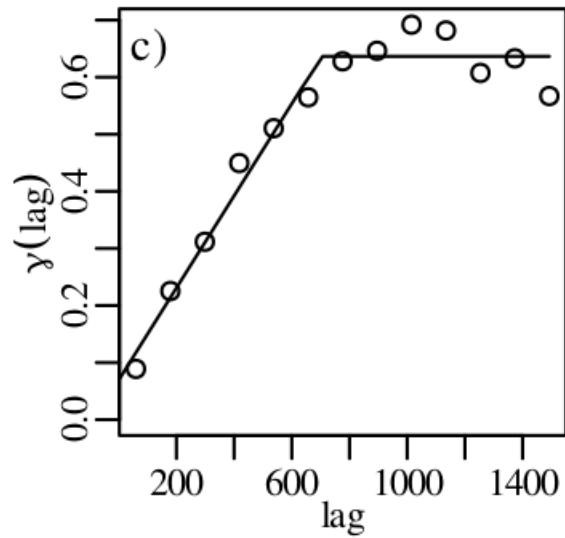
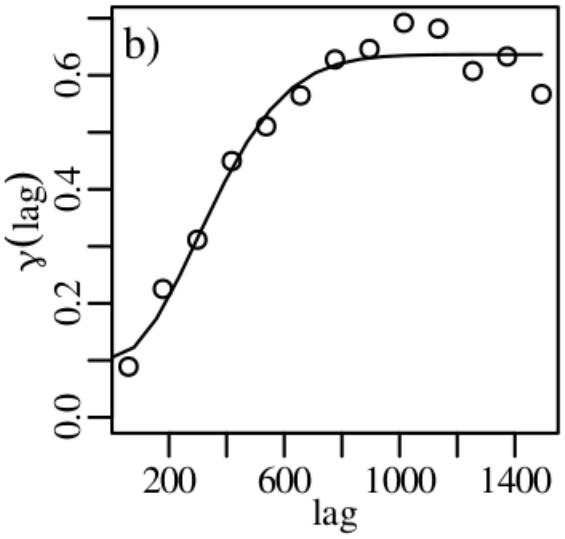
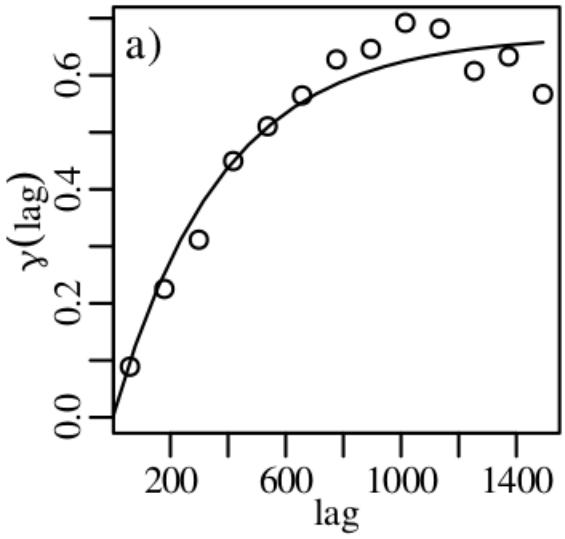
$$\gamma[0 < h \leq 150] = \frac{0.16^2 + 0.66^2 + 0.82^2 + \dots}{2 \times 166} = 0.10$$

|     | 1    | 2    | 3    | 4     | 5     | 6    | 7      | 8    | 9    | 10  | ... | 155  |
|-----|------|------|------|-------|-------|------|--------|------|------|-----|-----|------|
| 1   | 0    | 71   | 120  | 260   | 370   | 470  | 260    | 250  | 380  | 470 | ... | 3400 |
| 2   | 0.16 | 0    | 140  | 280   | 360   | 470  | 230    | 200  | 330  | 440 | ... | 3400 |
| 3   | 0.66 | 0.82 | 0    | 140   | 250   | 360  | 170    | 220  | 320  | 380 | ... | 3400 |
| 4   | 2.0  | 2.1  | 1.3  | 0     | 150   | 240  | 180    | 300  | 350  | 320 | ... | 3400 |
| 5   | 1.9  | 2.0  | 1.2  | 0.065 | 0     | 110  | 150    | 280  | 270  | 180 | ... | 3200 |
| 6   | 1.8  | 2.0  | 1.2  | 0.13  | 0.062 | 0    | 250    | 380  | 330  | 180 | ... | 3200 |
| 7   | 1.5  | 1.7  | 0.87 | 0.42  | 0.36  | 0.29 | 0      | 140  | 170  | 210 | ... | 3200 |
| 8   | 1.3  | 1.5  | 0.64 | 0.65  | 0.58  | 0.52 | 0.23   | 0    | 140  | 280 | ... | 3200 |
| 9   | 1.5  | 1.7  | 0.87 | 0.42  | 0.36  | 0.30 | 0.0041 | 0.22 | 0    | 180 | ... | 3100 |
| 10  | 2.4  | 2.6  | 1.8  | 0.48  | 0.54  | 0.61 | 0.90   | 1.1  | 0.90 | 0   | ... | 3000 |
| ⋮   | ⋮    | ⋮    | ⋮    | ⋮     | ⋮     | ⋮    | ⋮      | ⋮    | ⋮    | ⋮   | ⋮   | ⋮    |
| 155 | 1.4  | 1.6  | 0.76 | 0.53  | 0.47  | 0.41 | 0.11   | 0.11 | 0.11 | 1.0 | ... | 0    |

$$\gamma[150 < h \leq 300] = \frac{2.0^2 + 1.5^2 + 1.3^2 + 2.1^2 + \dots}{2 \times 530} = 0.28$$

|     | 1    | 2    | 3    | 4     | 5     | 6    | 7      | 8    | 9    | 10  | ... | 155  |
|-----|------|------|------|-------|-------|------|--------|------|------|-----|-----|------|
| 1   | 0    | 71   | 120  | 260   | 370   | 470  | 260    | 250  | 380  | 470 | ... | 3400 |
| 2   | 0.16 | 0    | 140  | 280   | 360   | 470  | 230    | 200  | 330  | 440 | ... | 3400 |
| 3   | 0.66 | 0.82 | 0    | 140   | 250   | 360  | 170    | 220  | 320  | 380 | ... | 3400 |
| 4   | 2.0  | 2.1  | 1.3  | 0     | 150   | 240  | 180    | 300  | 350  | 320 | ... | 3400 |
| 5   | 1.9  | 2.0  | 1.2  | 0.065 | 0     | 110  | 150    | 280  | 270  | 180 | ... | 3200 |
| 6   | 1.8  | 2.0  | 1.2  | 0.13  | 0.062 | 0    | 250    | 380  | 330  | 180 | ... | 3200 |
| 7   | 1.5  | 1.7  | 0.87 | 0.42  | 0.36  | 0.29 | 0      | 140  | 170  | 210 | ... | 3200 |
| 8   | 1.3  | 1.5  | 0.64 | 0.65  | 0.58  | 0.52 | 0.23   | 0    | 140  | 280 | ... | 3200 |
| 9   | 1.5  | 1.7  | 0.87 | 0.42  | 0.36  | 0.30 | 0.0041 | 0.22 | 0    | 180 | ... | 3100 |
| 10  | 2.4  | 2.6  | 1.8  | 0.48  | 0.54  | 0.61 | 0.90   | 1.1  | 0.90 | 0   | ... | 3000 |
| :   | :    | :    | :    | :     | :     | :    | :      | :    | :    | :   | ..  | :    |
| 155 | 1.4  | 1.6  | 0.76 | 0.53  | 0.47  | 0.41 | 0.11   | 0.11 | 0.11 | 1.0 | ... | 0    |





a) exponential

$$\gamma[h] = c_s + (c_n - c_s) \exp \left[ -\frac{h}{c_r} \right]$$

b) Gaussian

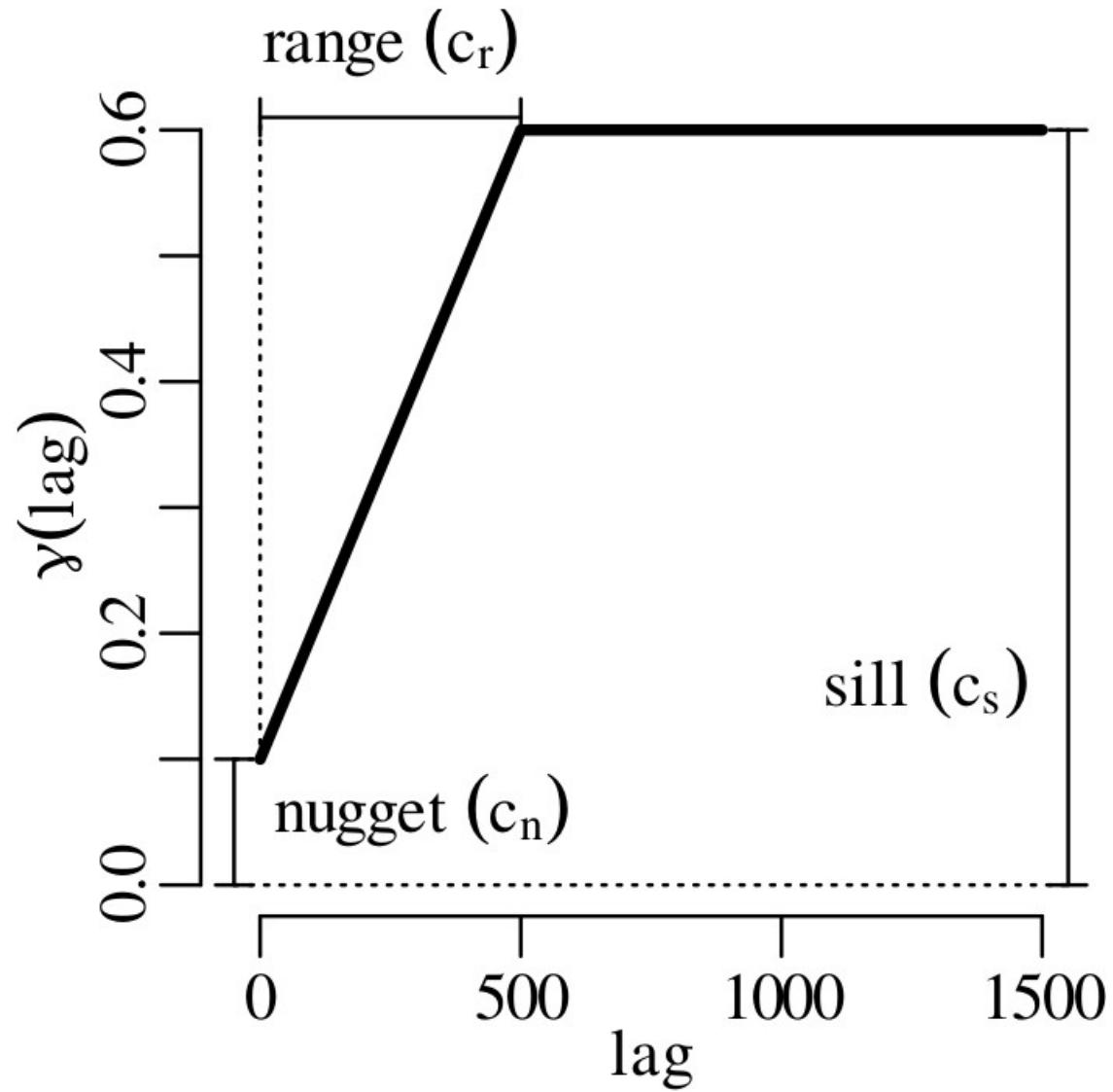
$$\gamma[h] = c_s + (c_n - c_s) \exp \left[ -\frac{h^2}{c_r^2} \right]$$

c) linear

$$\begin{cases} \gamma[h] = c_n + \frac{(c_s - c_n)h}{c_r} & \text{if } h \leq c_r \\ \gamma[h] = c_s & \text{if } h > c_r \end{cases}$$

d) spherical

$$\begin{cases} \gamma[h] = c_n + (c_s - c_n) \left( \frac{3}{2} \frac{h}{c_r} - \frac{1}{2} \left[ \frac{h}{c_r} \right]^3 \right) & \text{if } h \leq c_r \\ \gamma[h] = c_s & \text{if } h > c_r \end{cases}$$



# Kriging interpolation

$$z(x_o, y_o) = \sum_{i=1}^n w_i z(x_i, y_i)$$

$$\begin{bmatrix} w_1 \\ w_2 \\ \vdots \\ w_n \\ \lambda \end{bmatrix} = \begin{bmatrix} c_n & \gamma[h_{1,2}] & \gamma[h_{1,3}] & \dots & \gamma[h_{1,n}] & 1 \\ \gamma[h_{2,1}] & c_n & \gamma[h_{2,3}] & \dots & \gamma[h_{2,n}] & 1 \\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\ \gamma[h_{n,1}] & \gamma[h_{n,2}] & \gamma[h_{n,3}] & \dots & c_n & 1 \\ 1 & 1 & 1 & \dots & 1 & 0 \end{bmatrix}^{-1} \begin{bmatrix} \gamma[h_{1,o}] \\ \gamma[h_{2,o}] \\ \vdots \\ \gamma[h_{n,o}] \\ 1 \end{bmatrix}$$

$$\text{where } h_{i,j} = \sqrt{(x_i - x_j)^2 + (y_i - y_j)^2}$$

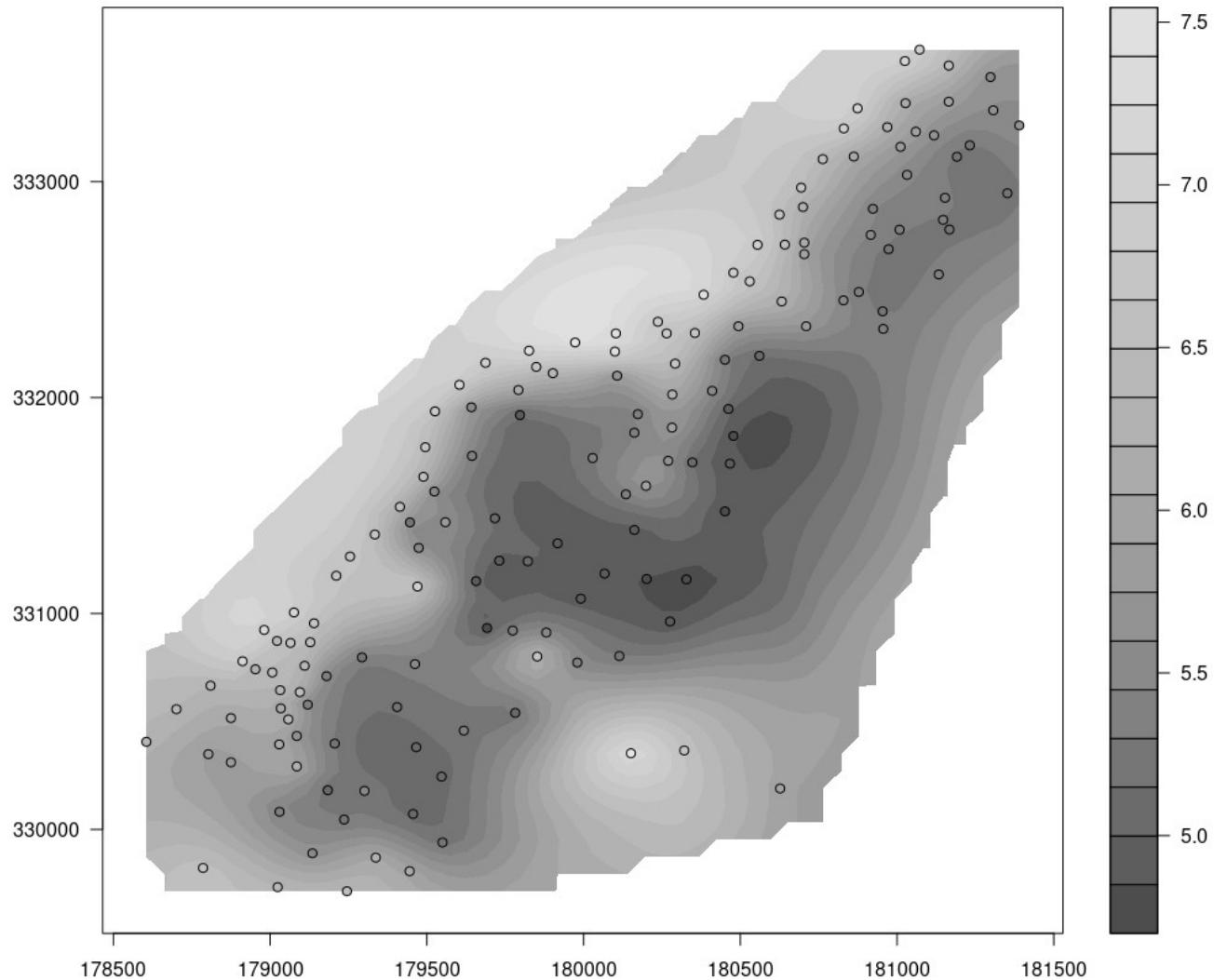
$$\begin{bmatrix} w_1 \\ w_2 \\ w_3 \\ \vdots \\ w_{50} \\ \vdots \\ w_{155} \\ \lambda \end{bmatrix} = \begin{bmatrix} 0.0022 \\ -0.00030 \\ 0.00022 \\ \vdots \\ -0.028 \\ \vdots \\ 0.0019 \\ 0.0015 \end{bmatrix} = \begin{bmatrix} 0.035 & 0.11 & 0.15 & \dots & 0.64 & \dots & 0.64 & 1 \\ 0.11 & 0.035 & 0.18 & \dots & 0.64 & \dots & 0.64 & 1 \\ 0.15 & 0.18 & 0.035 & \dots & 0.64 & \dots & 0.64 & 1 \\ \vdots & \vdots & \vdots & \ddots & \vdots & \ddots & \vdots & \vdots \\ 0.64 & 0.64 & 0.64 & \dots & 0.035 & \dots & 0.64 & 1 \\ \vdots & \vdots & \vdots & \ddots & \vdots & \ddots & \vdots & \vdots \\ 0.64 & 0.64 & 0.64 & \dots & 0.64 & \dots & 0.035 & 1 \\ 1 & 1 & 1 & \dots & 1 & \dots & 1 & 0 \end{bmatrix}_n^{-1} \begin{bmatrix} 0.64 \\ 0.64 \\ 0.64 \\ \vdots \\ 0.37 \\ \vdots \\ 0.64 \\ 1 \end{bmatrix}$$

$$z(x_o, y_o) = \sum_{i=1}^n w_i z(x_i, y_i)$$

$$\begin{aligned} z_{x_o, y_o} &= 0.0022 \times 6.93 - 0.00030 \times 7.04 + 0.00022 \times 6.46 + \dots \\ &\quad - 0.028 \times 5.93 + \dots + 0.0019 \times 5.93 = 5.01 \end{aligned}$$

$$\exp[5.01] = 150 \text{ ppm}$$

$\ln[\text{Zn}]$



$$s^2(x_o, y_o) = \lambda + \sum_{i=1}^n \gamma[h_{i,o}]w_i$$

For location  $\{x_o = 179850, y_o = 331650\}$ :

$$s^2(x_o, y_o) = 0.00029 + 0.64 \times 0.0022 - 0.64 \times 0.00030 + \dots = 0.22$$

**addition**       $z = a + bx + cy$        $s[z]^2 = b^2 s[x]^2 + c^2 s[y]^2 + 2bc s[x, y]$

**subtraction**       $z = ax - by$        $s[z]^2 = a^2 s[x]^2 + b^2 s[y]^2 - 2ab s[x, y]$

**multiplication**       $z = axy$        $\left(\frac{s[z]}{z}\right)^2 = \left(\frac{s[x]}{x}\right)^2 + \left(\frac{s[y]}{y}\right)^2 + 2\frac{s[x, y]}{xy}$

**division**       $z = a\frac{x}{y}$        $\left(\frac{s[z]}{z}\right)^2 = \left(\frac{s[x]}{x}\right)^2 + \left(\frac{s[y]}{y}\right)^2 - 2\frac{s[x, y]}{xy}$

**exponentiation**       $z = ae^{bx}$        $\left(\frac{s[z]}{z}\right)^2 = b^2 s[x]^2$

**logarithms**       $z = a \ln[bx]$        $s[z]^2 = a^2 \left(\frac{s[x]}{x}\right)^2$

**power**       $z = ax^b$        $\left(\frac{s[z]}{z}\right)^2 = b^2 \left(\frac{s[x]}{x}\right)^2$

$$s^2(x_{\circ}, y_{\circ}) = \lambda + \sum_{i=1}^n \gamma[h_{i,\circ}]w_i$$

For location  $\{x_{\circ} = 179850, y_{\circ} = 331650\}$ :

$$s^2(x_{\circ}, y_{\circ}) = 0.00029 + 0.64 \times 0.0022 - 0.64 \times 0.00030 + \dots = 0.22$$

$$s[z]/z = \sqrt{0.22} = 47\%$$

$s[\text{Zn}]/\text{Zn}$ 