

Applied Biostatistics I

slides based on a lecture by Dr. Alain Hauser

Stephan Peischl

Interfaculty Bioinformatics Unit, University of Bern

stephan.peischl@bioinformatics.unibe.ch

September 14, 2018

Part I

Introduction

Why (bio-)statistics??

- Statistics closely linked to probability theory
- Aim of probability theory: **modeling phenomena with uncertainty**
- Aim of statistics: performing **inference for probabilistic models**

Why (bio-)statistics??

- Statistics closely linked to probability theory
- Aim of probability theory: **modeling phenomena with uncertainty**
- Aim of statistics: performing **inference for probabilistic models**
- Sources of uncertainty:
 - ▶ variety of “samples” (e.g., individuals, cells, etc.)
 - ▶ missing control of variables influencing outcome
 - ▶ missing knowledge

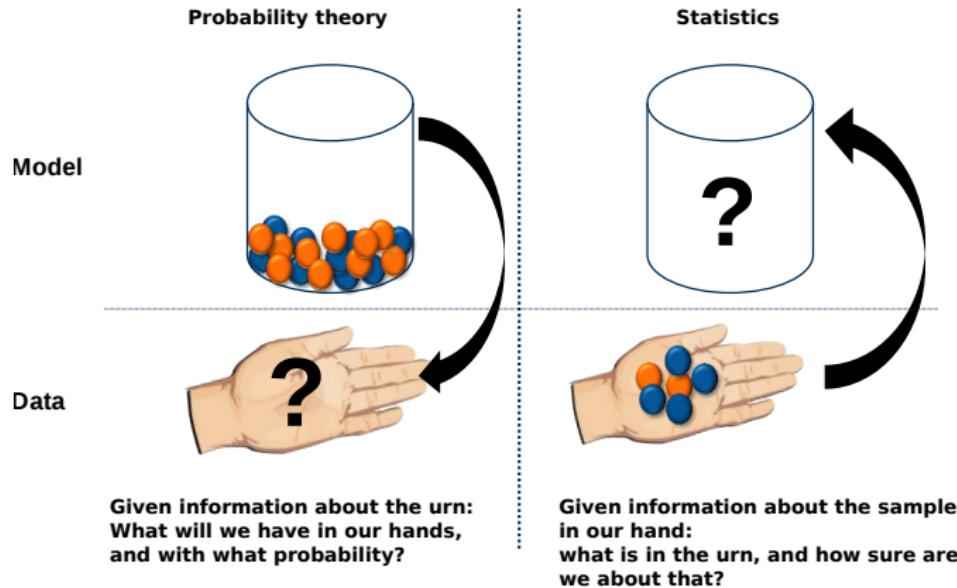
Statistics and probability theory

- Probability theory:
 - ▶ modeling systems with uncertainty
 - ▶ predicting data generated by such a system

Statistics and probability theory

- Probability theory:
 - ▶ modeling systems with uncertainty
 - ▶ predicting data generated by such a system
- Statistics:
 - ▶ collecting data
 - ▶ describing a model
 - ▶ inferring model parameters

Statistics and probability theory



(Source: Maier and Weiss (2013))

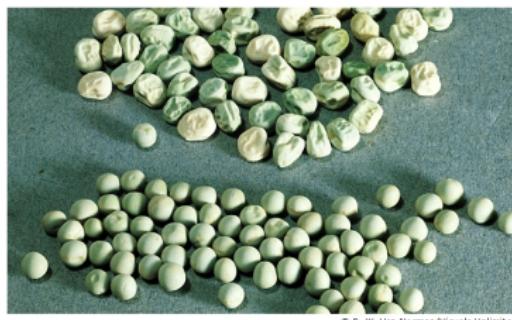
Examples of statistical questions in biology

- Have a look at 6 different examples of statistical questions in biology
- We will be able to tackle them at the end of the course

Example 1: hereditary traits in peas (Mendel's studies)

Mendel's experiment with peas:

- breeding pea plants that have either round or wrinkled peas
- Crossing plants from generation P ("parental generation") grown from round peas with plants grown from wrinkled peas \rightsquigarrow only round peas; generation F₁ ("filial generation")
- Crossing plants from generation F₁ \rightsquigarrow round *and* wrinkled peas; generation F₂



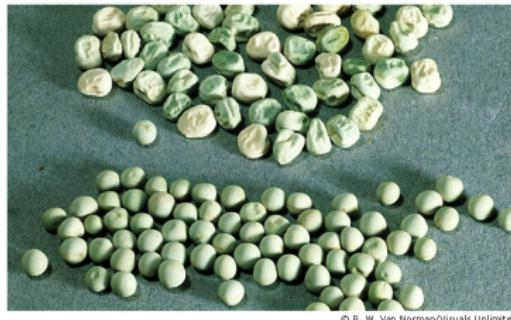
© R. W. Van Norman/Visuals Unlimited

(Source: Van Norman (1971))

Example 1: hereditary traits in peas (Mendel's studies)

Mendel's experiment with peas:

- breeding pea plants that have either round or wrinkled peas
- Crossing plants from generation P ("parental generation") grown from round peas with plants grown from wrinkled peas \rightsquigarrow only round peas; generation F₁ ("filial generation")
- Crossing plants from generation F₁ \rightsquigarrow round *and* wrinkled peas; generation F₂
- Experiment: counting round and wrinkled peas after crossing generation F₁



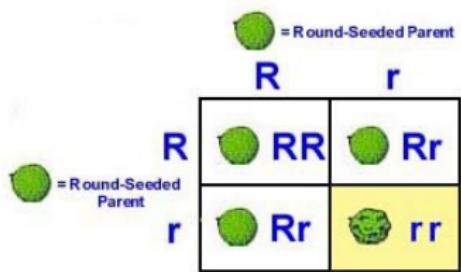
© R. W. Van Norman/Visuals Unlimited

(Source: Van Norman (1971))

Example 1: hereditary traits in peas (Mendel's studies)

Modern explanation:

- one gene with alleles coding for round (R, dominant) or wrinkled (r, recessive) pea



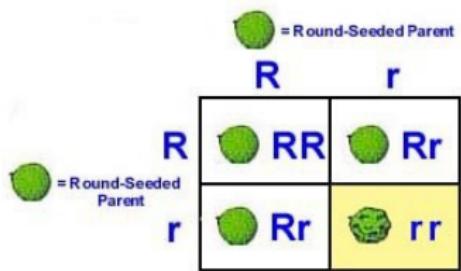
(Source:

<http://evolpsychology.blogspot.ch/>)

Example 1: hereditary traits in peas (Mendel's studies)

Modern explanation:

- one gene with alleles coding for round (R, dominant) or wrinkled (r, recessive) pea
- Generation P: homozygous, genotype either RR or rr



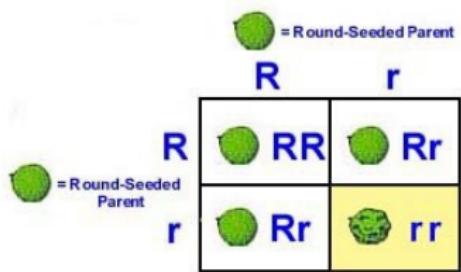
(Source:

<http://evolpsychology.blogspot.ch/>)

Example 1: hereditary traits in peas (Mendel's studies)

Modern explanation:

- one gene with alleles coding for round (R, dominant) or wrinkled (r, recessive) pea
- Generation P: homozygous, genotype either RR or rr
- Generation F₁: heterozygous, genotype Rr



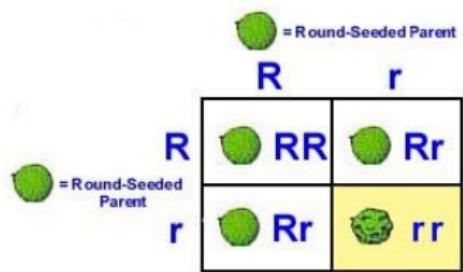
(Source:

<http://evolpsychology.blogspot.ch/>)

Example 1: hereditary traits in peas (Mendel's studies)

Modern explanation:

- one gene with alleles coding for round (R, dominant) or wrinkled (r, recessive) pea
- Generation P: homozygous, genotype either RR or rr
- Generation F₁: heterozygous, genotype Rr
- Generation F₂: genotypes RR, Rr and rr in ration 1 : 2 : 1



(Source:

<http://evolpsychology.blogspot.ch/>)

Example 1: hereditary traits in peas (Mendel's studies)

Data from Mendel's experiments: number of round and wrinkled peas (F_2) in 10 plants of generation F_1 :

Plant	1	2	3	4	5	6	7	8	9	10
round	45	27	24	19	32	26	88	22	28	25
wrinkled	12	8	7	10	11	6	24	10	6	7
ratio: $X : 1$	3.8	3.4	3.4	1.9	2.9	4.3	3.7	2.2	4.7	3.6

(Source: Stahel (2002))

Example 1: hereditary traits in peas (Mendel's studies)

Data from Mendel's experiments: number of round and wrinkled peas (F_2) in 10 plants of generation F_1 :

Plant	1	2	3	4	5	6	7	8	9	10
round	45	27	24	19	32	26	88	22	28	25
wrinkled	12	8	7	10	11	6	24	10	6	7
ratio: $X : 1$	3.8	3.4	3.4	1.9	2.9	4.3	3.7	2.2	4.7	3.6

(Source: Stahel (2002))

Do these numbers support Mendel's inheritance laws? Are the numbers really random deviations from the expected ratio 3 : 1?

Example 2: vaccine against Anthrax

- Anthrax: lethal bacterial disease of sheep and cattle
- Experiment of Louis Pasteur 1881: vaccinate 24 sheep, 24 unvaccinated sheep as control group
- Infect all 48 sheep with anthrax bacillus

Example 2: vaccine against Anthrax

- Anthrax: lethal bacterial disease of sheep and cattle
- Experiment of Louis Pasteur 1881: vaccinate 24 sheep, 24 unvaccinated sheep as control group
- Infect all 48 sheep with anthrax bacillus

Result:	Treatment	vaccinated	not vaccinated
		Died	24
	Survived	24	0

(Source: Samuels et al. (2012))

Example 3: influence of bacteria to cancer

- Experiment with strain of mice with high incidence of liver tumors
- One group maintained germ free, one group exposed to *Escherichia coli*

Example 3: influence of bacteria to cancer

- Experiment with strain of mice with high incidence of liver tumors
- One group maintained germ free, one group exposed to *Escherichia coli*

Result:

Treatment	<i>E. coli</i>	germ free
liver tumors	8	19
no liver tumors	5	30
Percent with lever tumors	62%	39%

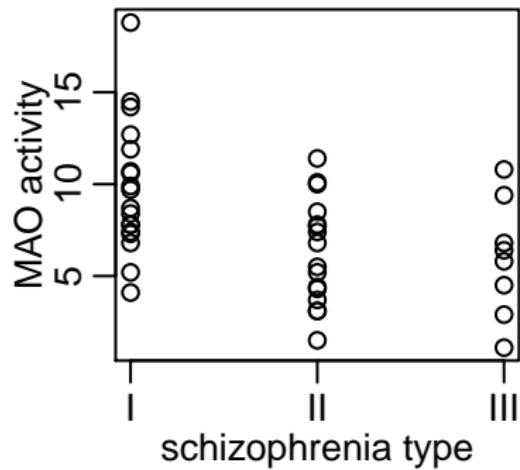
(Source: Mizutani and Mitsuoka (1979))

Can one conclude that *E. coli* have an influence on the incidence of liver cancer?

Example 4: Monoamine oxidase and schizophrenia

- Monoamine oxidase (MAO): enzyme thought to play a role in regulation of behavior
- Study: measured levels of MAO activity in 42 patients with different three types of schizophrenia

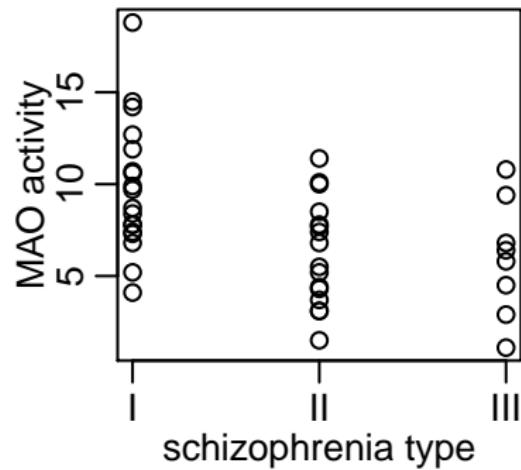
(Source: Potkin et al. (1978))



Example 4: Monoamine oxidase and schizophrenia

- Monoamine oxidase (MAO): enzyme thought to play a role in regulation of behavior
- Study: measured levels of MAO activity in 42 patients with different three types of schizophrenia

(Source: Potkin et al. (1978))

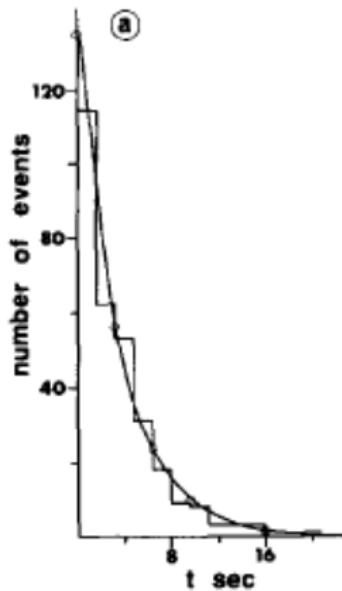


Are different types of schizophrenia associated with different levels of MAO activity?

Example 5: cell firing times of a neuron

- Analysis of measured time intervals between signals of neurons
- Distribution of intervals depicted to the right

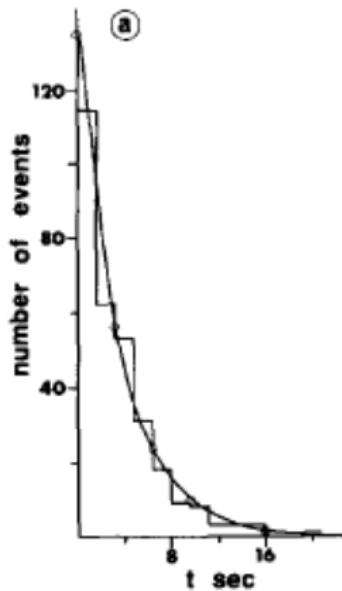
(Source: Nurse (1981))



Example 5: cell firing times of a neuron

- Analysis of measured time intervals between signals of neurons
- Distribution of intervals depicted to the right

(Source: Nurse (1981))

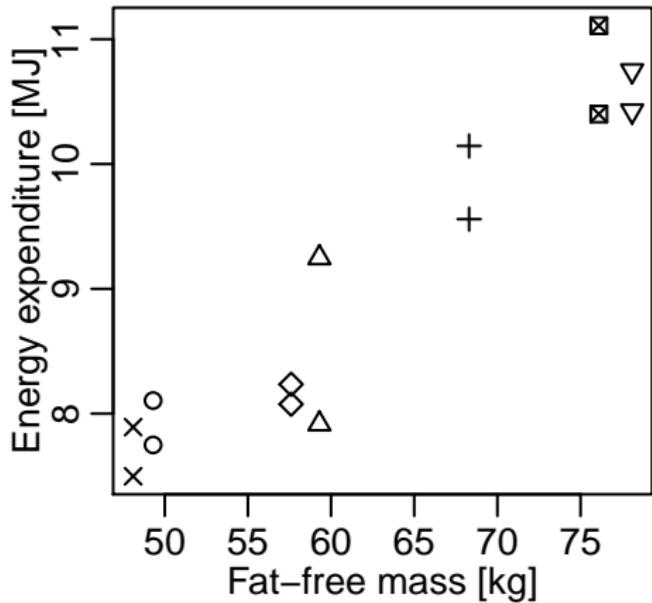


How to “describe the findings”? How to model the cell firing intervals?
Best model to predict or simulate cell firing times?

Example 6: body mass and energy expenditure

Measurements of fat-free mass
and energy expenditure in 24 h
of 7 subjects

(Source: Webb (1981))

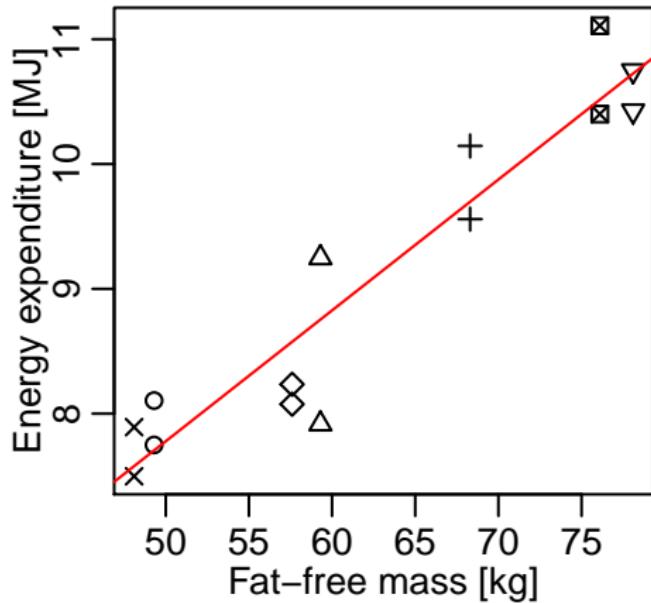


Example 6: body mass and energy expenditure

Measurements of fat-free mass
and energy expenditure in 24 h
of 7 subjects

(Source: Webb (1981))

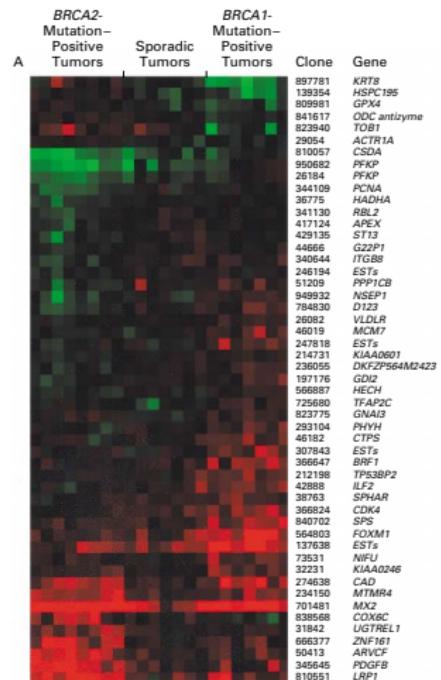
Description of the connection between both variables?
Model to predict energy expenditure based on mass?



Example 7: finding differentially expressed genes

- Measurements of expression levels of 3170 genes in two different types of breast cancer
- Goal: find genes that are differentially expressed in both types
- Heat map (green = high expression, red = low expression) of 51 genes with “strong differential expression” on the right

(Source: Hedenfalk et al. (2001))

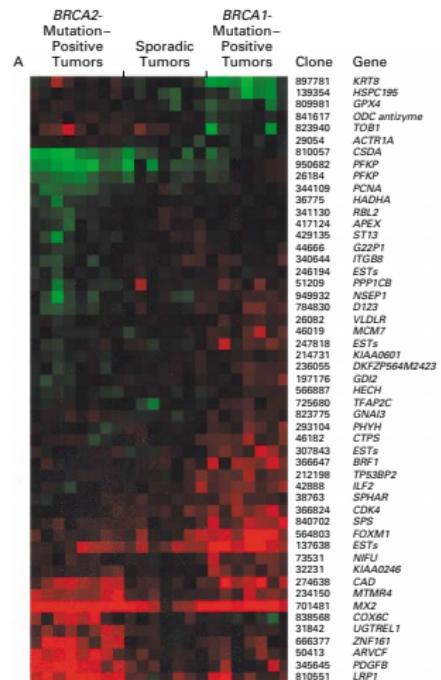


Example 7: finding differentially expressed genes

- Measurements of expression levels of 3170 genes in two different types of breast cancer
- Goal: find genes that are differentially expressed in both types
- Heat map (green = high expression, red = low expression) of 51 genes with “strong differential expression” on the right

(Source: Hedenfalk et al. (2001))

How to find genes that are differentially expressed? How to distinguish signal from noise?



Organization of lecture

For more details see organizational sheet.

- Lecture: Monday, 14:15 – 16:00, lecture room C 159, Baltzerstrasse 4
- Exercises: Thursday, 16:15 – 18:00, lecture room C 159, Baltzerstrasse 4
- Exam: Friday, 21.12.2018, time and place TBA
- Exercises will be graded based on participation and presentations.
More details on Thursday in the first session.
- Final grade: weighted mean of exercise grade (20%) and grade from written exam (80%)

Course contents and focus

Topics

- Probability and combinatorics
- Probability distributions
- Descriptive statistics
- Law of large numbers
- Frequentist inference
- Bayesian inference
- Hypothesis testing
- Linear regression

Course contents and focus

Topics

- Probability and combinatorics
- Probability distributions
- Descriptive statistics
- Law of large numbers
- Frequentist inference
- Bayesian inference
- Hypothesis testing
- Linear regression

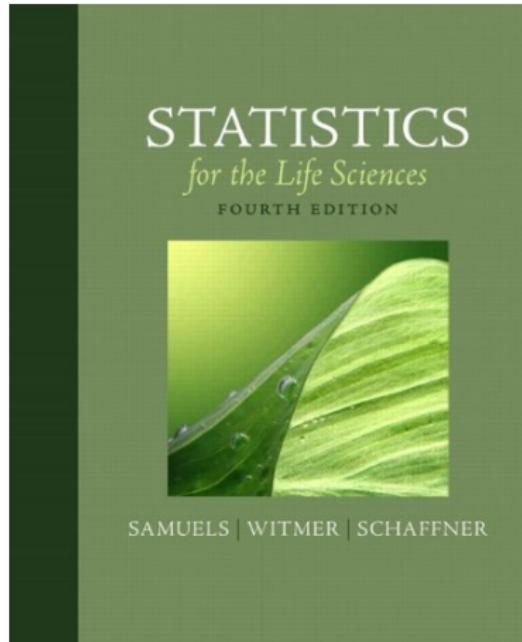
Focus: Applied Bio-Statistics

- Statistics: probabilistic models, inference methods
- Bio: biologically relevant models and methods, biological examples
- Applied: more and more R during course, hands-on exercises
- Wrap-up of introductory course in statistics

Literature

Large parts of the lectures are based on

M. L. Samuels, J. A. Witmer, A. Schaffner: *Statistics for the Life Sciences*, Pearson, 2012



SAMUELS | WITMER | SCHAFFNER

Other courses in biostatistics at UniBE

Introductory

- Prof. Dr. L. Dümbgen: *Statistik für Naturwissenschaften*, AS
- Prof. Dr. L. Excoffier and Dr. V. Sousa: *Introduction to R Programming and Analyses*, AS
- Dr. D. Prati: *Statistical Analysis of Experiments for Ecologists*, AS

Other courses in biostatistics at UniBE

Introductory

- Prof. Dr. L. Dümbgen: *Statistik für Naturwissenschaften*, AS
- Prof. Dr. L. Excoffier and Dr. V. Sousa: *Introduction to R Programming and Analyses*, AS
- Dr. D. Prati: *Statistical Analysis of Experiments for Ecologists*, AS

Advanced

- Prof. Dr. L. Excoffier, Prof. Dr. W. Nentwig: *Ökologie und Populationsgenetik*, AS
- Dr. S. Peischl: *Applied Biostatistics II*, SS
- Prof. Dr. D. Wegmann (Uni Fribourg): *Machine learning in bioinformatics*, SS
- Dr. D. Prati: course on mixed effects models
- Prof. Dr. D. Newbery: course on spacial statistics

Part II

Probability

Learning objectives

- Know the basic concepts of probability: event, sample space, probability measure, independence, conditional probability
- Explain the frequentist and Bayesian interpretation of probability
- Draw and read Venn diagrams
- Draw and read probability trees
- Calculate conditional probabilities using Bayes' theorem

Suggested literature

This lecture is partly based on the following source:

- Samuels et al. (2012), Chapters 3.2 and 3.3

Probability theory

- Experiments are always “random”: under “same conditions”, we get different results
- Reasons for randomness:
 - ▶ Inherent randomness: some processes in biology are simply not deterministic. Example: segregation of chromosomes in the formation of gametes.
 - ▶ Incomplete control of experimental conditions: e.g. due to genetic variations of samples, not fully controllable experimental environment
- Aim of probability theory: describe variability of the results

Important concepts

Definition (Events, sample space)

An **elementary event** ω is a possible outcome of an experiment.

The **sample space** Ω is the set of all elementary events.

An **event** A is a subset of the sample space ($A \subset \Omega$).

Examples:

Important concepts

Definition (Events, sample space)

An **elementary event** ω is a possible outcome of an experiment.

The **sample space** Ω is the set of all elementary events.

An **event** A is a subset of the sample space ($A \subset \Omega$).

Examples:

- ① Tossing a die. Sample space: all possible numbers we can get, hence $\Omega = \{1, 2, \dots, 6\}$. Possible event “die shows an even number”: $A = \{2, 4, 6\}$.

Important concepts

Definition (Events, sample space)

An **elementary event** ω is a possible outcome of an experiment.

The **sample space** Ω is the set of all elementary events.

An **event** A is a subset of the sample space ($A \subset \Omega$).

Examples:

- ① Tossing a die. Sample space: all possible numbers we can get, hence $\Omega = \{1, 2, \dots, 6\}$. Possible event “die shows an even number”: $A = \{2, 4, 6\}$.
- ② Sampling *D. melanogaster* from a large population (n flies). Elementary event: sampling one particular fly. Example of an event: sampling a fly with vestigial wings.

Probability

Definition (Probability)

Let Ω be a sample space. A **probability measure** is a function that assigns a value between 0 and 1 to an event $A \subset \Omega$: $P[A] \in [0, 1]$. It obeys the following properties (axioms of Kolmogorov):

- i) $0 \leq P[A] \leq 1$ for every event $A \subset \Omega$
- ii) $P[\Omega] = 1$
- iii) $P[A \cup B] = P[A] + P[B]$ for *disjoint* event A and B .

Examples (continued):

Probability

Definition (Probability)

Let Ω be a sample space. A **probability measure** is a function that assigns a value between 0 and 1 to an event $A \subset \Omega$: $P[A] \in [0, 1]$. It obeys the following properties (axioms of Kolmogorov):

- i) $0 \leq P[A] \leq 1$ for every event $A \subset \Omega$
- ii) $P[\Omega] = 1$
- iii) $P[A \cup B] = P[A] + P[B]$ for *disjoint* event A and B .

Examples (continued):

- ① Tossing a die: for each number $\omega \in \{1, \dots, 6\}$, the probability is $P[\{\omega\}] = \frac{1}{6}$. The probability to get an even number is $P[A] = P[\{2\}] + P[\{4\}] + P[\{6\}] = \frac{1}{2}$.

Probability

Definition (Probability)

Let Ω be a sample space. A **probability measure** is a function that assigns a value between 0 and 1 to an event $A \subset \Omega$: $P[A] \in [0, 1]$. It obeys the following properties (axioms of Kolmogorov):

- i) $0 \leq P[A] \leq 1$ for every event $A \subset \Omega$
- ii) $P[\Omega] = 1$
- iii) $P[A \cup B] = P[A] + P[B]$ for *disjoint* event A and B .

Examples (continued):

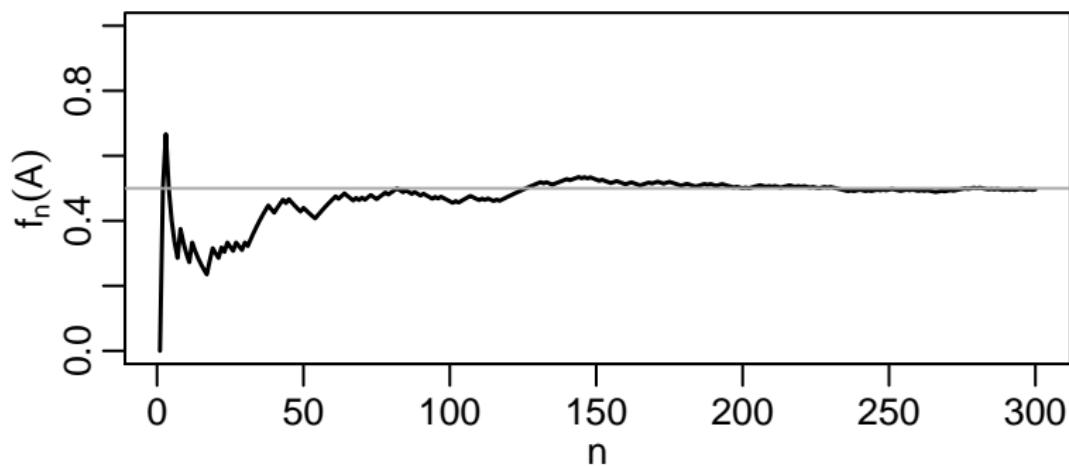
- ① Tossing a die: for each number $\omega \in \{1, \dots, 6\}$, the probability is $P[\{\omega\}] = \frac{1}{6}$. The probability to get an even number is $P[A] = P[\{2\}] + P[\{4\}] + P[\{6\}] = \frac{1}{2}$.
- ② Sampling one fruit fly out of a population of n : the probability of sampling one specific fly is $P[\{\omega\}] = \frac{1}{n}$. The probability of sampling a fly with vestigial wings depends on the number of such flies.

Interpretation of probabilities

- **Frequentist interpretation:** $P[A]$ is the relative frequency of event A in “infinitely many” experiments
- **Bayesian interpretation:** $P[A]$ is a measure for the belief that A will be the outcome of an experiment

Interpretation of probabilities

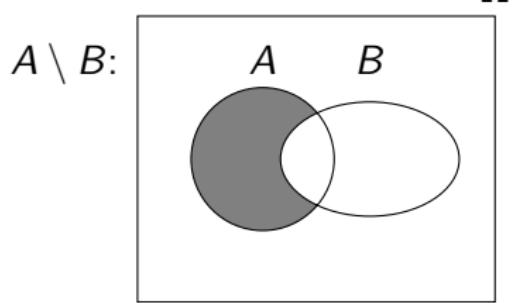
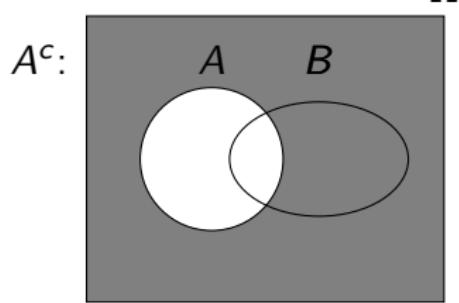
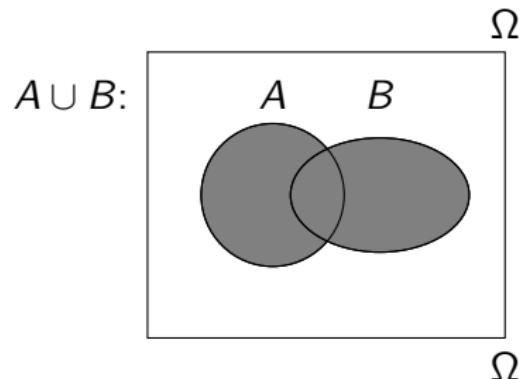
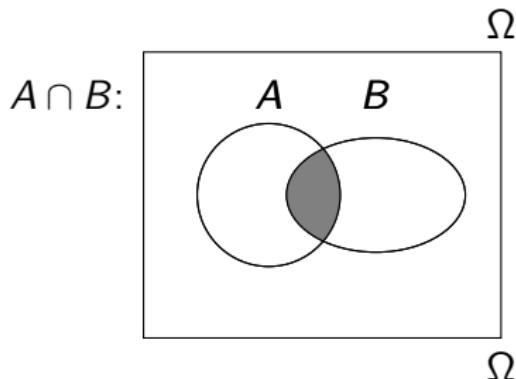
- **Frequentist interpretation:** $P[A]$ is the relative frequency of event A in “infinitely many” experiments
- **Bayesian interpretation:** $P[A]$ is a measure for the belief that A will be the outcome of an experiment



Relative frequency of the event $A = \text{"head"}$ after tossing a coin n times

Venn diagram: visualization of events

Events and set operations on them can be visualized with **Venn diagrams**:



Calculation rules for events and probabilities

Proposition (De Morgan's laws)

Let A and B be events. Then, $(A \cap B)^c = A^c \cup B^c$ and $(A \cup B)^c = A^c \cap B^c$.

Exercise: prove these rules using Venn diagrams!

Calculation rules for events and probabilities

Proposition (De Morgan's laws)

Let A and B be events. Then, $(A \cap B)^c = A^c \cup B^c$ and $(A \cup B)^c = A^c \cap B^c$.

Exercise: prove these rules using Venn diagrams!

Proposition (Probability of unions)

Let A and B be events. Then, $P[A \cup B] = P[A] + P[B] - P[A \cap B]$.

More general: let A_1, A_2, \dots, A_n be events. Then,

$$P[A_1 \cup A_2 \cup \dots \cup A_n] = \sum_{i_1=1}^n P[A_{i_1}] - \sum_{i_1=1}^{n-1} \sum_{i_2=i_1+1}^n P[A_{i_1} \cap A_{i_2}] + \\ \sum_{i_1=1}^{n-2} \sum_{i_2=i_1+1}^{n-1} \sum_{i_3=i_2+1}^n P[A_{i_1} \cap A_{i_2} \cap A_{i_3}] - \dots$$

Discrete probability spaces

- Assume finite (or countable) sample space: $\Omega = \{\omega_1, \omega_2, \dots\}$
- Probability of an event $A \subset \Omega$: $P[A] = \sum_{i:\omega_i \in A} P[\{\omega_i\}]$
- Normalization: $P[\Omega] = \sum_{i \geq 1} P[\{\omega_i\}]$

Discrete probability spaces

- Assume finite (or countable) sample space: $\Omega = \{\omega_1, \omega_2, \dots\}$
- Probability of an event $A \subset \Omega$: $P[A] = \sum_{i:\omega_i \in A} P[\{\omega_i\}]$
- Normalization: $P[\Omega] = \sum_{i \geq 1} P[\{\omega_i\}]$
- If Ω is *finite*, we often have $P[\{\omega_i\}] = 1/|\Omega|$. (Examples: tossing a die, sampling fruit flies)

Independence

Definition (Independence)

Two events A and B are called **independent** if $P[A \cap B] = P[A] \cdot P[B]$.

Example: independent assortment in *D. melanogaster* I

- Consider genes "h" (hairy body), "b" (black body), and "cn" (cinnabar eyes)
- b and cn lie on chromosome II, h on chromosome III
- Consider a hybrid with genotype $h^+h^-b^+b^-cn^+cn^-$.
- Assume we examine a gamete of this hybrid. Notation for events: H: gamete has allele h, B: gamete has allele b, C: gamete has allele cn.

Example: independent assortment in *D. melanogaster* II

- By Mendel's **law of independent assortment**, events H and B occur together with probability

$$P[H \cap B] = P[H] \cdot P[B] = \frac{1}{2} \cdot \frac{1}{2} = \frac{1}{4}$$

(h and b are inherited **independently**)

Example: independent assortment in *D. melanogaster* II

- By Mendel's **law of independent assortment**, events H and B occur together with probability

$$P[H \cap B] = P[H] \cdot P[B] = \frac{1}{2} \cdot \frac{1}{2} = \frac{1}{4}$$

(h and b are inherited **independently**)

- Mendel's law of independent assortment is not valid for b and cn because of **genetic linkage**. Events B and C occur together with probability

$$P[B \cap C] = \frac{1}{2}(1 - r) \text{ , where } r : \text{recombination frequency}$$

(b and cn are *not* inherited independently)

Conditional probabilities

Definition (Conditional probability)

Let A and B be events (with $P[B] > 0$). The **conditional probability of A given B** is defined as

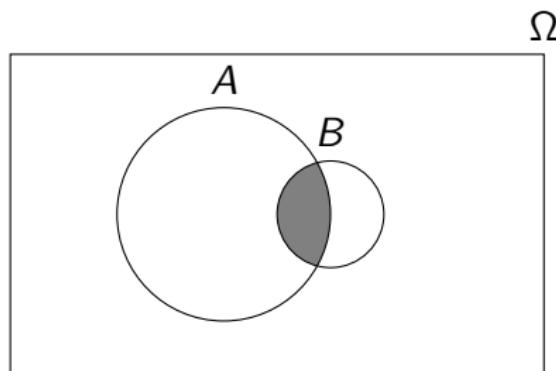
$$P[A|B] = \frac{P[A \cap B]}{P[B]} .$$

Conditional probabilities

Definition (Conditional probability)

Let A and B be events (with $P[B] > 0$). The **conditional probability of A given B** is defined as

$$P[A|B] = \frac{P[A \cap B]}{P[B]} .$$



(Source: Maier and Weiss (2013))

Example (continued): inheritance in *D. melanogaster*

- Probability that the gamete has allele h *given that* it has allele b:

$$P[H \mid B] = \frac{P[H \cap B]}{P[B]} = P[H]$$

Example (continued): inheritance in *D. melanogaster*

- Probability that the gamete has allele h *given that* it has allele b:

$$P[H \mid B] = \frac{P[H \cap B]}{P[B]} = P[H]$$

- Probability that the gamete has allele cn *given that* it has allele b:

$$P[C \mid B] = \frac{P[C \cap B]}{P[B]} = 1 - r$$

Example (continued): inheritance in *D. melanogaster*

- Probability that the gamete has allele h *given that* it has allele b:

$$P[H \mid B] = \frac{P[H \cap B]}{P[B]} = P[H]$$

- Probability that the gamete has allele cn *given that* it has allele b:

$$P[C \mid B] = \frac{P[C \cap B]}{P[B]} = 1 - r$$

General rules: let A and B be events with $P[A] > 0$ and $P[B] > 0$. Then:

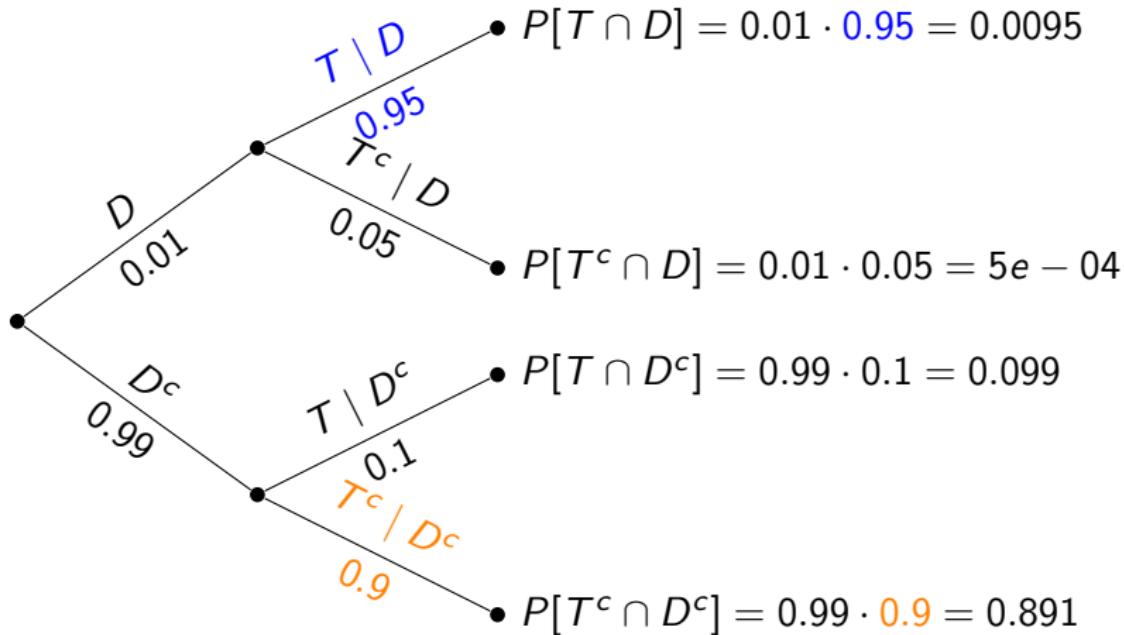
- $P[A \cap B] = P[A] \cdot P[B|A] = P[B] \cdot P[A|B]$
- A and B are independent $\Leftrightarrow P[A|B] = P[A] \Leftrightarrow P[B|A] = P[B]$

Example: medical test

- Consider a medical test for a rare disease. Test quite accurate: it detects the disease with 95% probability (**sensitivity** of the test), and indicates the *absence* of the disease with 90% probability (**specificity** of the test).
- Notation: event D : person has the disease; event T : test is positive (i.e., indicates disease)
- Disease has incidence of 1%: $P[D] = 0.01$.
- What is the probability that a random person gets a positive test result?

Probability trees

Medical test example can easily be visualized by a **probability tree**:



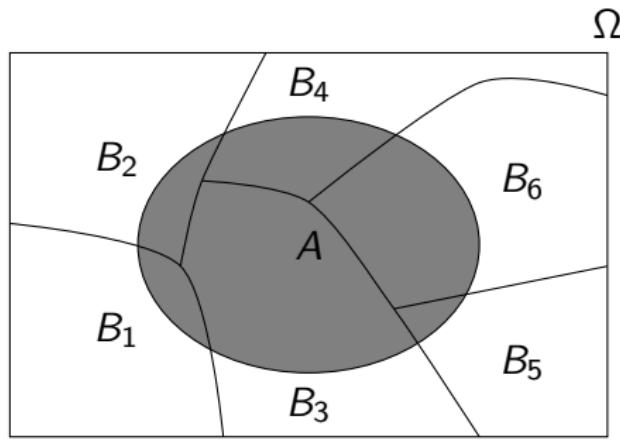
$$P[T] = P[T | D]P[D] + P[T | D^c]P[D^c] = 0.0095 + 0.099 = 0.1085$$

Law of total probability

Proposition (Law of total probability)

Assume B_1, B_2, \dots, B_k are disjoint events with $B_1, B_2, \dots, B_k = \Omega$. Then we can calculate the probability of any event A as

$$P[A] = \sum_{i=1}^k P[A \cap B_i] = \sum_{i=1}^k P[A|B_i]P[B_i].$$



(Source: Maier and Weiss (2013))

Bayes' theorem

Theorem (Bayes' theorem)

Let A and B be events with $P[A] > 0$ and $P[B] > 0$. Then we have:

$$P[B|A] = \frac{P[A \cap B]}{P[A]} = \frac{P[A|B]P[B]}{P[A]} .$$

In the setting of the law of total probability, we have

$$P[B_i|A] = \frac{P[A \cap B_i]}{P[A]} = \frac{P[A|B_i]P[B_i]}{\sum_{j=1}^k P[A|B_j]P[B_j]} .$$

Bayes' theorem

Theorem (Bayes' theorem)

Let A and B be events with $P[A] > 0$ and $P[B] > 0$. Then we have:

$$P[B|A] = \frac{P[A \cap B]}{P[A]} = \frac{P[A|B]P[B]}{P[A]} .$$

In the setting of the law of total probability, we have

$$P[B_i|A] = \frac{P[A \cap B_i]}{P[A]} = \frac{P[A|B_i]P[B_i]}{\sum_{j=1}^k P[A|B_j]P[B_j]} .$$

Example: medical test. The medical test from the previous example gives a positive result (i.e. indicates the disease). **What is the probability that you actually have the disease?**

Monty Hall problem

US American TV show: "Let's make a deal", moderated by Monty Hall:



- Main prize is hidden behind one of 3 doors.
- As a candidate, you choose one of the 3 doors.
- The show master opens a *different* door; the main prize is *not* there
- **Do you change the door, or do you stick with your initial choice?**

Part III

Probability Distributions

Learning objectives

- Know the definition of a random variable
- Model a measurement correctly with a discrete or continuous random variable
- Know how to specify the distribution of a random variable
- Recognize situations that must be modeled by binomial, Poisson, uniform, normal or exponential distributions

Suggested literature

This lecture is partly based on the following source:

- Samuels et al. (2012), Chapters 3.4, 3.5, 3.6; 4.1, 4.2, 4.3

Random variables

A **random variable** is a variable that takes numerical values which depend on the outcome of a random experiment.

Random variables

A **random variable** is a variable that takes numerical values which depend on the outcome of a random experiment.

More mathematical definition:

Definition (Random variable)

A **random variable** X is a function mapping a sample space Ω to \mathbb{R} (or a subset): $X : \Omega \rightarrow \mathbb{R}$.

Random variables

A **random variable** is a variable that takes numerical values which depend on the outcome of a random experiment.

More mathematical definition:

Definition (Random variable)

A **random variable** X is a function mapping a sample space Ω to \mathbb{R} (or a subset): $X : \Omega \rightarrow \mathbb{R}$.

A random variable X induces a probability measure on \mathbb{R} : for an event (set) $A \subset \mathbb{R}$, we have $P_X[A] = P[X^{-1}(A)]$.

Shorter and more intuitive notation:

- events: $\{\omega \in \Omega | X(\omega) \in A\} \equiv \{X \in A\}$;
- probabilities: $P_X[A] \equiv P[X \in A]$.

Random variables: examples

- ① **Family size:** choose a family at random from a population, let X be the number of children. Possible values: $X \in \mathbb{N} = \{0, 1, 2, 3, \dots\}$. If e.g. 23% of the families have 2 children, then $P[X = 2] = 0.23$; if 72% of the families have *at most* 2 children, then $P[X \in \{0, 1, 2\}] = P[X \leq 2] = 0.72$.

Random variables: examples

- ① **Family size:** choose a family at random from a population, let X be the number of children. Possible values: $X \in \mathbb{N} = \{0, 1, 2, 3, \dots\}$. If e.g. 23% of the families have 2 children, then $P[X = 2] = 0.23$; if 72% of the families have *at most* 2 children, then $P[X \in \{0, 1, 2\}] = P[X \leq 2] = 0.72$.
- ② **Length of fish:** measure the length of a fish randomly sampled from a population, let X be the length in cm. If e.g. 64% of the fish have a length between 11.5 cm and 16.2 cm, then $P[X \in [11.5, 16.2]] = P[11.5 \leq X \leq 16.2] = 0.64$.
What's the probability that the fish has *exactly* a length of 14 cm?

Random variables: notation

- Capital letter, e.g. X : random variable; lower case letter, e.g. x : realized value.
- $\{X = x\}$: elementary event that random variable X takes value x .

Random variables: notation

- Capital letter, e.g. X : random variable; lower case letter, e.g. x : realized value.
- $\{X = x\}$: elementary event that random variable X takes value x .

In words:

- Capital letter: description of an experiment (e.g., “measurement of the length of a fish”)
- Lower case letter: outcome of the experiment (e.g., 135 mm)

Probability distributions

Next goal: describe “distributions” of random variables, i.e., indicate how likely it is that they lie in a certain range

Several quantities of interest:

- Cumulative distribution function
- Probability mass function, probability density
- Expectation value
- Variance
- ...

Cumulative distribution function

Definition (Cumulative distribution function)

The **cumulative distribution function** (CDF) of a random variable X is defined as $F_X(x) := P[X \leq x]$.

Cumulative distribution function

Definition (Cumulative distribution function)

The **cumulative distribution function** (CDF) of a random variable X is defined as $F_X(x) := P[X \leq x]$.

Properties of a CDF F_X :

- F_X is monotonically increasing
- $\lim_{x \rightarrow -\infty} F_X(x) = 0, \lim_{x \rightarrow \infty} F_X(x) = 1$
- $P[a < X \leq b] = F_X(b) - F_X(a)$

Discrete random variables

- Discrete random variable: random variable with finite (or countable) image (i.e., set of possible values): $X : \Omega \rightarrow \{x_1, x_2, \dots\}$
- Characterized by **probability mass function** $p(x_k) := P[X = x_k]$
- Properties:
 - ▶ For each set $A \subset \{x_1, x_2, \dots\}$, we have

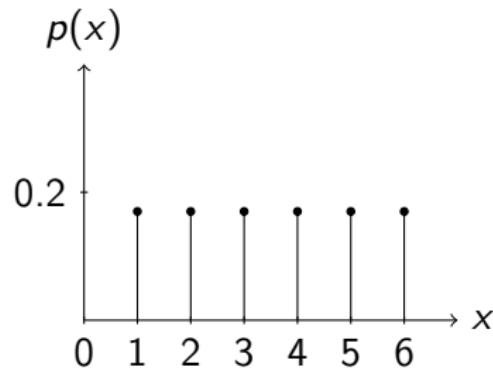
$$P[X \in A] = \sum_{k: x_k \in A} p(x_k)$$

- ▶ Normalization: $\sum_k p(x_k) = 1$
- ▶ Connection to CDF: $F_X(x) = P[X \leq x] = \sum_{k: x_k \leq x} p(x_k)$

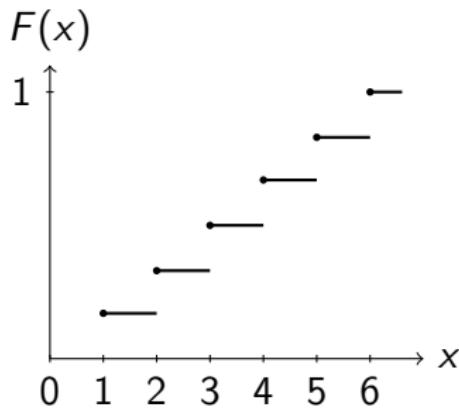
Example: fair die

A die can take values in $\{1, 2, \dots, 6\}$; if it is fair, it takes all values with the same probability. Its probability mass function and CDF look as follows:

Probability mass function



Cumulative distribution function



Expectation value and variance

Definition (Expectation value)

The **expectation value** of a discrete random variable X is defined as

$$E[X] := \sum_k x_k p(x_k) .$$

Expectation value and variance

Definition (Expectation value)

The **expectation value** of a discrete random variable X is defined as

$$E[X] := \sum_k x_k p(x_k) .$$

Definition (Variance)

The **variance** of a discrete random variable X is defined as

$$\text{Var}(X) := \sum_k (x_k - E[X])^2 p(x_k) .$$

Expectation value and variance

Definition (Expectation value)

The **expectation value** of a discrete random variable X is defined as

$$E[X] := \sum_k x_k p(x_k) .$$

Definition (Variance)

The **variance** of a discrete random variable X is defined as

$$\text{Var}(X) := \sum_k (x_k - E[X])^2 p(x_k) .$$

Example: fair die. Let X be the result of a fair die. X has expectation value $E[X] = 3.5$ and variance $\text{Var}(X) = 2.917$. (**Exercise: prove this!**)

Discrete probability distributions

We will now consider three discrete probability distributions widely used in biology (or in natural sciences in general):

- Bernoulli distribution
- binomial distribution
- Poisson distribution

Bernoulli distribution

The Bernoulli distribution is the simplest non-trivial discrete probability distribution:

Definition (Bernoulli distribution)

A discrete random variable X that can only take the values 0 and 1 is said to have **Bernoulli distribution**. The distribution is specified by the probability $\pi := P[X = 1]$.

We write $X \sim \text{Bernoulli}(\pi)$.

Binomial distribution

- Distribution of the sum of independent Bernoulli random variables
- Distribution of the number of “successes” of n independent trials with individual success probability π

Definition (Binomial distribution)

A discrete random variable $X \in \{0, 1, \dots, n\}$ has **Binomial distribution** if

$$p(x) = P[X = x] = \binom{n}{x} \pi^x (1 - \pi)^{n-x}.$$

We write $X \sim \text{Bin}(n, \pi)$, $n \in \mathbb{N}$, $\pi \in (0, 1)$.

Binomial distribution

- Distribution of the sum of independent Bernoulli random variables
- Distribution of the number of “successes” of n independent trials with individual success probability π

Definition (Binomial distribution)

A discrete random variable $X \in \{0, 1, \dots, n\}$ has **Binomial distribution** if

$$p(x) = P[X = x] = \binom{n}{x} \pi^x (1 - \pi)^{n-x}.$$

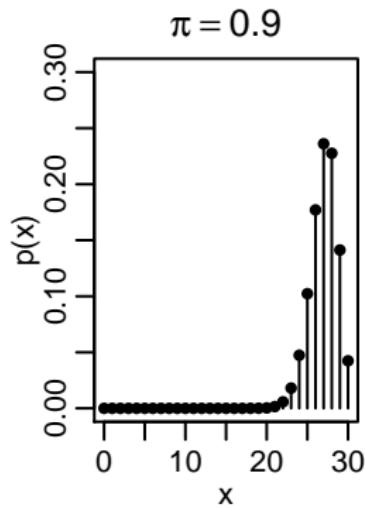
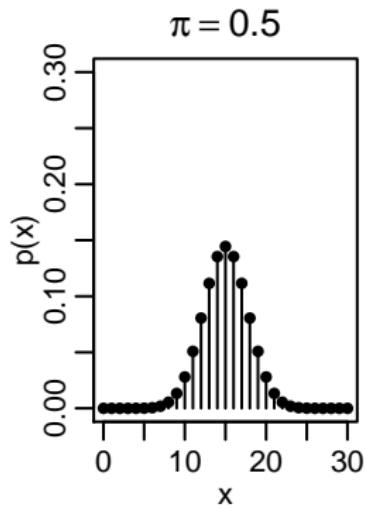
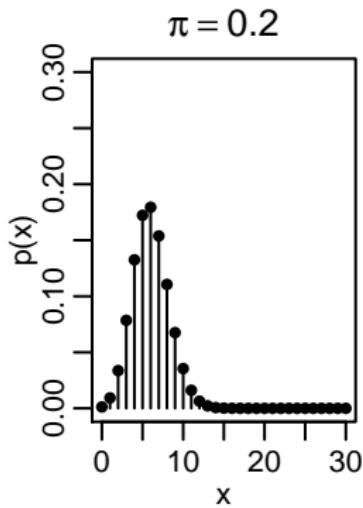
We write $X \sim \text{Bin}(n, \pi)$, $n \in \mathbb{N}$, $\pi \in (0, 1)$.

Expectation value: $E[X] = n\pi$, variance: $\text{Var}(X) = n\pi(1 - \pi)$

R functions: `dbinom` (probability mass function), `pbinom` (CDF)

Binomial distribution

Probability mass function of binomial distributions $\text{Bin}(50, \pi)$ for different probabilities π :



Example: test of a new drug

- A new drug is tested on $n = 200$ patients. Subjects with a rare genetic disposition (incidence of $\pi = \frac{1}{1000}$) may have severe side effects.
- What's the probability that at least one patient in the study has this genetic disposition?
- What's the probability that at least 3 patients have the disposition?

Poisson distribution

- Binomial distribution: range of possible values limited
- What if we consider the number of successes in potentially *infinitely* many trials?

Definition (Poisson distribution)

A discrete random variable $X \in \mathbb{N}$ has **Poisson distribution** with parameter λ if

$$p(x) = P[X = x] = e^{-\lambda} \frac{\lambda^x}{x!} .$$

We write $X \sim \text{Pois}(\lambda)$, $\lambda > 0$.

Poisson distribution

- Binomial distribution: range of possible values limited
- What if we consider the number of successes in potentially *infinitely* many trials?

Definition (Poisson distribution)

A discrete random variable $X \in \mathbb{N}$ has **Poisson distribution** with parameter λ if

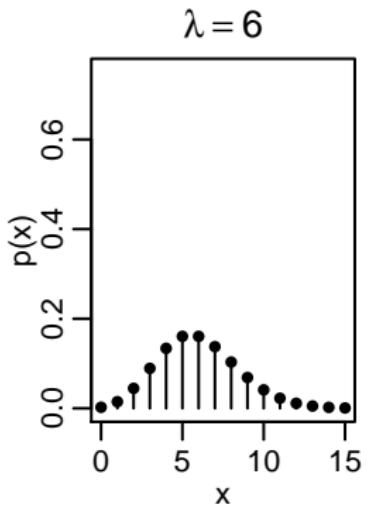
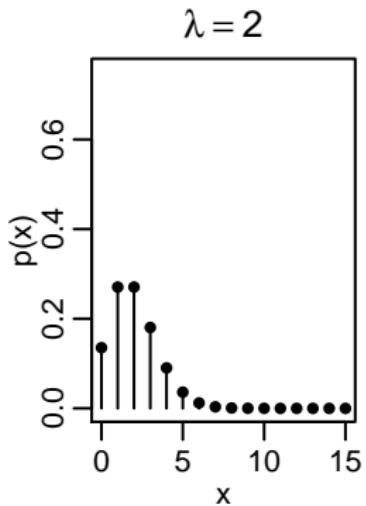
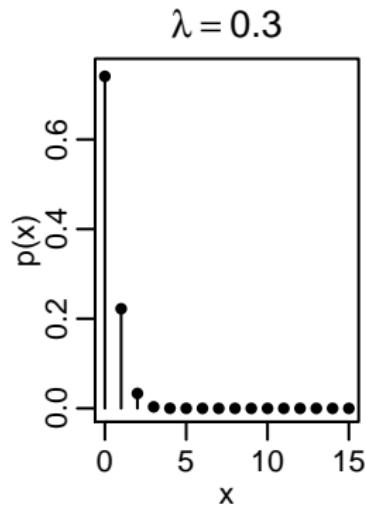
$$p(x) = P[X = x] = e^{-\lambda} \frac{\lambda^x}{x!} .$$

We write $X \sim \text{Pois}(\lambda)$, $\lambda > 0$.

Expectation value: $E[X] = \lambda$, variance: $\text{Var}(X) = \lambda$

R functions: `dpois` (probability mass function), `ppois` (CDF)

Poisson distribution



Poisson distribution

- Poisson distribution models the number of *rare* “successes” (or “events”) happening in a given time interval and/or space range

Poisson distribution

- Poisson distribution models the number of *rare* “successes” (or “events”) happening in a given time interval and/or space range
- More formally: a sequence of binomial distributions with $n\pi = \lambda$ fix converges to a Poisson distribution.

Poisson distribution

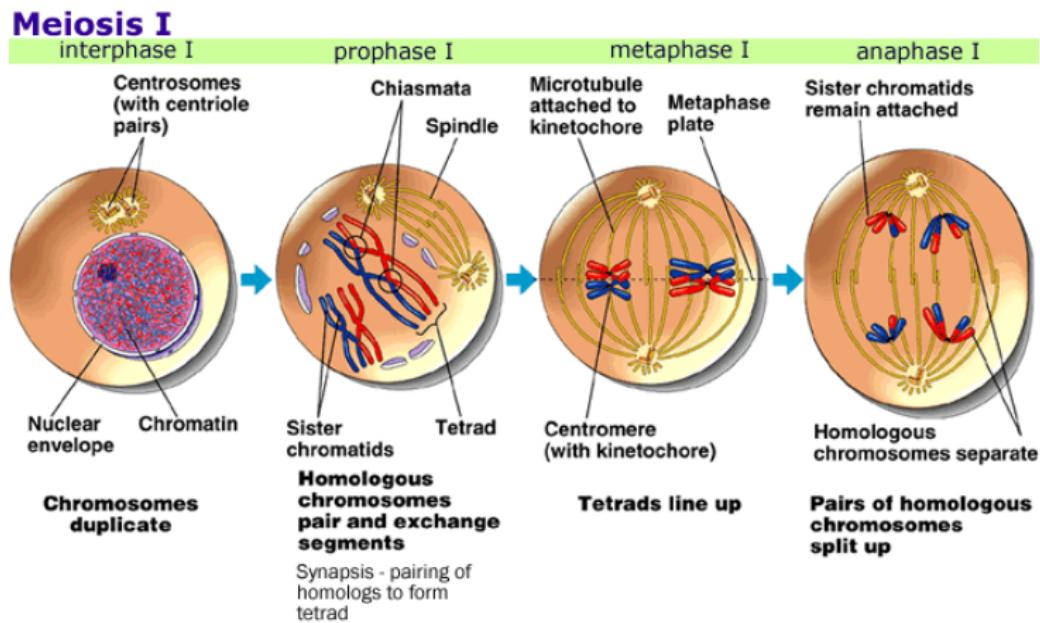
- Poisson distribution models the number of *rare* “successes” (or “events”) happening in a given time interval and/or space range
- More formally: a sequence of binomial distributions with $n\pi = \lambda$ fix converges to a Poisson distribution.
- Even more formally: consider a sequence of binomial random variables $X_n \sim \text{Bin}(n, \pi)$ with $n\pi = \lambda$ (fix!). Then, as $n \rightarrow \infty$,

$$P[X_n = x] \xrightarrow{n \rightarrow \infty} e^{-\lambda} \frac{\lambda^x}{x!} = p_{\text{Pois}}(x)$$

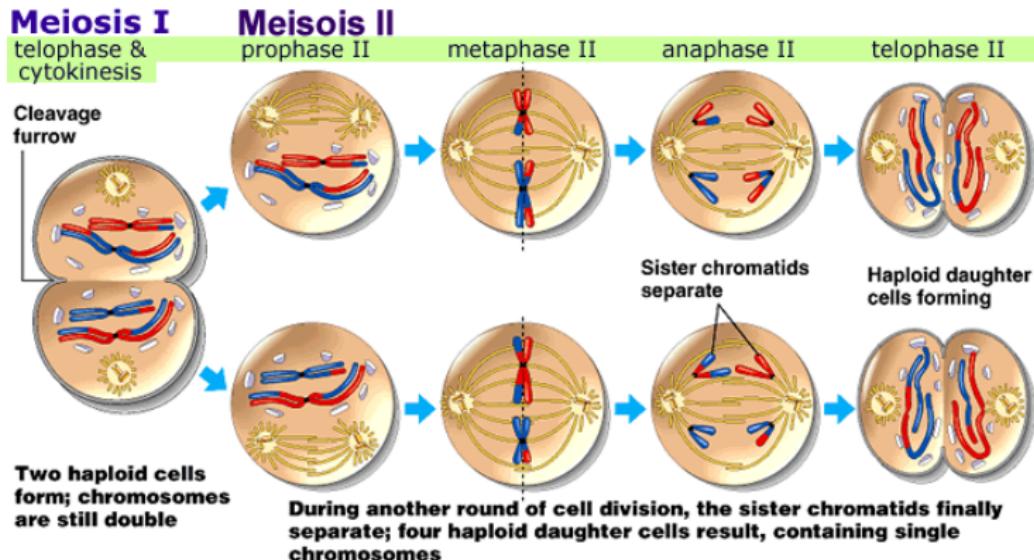
Haldane's model

- Simple mathematical model for crossovers in meiosis
- Model developed by John B. S. Haldane in 1919
- Model quite good for chromosomes that are not too short
- Model incorporates a lot of different probability distributions we will encounter in the lecture

Meiosis I



Meiosis II



Copyright © Pearson Education, Inc., publishing as Benjamin Cummings.

Haldane's model

- **Morgan:** distance measure for genes on a chromosome
- Two genes on a chromosome have a distance of d M = d Morgan if there are on average d crossovers between them during meiosis 1

Haldane's model

- **Morgan:** distance measure for genes on a chromosome
- Two genes on a chromosome have a distance of $d \text{ M} = d \text{ Morgan}$ if there are on average d crossovers between them during meiosis 1

Haldane's model

Let M be the number of crossovers between two genes with distance $d \text{ M}$ on a chromosome. M can be modeled as a Poisson distributed random variable: $M \sim \text{Pois}(d)$.

Note: in accordance with the definition of a Morgan, we have $E[M] = d$ in Haldane's model

Properties of Poisson distributions

Proposition (Sum of Poisson random variables)

Let $X \sim \text{Pois}(\lambda_1)$ and $Y \sim \text{Pois}(\lambda_2)$ be independent. Then,
 $X + Y \sim \text{Pois}(\lambda_1 + \lambda_2)$.

Properties of Poisson distributions

Proposition (Sum of Poisson random variables)

Let $X \sim \text{Pois}(\lambda_1)$ and $Y \sim \text{Pois}(\lambda_2)$ be independent. Then,
 $X + Y \sim \text{Pois}(\lambda_1 + \lambda_2)$.

Is $\frac{1}{2}(X + Y)$ also Poisson distributed?

Continuous random variables

- Continuous random variable: random variable whose image contains an interval, or \mathbb{R} .

Continuous random variables

- Continuous random variable: random variable whose image contains an interval, or \mathbb{R} .
- Better definition: random variable whose CDF F is differentiable (almost everywhere)

Continuous random variables

- Continuous random variable: random variable whose image contains an interval, or \mathbb{R} .
- Better definition: random variable whose CDF F is differentiable (almost everywhere)
- Characterized by **probability density** $f(x) := \frac{d}{dx} F(x)$
- For any set $A \subset \mathbb{R}$, $P[X \in A] = \int_A f(x) dx$; for $a < b$,
 $P[a < X < b] = \int_a^b f(x) dx$ (**area under the density curve**)

Continuous random variables

- Continuous random variable: random variable whose image contains an interval, or \mathbb{R} .
- Better definition: random variable whose CDF F is differentiable (almost everywhere)
- Characterized by **probability density** $f(x) := \frac{d}{dx} F(x)$
- For any set $A \subset \mathbb{R}$, $P[X \in A] = \int_A f(x) dx$; for $a < b$,
 $P[a < X < b] = \int_a^b f(x) dx$ (**area under the density curve**)
- Expectation value: $E[X] = \int_{\mathbb{R}} xf(x) dx$
- Variance: $\text{Var}(X) = \int_{\mathbb{R}} (x - E[X])^2 f(x) dx$

Linear transformations of random variables

Let X be a (continuous or discrete) random variable, and a and b two real numbers.

- $E[aX + b] = aE[X] + b$
- $\text{Var}(aX + b) = a^2 \text{Var}(X)$

Uniform distribution

Definition (Uniform distribution)

A continuous random variable X has **uniform distribution** in $[a, b]$ if

$$f(x) = \begin{cases} \frac{1}{b-a}, & x \in [a, b], \\ 0, & \text{otherwise.} \end{cases}$$

We write $X \sim \mathcal{U}([a, b])$.

Uniform distribution

Definition (Uniform distribution)

A continuous random variable X has **uniform distribution** in $[a, b]$ if

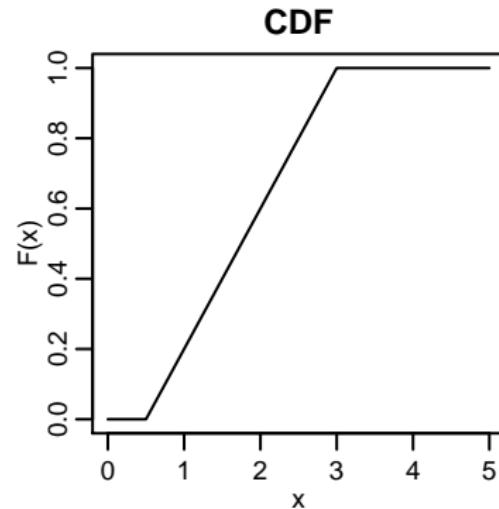
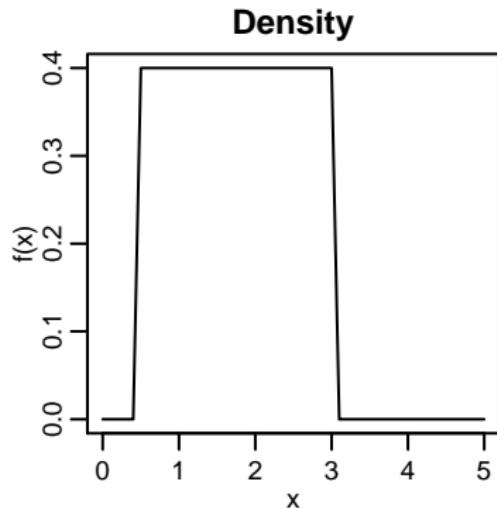
$$f(x) = \begin{cases} \frac{1}{b-a}, & x \in [a, b], \\ 0, & \text{otherwise.} \end{cases}$$

We write $X \sim \mathcal{U}([a, b])$.

Expectation value: $E[X] = \frac{b+a}{2}$, variance: $\text{Var}(X) = \frac{(b-a)^2}{12}$
R functions: `dunif` (probability density), `punif` (CDF)

Uniform distribution

Uniform distribution on the interval $[0.5, 3]$:



Example: position of crossover in Haldane's model

- Consider chromosome of length d (in Morgan). Assume that we know that exactly one crossover occurred ($M = 1$) on this chromosome during meiosis; let T denote the position of this crossover on the chromosome.
- Distribution of T : $T \mid M = 1 \sim \mathcal{U}([0, d])$ (mathematical derivation difficult)
- In words: given that one crossover occurred, every possible position of this crossover is equally likely. There is no preferred place for crossovers on a chromosome.

Normal distribution

Definition (Normal distribution)

A continuous random variable X has **normal distribution** with mean μ and variance σ^2 if

$$f(x) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left\{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2\right\}, \quad x \in \mathbb{R}.$$

We write $X \sim \mathcal{N}(\mu, \sigma^2)$, $\mu \in \mathbb{R}$, $\sigma^2 > 0$.

Normal distribution

Definition (Normal distribution)

A continuous random variable X has **normal distribution** with mean μ and variance σ^2 if

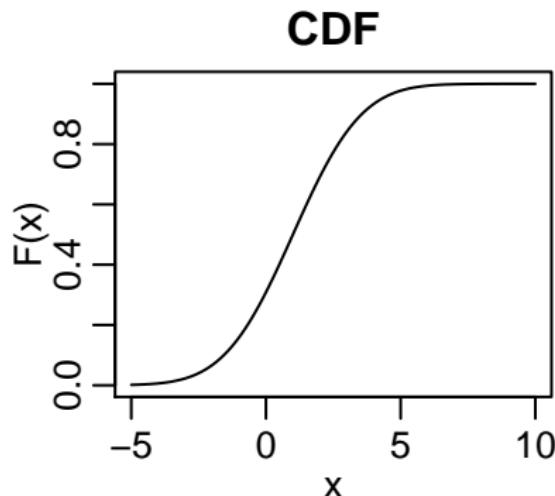
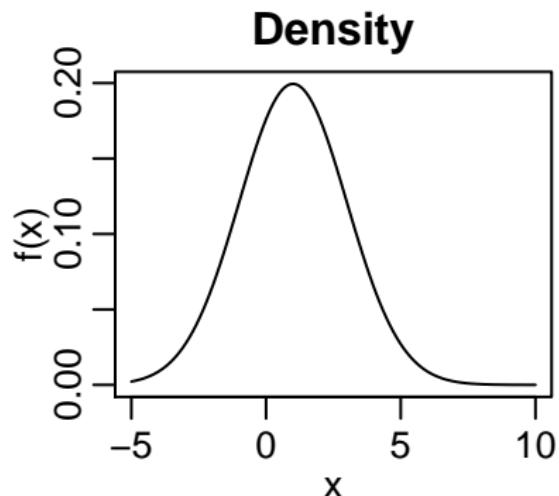
$$f(x) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left\{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2\right\}, \quad x \in \mathbb{R}.$$

We write $X \sim \mathcal{N}(\mu, \sigma^2)$, $\mu \in \mathbb{R}$, $\sigma^2 > 0$.

- Examples of normally distributed random variables: measurement errors, size of individuals, certain measurements of physiological quantities, ...
- Most important continuous probability distribution due to central limit theorem (see later): empirical mean of independent random variables of *any* probability distribution is approximately normally distributed

Normal distribution

Normal distribution with mean $\mu = 1$ and variance $\sigma^2 = 4$:



R functions: `dnorm` (probability density), `pnorm` (CDF)

Standard normal distribution

- **Standard normal distribution** $Z \sim \mathcal{N}(0, 1)$: very important in hypothesis testing, calculation of confidence intervals, etc.
- Special symbols reserved for its density and CDF:

$$\varphi(z) = \frac{1}{\sqrt{2\pi}} e^{-z^2/2}, \quad \Phi(z) = \int_{-\infty}^z \varphi(t) dt .$$

- No closed form for $\Phi(x)$ in terms of elementary functions

Standard normal distribution

- **Standard normal distribution** $Z \sim \mathcal{N}(0, 1)$: very important in hypothesis testing, calculation of confidence intervals, etc.
- Special symbols reserved for its density and CDF:

$$\varphi(z) = \frac{1}{\sqrt{2\pi}} e^{-z^2/2}, \quad \Phi(z) = \int_{-\infty}^z \varphi(t) dt .$$

- No closed form for $\Phi(x)$ in terms of elementary functions
- If $X \sim \mathcal{N}(\mu, \sigma^2)$ is an arbitrary, normally distributed random variable, $Z = aX + b$ is also normally distributed for any numbers a and b :

$$Z = aX + b \sim \mathcal{N}(a\mu + b, a^2\sigma^2)$$

- With a special choice of a and b , we can **standardize** the normal distribution:

$$a = \frac{1}{\sigma}, b = -\frac{\mu}{\sigma} \Rightarrow Z = \frac{X - \mu}{\sigma} \sim \mathcal{N}(0, 1)$$

Area under the curve

- Recall: for any continuous random variable with probability density f , the area under the curve encodes the probability that the variable takes a certain range of values: for $a < b$, $P[a \leq X \leq b] = \int_a^b f(x) dx$
- Area under the curve for $Z \sim \mathcal{N}(0, 1)$:

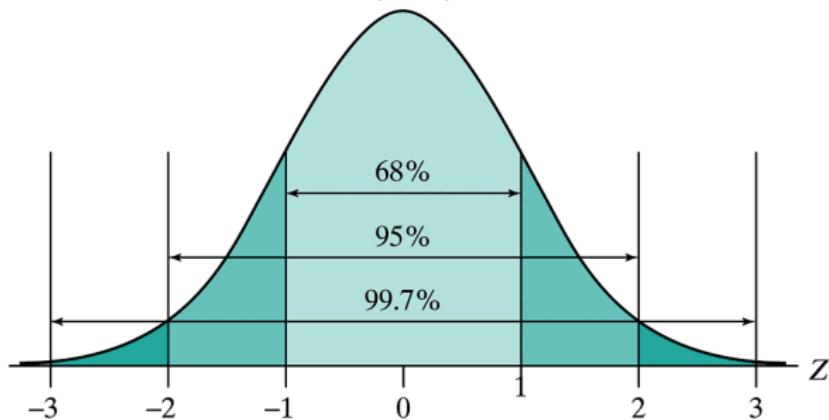


Figure 4.3.5 Areas under a standard normal curve between -1 and $+1$, between -2 and $+2$, and between -3 and $+3$

(Source: Samuels et al. (2012))

Example: lengths of fish

- The individuals of a fish population have a normally distributed length with mean 54.0 mm and standard deviation 4.5 mm.
- What percentage of the fish are less than 60 mm long?
- What percentage of the fish are more than 51 mm long?

Exponential distribution

Definition (Exponential distribution)

A random variable $X \geq 0$ has **exponential distribution** with parameter ("rate") λ if

$$f(x) = \begin{cases} \lambda e^{-\lambda x}, & x \geq 0, \\ 0, & \text{otherwise.} \end{cases}$$

We write $X \sim \text{Exp}(\lambda)$, $\lambda > 0$.

Exponential distribution

Definition (Exponential distribution)

A random variable $X \geq 0$ has **exponential distribution** with parameter ("rate") λ if

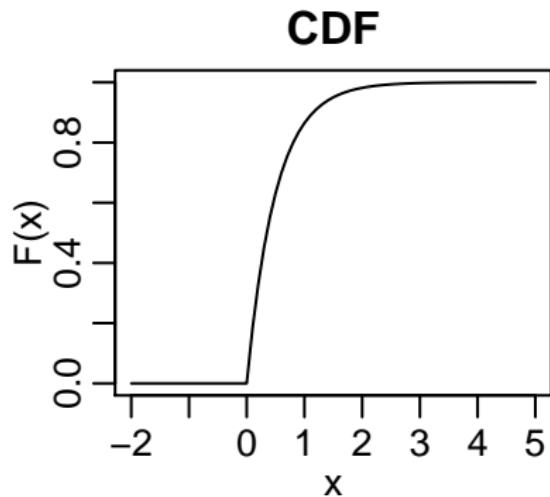
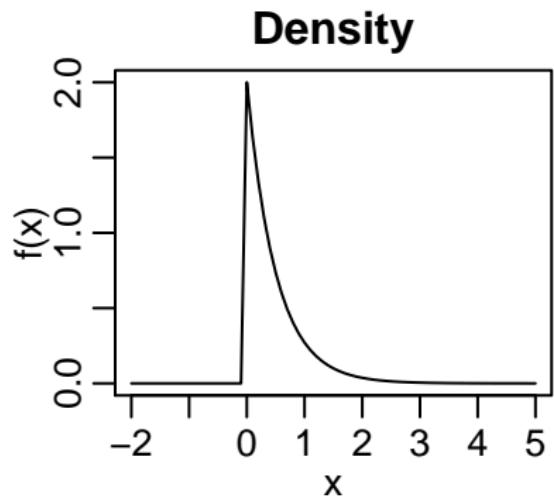
$$f(x) = \begin{cases} \lambda e^{-\lambda x}, & x \geq 0, \\ 0, & \text{otherwise.} \end{cases}$$

We write $X \sim \text{Exp}(\lambda)$, $\lambda > 0$.

Expectation value: $E[X] = \frac{1}{\lambda}$, variance: $\text{Var}(X) = \frac{1}{\lambda^2}$
R functions: `dexp` (probability density), `pexp` (CDF)

Exponential distribution

Exponential distribution for $\lambda = 2$:



Example: position of first crossover (Haldane's model)

- Consider a “very long” chromosome; on such, it is very probable that crossover occurs during meiosis.
- Let X be the position of the first crossover. How is X distributed?

Example: position of first crossover (Haldane's model)

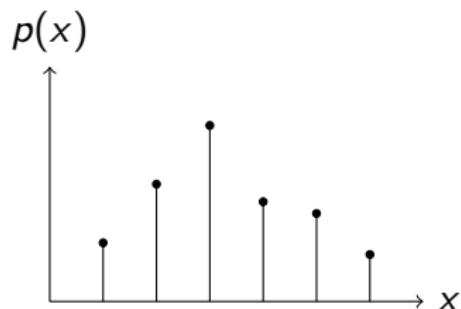
- Consider a “very long” chromosome; on such, it is very probable that crossover occurs during meiosis.
- Let X be the position of the first crossover. How is X distributed?
- $X \sim \text{Exp}(1)$: exponential distribution (the actual length of the chromosome does not matter)
- Derivation: see blackboard

Summary: discrete and continuous distributions

discrete

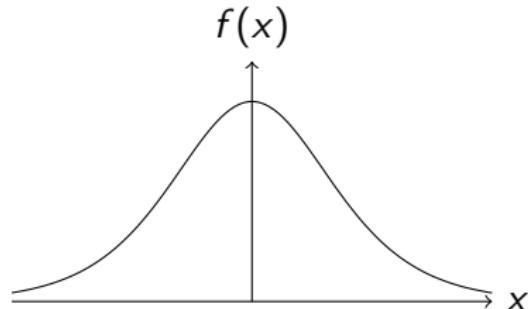
continuous

Probability mass function



$$P[X = x_k] = p(x_k) \in [0, 1], x_k \in W$$

Probability density



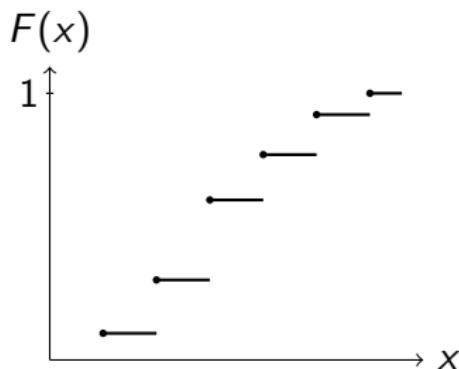
$$P[X = x] = 0, x \in \mathbb{R}$$

(Source: Maier and Weiss (2013))

Summary: discrete and continuous distributions

discrete

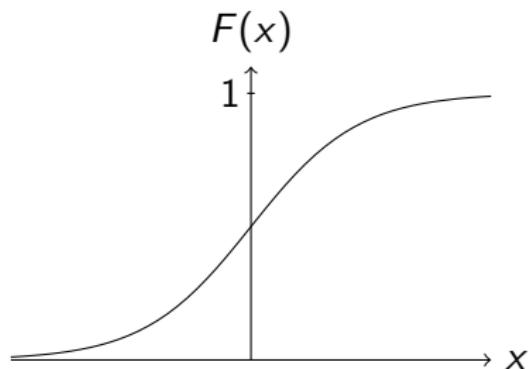
Cumulative distribution function



$$F(x) = \sum_{k : x_k \leq x} p(x_k)$$

continuous

Cumulative distribution function



$$F(x) = \int_{-\infty}^x f(u) du$$

Summary: discrete and continuous distributions

discrete	continuous
Expectation value $E[X] = \sum_{k \geq 1} x_k p(x_k)$	Expectation value $E[X] = \int_{-\infty}^{\infty} xf(x) dx$
Variance $\text{Var}(X) = \sum_{k \geq 1} (x_k - E[X])^2 p(x_k)$	Variance $\text{Var}(X) = \int_{-\infty}^{\infty} (x - E[X])^2 f(x) dx$

Part IV

Multivariate Distributions

Learning targets

- Writing down the joint probability distribution of two random variables
- Calculating the marginal and conditional probability distributions when given the joint distribution
- Recognizing independence from joint probability distribution
- Knowing the definition of covariance
- Calculating expectation and variance of a sum of random variables

Multivariate distributions

- Up to now: considered distribution of *single* random variable
- Next goal: modeling 2 (or more) random variables that are *not* independent

Discrete multivariate distributions

Let $X : \Omega \rightarrow W_X$ and $Y : \Omega \rightarrow W_Y$ be discrete random variables

Definition

The **joint cumulative distribution function** of X and Y is given by

$$F_{X,Y}(x,y) := P[X \leq x, Y \leq y] .$$

The **joint probability mass function** of X and Y is

$$p_{X,Y}(x,y) := P[X = x, Y = y], x \in W_X, y \in W_Y .$$

Example: joint probability mass function

- A joint probability mass function can be indicated as a table
- Example: discretized expression levels of two genes (X and Y): 1 = low expression, 2 = medium expression, 3 = high expression

X/Y	1	2	3
1	0.31	0.08	0.02
2	0.11	0.18	0.06
3	0.04	0.09	0.11

Marginal distribution

Let X and Y be two discrete random variables, and $p_{X,Y}$ their joint probability mass function.

Definition (Marginal probability mass function)

The **marginal** probability mass function of X is given by

$$p_X(x) = P[X = x] = \sum_{y \in W_Y} p_{X,Y}(x,y).$$

Marginal distribution

Let X and Y be two discrete random variables, and $p_{X,Y}$ their joint probability mass function.

Definition (Marginal probability mass function)

The **marginal** probability mass function of X is given by

$$p_X(x) = P[X = x] = \sum_{y \in W_Y} p_{X,Y}(x, y).$$

Definition

X and Y are **independent** if $p_{X,Y}(x, y) = p_X(x)p_Y(y)$.

Conditional distribution

Let X and Y be two discrete random variables, and $p_{X,Y}$ their joint probability mass function.

Definition (Conditional probability mass function)

The **conditional probability mass function** of X given $Y = y$ is

$$p_{X|Y=y}(x) = \frac{p_{X,Y}(x,y)}{p_Y(y)}$$

Continuous multivariate distributions

Let $X \rightarrow \mathbb{R}$ and $Y \rightarrow \mathbb{R}$ be continuous random variables.

Definition

The **joint cumulative distribution function** of X and Y is given by

$$F_{X,Y}(x,y) := P[X \leq x, Y \leq y].$$

Their **joint probability density** is

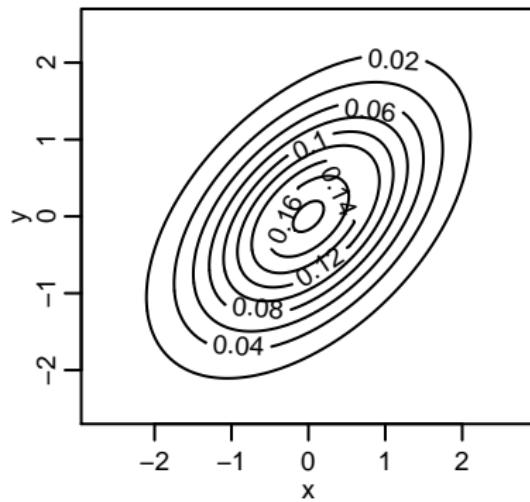
$$f_{X,Y}(x,y) := \frac{\partial}{\partial x} \frac{\partial}{\partial y} F_{X,Y}(x,y)$$

Usage: for real number $a < b$, $c < d$, we can calculate

$$P[a \leq X \leq b, c \leq Y \leq d] = \int_a^b \int_c^d f_{X,Y}(x,y) \, dy \, dx$$

Contour plots

Visualization of joint probability densities by contour plots:



Marginal density

Let X and Y be continuous random variables with joint density $p_{X,Y}$. In analogy to the discrete case, we define:

Definition (Marginal probability density)

The **marginal** probability density of X is given by

$$f_X(x) := \int_{-\infty}^{\infty} f_{X,Y}(x,y) dy .$$

Marginal density

Let X and Y be continuous random variables with joint density $p_{X,Y}$. In analogy to the discrete case, we define:

Definition (Marginal probability density)

The **marginal** probability density of X is given by

$$f_X(x) := \int_{-\infty}^{\infty} f_{X,Y}(x,y) dy .$$

Definition (Independence)

X and Y are independent if $f_{X,Y}(x,y) = f_X(x)f_Y(y)$.

Marginal and conditional density

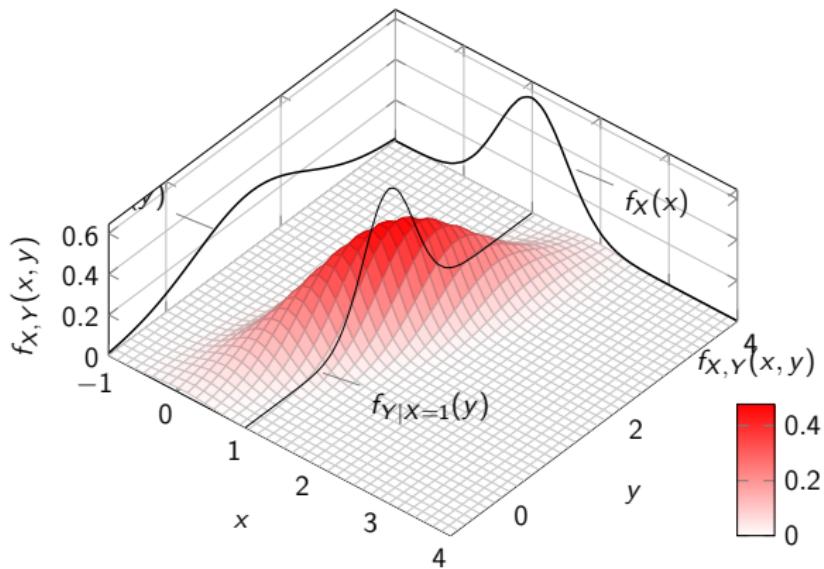
Let X and Y be continuous random variables with joint density $p_{X,Y}$.

Definition (Conditional probability density)

The **conditional probability density** of X given $Y = y$ is

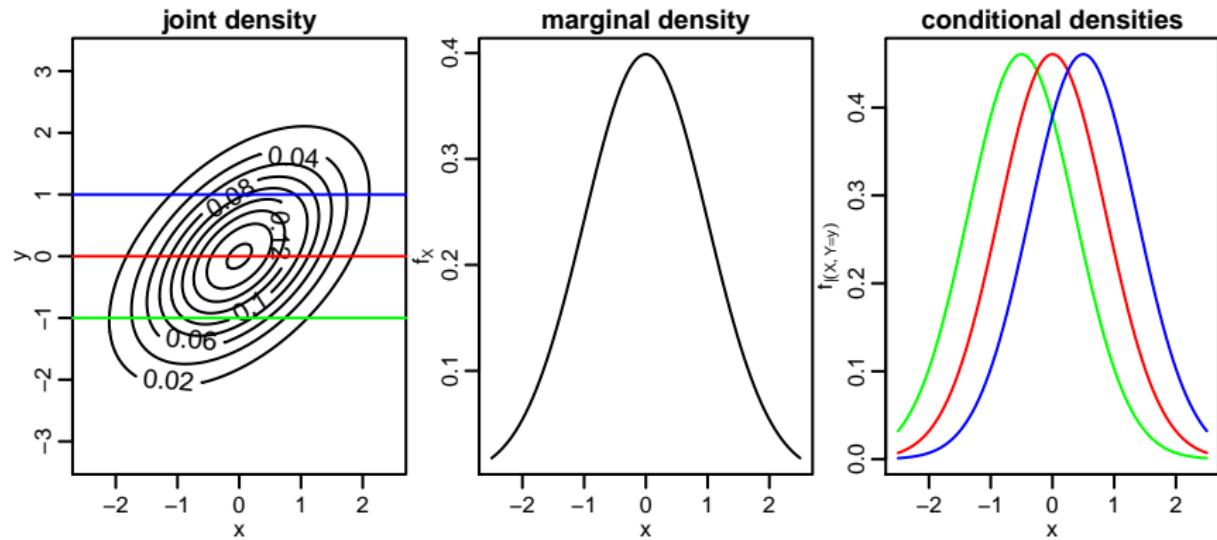
$$f_{X|Y=y}(x) = \frac{f_{X,Y}(x,y)}{f_Y(y)}$$

Marginal and conditional density: visualization



Source: <http://tex.stackexchange.com/questions/31708/draw-a-bivariate-normal-distribution-in-tikz>

Contour plots and conditional densities



Covariance and correlation

Let X and Y be two (discrete or continuous) random variables.

Definition

The **covariance** between X and Y is defined as

$$\text{Cov}(X, Y) := E \left[(X - E[X]) (Y - E[Y]) \right].$$

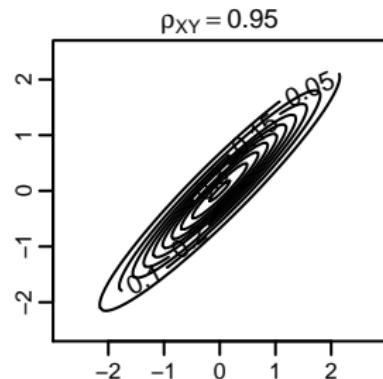
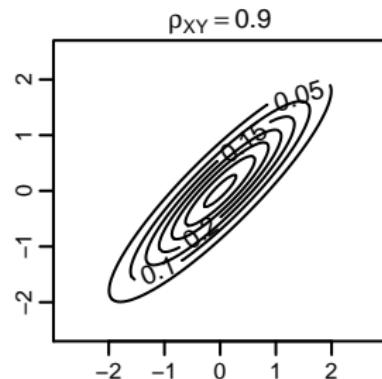
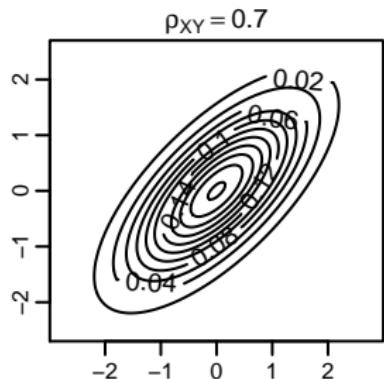
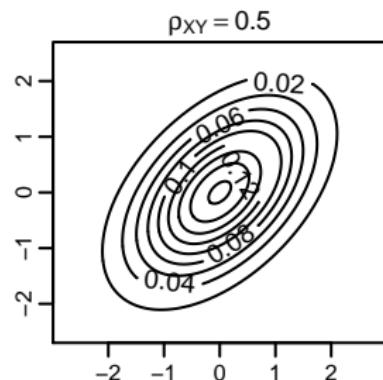
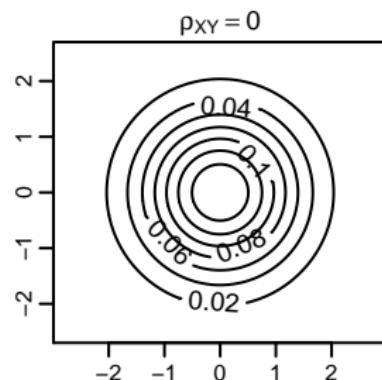
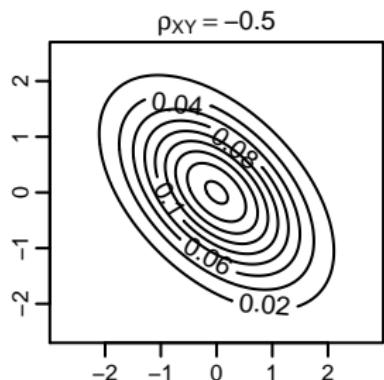
Their **correlation** is

$$\rho_{XY} := \frac{\text{Cov}(X, Y)}{\sqrt{\text{Var}(X) \text{Var}(Y)}}.$$

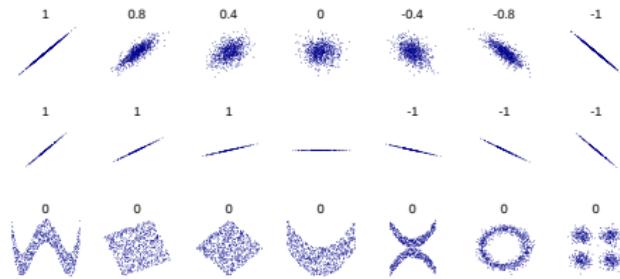
Interpretation:

- If X and Y are independent, $\text{Cov}(X, Y) = 0$ and $\rho_{XY} = 0$ (the other direction is *not* true!)
- $-1 \leq \rho_{XY} \leq 1$
- $\rho_{XY} = 1$ if $Y = a + bX$ for some $b > 0$
- $\rho_{XY} = -1$ if $Y = a + bX$ for some $b < 0$

Correlation: examples I



Correlation: examples II



Several data sets with the correlation coefficient of the underlying distribution. (Source: <http://en.wikipedia.org/wiki/Correlation>)

Calculation rules for $E[\cdot]$ and $\text{Var}(\cdot)$

Let X and Y be (discrete or continuous) random variables, and $a \in \mathbb{R}$ a number.

- $E[X + Y] = E[X] + E[Y]$
- If X and Y are independent, $E[XY] = E[X] \cdot E[Y]$
- $E[aX] = aE[X]$

Calculation rules for $E[\cdot]$ and $\text{Var}(\cdot)$

Let X and Y be (discrete or continuous) random variables, and $a \in \mathbb{R}$ a number.

- $E[X + Y] = E[X] + E[Y]$
- If X and Y are independent, $E[XY] = E[X] \cdot E[Y]$
- $E[aX] = aE[X]$
- $\text{Var}(X + Y) = \text{Var}(X) + \text{Var}(Y) + 2\text{Cov}(X, Y)$
- If X and Y are independent, $\text{Cov}(X, Y) = 0$
- $\text{Var}(aX) = a^2 \text{Var}(X)$

Part V

Descriptive Statistics

Learning objectives

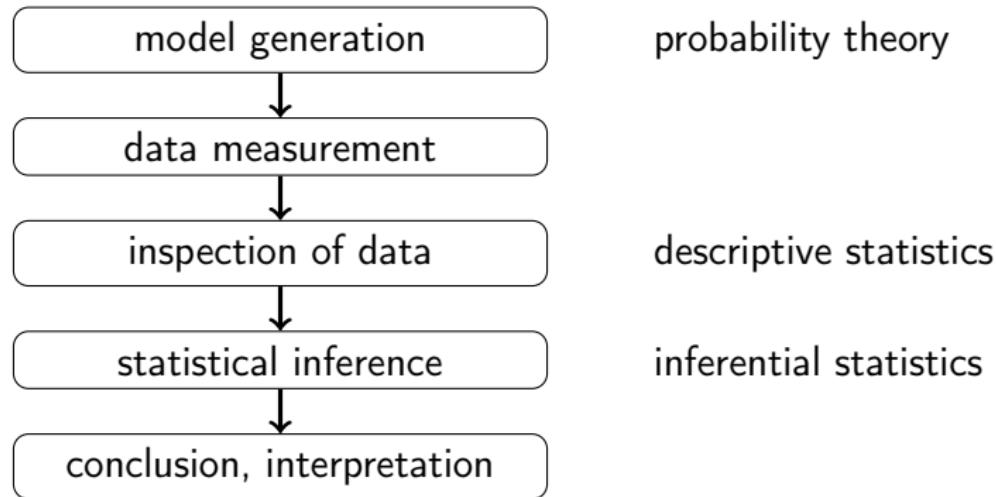
- Know descriptive statistics characterizing a sample: sample mean, sample variance, median, sample quantile, empirical correlation
- Explain the difference between quantities defined for probability distributions and quantities defined for samples
- Draw and read plots visualizing a sample: histogram, density curve, box plot, ECDF
- Know advantages and disadvantages of the plots mentioned before

Suggested literature

This lecture is partly based on the following source:

- Samuels et al. (2012), Chapters 2.3, 2.4

Pipeline of data analysis



Descriptive vs. inferential statistics

Descriptive statistics

- overview over data set
- visualize distribution of data
- find striking features
- describe distribution by few quantities

Descriptive vs. inferential statistics

Descriptive statistics

- overview over data set
- visualize distribution of data
- find striking features
- describe distribution by few quantities

Inferential statistics

- draw conclusions from data
- estimation
- hypothesis testing

From models to data

- Up to now: considered models described by probability theory
- Now: consider data generated by probabilistic model
- Usual assumption: n independent measurements from same probability distribution: X_1, X_2, \dots, X_n are independent copies of random variable X with expectation value μ and variance σ^2 .

From models to data

- Up to now: considered models described by probability theory
- Now: consider data generated by probabilistic model
- Usual assumption: n independent measurements from same probability distribution: X_1, X_2, \dots, X_n are independent copies of random variable X with expectation value μ and variance σ^2 .
- Formally:

Model: $X_1, X_2, \dots, X_n \stackrel{\text{i.i.d.}}{\sim} F_X(\cdot)$,

Sample: x_1, x_2, \dots, x_n

i.i.d.: **independent and identically distributed**

The i.i.d. assumption

- **Sample** x_1, x_2, \dots, x_n is assumed to consist of n **realizations** of the random variable X
- Model for repeated, independent measurements of “the same” quantity
- Examples: measurement from different cells/subjects that were not treated differently

Characteristic numbers for univariate data

- Mean
- Variance and standard deviation
- Median
- Quantiles

Descriptive statistics for univariate data

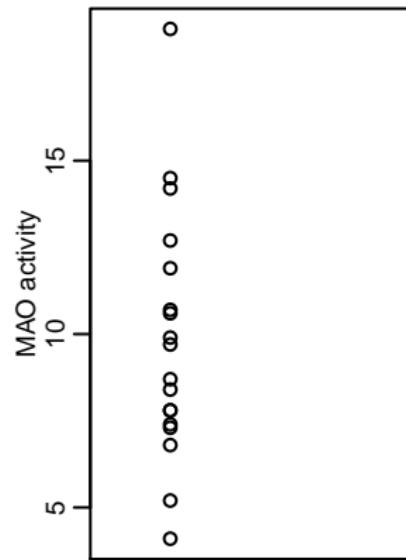
- Situation: multiple measurements of a single quantity
- Example data set: activity level of monoamine oxidase (MAO) in 18 patients with type I schizophrenia
- MAO: enzyme thought to play a role in regulation of behavior

(Source: Potkin et al. (1978))

Descriptive statistics for univariate data

- Situation: multiple measurements of a single quantity
- Example data set: activity level of monoamine oxidase (MAO) in 18 patients with type I schizophrenia
- MAO: enzyme thought to play a role in regulation of behavior

(Source: Potkin et al. (1978))



Sample mean

- **Sample mean:**

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

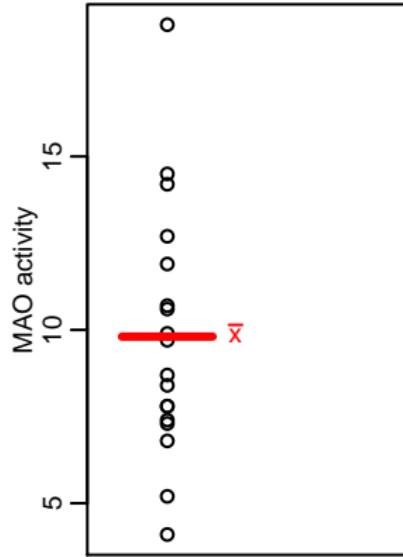
- R function: `mean`

Sample mean

- Sample mean:

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

- R function: `mean`



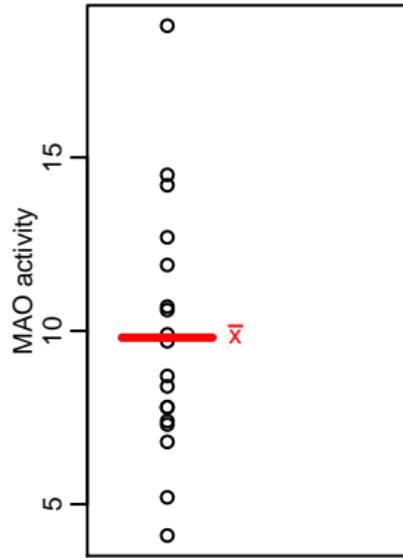
Sample mean

- Sample mean:

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

- R function: `mean`
- Sample mean is **consistent** estimator for $\mu = E[X]$:

$$\bar{X} = \frac{1}{n} \sum_{i=1}^n X_i \rightarrow \mu \text{ if } n \rightarrow \infty$$



Sample mean

- Sample mean:

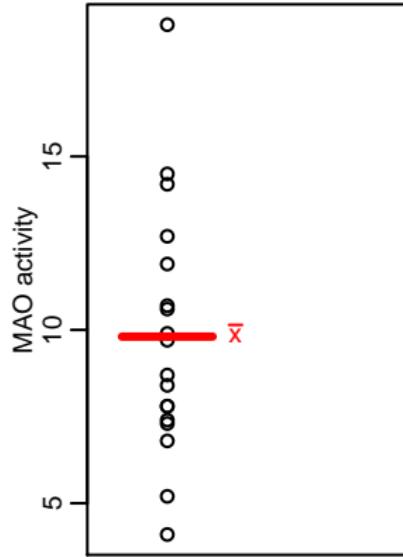
$$\bar{x} = \frac{x_1 + \dots + x_n}{n} = \frac{1}{n} \sum_{i=1}^n x_i$$

- R function: `mean`
- Sample mean is **consistent** estimator for $\mu = E[X]$:

$$\bar{X} = \frac{1}{n} \sum_{i=1}^n X_i \rightarrow \mu \text{ if } n \rightarrow \infty$$

- Sample mean is **unbiased** estimator for $\mu = E[X]$:

$$E[\bar{X}] = \mu$$



Sample variance

- Sample variance:

$$s_x^2 = \frac{1}{n-1} \sum_{i=1}^n (x_i - \bar{x})^2 \quad (s_x: \text{sample standard deviation})$$

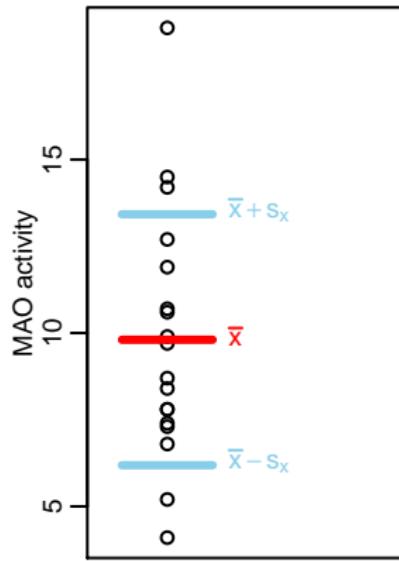
- R function: var

Sample variance

- Sample variance:

$$s_x^2 = \frac{1}{n-1} \sum_{i=1}^n (x_i - \bar{x})^2 \quad (s_x: \text{sample standard deviation})$$

- R function: var



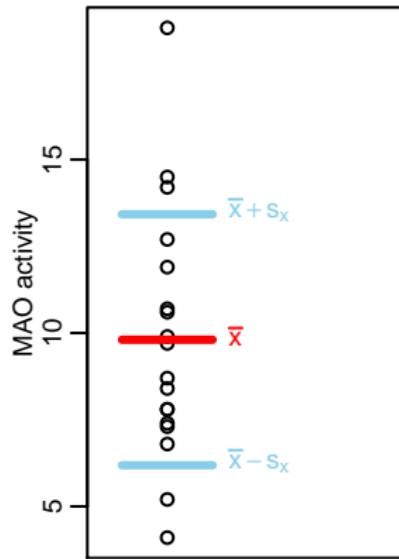
Sample variance

- Sample variance:

$$s_x^2 = \frac{1}{n-1} \sum_{i=1}^n (x_i - \bar{x})^2 \quad (s_x: \text{sample standard deviation})$$

- R function: var
- Sample variance is **consistent** estimator for $\sigma^2 = \text{Var}(X)$:

$$s_x^2 \rightarrow \sigma^2 \text{ if } n \rightarrow \infty$$



Sample variance

- Sample variance:

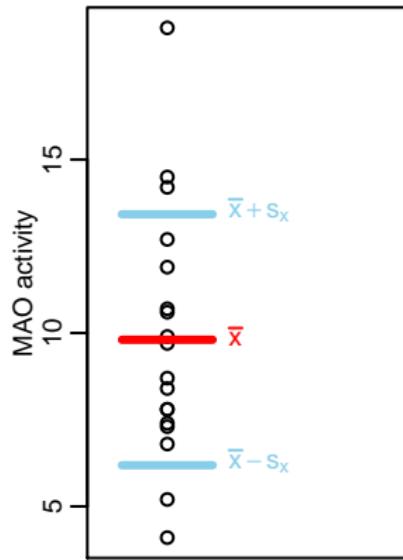
$$s_x^2 = \frac{1}{n-1} \sum_{i=1}^n (x_i - \bar{x})^2 \quad (s_x: \text{sample standard deviation})$$

- R function: var
- Sample variance is **consistent** estimator for $\sigma^2 = \text{Var}(X)$:

$$s_x^2 \rightarrow \sigma^2 \text{ if } n \rightarrow \infty$$

- Sample variance is **unbiased** estimator for $\sigma^2 = \text{Var}(X)$:

$$E[s_x^2] = \sigma^2$$



Median

- **Median:** value that separates the higher half from the lower half of the data
- Calculation: order data

$x_{(1)} \leq x_{(2)} \leq \dots \leq x_{(n)}$. Median:

$$m = \begin{cases} x_{((n+1)/2)}, & n \text{ is odd}, \\ \frac{1}{2}(x_{(n/2)} + x_{(n/2+1)}), & \text{otherwise} \end{cases}$$

- R function: `median`

Median

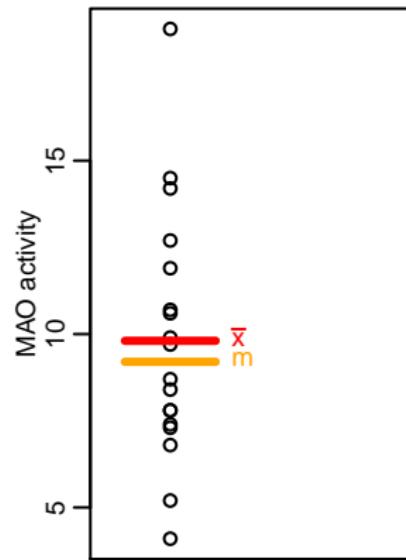
- **Median:** value that separates the higher half from the lower half of the data

- Calculation: order data

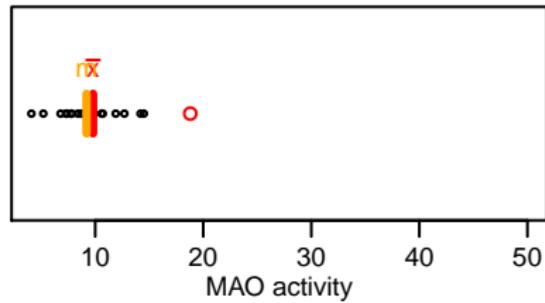
$x_{(1)} \leq x_{(2)} \leq \dots \leq x_{(n)}$. Median:

$$m = \begin{cases} x_{((n+1)/2)}, & n \text{ is odd}, \\ \frac{1}{2}(x_{(n/2)} + x_{(n/2+1)}), & \text{otherwise} \end{cases}$$

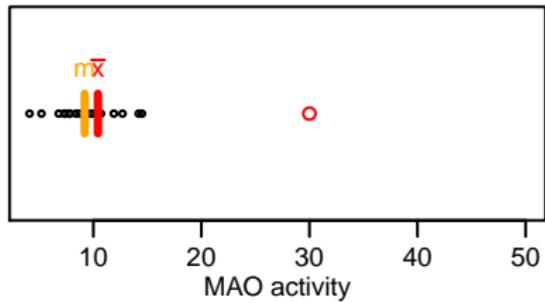
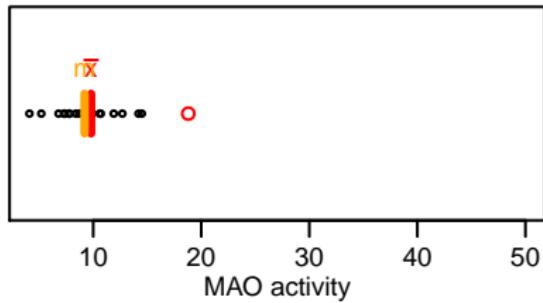
- R function: `median`



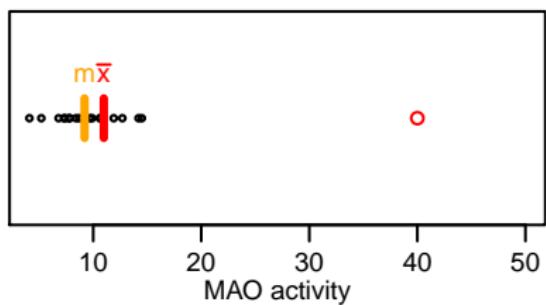
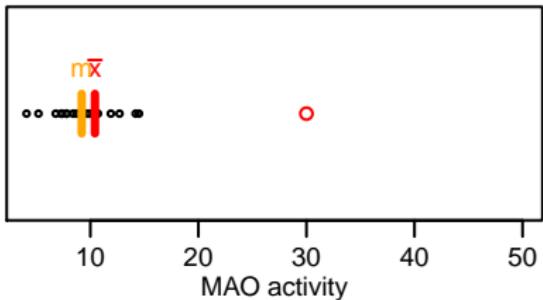
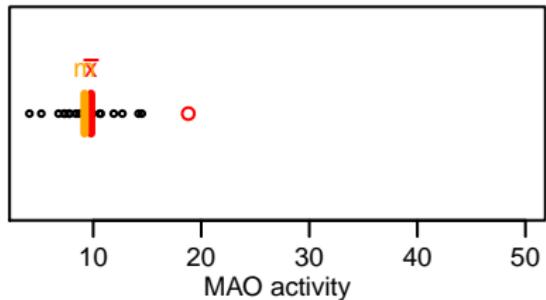
Comparison of mean and median



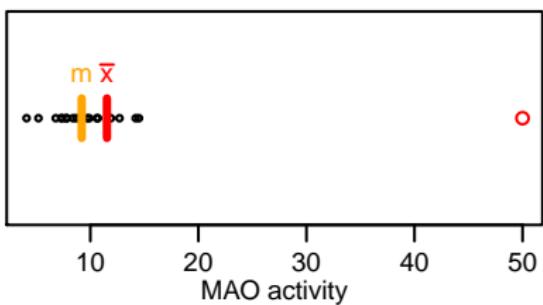
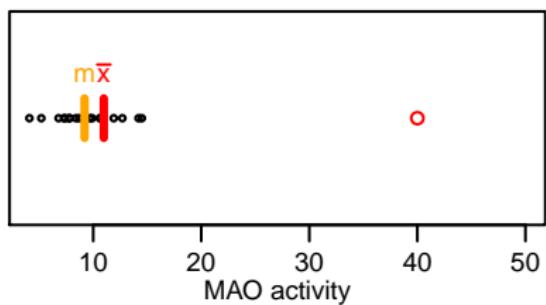
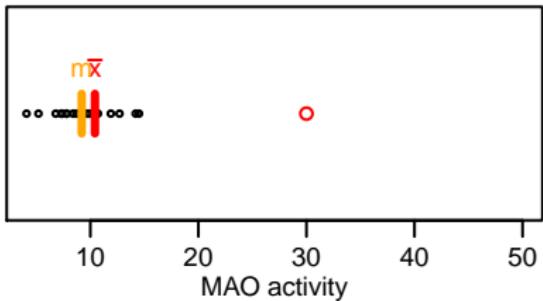
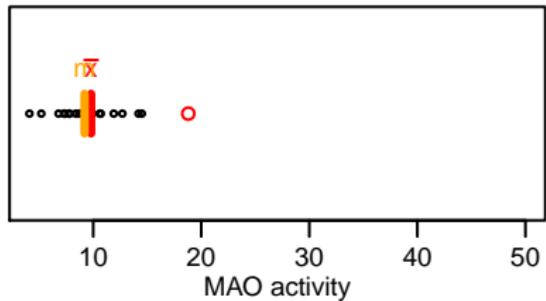
Comparison of mean and median



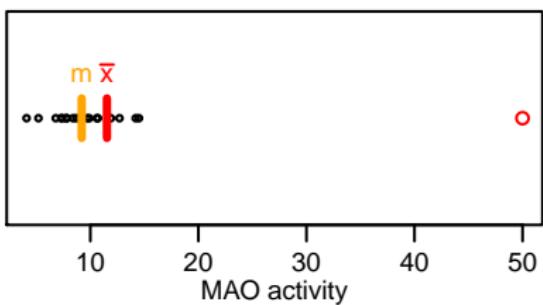
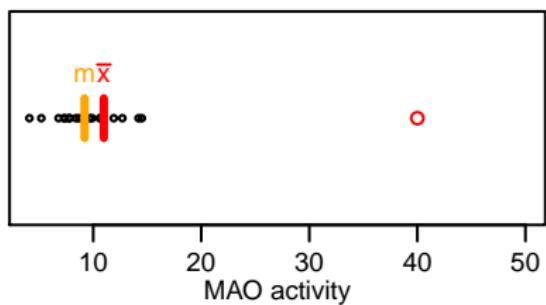
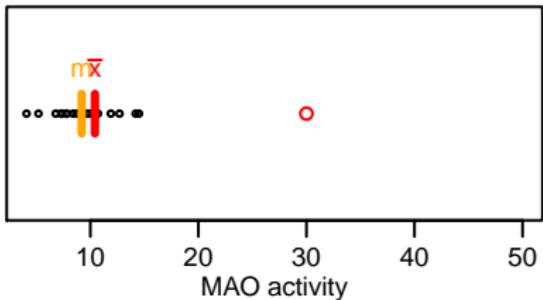
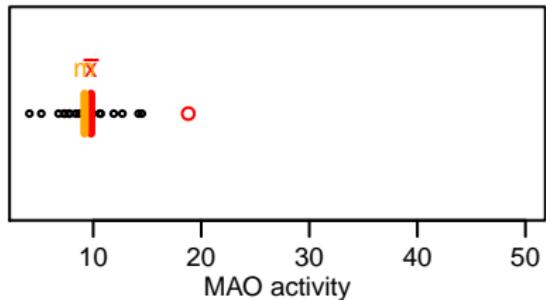
Comparison of mean and median



Comparison of mean and median



Comparison of mean and median



Median is robust, mean is not!

Quantiles

- Generalization of the concept of the median
- **Empirical α quantile:** value that is larger than a fraction of α of the data and smaller than a fraction of $1 - \alpha$ of the data
- Calculation:
 - ▶ Order data $x_{(1)} \leq x_{(2)} \leq \dots \leq x_{(n)}$
 - ▶ If $\alpha \cdot (n - 1)$ is an integer, take $q_\alpha = x_{(\alpha(n-1)+1)}$; otherwise, interpolate between $x_{(\lfloor \alpha(n-1) \rfloor + 1)}$ and $x_{(\lceil \alpha(n-1) \rceil + 1)}$
- R function: `quantile`

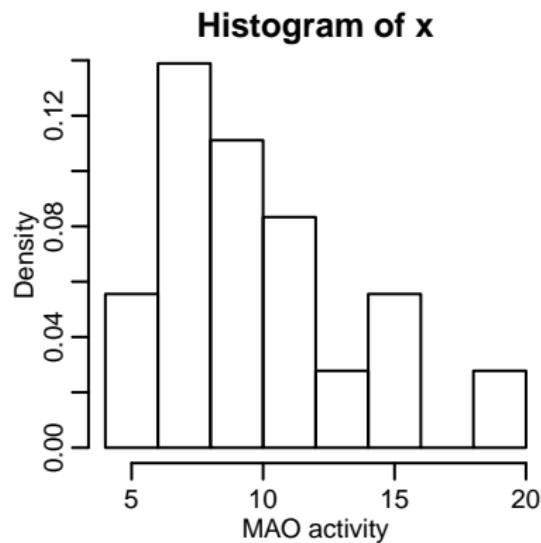
Quantiles

- Generalization of the concept of the median
- **Empirical α quantile:** value that is larger than a fraction of α of the data and smaller than a fraction of $1 - \alpha$ of the data
- Calculation:
 - ▶ Order data $x_{(1)} \leq x_{(2)} \leq \dots \leq x_{(n)}$
 - ▶ If $\alpha \cdot (n - 1)$ is an integer, take $q_\alpha = x_{(\alpha(n-1)+1)}$; otherwise, interpolate between $x_{(\lfloor \alpha(n-1) \rfloor + 1)}$ and $x_{(\lceil \alpha(n-1) \rceil + 1)}$
- R function: `quantile`
- α quantile of random variable X : value m such that $P[X \leq m] \geq \alpha$ and $P[X \geq m] \geq 1 - \alpha$

Graphical representations for univariate data

- Histogram
- Boxplot
- Empirical cumulative distribution function
- later: Q-Q (quantile-quantile) plot

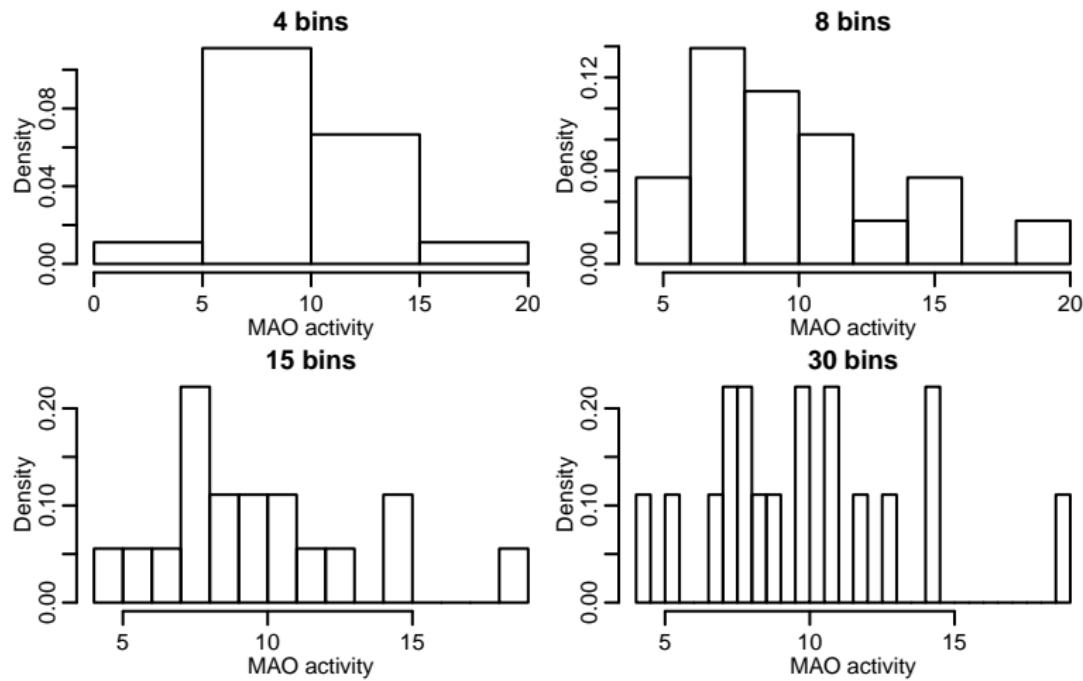
Histogram



- Divide range of measured values into bins $(c_{k-1}, c_k]$
E.g.
 $c_k = 4, 6, 8, 10, 12, 14, 16, 18, 20$
- Calculate number of data points in bins:
 $h_k := \#\{i | x_i \in (c_{k-1}, c_k]\}$
E.g. $h_k = 2, 5, 4, 3, 1, 2, 0, 1$
- Draw density $\frac{h_k}{n(c_k - c_{k-1})}$ over bin $(c_{k-1}, c_k]$
(Or: draw counts h_k over bin $(c_{k-1}, c_k]$)

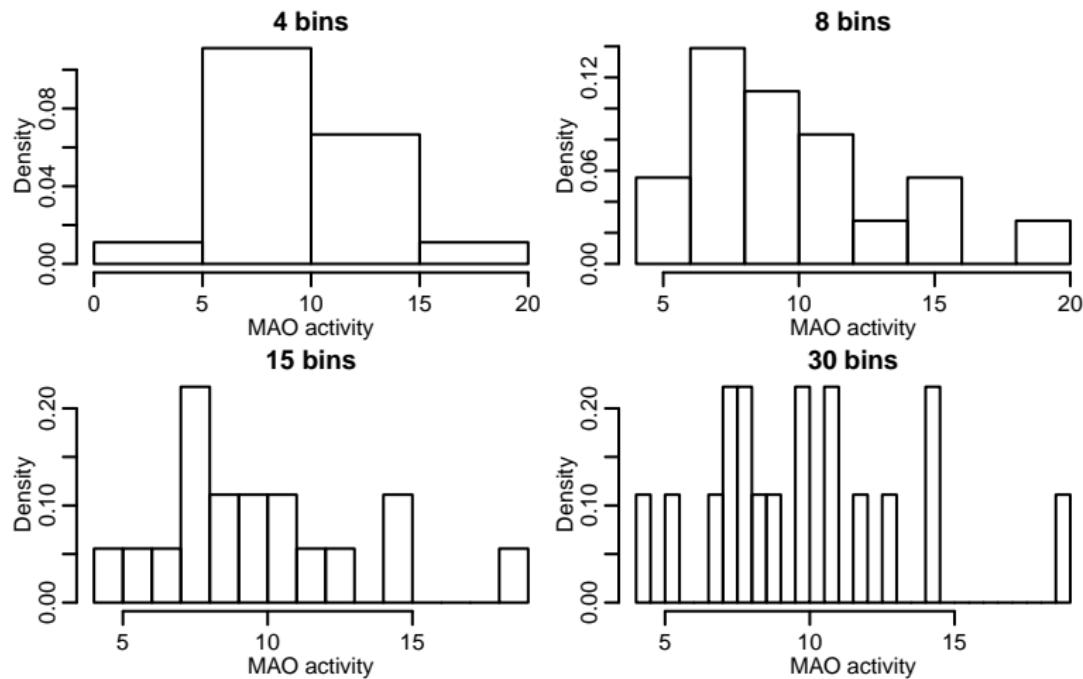
Histogram: choice of a bin width

How should we choose the width of the bins of a histogram?



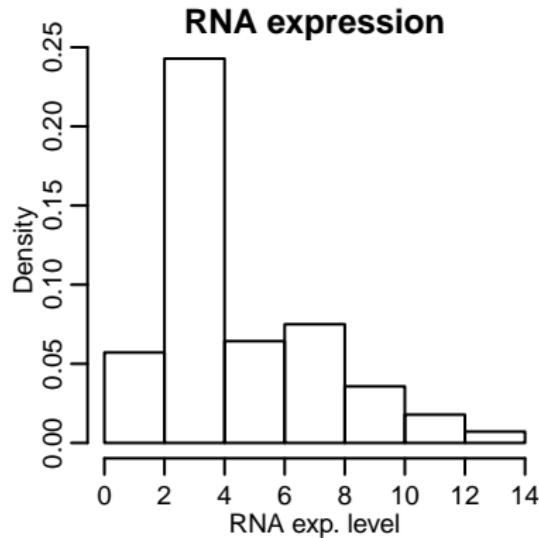
Histogram: choice of a bin width

How should we choose the width of the bins of a histogram?

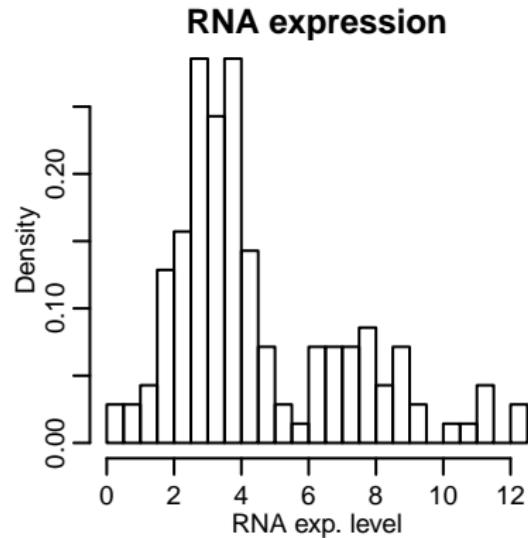
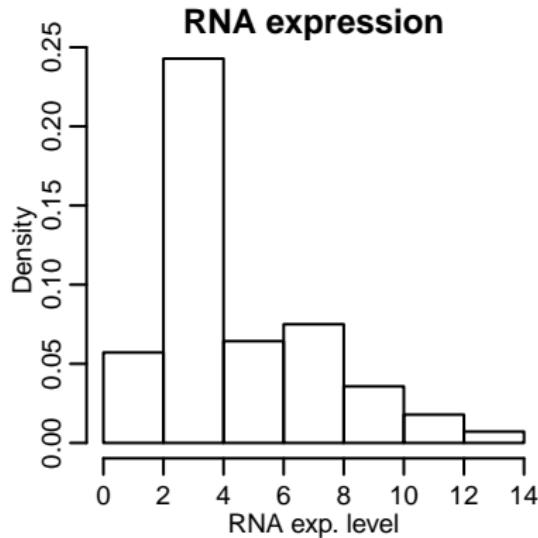


Play around, or let do R an automatic choice

Histograms for bimodal distributions



Histograms for bimodal distributions



Caveat: wrongly chosen bin size can hide features (here: bimodality)

Nonparametric density estimation

- Drawbacks of histogram: depends on arbitrary choice of bin borders; density jump at bin borders is not realistic.
- Alternative: non-parametric density estimation.

Nonparametric density estimation

- Drawbacks of histogram: depends on arbitrary choice of bin borders; density jump at bin borders is not realistic.
- Alternative: non-parametric density estimation.
- First improvement over histogram: do not count data points in predefined bins, but in a **sliding window** of fix width centered around *any* point of the x-line.
- Second improvement: give points near the center of the “sliding window” more weight than points far apart from the center: idea of **kernel density estimation**

Kernel density estimation

Given a set of points x_1, x_2, \dots, x_n , the **kernel density estimator** for the generating distribution is

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

K is called **kernel function** and stands for an arbitrary positive symmetric function that integrates to 1; h is called **bandwidth**. Typical kernel functions are:

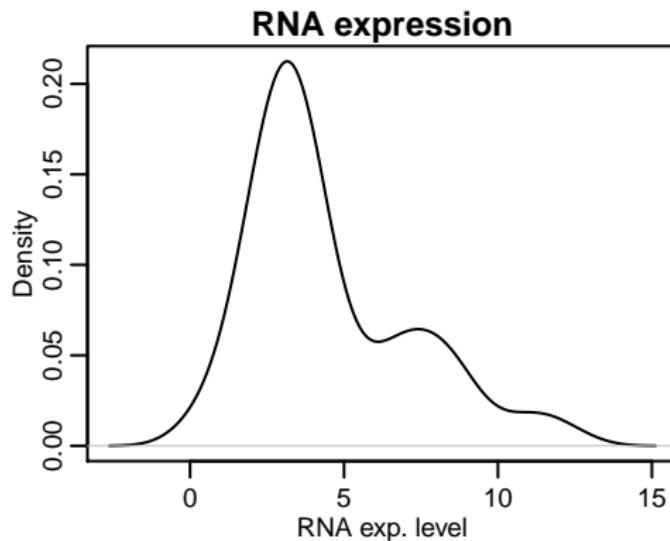
- uniform or rectangular kernel: K is the density of $\mathcal{U}([- \frac{1}{2}, \frac{1}{2}])$; gives the same weight to all data points in $[x - h, x + h]$.
- Gaussian kernel: K is the density of $\mathcal{N}(0, 1)$; gives less weight to points that are far apart from x .

Choosing a bandwidth

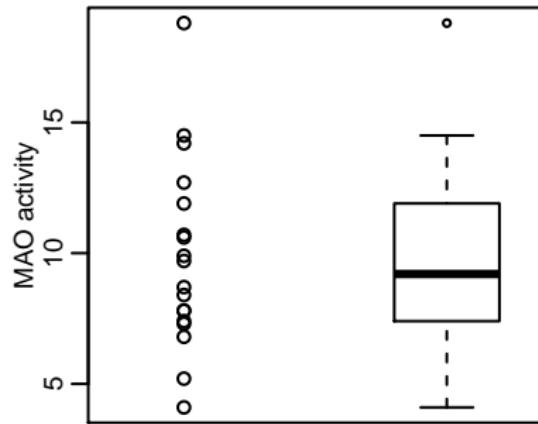
- The bandwidth h has a large influence on the estimator.
- A small bandwidth makes the estimator wiggly.
- A large bandwidth makes the estimator flat.
- (Very rough) rule of thumb for manual choice of bandwidth: choose bandwidth such that it typically covers at least 5 to 10 data points, and not larger than 20% of the whole range of the data set.
- Better alternative: let R choose the bandwidth automatically.

Example: density estimation for RNA data set

The kernel estimator is implemented in the R function `density`. The bandwidth can be specified via argument `bw`; if left out, the bandwidth is chosen automatically.

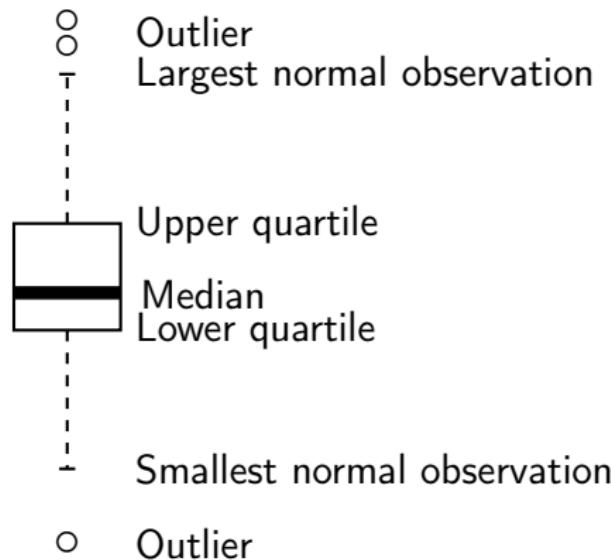


Box plot



R function: `boxplot`

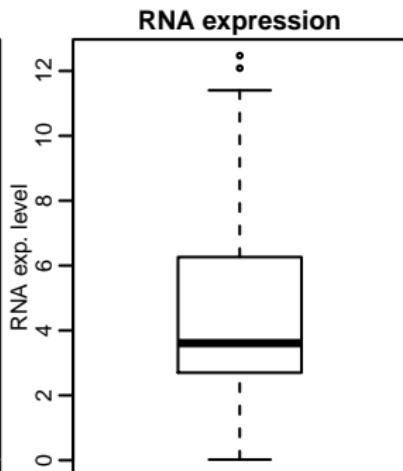
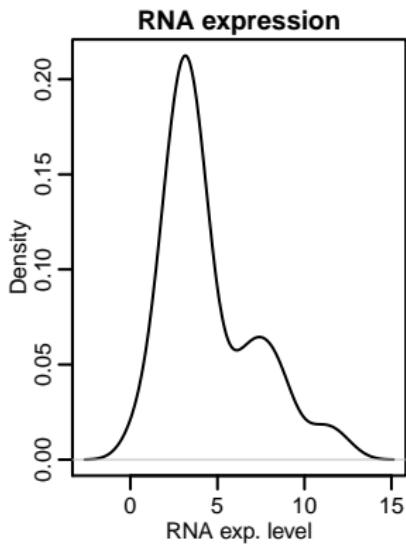
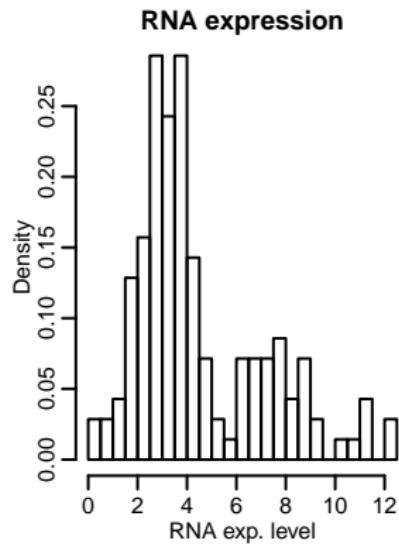
Box plot



“Normal values”: values separated from the quartiles by no more than 1.5 IQR
IQR (“interquartile range”):
 $q_{0.75} - q_{0.25}$

Box plot for bimodal data

Comparison: histogram, non-parametric density estimation, and box plot for bimodal data set:



Box plot completely hides bimodality!

Empirical cumulative distribution function

- Remember: random variable X ,
cumulative distribution function

$$F_X = P[X \leq x]$$

Empirical cumulative distribution function

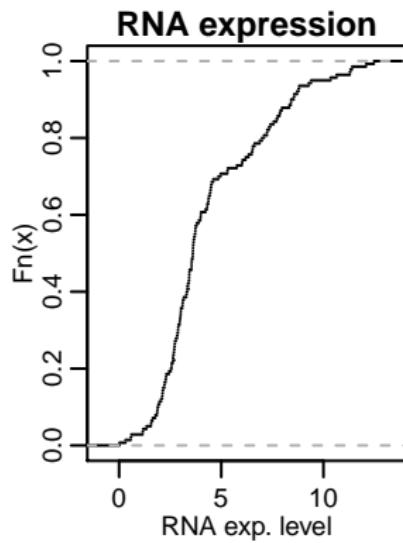
- Remember: random variable X ,
cumulative distribution function
 $F_X = P[X \leq x]$
- Now: sample x_1, x_2, \dots, x_n . **Empirical cumulative distribution function** (ECDF):

$$\hat{F}(x) = \frac{\#\{k | x_k \leq x\}}{n}$$

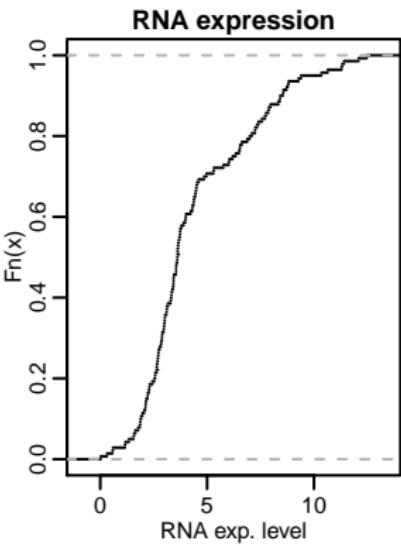
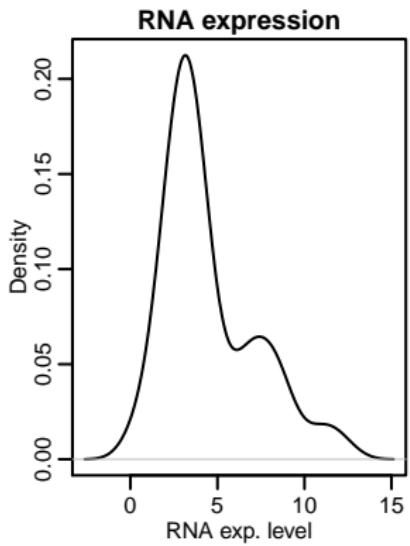
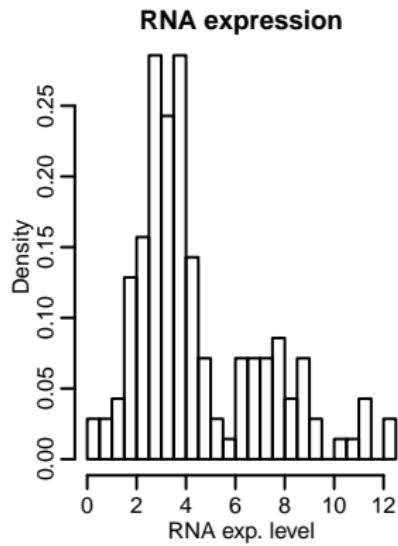
Empirical cumulative distribution function

- Remember: random variable X , cumulative distribution function $F_X = P[X \leq x]$
- Now: sample x_1, x_2, \dots, x_n . **Empirical cumulative distribution function** (ECDF):

$$\hat{F}(x) = \frac{\#\{k | x_k \leq x\}}{n}$$

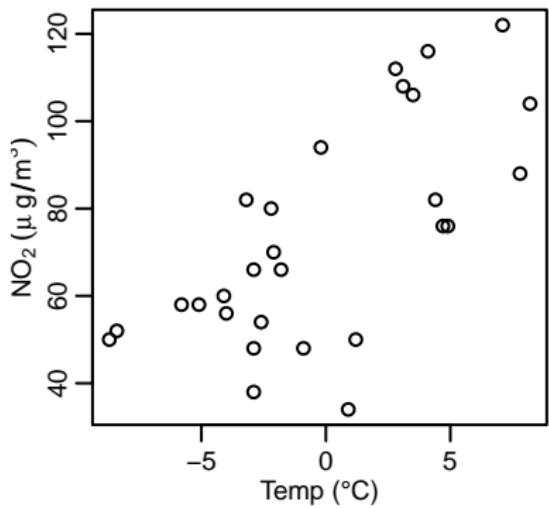


Different ways of visualizing bimodal data set



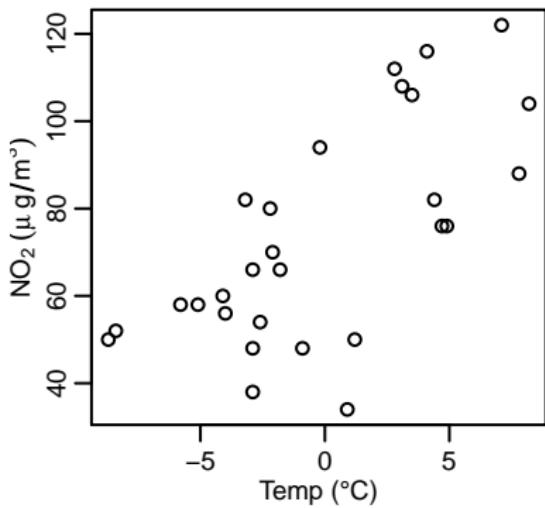
Descriptive statistics for multivariate data

Scatter plot:



Descriptive statistics for multivariate data

Scatter plot:



Empirical correlation:

$$r = \frac{s_{xy}}{s_x s_y} \in [-1, 1],$$

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

In R: > cor(no2NO2, no2Temp) [1]
0.6799612

Empirical correlation

Empirical correlation r indicates strength of linear dependence between 2 samples $\{x_i\}$ and $\{y_i\}$:

- $r = +1$ if $y_i = a + bx_i$ for some $b > 0$
- $r = -1$ if $y_i = a + bx_i$ for some $b < 0$

Empirical correlation

Empirical correlation r indicates strength of linear dependence between 2 samples $\{x_i\}$ and $\{y_i\}$:

- $r = +1$ if $y_i = a + bx_i$ for some $b > 0$
- $r = -1$ if $y_i = a + bx_i$ for some $b < 0$

Caveat: many (non-linear) structures can lead to the same value of r !

Summary: descriptive statistics for univariate data

Random variable X	Sample x_1, x_2, \dots, x_n
Expectation value $E[X] = \sum_{k \geq 1} x_k p(x_k)$, or $E[X] = \int_{-\infty}^{\infty} xf(x) dx$	Sample mean $\bar{x} = \frac{1}{n} \sum_{i=1}^n x_i$
Variance $\text{Var}(X) = \sum_{k \geq 1} (x_k - E[X])^2 p(x_k)$, or $\text{Var}(X) = \int_{-\infty}^{\infty} (x - E[X])^2 f(x) dx$	Sample variance $s_x^2 = \frac{1}{n-1} \sum_{i=1}^n (x_i - \bar{x})^2$

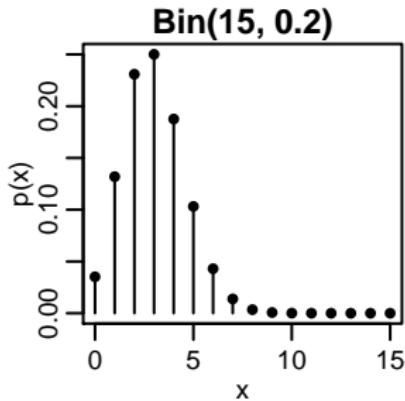
Summary: descriptive statistics for univariate data

Random variable X	Sample x_1, x_2, \dots, x_n
α quantile m such that $P[X \leq m] \geq \alpha$ and $P[X \geq m] \geq 1 - \alpha$	Empirical α quantile m such that $\hat{F}(m) \geq \alpha$ and $1 - \hat{F}(m) \geq 1 - \alpha$

Summary: descriptive statistics for univariate data

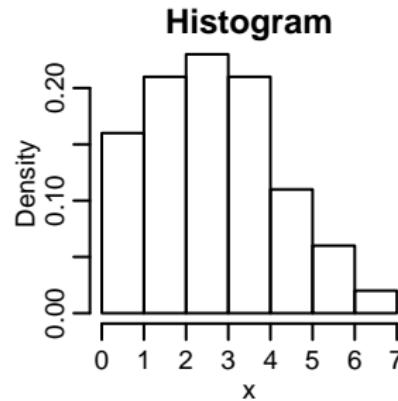
Discrete random variable X

Probability mass function



Sample x_1, x_2, \dots, x_n

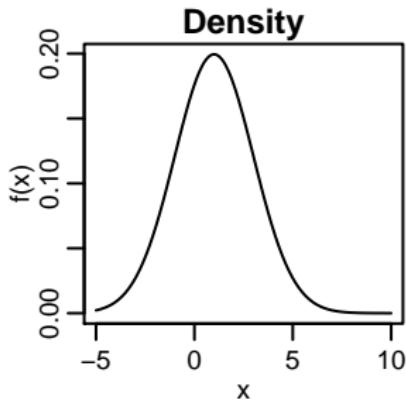
Histogram



Summary: descriptive statistics for univariate data

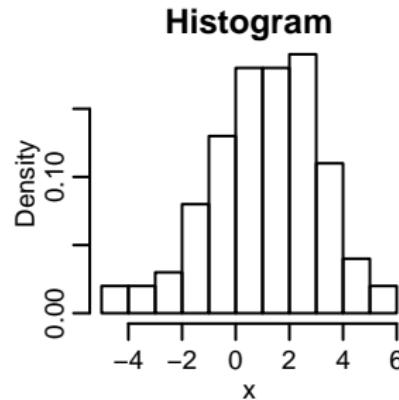
Continuous random variable X

Probability density



Sample x_1, x_2, \dots, x_n

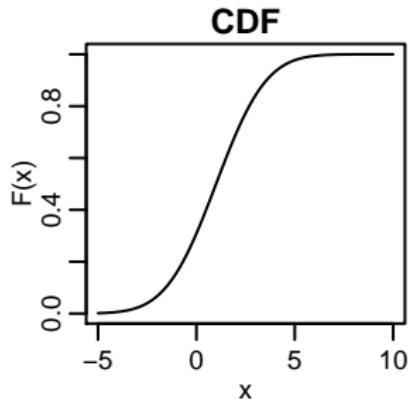
Histogram



Summary: descriptive statistics for univariate data

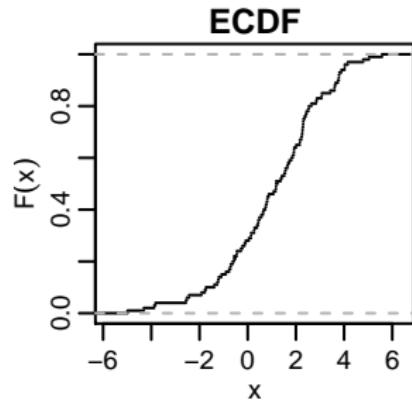
Random variable X

Cumulative distribution function



Sample x_1, x_2, \dots, x_n

ECDF



Part VI

Law of Large Numbers and Central Limit Theorem

Learning objectives

- Know the law of large numbers mathematically and in words
- Know the central limit theorem
- Explain the connection between the central limit theorem and the law of large numbers
- Calculate the standard error of the mean
- Explain differences between the standard error of the mean and the standard deviation
- Calculate an approximate confidence interval for a mean
- Approximate a binomial distribution by a normal one

Suggested literature

This lecture is partly based on the following source:

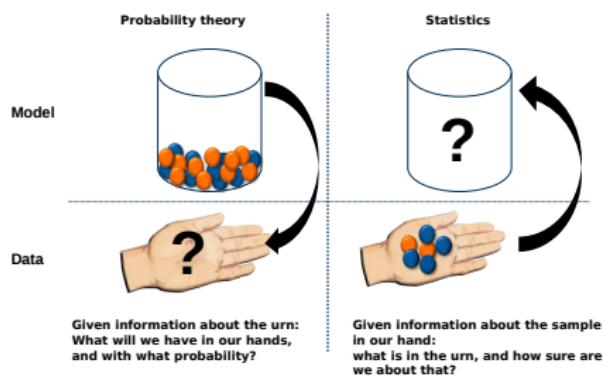
- Samuels et al. (2012), Chapters 5.1, 5.2, 5.3, 5.4; 6.1, 6.2, 6.3

Overview

- I.i.d. assumption throughout chapter
- Rough idea: as $n \rightarrow \infty$, characteristics of i.i.d. sample x_1, \dots, x_n resemble more and more those of the underlying probability distribution
- Basis for many statistical tests and estimators

Overview

- I.i.d. assumption throughout chapter
- Rough idea: as $n \rightarrow \infty$, characteristics of i.i.d. sample x_1, \dots, x_n resemble more and more those of the underlying probability distribution
- Basis for many statistical tests and estimators



(Source: Maier and Weiss (2013))

Properties of sample mean

- Let X be random variable with expectation value μ and standard deviation σ
- Consider n i.i.d. copies X_1, X_2, \dots, X_n of X
- Define sum $S_n = \sum_{i=1}^n X_i$ and mean $\bar{X}_n = S_n/n$
- Properties: $E[\bar{X}_n] = \mu$, $\sigma(\bar{X}_n) = \frac{\sigma}{\sqrt{n}} \rightarrow 0$ as $n \rightarrow \infty$

Properties of sample mean

- Let X be random variable with expectation value μ and standard deviation σ
- Consider n i.i.d. copies X_1, X_2, \dots, X_n of X
- Define sum $S_n = \sum_{i=1}^n X_i$ and mean $\bar{X}_n = S_n/n$
- Properties: $E[\bar{X}_n] = \mu$, $\sigma(\bar{X}_n) = \frac{\sigma}{\sqrt{n}} \rightarrow 0$ as $n \rightarrow \infty$
- In words: the sample mean is centered around the expectation value of X , and its distribution becomes more and more narrow the larger the sample size is.

Properties of sample mean

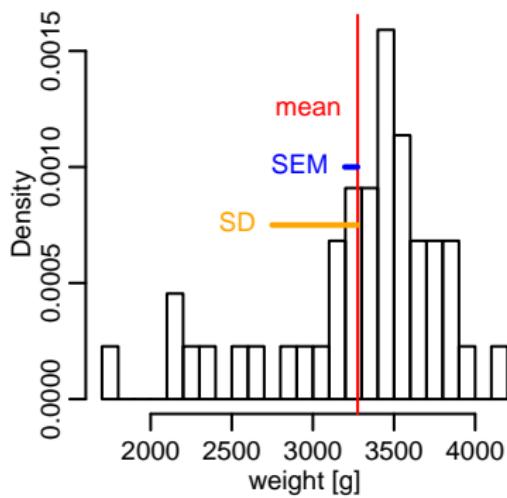
- Let X be random variable with expectation value μ and standard deviation σ
- Consider n i.i.d. copies X_1, X_2, \dots, X_n of X
- Define sum $S_n = \sum_{i=1}^n X_i$ and mean $\bar{X}_n = S_n/n$
- Properties: $E[\bar{X}_n] = \mu$, $\sigma(\bar{X}_n) = \frac{\sigma}{\sqrt{n}} \rightarrow 0$ as $n \rightarrow \infty$
- In words: the sample mean is centered around the expectation value of X , and its distribution becomes more and more narrow the larger the sample size is.

Definition (Standard error of the mean)

A natural estimator for $\sigma(\bar{X}_n)$ is given by the **standard error of the mean (SEM)**: $s_{\bar{x}} = \frac{s_x}{\sqrt{n}}$ (s_x is the *empirical standard deviation*)

Example: SEM of birth weights

Data set with birth weights of 44 babies (in g):



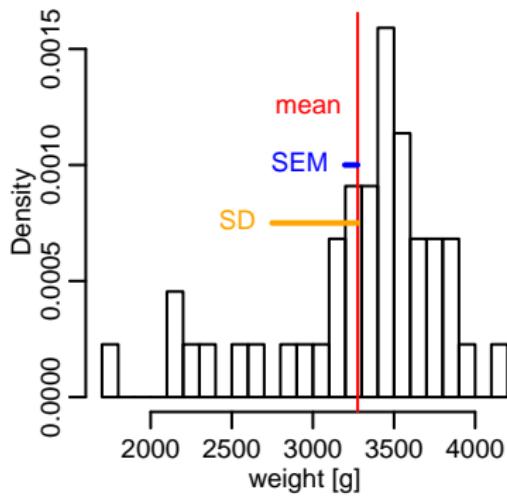
Mean = 3276 g

emp. SD = 528 g

SEM = 80 g

Example: SEM of birth weights

Data set with birth weights of 44 babies (in g):



Mean = 3276 g

emp. SD = 528 g

SEM = 80 g

When the sample size n grows, ...

- ... the mean converges to the expectation value,
- ... the empirical SD converges to the true standard deviation,
- ... the SEM converges to 0

Law of large numbers

The finding $E[\bar{X}_n] = \mu$, $\sigma(\bar{X}_n) = \frac{\sigma}{\sqrt{n}}$ leads to the following theorem:

Theorem (Law of large numbers)

Let X be random variable with expectation value μ , and X_1, X_2, \dots, X_n i.i.d. copies of X . Then,

$$\bar{X}_n \rightarrow \mu \text{ as } n \rightarrow \infty$$

(Mathematically not very precise: should first define convergence for random variables...)

Law of large numbers

The finding $E[\bar{X}_n] = \mu$, $\sigma(\bar{X}_n) = \frac{\sigma}{\sqrt{n}}$ leads to the following theorem:

Theorem (Law of large numbers)

Let X be random variable with expectation value μ , and X_1, X_2, \dots, X_n i.i.d. copies of X . Then,

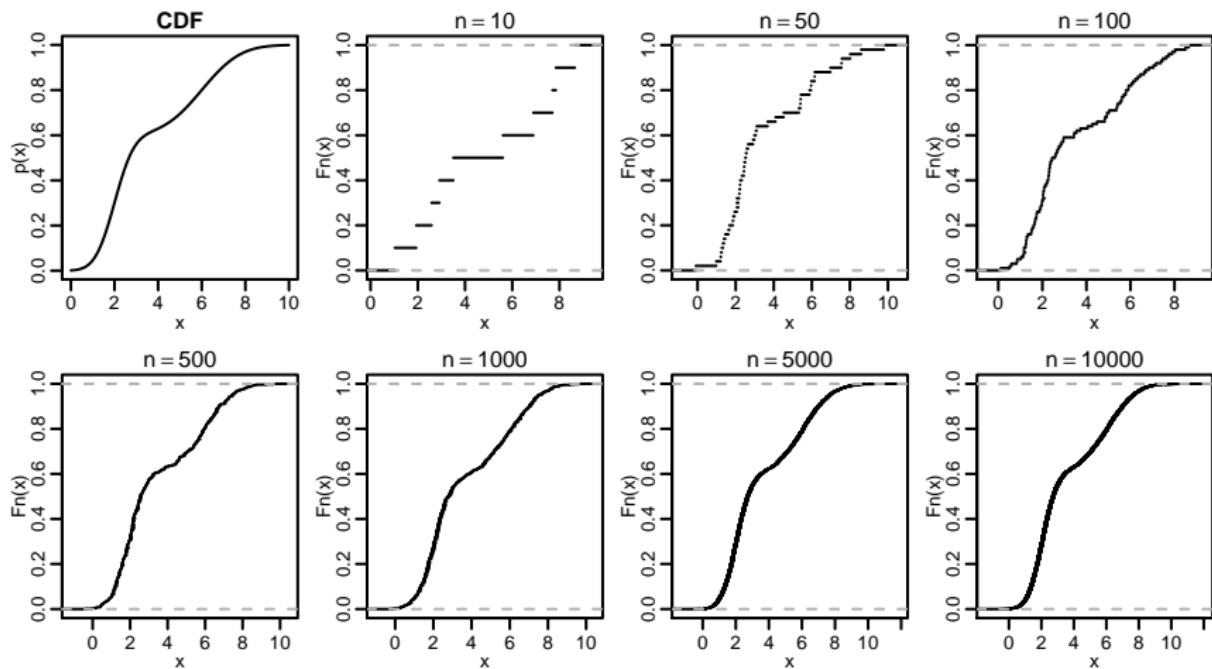
$$\bar{X}_n \rightarrow \mu \text{ as } n \rightarrow \infty$$

(Mathematically not very precise: should first define convergence for random variables...)

Consequence of the law of large numbers: empirical cumulative distribution function converges to the true CDF

ECDF

Consequence of the law of large numbers: empirical cumulative distribution function converges to the true CDF



Central limit theorem

The **central limit theorem** (CLT) strengthens the result of the law of large numbers:

Theorem (Central limit theorem)

Let X be random variable with expectation value μ and variance σ^2 , and X_1, X_2, \dots, X_n i.i.d. copies of X . Then,

$$\bar{X}_n \approx \mathcal{N} \left(\mu, \frac{\sigma^2}{n} \right) \text{ for large } n .$$

Goodness of approximation depends on n and the distribution of X .

Recall: transformations of the normal distribution

- Let $X \sim \mathcal{N}(\mu, \sigma^2)$ be a normally distributed random variable
- Define $Y = a + bX$ for real numbers a and b
- Y is also normally distributed: $Y \sim \mathcal{N}(a + b\mu, b^2\sigma^2)$

Recall: transformations of the normal distribution

- Let $X \sim \mathcal{N}(\mu, \sigma^2)$ be a normally distributed random variable
- Define $Y = a + bX$ for real numbers a and b
- Y is also normally distributed: $Y \sim \mathcal{N}(a + b\mu, b^2\sigma^2)$
- Application: **normalization** of X :

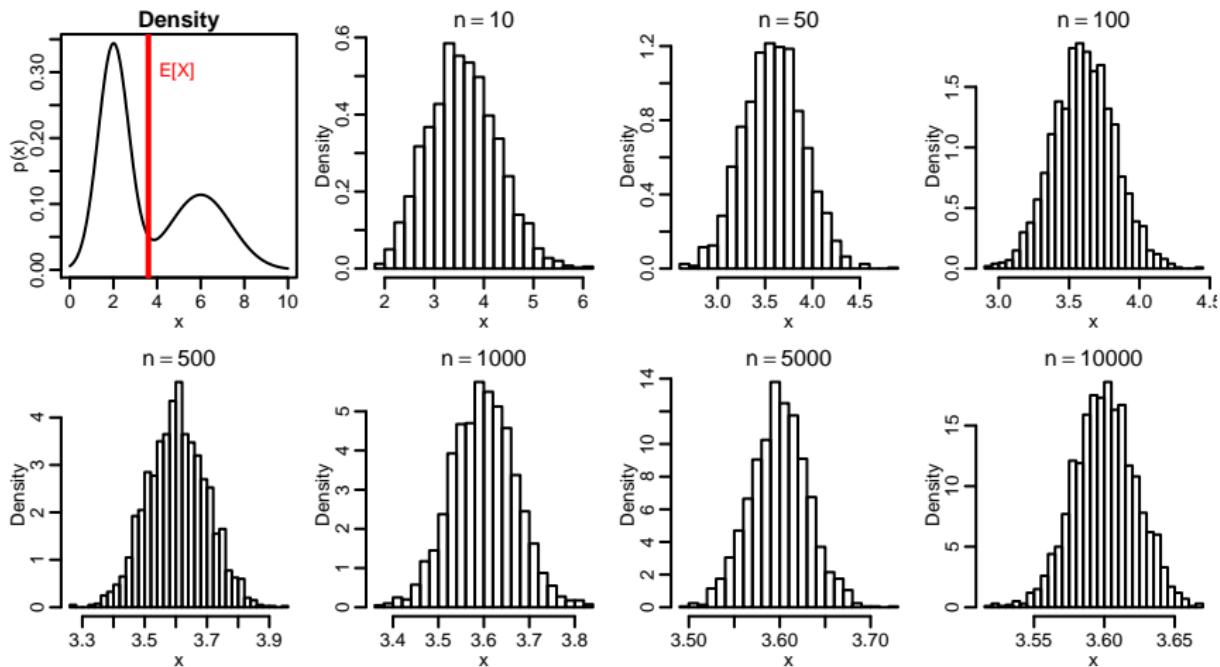
$$Z := \frac{X - \mu}{\sigma} \sim \mathcal{N}(0, 1)$$

- The central limit theorem can hence also be stated as

$$\frac{\sqrt{n}(\bar{X}_n - \mu)}{\sigma} \approx \mathcal{N}(0, 1)$$

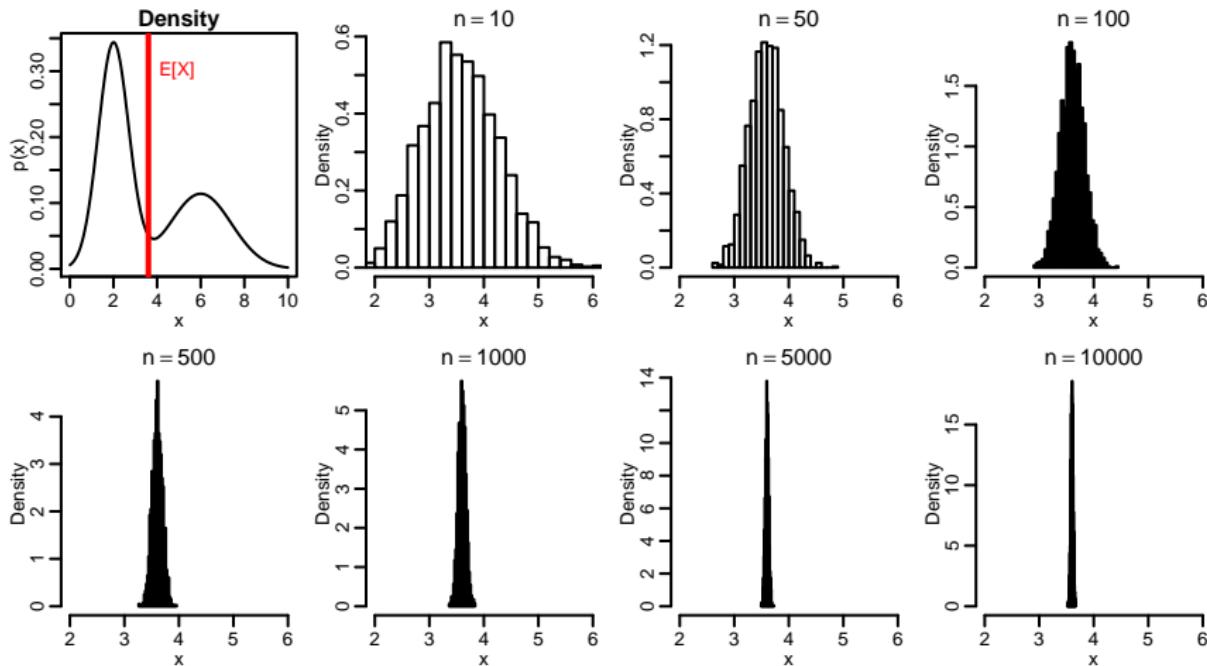
Central limit theorem: illustration

The distribution of the sample mean (!) resembles a normal distribution more and more:



Central limit theorem: illustration

The distribution of the sample mean (!) resembles a normal distribution more and more:



Applications of the central limit theorem

- (Approximate) confidence intervals for the expectation value
- Normal approximation of a binomial distribution

Approximate confidence interval for the expectation value I

- Consider an i.i.d. sample x_1, x_2, \dots, x_n from a distribution with expectation value μ and standard deviation σ .
- Central limit theorem: $\frac{\sqrt{n}(\bar{X}_n - \mu)}{\sigma} \approx \mathcal{N}(0, 1)$ for large n

Approximate confidence interval for the expectation value I

- Consider an i.i.d. sample x_1, x_2, \dots, x_n from a distribution with expectation value μ and standard deviation σ .
- Central limit theorem: $\frac{\sqrt{n}(\bar{X}_n - \mu)}{\sigma} \approx \mathcal{N}(0, 1)$ for large n
- Consequence: $\frac{\sqrt{n}(\bar{X}_n - \mu)}{\sigma}$ lies between -1.96 and 1.96 (2.5% and 97.5% -quantile of $\mathcal{N}(0, 1)$) with probability 95%

Approximate confidence interval for the expectation value I

- Consider an i.i.d. sample x_1, x_2, \dots, x_n from a distribution with expectation value μ and standard deviation σ .
- Central limit theorem: $\frac{\sqrt{n}(\bar{X}_n - \mu)}{\sigma} \approx \mathcal{N}(0, 1)$ for large n
- Consequence: $\frac{\sqrt{n}(\bar{X}_n - \mu)}{\sigma}$ lies between -1.96 and 1.96 (2.5% and 97.5% -quantile of $\mathcal{N}(0, 1)$) with probability 95%
- Consequence 2: $\mu \in \left[\bar{X}_n - 1.96 \cdot \frac{\sigma}{\sqrt{n}}, \bar{X}_n + 1.96 \cdot \frac{\sigma}{\sqrt{n}} \right]$ with probability 95% .
- Problem: σ is unknown; we can estimate it with s_x .

Approximate confidence interval for the expectation value II

- The interval $\left[\bar{X}_n - 1.96 \cdot \frac{s_x}{\sqrt{n}}, \bar{X}_n + 1.96 \cdot \frac{s_x}{\sqrt{n}} \right]$ is an (approximate) **95%-confidence interval** for the expectation value.
- More general: $\left[\bar{X}_n - \Phi^{-1}(1 - \alpha/2) \cdot \frac{s_x}{\sqrt{n}}, \bar{X}_n + \Phi^{-1}(1 - \alpha/2) \cdot \frac{s_x}{\sqrt{n}} \right]$ is a confidence interval to the **confidence level** $1 - \alpha$, $\frac{1}{2} < \alpha < 1$.
 $\Phi^{-1}(1 - \alpha/2)$: $1 - \alpha/2$ -quantile of the standard normal distribution.

Definition (Confidence interval for the expectation value)

A **confidence interval** to a **confidence level** $1 - \alpha$ for the expectation value μ is an interval I with the property that $P[\mu \in I] = 1 - \alpha$.

NOTE: the borders of the interval are random, not μ !!! μ is fix; the probability refers to the randomness in the *sample*.

Comments on the confidence intervals

- The confidence intervals for μ from the previous slide are only approximate because:
 - ▶ normal distribution of central limit theorem is only valid for large samples
 - ▶ standard deviation of underlying distribution is estimated, not known
- When we *know* that the distribution generating the sample is normal (with unknown SD), we get better (exact!) confidence intervals based on Student's t -distribution (see later)

Normal approximation of a binomial distribution

- Let $X \sim \text{Bin}(n, \pi)$ be a binomially distributed random variable
- If n is large enough, we can approximate its CDF by a normal one:

$$X \approx \mathcal{N}(n\pi, n\pi(1 - \pi))$$

- Rule of thumb: this can be done if $n\pi > 5$ and $n(1 - \pi) > 5$

Normal approximation of a binomial distribution

- Let $X \sim \text{Bin}(n, \pi)$ be a binomially distributed random variable
- If n is large enough, we can approximate its CDF by a normal one:

$$X \approx \mathcal{N}(n\pi, n\pi(1 - \pi))$$

- Rule of thumb: this can be done if $n\pi > 5$ and $n(1 - \pi) > 5$
- Note: this makes only sense if one is interested in the CDF. The probability mass function of X can **not** be replaced by a normal density!

Part VII

Fitting Distributions to Data: Maximum Likelihood Estimation

Learning objectives

- Know the definition of a likelihood
- Analytically derive the maximum likelihood estimator for a simple density
- Know the maximum likelihood estimators for the distributions learned so far
- Generate, read and interpret a Q-Q plot

Fitting distributions to data

- Goal: given data set, find parametric distribution that “explains the data set”

Fitting distributions to data

- Goal: given data set, find parametric distribution that “explains the data set”
- Concretely: find family of distribution (binomial, Poisson, normal, etc.) *and corresponding parameters* (binomial: π , Poisson: λ , normal: μ and σ^2 , etc.) that fits the data well

Example: estimate a probability

- Situation: population of *Drosophila melanogaster* that is not homozygous for a certain trait, e.g. “vestigial wings”
- Aim: estimate probability π of an eclosed fruit fly *with unknown parents* to have vestigial wings
- Intuitive solution? Justification?

Drosophila example: probabilistic model

- Bernoulli variable indicating whether a fruit fly has vestigial wings:

$$X = \begin{cases} 1, & \text{fly has vestigial wings,} \\ 0, & \text{otherwise.} \end{cases} \quad P[X = 1] = \pi$$

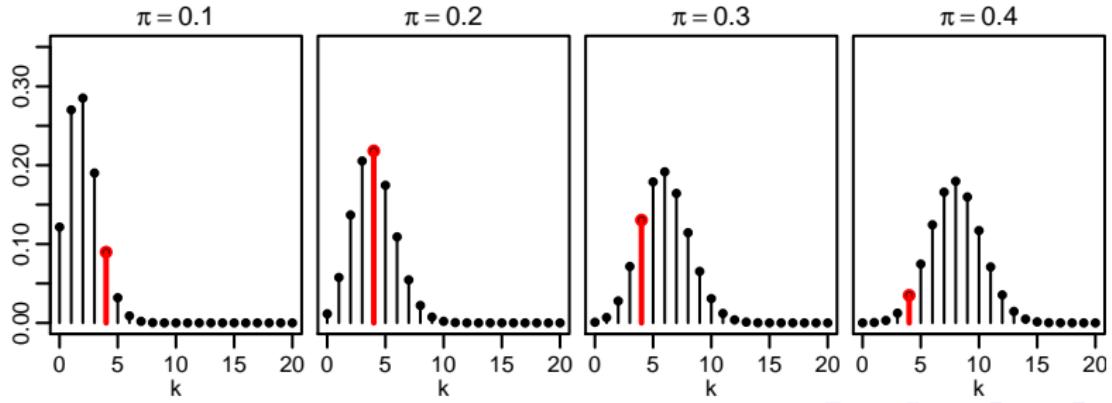
- Collect i.i.d. sample of 20 flies, count number of flies with vestigial wings: e.g. $k = 4$.

Drosophila example: probabilistic model

- Bernoulli variable indicating whether a fruit fly has vestigial wings:

$$X = \begin{cases} 1, & \text{fly has vestigial wings,} \\ 0, & \text{otherwise.} \end{cases} \quad P[X = 1] = \pi$$

- Collect i.i.d. sample of 20 flies, count number of flies with vestigial wings: e.g. $k = 4$.
- How probable is it to find k flies with vestigial wings? As a random variable, the quantity has binomial distribution: $K \sim \text{Bin}(20, \pi)$



Drosophila example: probabilistic model

- Bernoulli variable indicating whether a fruit fly has vestigial wings:

$$X = \begin{cases} 1, & \text{fly has vestigial wings,} \\ 0, & \text{otherwise.} \end{cases} \quad P[X = 1] = \pi$$

- Collect i.i.d. sample of flies, determine their wings: x_1, x_2, \dots, x_n .
Count number of flies with vestigial wings: $k = \#\{i | x_i = 1\}$
- The probability mass function of the random variable X_1 at its value x_1 can be written as

$$P[X_1 = x_1] = \pi^{x_1}(1 - \pi)^{1-x_1}$$

- The **joint probability mass function** of the sample x_1, \dots, x_n is hence

$$\left[\pi^{x_1}(1 - \pi)^{1-x_1}\right] \cdot \dots \cdot \left[\pi^{x_n}(1 - \pi)^{1-x_n}\right] = \pi^k(1 - \pi)^{n-k}$$

Drosophila example: likelihood

- The **likelihood** of the sample is its joint probability mass function, viewed as a function of the parameter π :

$$L(\pi; x_1, \dots, x_n) = \pi^k (1 - \pi)^{n-k}$$

- The likelihood is no new concept or definition: it's just the joint probability mass function. We speak of the *likelihood* when we consider the sample as *fix* and the parameter as *variable*. Hence we sometimes do not write the sample explicitly:

$$L(\pi) = L(\pi; x_1, \dots, x_n)$$

Drosophila example: likelihood

- The **likelihood** of the sample is its joint probability mass function, viewed as a function of the parameter π :

$$L(\pi; x_1, \dots, x_n) = \pi^k (1 - \pi)^{n-k}$$

- The likelihood is no new concept or definition: it's just the joint probability mass function. We speak of the *likelihood* when we consider the sample as *fix* and the parameter as *variable*. Hence we sometimes do not write the sample explicitly:

$$L(\pi) = L(\pi; x_1, \dots, x_n)$$

- Log-likelihood** of the sample:

$$\ell(\pi) := \log(L(\pi)) = k \log(\pi) + (n - k) \log(1 - \pi)$$

- Parameter π that maximizes the likelihood (and hence the log-likelihood: $\hat{\pi} = \frac{k}{n}$ (**maximum likelihood estimator**)

Maximum likelihood estimation for discrete distributions

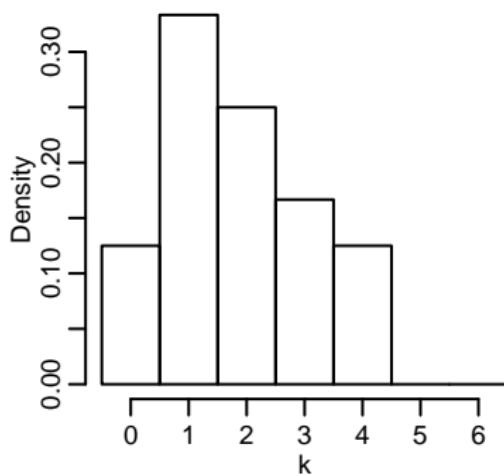
- Measurements X_1, X_2, \dots, X_n : i.i.d copies of discrete random variable X with probability mass function $p(x; \theta)$: **parameterized by θ**
- **Likelihood** $L(\theta) := p(x_1; \theta) \cdot p(x_2; \theta) \cdot \dots \cdot p(x_n; \theta) = \prod_{i=1}^n p(x_i; \theta)$
- **Log-likelihood** $\ell(\theta) := \log(L(\theta)) = \log(p(x_1; \theta)) + \log(p(x_2; \theta)) + \dots + \log(p(x_n; \theta)) = \sum_{i=1}^n \log(p(x_i; \theta))$
- **Maximum likelihood estimator (MLE)** for θ : $\hat{\theta}$ = value of θ for which ℓ attains its maximum
- Calculation: derive $\ell(\theta)$ w.r.t. θ , set derivation to 0, solve for θ

Maximum likelihood estimation for continuous distributions

- Measurements X_1, X_2, \dots, X_n : i.i.d copies of continuous random variable X with probability density $f(x; \theta)$: **parameterized by θ**
- **Likelihood** $L(\theta) := f(x_1; \theta) \cdot f(x_2; \theta) \cdot \dots \cdot f(x_n; \theta) = \prod_{i=1}^n f(x_i; \theta)$
- **Log-likelihood** $\ell(\theta) := \log(L(\theta)) = \log(f(x_1; \theta)) + \log(f(x_2; \theta)) + \dots + \log(f(x_n; \theta)) = \sum_{i=1}^n \log(f(x_i; \theta))$
- **Maximum likelihood estimator (MLE)** for θ : $\hat{\theta}$ = value of θ for which ℓ attains its maximum
- Calculation: derive $\ell(\theta)$ w.r.t. θ , set derivation to 0, solve for θ

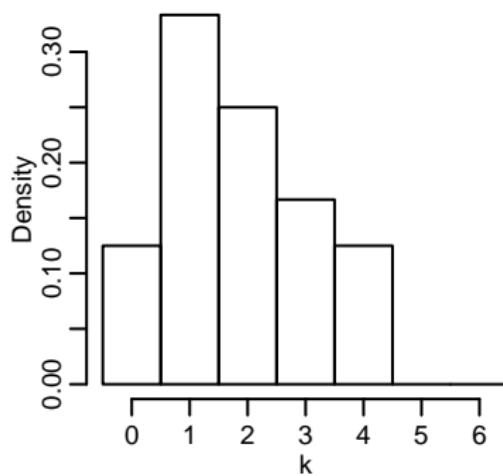
Example: birth statistic

- Birth time, sex and weight of 44 babies born within 24 h at a hospital in Brisbane, Australia, collected
- Histogram: number of births per hour:



Example: birth statistic

- Birth time, sex and weight of 44 babies born within 24 h at a hospital in Brisbane, Australia, collected
- Histogram: number of births per hour:



- Appropriate distribution to fit data?

MLE for Poisson distribution

- Sample from Poisson distribution: x_1, x_2, \dots, x_n
- Likelihood of sample: $L(\lambda) = \prod_{i=1}^n e^{-\lambda} \frac{\lambda^{x_i}}{x_i!}$
- Log-likelihood of sample: $\ell(\lambda) = \sum_{i=1}^n [x_i \log(\lambda) - \lambda - \log(x_i!)]$
- **Maximum likelihood estimator:**
$$\hat{\lambda} = \frac{1}{n}(x_1 + x_2 + \dots + x_n) = \frac{1}{n} \sum_{i=1}^n x_i = \bar{x}$$

Example: birth statistic

Data set:

Hour i :	0	1	2	3	4	5	6	7	8	9	...
Births x_i :	1	3	1	0	4	0	0	2	2	1	...

Example: birth statistic

Data set:

Hour i :	0	1	2	3	4	5	6	7	8	9	...
Births x_i :	1	3	1	0	4	0	0	2	2	1	...

- MLE for Poisson parameter λ :
 $\hat{\lambda} = \bar{x} = 1.833$
- \rightsquigarrow Model: $X \sim \text{Pois}(\hat{\lambda})$ with $\hat{\lambda} = 1.833$

Example: birth statistic

Data set:

Hour i :	0	1	2	3	4	5	6	7	8	9	...
Births x_i :	1	3	1	0	4	0	0	2	2	1	...

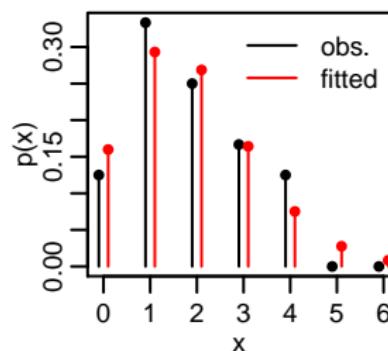
- MLE for Poisson parameter λ :
 $\hat{\lambda} = \bar{x} = 1.833$
- \rightsquigarrow Model: $X \sim \text{Pois}(\hat{\lambda})$ with $\hat{\lambda} = 1.833$
- Does our data set fit this distribution well?

Example: birth statistic

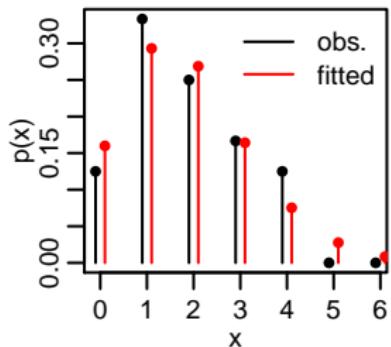
Data set:

Hour i :	0	1	2	3	4	5	6	7	8	9	...
Births x_i :	1	3	1	0	4	0	0	2	2	1	...

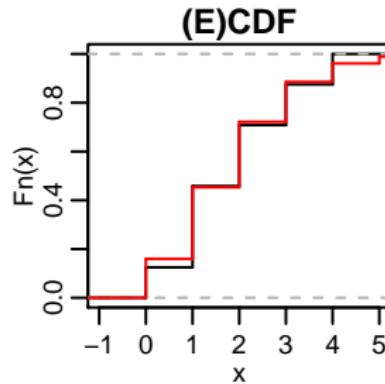
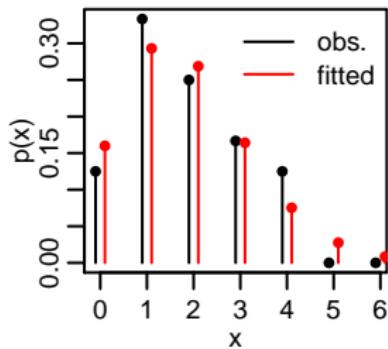
- MLE for Poisson parameter λ :
 $\hat{\lambda} = \bar{x} = 1.833$
- Model: $X \sim \text{Pois}(\hat{\lambda})$ with $\hat{\lambda} = 1.833$
- Does our data set fit this distribution well?



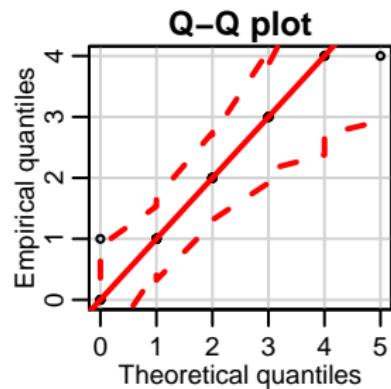
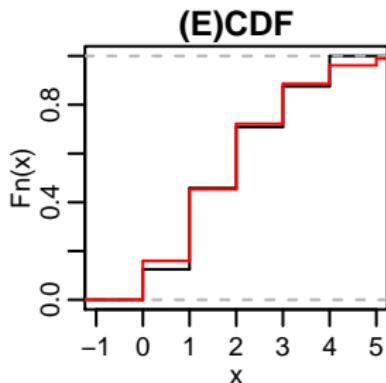
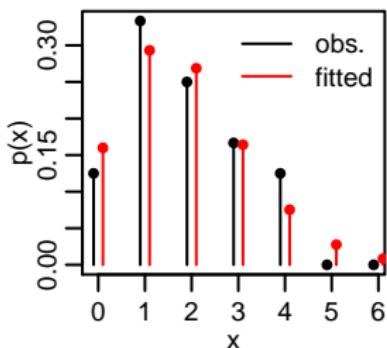
Q-Q (quantile-quantile) plot



Q-Q (quantile-quantile) plot

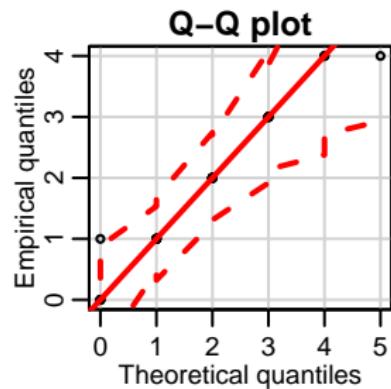
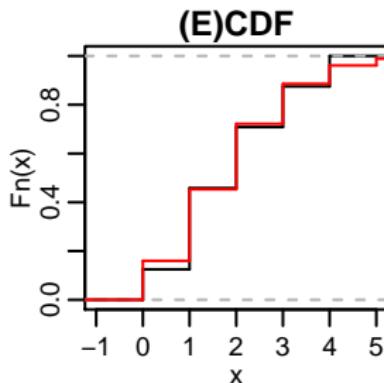
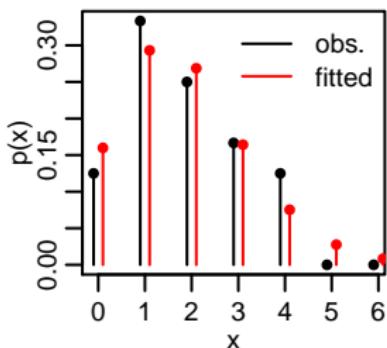


Q-Q (quantile-quantile) plot



The closer the points are on the diagonal of the Q-Q plot, the better the data fits the chosen distribution.

Q-Q (quantile-quantile) plot



The closer the points are on the diagonal of the Q-Q plot, the better the data fits the chosen distribution.

Next question: how precise is the estimated $\hat{\lambda}$?

Confidence interval for estimated parameter

- Since $\hat{\lambda}$ was calculated as a *mean* of the sample x_1, \dots, x_n , we can use the (approximate) confidence intervals derived in the last chapter:

$$\left[\hat{\lambda} - \Phi^{-1}(0.975) \frac{s_x}{\sqrt{n}}, \hat{\lambda} + \Phi^{-1}(0.975) \frac{s_x}{\sqrt{n}} \right]$$

Confidence interval for estimated parameter

- Since $\hat{\lambda}$ was calculated as a *mean* of the sample x_1, \dots, x_n , we can use the (approximate) confidence intervals derived in the last chapter:

$$\left[\hat{\lambda} - \Phi^{-1}(0.975) \frac{s_x}{\sqrt{n}}, \hat{\lambda} + \Phi^{-1}(0.975) \frac{s_x}{\sqrt{n}} \right]$$

- With (approximately) 95% probability, this interval contains the true parameter λ
- Note** (again): the interval is random, not the true parameter!

Example: birth statistic

- (Approximate) 95% confidence interval for true parameter λ :

$$\left[\hat{\lambda} - \Phi^{-1}(0.975) \frac{s_x}{\sqrt{n}}, \hat{\lambda} + \Phi^{-1}(0.975) \frac{s_x}{\sqrt{n}} \right]$$

$\Phi^{-1}(0.975)$: 97.5% quantile of standard normal distribution

- In our example: > mean(x)

```
[1] 1.833333
```

```
> sd(x)
```

```
[1] 1.239448
```

```
> qnorm(0.975)
```

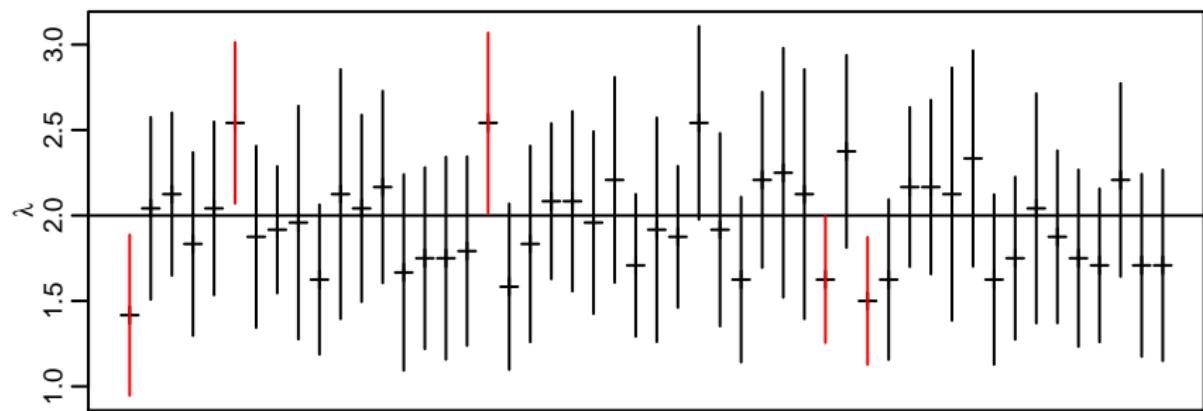
```
[1] 1.959964
```

- 95% confidence interval for λ : [1.337, 2.329]

Confidence intervals are random

Simulation: repeat 50 times:

- Draw a sample of size 24 from $\text{Pois}(\lambda = 2)$
- Calculate the corresponding confidence interval for λ



Always publish estimated numbers together with confidence intervals!

Fitting a normal distribution

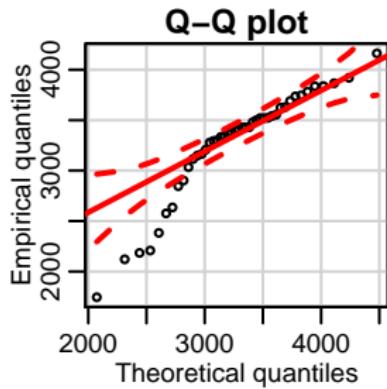
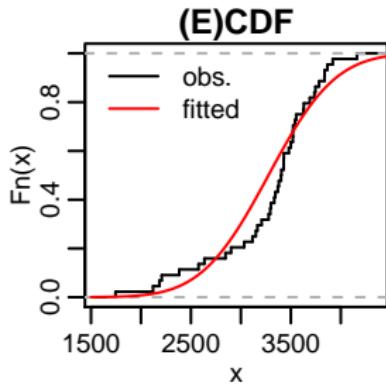
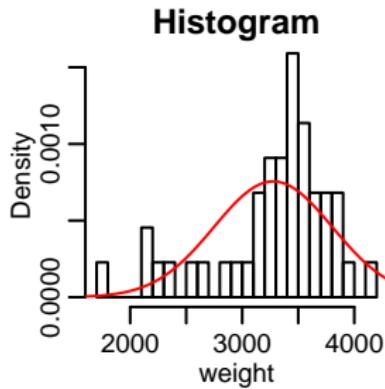
- Normal distribution $\mathcal{N}(\mu, \sigma^2)$: characterized by mean (μ) and variance (σ^2)
- Consider i.i.d. sample x_1, x_2, \dots, x_n from $\mathcal{N}(\mu, \sigma^2)$
- Already seen unbiased estimators: \bar{X} for μ , s_x^2 for σ^2
- Maximum likelihood estimators:

$$\hat{\mu} = \bar{X}, \quad \hat{\sigma}^2 = \frac{1}{n} \sum_{i=1}^n (x_i - \bar{x})^2$$

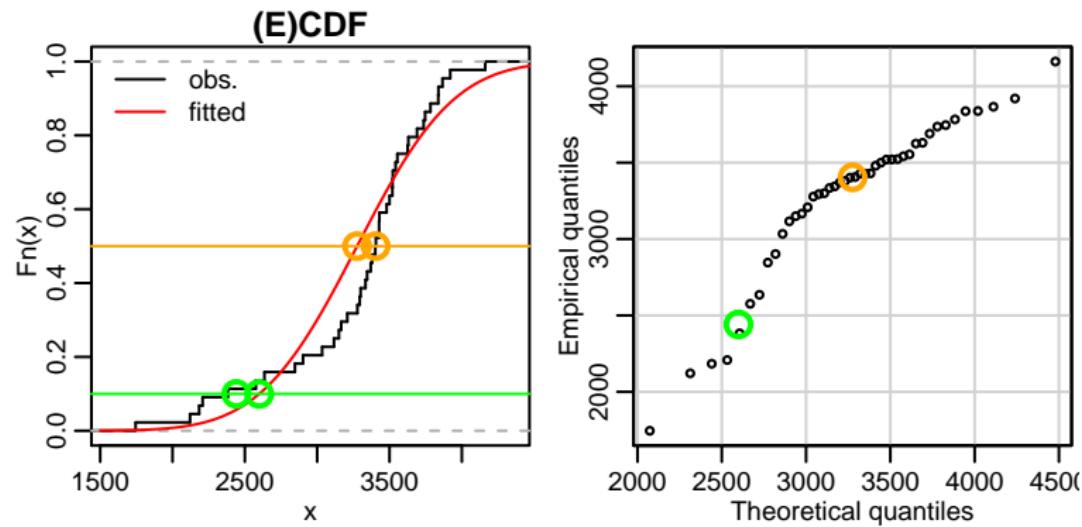
(note: $\hat{\sigma}^2$ is [slightly] biased.)

Example: birth statistic

- Back to data set of babies: consider $X = \text{weight of babies}$
- Is the weight normally distributed?



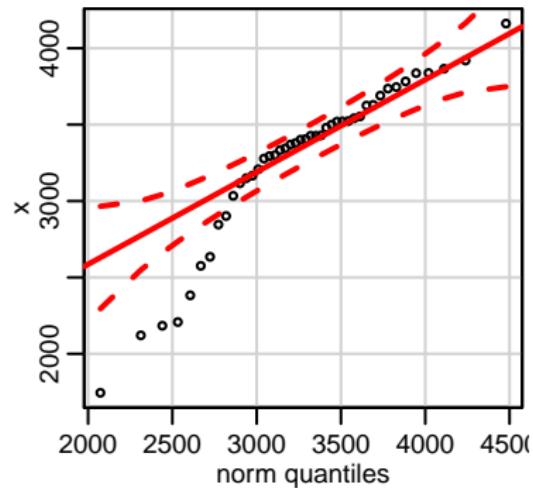
Q-Q plot: a closer look



Q-Q plot: R commands

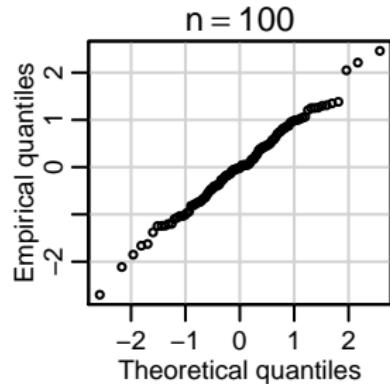
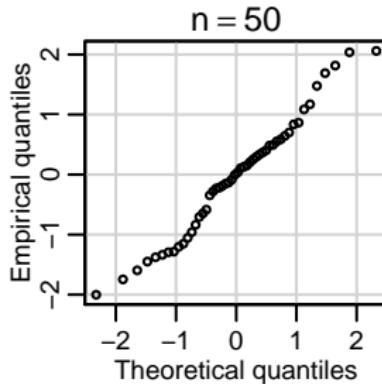
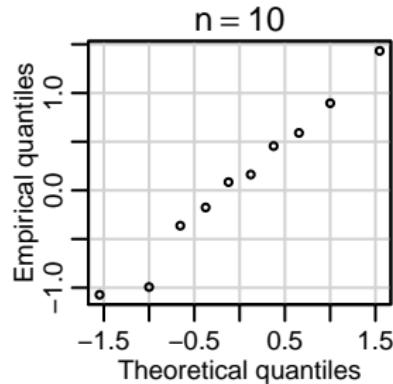
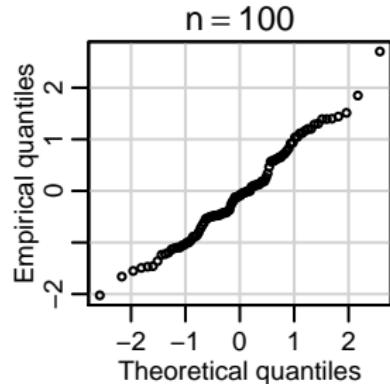
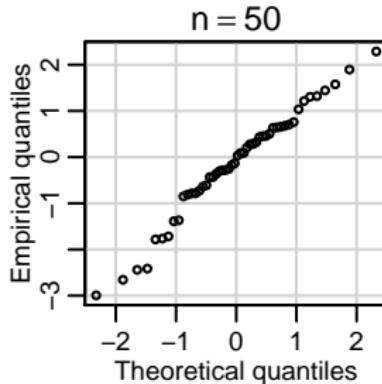
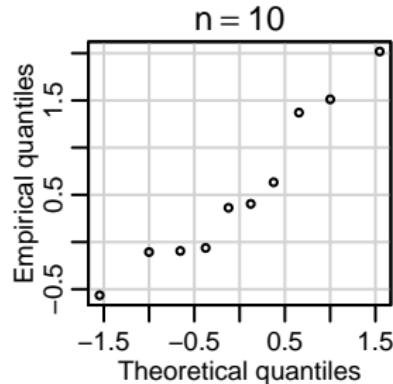
Use R package car ("Companion to Applied Regression") for Q-Q plots:

```
> library(car) > (est.mean <-  
mean(x))  
[1] 3275.955  
> (est.sd <- sd(x))  
[1] 528.0325  
> qqPlot(x, dist = "norm", mean =  
est.mean, sd = est.sd)
```

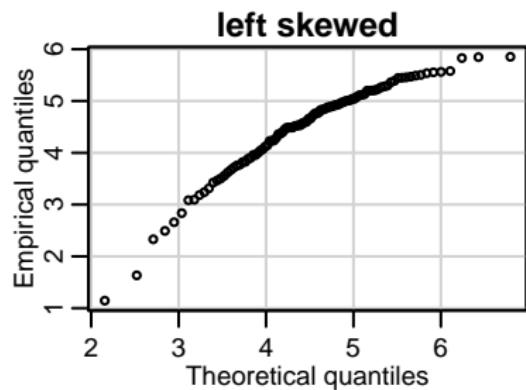
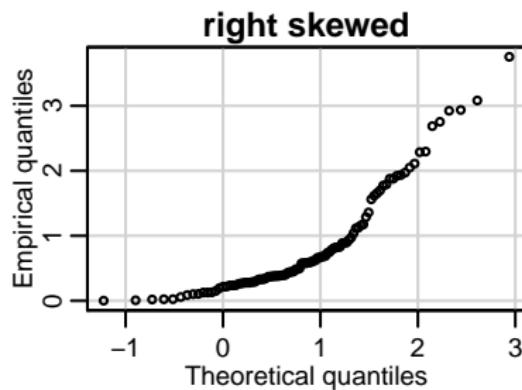
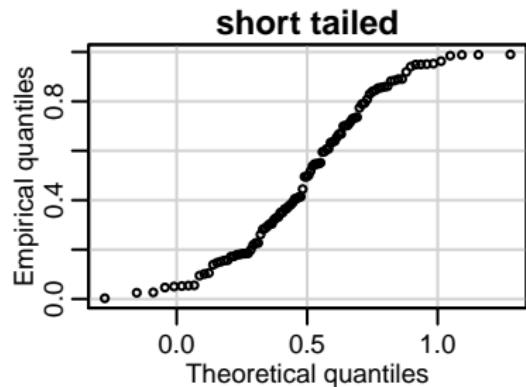
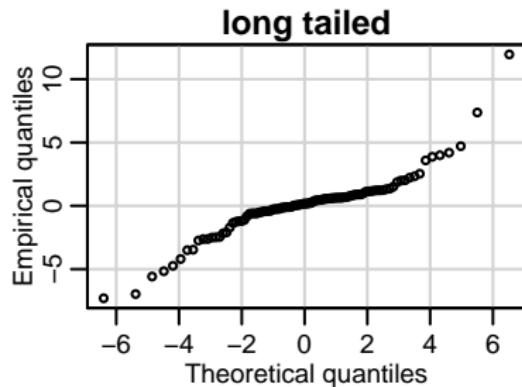


Q-Q plots of normal samples

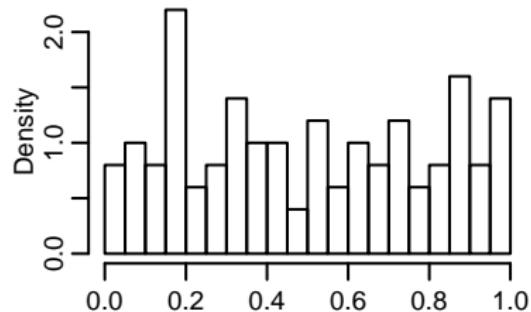
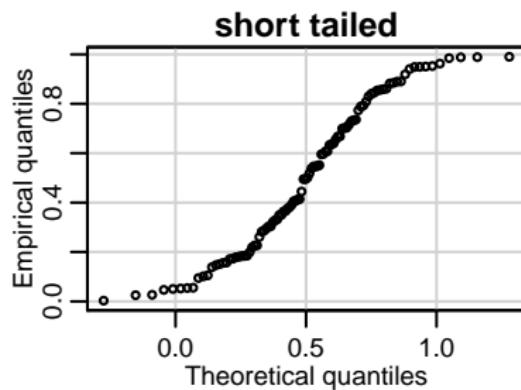
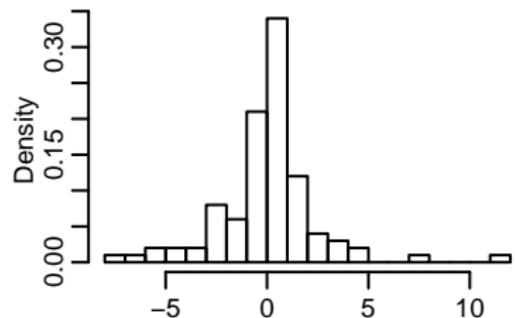
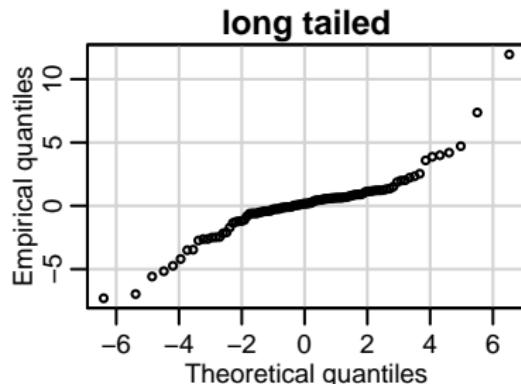
Q-Q plots of samples simulated from the normal distribution:



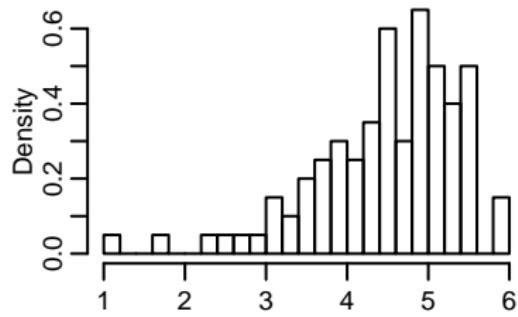
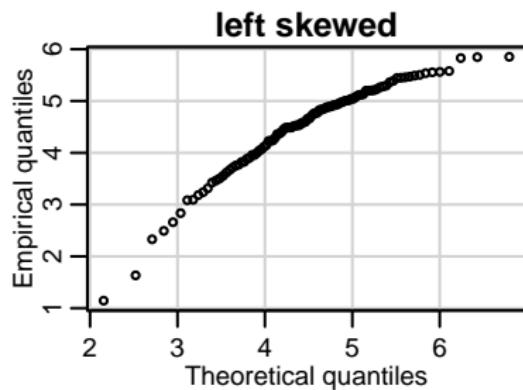
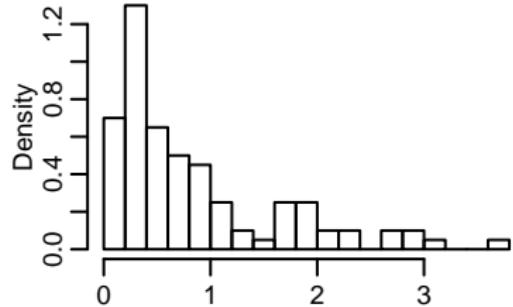
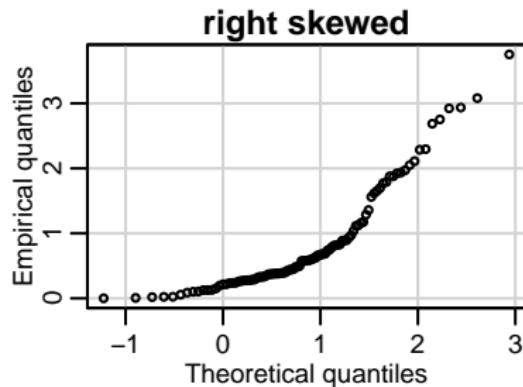
Q-Q plots: deviations from normality



Q-Q plots: deviations from normality

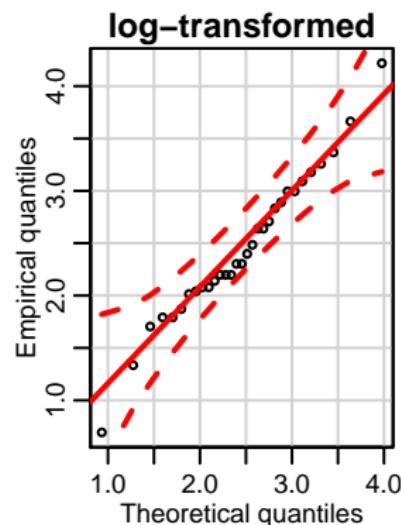
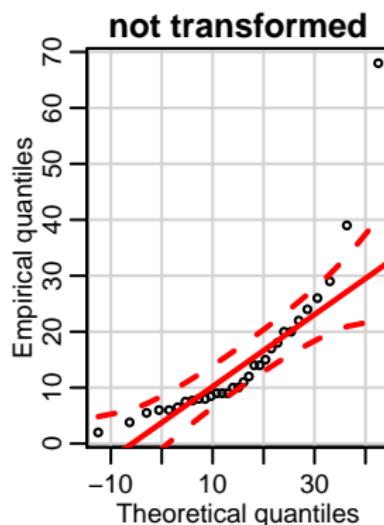
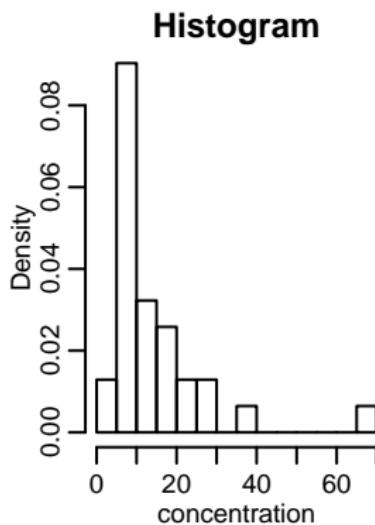


Q-Q plots: deviations from normality



Right-skewness and transformations

- Right-skewed data sometimes fits a normal distribution after a transformation
- Often used transformations: logarithm, square root
- Example: set of measurements of plumb concentration in granite



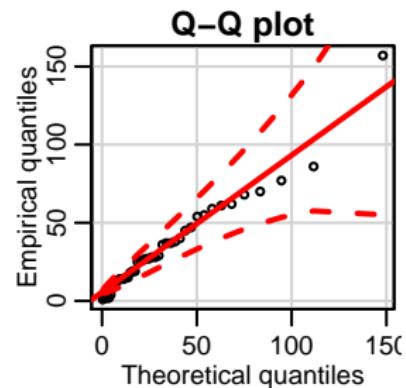
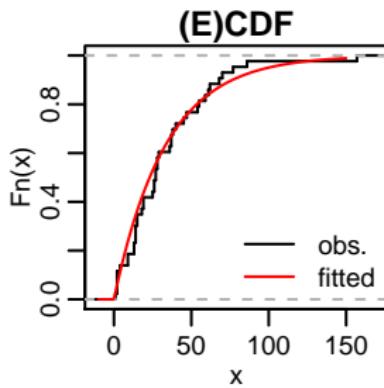
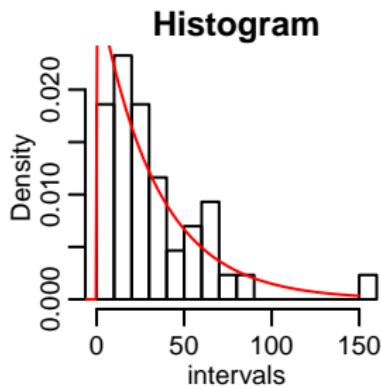
Fitting an exponential distribution

- Exponential distribution $\text{Exp}(\lambda)$: characterized by rate parameter λ
- Consider i.i.d. sample x_1, x_2, \dots, x_n from $\text{Exp}(\lambda)$
- Maximum likelihood estimator: $\hat{\lambda} = \frac{1}{\bar{x}}$
- Approximate 95% confidence interval:

$$\left[\hat{\lambda} \left(1 - \frac{\Phi^{-1}(0.975)}{\sqrt{n}} \right), \hat{\lambda} \left(1 + \frac{\Phi^{-1}(0.975)}{\sqrt{n}} \right) \right]$$

Example: birth statistic

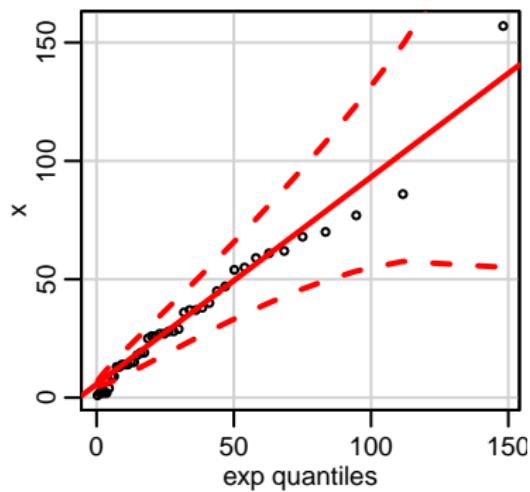
- Data set of babies: consider t_i = time intervals between births (in minutes)
- When number of births in time interval is Poisson distributed, time interval between births should be exponentially distributed (remember Haldane's model!)



Example: birth statistic

R commands to fit distribution:

```
> (rate <- 1/mean(x))  
[1] 0.03006993  
> qqPlot(x, dist = "exp", rate = rate)
```



Part VIII

A Glimpse on Bayesian Parameter Estimation

Learning targets

- Recognize and explain categorical random variables
- Apply Bayes' formula
- Explain the difference between maximum likelihood (ML) and maximum a posteriori (MAP) estimation
- Recognize situations in which one or the other approach is more appropriate
- Explain the influence of the sample size on the MAP estimator

ABO blood types

- ABO blood type is controlled by a single gene with three alleles: i , I^A and I^B
- Allele I^A yields type A, I^B type B, i type O
- Alleles I^A and I^B are *codominant* among each other, i.e., phenotype $I^A|I^B$ gives blood type AB
- Allele i is recessive against I^A and I^B
- Allele frequencies of I^A , I^B , i in European population $\pi_A = 0.3$, $\pi_B = 0.1$, $\pi_i = 0.6$

Example 1: probability of genotype of random person I

- What's the probability that a *random person* has genotype $I^A I^A$, $I^A i$, or $I^B i$?

Example 1: probability of genotype of random person I

- What's the probability that a *random person* has genotype $I^A I^A$, $I^A i$, or $I^B i$?
- Notation: X : random variable encoding for genotype. X is a **categorical** variable (also called **factor**), i.e. a variable that can take discrete values that are *not ordered* (e.g., genotype $I^A i$ is not “larger or smaller” than genotype $I^A I^A$)
- Categorical random variables have no *expectation value* or *variance*; but *events* such as $X = I^A I^A$ are well-defined.

Example 1: probability of genotype of random person II

- What's the probability that a *random person* has genotype $I^A I^A$, $I^A i$, or $I^B i$?

$$P[X = I^A I^A] = \pi_A^2 = (0.3)^2 = 0.09$$

$$P[X = I^A i] = 2\pi_A \pi_i = 2 \cdot 0.3 \cdot 0.6 = 0.36$$

$$P[X = I^B i] = 2\pi_B \pi_i = 2 \cdot 0.1 \cdot 0.6 = 0.12$$

Example 2: genotype of a person with known phenotype I

- What's the probability that a person *with blood type A* has genotype $I^A I^A$, $I^A i$, or $I^B i$?

Example 2: genotype of a person with known phenotype I

- What's the probability that a person *with blood type A* has genotype $I^A I^A$, $I^A i$, or $I^B i$?
- Introduce new categorical variable: Y : phenotype of the person. Y can have “values” A, B, AB, or 0.

Example 2: genotype of a person with known phenotype I

- What's the probability that a person *with blood type A* has genotype $I^A I^A$, $I^A i$, or $I^B i$?
- Introduce new categorical variable: Y : phenotype of the person. Y can have “values” A, B, AB, or 0.
- We are now not interested in $P[X = I^A I^A]$ etc. any more, but in

$$P[X = I^A I^A \mid Y = A]$$

Example 2: genotype of a person with known phenotype I

- What's the probability that a person *with blood type A* has genotype $I^A I^A$, $I^A i$, or $I^B i$?
- Introduce new categorical variable: Y : phenotype of the person. Y can have “values” A, B, AB, or 0.
- We are now not interested in $P[X = I^A I^A]$ etc. any more, but in

$$P[X = I^A I^A \mid Y = A]$$

- Bayes' theorem:

$$P[X = I^A I^A \mid Y = A] = \frac{P[Y = A \mid X = I^A I^A] \cdot P[X = I^A I^A]}{P[Y = A]}$$

Example 2: genotype of a person with known phenotype II

Terms in Bayes' formula:

- $P[X = I^A | A]$: calculated in example before

Example 2: genotype of a person with known phenotype II

Terms in Bayes' formula:

- $P[X = I^A | A]$: calculated in example before
- $P[Y = A | X = I^A | A] = 1$

Example 2: genotype of a person with known phenotype II

Terms in Bayes' formula:

- $P[X = I^A I^A]$: calculated in example before
- $P[Y = A | X = I^A I^A] = 1$
- Difficult term: $P[Y = A]$. By the **law of total probability** (see Part II of the lecture), we can write

$$P[Y = A] = \sum_{x: \text{genotype}} P[Y = A | X = x] \cdot P[X = x]$$

- If a genotype x leads to blood type A, $P[Y = A | X = x] = 1$; otherwise $P[Y = A | X = x] = 0$
- We finally find

$$P[Y = A] = P[X = I^A I^A] + P[X = I^A i] = 0.09 + 0.36 = 0.45$$

Example 2: genotype of a person with known phenotype III

Summary: what's the probability that a person *with blood type A* has genotype $I^A I^A$, $I^A i$, or $I^B i$?

$$P[X = I^A I^A \mid Y = A] = 0.2$$

$$P[X = I^A i \mid Y = A] = 0.8$$

$$P[X = I^B i \mid Y = A] = 0.0$$

Example 3: estimate genotype of unknown father I

Situation:

- Child has blood type AB
- Mother has blood type A
- Father unknown

Goal: estimate genotype of father

Example 3: estimate genotype of unknown father I

Situation:

- Child has blood type AB
- Mother has blood type A
- Father unknown

Goal: estimate genotype of father

First approach: maximum likelihood estimation

Example 3: estimate genotype of unknown father II

Likelihood approach:

- Observed variable X : child's allele inherited from father (I^B)
- Probability of inherited allele depends on genotype of father
- Formally: genotype of father θ "parameterizes" probability for inherited allele
- Possible genotypes of father: $I^A|I^B$, $I^B|I^B$, $I^B|i$

Example: estimate genotype of unknown father

Likelihood approach:

- Observed variable X : child's allele inherited from father (I^B)
- Probability of inherited allele depends on genotype of father
- Formally: genotype of father θ “parameterizes” probability for inherited allele
- Possible genotypes of father: $I^A|I^B$, $I^B|I^B$, $I^B|i$

Example: estimate genotype of unknown father

Likelihood approach:

- Observed variable X : child's allele inherited from father (I^B)
- Probability of inherited allele depends on genotype of father
- Formally: genotype of father θ "parameterizes" probability for inherited allele
- Possible genotypes of father: $I^A|I^B$, $I^B|I^B$, $I^B|i$
- Likelihoods:

$$P_{\theta=I^A|I^B}[X = I^B] = 0.5$$

$$P_{\theta=I^B|I^B}[X = I^B] = 1$$

$$P_{\theta=I^B|i}[X = I^B] = 0.5$$

Example: estimate genotype of unknown father

- Likelihoods:

$$P_{\theta=I^A I^B}[X = I^B] = 0.5$$

$$P_{\theta=I^B I^B}[X = I^B] = 1$$

$$P_{\theta=I^B i}[X = I^B] = 0.5$$

- Maximum likelihood estimate for genotype of father: $I^B I^B$

Example: estimate genotype of unknown father

- Likelihoods:

$$P_{\theta=I^A I^B}[X = I^B] = 0.5$$

$$P_{\theta=I^B I^B}[X = I^B] = 1$$

$$P_{\theta=I^B i}[X = I^B] = 0.5$$

- Maximum likelihood estimate for genotype of father: $I^B I^B$
- Knowing that frequency of I^B in population is $\pi_B = 0.1$, only a fraction of $\pi_B^2 = 0.01$ (1%) of the population has genotype $I^B I^B$
- Taking frequency of I^B into account, there are more probable “sources” of inheriting an I^B allele

Bayesian estimation: approach

Assuming discrete variable and parameter, as in example:

- Consider parameter θ as *random* rather than fix
- Likelihood is then a conditional probability:

$$L(\theta) = p_{X|\Theta=\theta}(x) = P[X = x | \Theta = \theta]$$

- Bayes' theorem:

$$P[\Theta = \theta | X = x] = \frac{P[X = x | \Theta = \theta] \cdot P[\Theta = \theta]}{P[X = x]}$$

Bayesian estimation: approach

Assuming discrete variable and parameter, as in example:

- Consider parameter θ as *random* rather than fix
- Likelihood is then a conditional probability:

$$L(\theta) = p_{X|\Theta=\theta}(x) = P[X = x | \Theta = \theta]$$

- Bayes' theorem:

$$\underbrace{P[\Theta = \theta | X = x]}_{\text{posterior}} = \frac{\underbrace{P[X = x | \Theta = \theta] \cdot P[\Theta = \theta]}_{\text{likelihood} \cdot \text{prior}}}{\underbrace{P[X = x]}_{\text{evidence}}}$$

Bayesian estimation: approach

Assuming discrete variable and parameter, as in example:

- Consider parameter θ as *random* rather than fix
- Likelihood is then a conditional probability:

$$L(\theta) = p_{X|\Theta=\theta}(x) = P[X = x | \Theta = \theta]$$

- Bayes' theorem:

$$\overbrace{P[\Theta = \theta | X = x]}^{\text{posterior}} = \frac{\underbrace{P[X = x | \Theta = \theta] \cdot P[\Theta = \theta]}_{\text{evidence}}}{\underbrace{P[X = x]}_{\text{prior}}}$$

- **Maximum a posteriori (MAP) estimator:** $\hat{\theta}$ that maximizes $P[\Theta = \theta | X = x]$.

Maximum a posteriori estimator

- Evidence $P[X = x]$ does not depend on parameter
- MAP estimator $\hat{\theta}$ hence maximizes *likelihood weighted by prior*:

$$P[X = x | \Theta = \theta] \cdot P[\Theta = \theta]$$

- In contrast, maximum likelihood (ML) estimator maximizes likelihood alone, ignoring prior

Example: AB0 blood types

Back to the blood types example: θ = genotype of father, X = allele inherited from father

- Likelihoods:

$$P[X = I^B | \Theta = I^A I^B] = 0.5$$

$$P[X = I^B | \Theta = I^B I^B] = 1$$

$$P[X = I^B | \Theta = I^B i] = 0.5$$

- Priors:

$$P[\Theta = I^A I^B] = 2\pi_A \pi_B = 0.06$$

$$P[\Theta = I^B I^B] = \pi_B^2 = 0.01$$

$$P[\Theta = I^B i] = 2\pi_B \pi_i = 0.12$$

Example: AB0 blood types

θ = genotype of father, X = allele inherited from father

- Likelihood times prior:

$$P[X = I^B | \Theta = I^A I^B] \cdot P[\Theta = I^A I^B] = 0.03$$

$$P[X = I^B | \Theta = I^B I^B] \cdot P[\Theta = I^B I^B] = 0.01$$

$$P[X = I^B | \Theta = I^B i] \cdot P[\Theta = I^B i] = 0.06$$

- Maximum a posteriori (MAP) estimator: $\hat{\theta} = I^B i$
- In Bayesian context, one is usually interested in complete posterior distribution, not only its maximum: wider view on different hypothesis consistent with the data

Bayesian estimation of continuous parameter

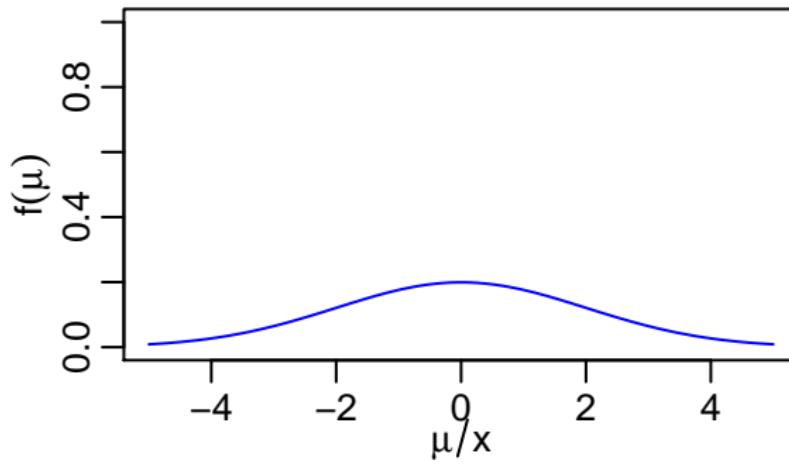
- Setting: continuous random variable X with density depending on parameter θ
- Parameter θ : realization of random variable Θ with density $f_\Theta(\theta)$
- For a given value of θ , distribution of X is described by *conditional* density $f_{X|\Theta=\theta}(x)$.
- Bayes' theorem for densities:

$$\underbrace{f_{\Theta|X=x}(\theta)}_{\text{posterior}} = \frac{\overbrace{f_{X|\Theta=\theta}(x) \cdot f_\Theta(\theta)}^{\text{likelihood prior}}}{\underbrace{f_X(x)}_{\text{evidence}}}$$

- MAP estimator for Θ : value $\hat{\theta}$ that maximizes posterior density $f_{\Theta|X=x}(\theta)$ (and hence product $f_{X|\Theta=\theta}(x) \cdot f_\Theta(\theta)$)

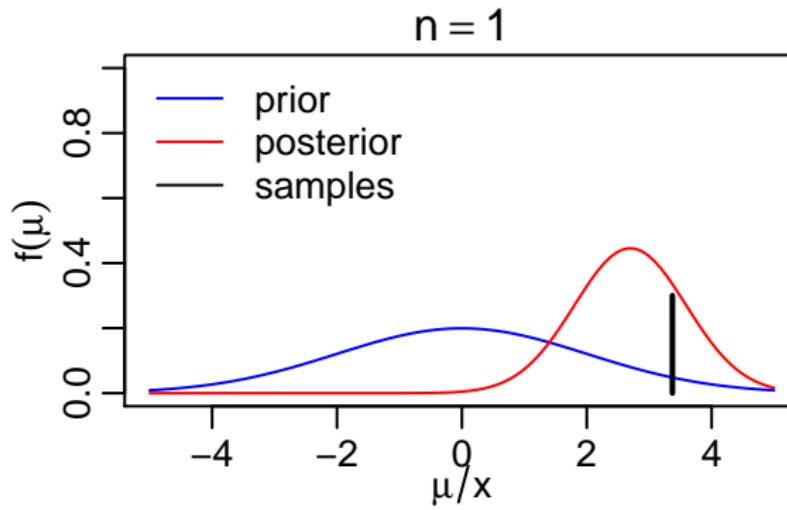
Example: MAP estimator for larger sample

- MAP estimator from previous slide can be generalized to larger samples
- Example: normal distribution with unknown mean μ and known variance $\sigma^2 = 1$
- Prior density of mean μ : $\mu \sim \mathcal{N}(0, 2)$
- The more samples we collect, the narrower becomes the posterior density:



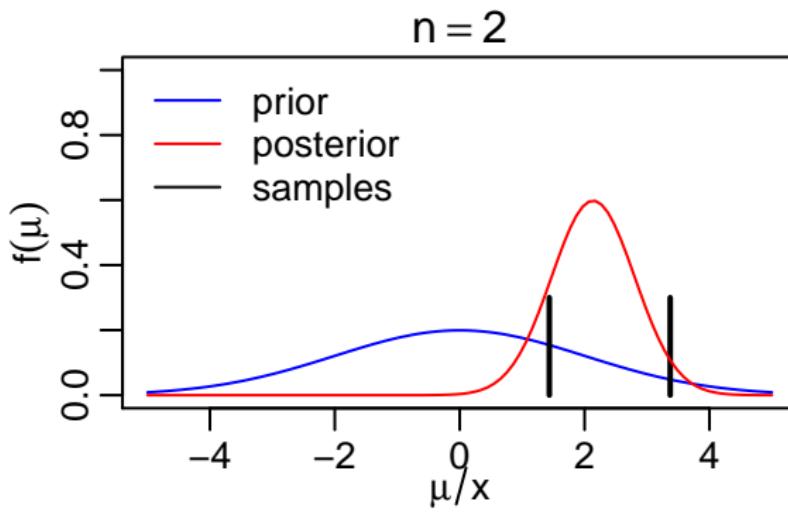
Example: MAP estimator for larger sample

- MAP estimator from previous slide can be generalized to larger samples
- Example: normal distribution with unknown mean μ and known variance $\sigma^2 = 1$
- Prior density of mean μ : $\mu \sim \mathcal{N}(0, 2)$
- The more samples we collect, the narrower becomes the posterior density:



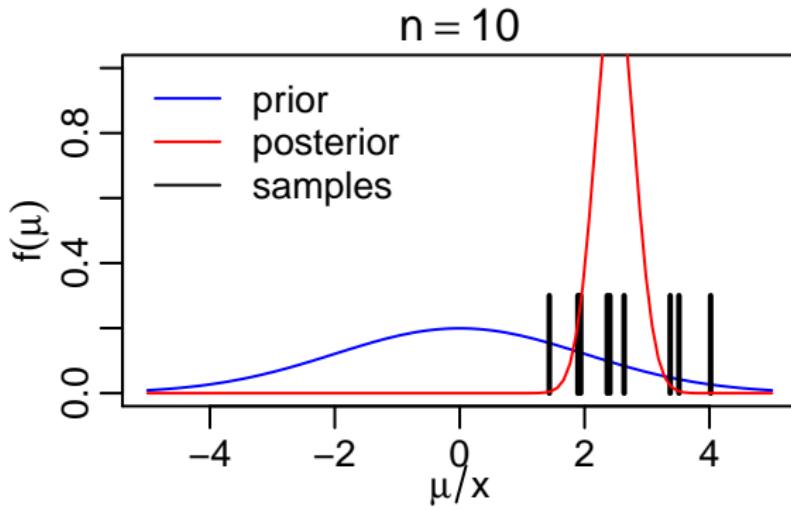
Example: MAP estimator for larger sample

- MAP estimator from previous slide can be generalized to larger samples
- Example: normal distribution with unknown mean μ and known variance $\sigma^2 = 1$
- Prior density of mean μ : $\mu \sim \mathcal{N}(0, 2)$
- The more samples we collect, the narrower becomes the posterior density:



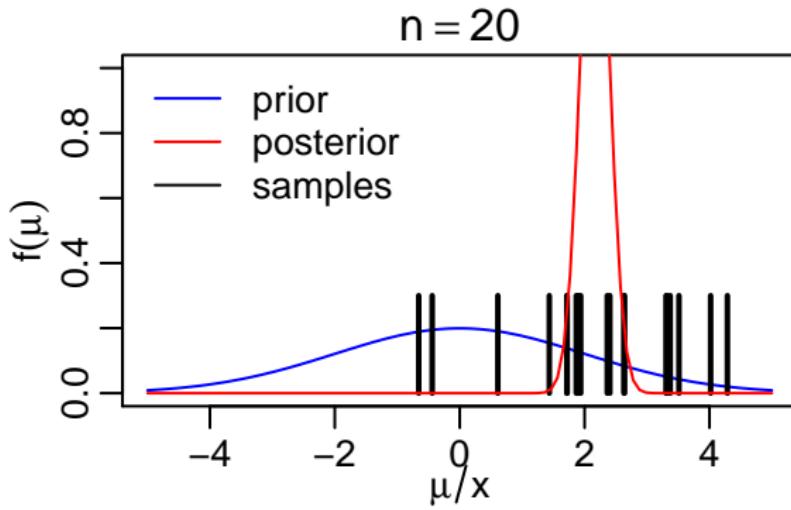
Example: MAP estimator for larger sample

- MAP estimator from previous slide can be generalized to larger samples
- Example: normal distribution with unknown mean μ and known variance $\sigma^2 = 1$
- Prior density of mean μ : $\mu \sim \mathcal{N}(0, 2)$
- The more samples we collect, the narrower becomes the posterior density:



Example: MAP estimator for larger sample

- MAP estimator from previous slide can be generalized to larger samples
- Example: normal distribution with unknown mean μ and known variance $\sigma^2 = 1$
- Prior density of mean μ : $\mu \sim \mathcal{N}(0, 2)$
- The more samples we collect, the narrower becomes the posterior density:



Calculation of posterior densities

- Small sample size: posterior distribution strongly influenced by prior distribution
- Large sample size: posterior distribution converges to maximum likelihood estimator
- Computation of posterior distribution often cumbersome:
 - ▶ choice of prior distribution (also) influenced by practical considerations (choosing distributions that are easier to handle)
 - ▶ analytical computation often impossible \rightsquigarrow requirement for advanced computational techniques such as Monte Carlo
- Fields of application for Bayesian methods:
 - ▶ calculation of risks for hereditary diseases (which could indicate necessity of prenatal tests)
 - ▶ phylogenetics
 - ▶ error correction, e.g. for PacBio reads of DNA

More in-depth treatment: see lectures of L. Excoffier

Comparison: ML and MAP estimation

ML estimation:

- Consider parameter of interest as fix, but unknown
- Underlying interpretation of probability: frequentist
- Maximizes likelihood of data
- Difficult to use for small samples
- Inclusion of prior knowledge via model class

Comparison: ML and MAP estimation

ML estimation:

- Consider parameter of interest as fix, but unknown
- Underlying interpretation of probability: frequentist
- Maximizes likelihood of data
- Difficult to use for small samples
- Inclusion of prior knowledge via model class

MAP estimation:

- Consider parameter of interest as random
- Underlying interpretation of probability: Bayesian
- Maximizes posterior probability/density, or equivalently: likelihood weighted by prior
- Suitable for small samples
- Inclusion of prior knowledge via prior probabilities

Comparison: ML and MAP estimation

ML estimation:

- Consider parameter of interest as fix, but unknown
- Underlying interpretation of probability: frequentist
- Maximizes likelihood of data
- Difficult to use for small samples
- Inclusion of prior knowledge via model class

MAP estimation:

- Consider parameter of interest as random
- Underlying interpretation of probability: Bayesian
- Maximizes posterior probability/density, or equivalently: likelihood weighted by prior
- Suitable for small samples
- Inclusion of prior knowledge via prior probabilities

In large sample limit, $n \rightarrow \infty$: MAP estimate converges to ML estimate

Review: AB0 blood types and probability of trait in *D. melanogaster*

Calculating probability for blood type alleles of a father given information about phenotype of mother and/or child (see also series 7, exercise 2):

- Interested in probability affecting *one* specific individual
- Repeated measurements impossible
- Probability only interpretable in Bayesian way

Estimating probability of a eclosed *Drosophila melanogaster* to have vestigial wings (lecture slides, part VII):

- Interested in probability affecting *all* individuals of a population
- No knowledge about single fly
- Frequentist understanding of probability appropriate

Part IX

Hypothesis Tests for One Sample

Learning objectives

- Know the formal six-step approach of hypothesis testing
- Conduct a binomial and a t-test
- Conduct a sign test and a Wilcoxon signed-rank test
- Choose an appropriate test
- Know the differences between parametric and non-parametric tests
- Calculate a p-value and explain its meaning
- Explain the trade-off between type I and type II errors

Suggested literature

This lecture is partly based on the following source:

- Samuels et al. (2012), Chapters 8.1, 8.2, 8.4, 8.5

Hypothesis testing

- Aim: test whether a hypothesis, formulated as a probabilistic model, is compatible with data
- By the nature of probabilistic models, it is in general impossible to exclude *for sure* that a model is “true” (i.e., generated the data)
- We can assert that data is *likely* or *unlikely* to be generated by a model
- With statistical hypothesis tests, we can *reject* hypothesis that are *not plausible*

Example: ametropia I

- Claim: “University graduates are less likely to be ametropic than the rest of the population.”
- Define “ametropic” as “wearing glasses or contact lenses.” In the whole population, the fraction of ametropics is $\pi_0 = 63.1\%$ (data for Germany; Bra, 2011)
- Data collection: ask n randomly chosen university graduates whether they wear glasses or contact lenses (assume $n = 28$)
- Outcome (example): $x = 13$ persons out of $n = 28$ wearing glasses or contact lenses

Example: ametropia I

- Claim: “University graduates are less likely to be ametropic than the rest of the population.”
- Define “ametropic” as “wearing glasses or contact lenses.” In the whole population, the fraction of ametropics is $\pi_0 = 63.1\%$ (data for Germany; Bra, 2011)
- Data collection: ask n randomly chosen university graduates whether they wear glasses or contact lenses (assume $n = 28$)
- Outcome (example): $x = 13$ persons out of $n = 28$ wearing glasses or contact lenses

Intuitive question for hypothesis tests

How probable is it to find a sample that is *at least as extreme* as the actual one *by pure chance*?

Example: ametropia II

- What means “at least as extreme” and “by pure chance”?
- *Null hypothesis*: university graduates are on average as ametropic as the rest of the population.
Alternative hypothesis: university graduates are on average *less* ametropic than the rest of the population.

Example: ametropia II

- What means “at least as extreme” and “by pure chance”?
- *Null hypothesis*: university graduates are on average as ametropic as the rest of the population.
Alternative hypothesis: university graduates are on average *less* ametropic than the rest of the population.
- Question for testing our hypothesis: how probable is it *under the null hypothesis* to find 13 or less ametropic persons in a sample of size 28?

k	10	11	12	13	14	15
$P[X \leq k]$	0.003	0.009	0.023	0.053	0.109	0.197

Basic principle behind statistical hypothesis testing

- Formulate claim you want to prove: “surprising” finding, extending current knowledge, new model
- Formulate **null hypothesis**: simpler model based on current knowledge which you hope does *not* explain your data
- Show that it is very unlikely to find your measured result or a more extreme one
- General line of action consists of **6 steps**
- As an example, consider binomial test

Example: binomial test

- ① **Model:** X : number of questioned university graduates being ametropic; $X \sim \text{Bin}(n, \pi)$
- ② **Null hypothesis:** $H_0 : \pi = \pi_0 = 0.631$
Alternative hypothesis: $H_A : \pi < \pi_0$
- ③ **Test statistic:** $X = 13$
- ④ **Choose significance level:** e.g. $\alpha = 5\%$

Example: binomial test

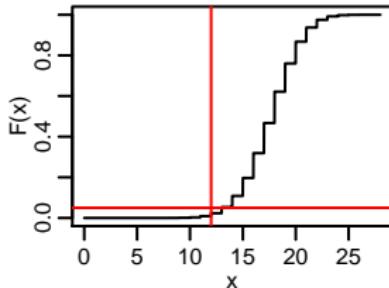
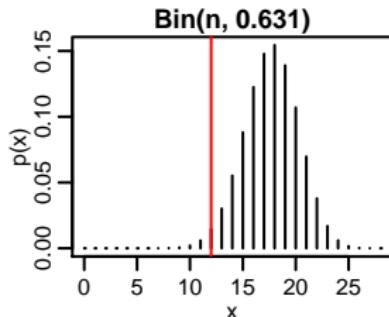
- ① **Range of rejection K :** range of values of the test statistic such that

$$P[X \in K] \lesssim \alpha \text{ under } H_0$$

Here: $K = [0, c]$ with c such that

$P_{H_0}[X \leq c] \lesssim \alpha$; from table before, we know that the largest number c with

$$P_{H_0}[X \leq c] \leq \alpha \text{ is } c = 12$$



Example: binomial test

- ① **Range of rejection K :** range of values of the test statistic such that

$$P[X \in K] \lesssim \alpha \text{ under } H_0$$

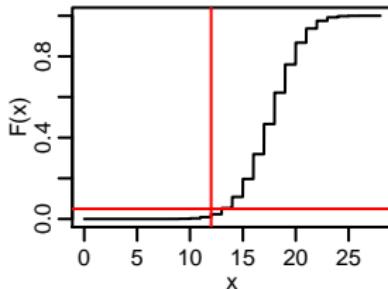
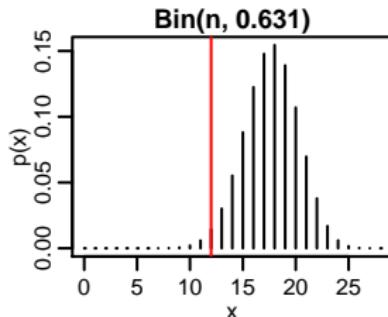
Here: $K = [0, c]$ with c such that

$P_{H_0}[X \leq c] \lesssim \alpha$; from table before, we know that the largest number c with

$$P_{H_0}[X \leq c] \leq \alpha \text{ is } c = 12$$

- ② **Test decision:** reject H_0 if $X \in K$, otherwise keep it.

Here: $X = 13$, $K = [0, 12]$; $X \notin K$, hence keep H_0



A few conceptual questions about hypothesis testing

- ① Do we prove the alternative hypothesis by rejecting the null hypothesis?
- ② Do we prove the null hypothesis by not rejecting it?
- ③ Is the rejection of the null hypothesis really indicative for the alternative hypothesis?
- ④ Can our sampling procedure influence our result?

Type I and type II errors

		Decision	
		H_0	H_A
Truth	H_0	true negative	type I error
	H_A	type II error	true positive

- **Significance level α :** probability of type I error *given that H_0 is true*
- **Power $1 - \beta$:** β is probability of type II error *given that H_1 is true*
- By reducing the probability of type I errors, we increase the probability of type II errors. “Higher significance implies lower power.”

Example: power of binomial test

Back to the ametropia example:

- What's the power of the binomial test if the true parameter is $\pi = 0.5$ (fraction of ametropia among university graduates)?

$$P_{\pi=0.5}[X \leq c = 12] = 0.286$$

Example: power of binomial test

Back to the ametropia example:

- What's the power of the binomial test if the true parameter is $\pi = 0.5$ (fraction of ametropia among university graduates)?

$$P_{\pi=0.5}[X \leq c = 12] = 0.286$$

- What's the power of the binomial test if the true parameter is $\pi = 0.3$ (fraction of ametropia among university graduates)?

$$P_{\pi=0.3}[X \leq c = 12] = 0.951$$

Example: power of binomial test

Back to the ametropia example:

- What's the power of the binomial test if the true parameter is $\pi = 0.5$ (fraction of ametropia among university graduates)?

$$P_{\pi=0.5}[X \leq c = 12] = 0.286$$

- What's the power of the binomial test if the true parameter is $\pi = 0.3$ (fraction of ametropia among university graduates)?

$$P_{\pi=0.3}[X \leq c = 12] = 0.951$$

- How can we increase the power of the test?

P-value

Definition (p-value)

The **p-value** is the smallest significance level α for which we reject a null hypothesis for the given data set.

P-value

Definition (p-value)

The **p-value** is the smallest significance level α for which we reject a null hypothesis for the given data set.

Definition (p-value: alternative definition)

The p-value is the probability under the null hypothesis to find the actual outcome or a more extreme one.

P-value

Definition (p-value)

The **p-value** is the smallest significance level α for which we reject a null hypothesis for the given data set.

Definition (p-value: alternative definition)

The p-value is the probability under the null hypothesis to find the actual outcome or a more extreme one.

In ametropia example: $p = P_{H_0}[X \leq x] = 0.053$.

Test using the normal approximation

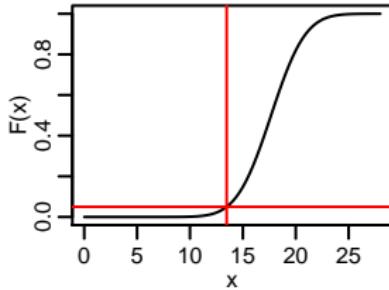
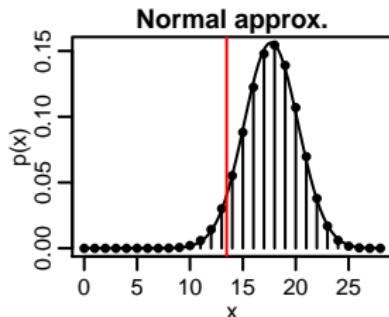
- Remember (slides part VI): if $n\pi_0 > 5$ and $n(1 - \pi_0) > 5$, we can approximate the CDF of a binomially distributed variable by a normal one:

$$X \approx \mathcal{N}(n\pi_0, n\pi_0(1 - \pi_0))$$

- For large samples, we can replace binomial test by a normal test:
- Model:** X : number of questioned university graduates being ametropic; $X \approx \mathcal{N}(n\pi_0, n\pi_0(1 - \pi_0))$
 - Null hypothesis:** $H_0 : \pi = \pi_0 = 0.631$
Alternative hypothesis: $H_A : \pi < \pi_0$
 - Test statistic:** $Z = \frac{X - n\pi_0}{\sqrt{n\pi_0(1 - \pi_0)}} = -1.828$
Distribution of Z under H_0 : $Z \approx \mathcal{N}(0, 1)$
 - Choose significance level:** e.g. $\alpha = 5\%$

Z-test: test using the normal approximation

- ⑤ Range of rejection $K = (-\infty, c]$ with c such that $P_{H_0}[Z \leq c] = \alpha$:
 $c = \Phi^{-1}(\alpha) = -\Phi^{-1}(1 - \alpha) = -1.645$



Z-test: test using the normal approximation

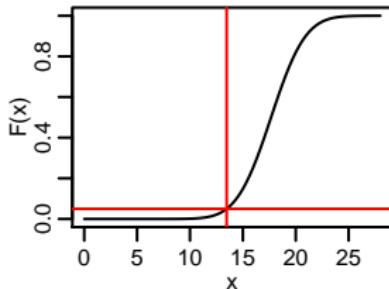
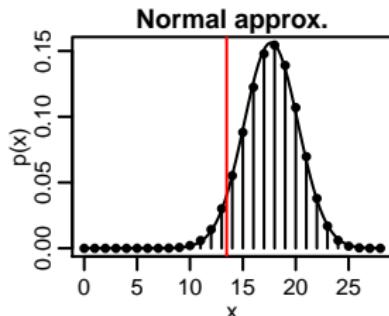
- ⑤ **Range of rejection** $K = (-\infty, c]$ with c such that $P_{H_0}[Z \leq c] = \alpha$:
 $c = \Phi^{-1}(\alpha) = -\Phi^{-1}(1 - \alpha) = -1.645$

- ⑥ **Test decision:** reject H_0 if $Z \in K$, otherwise keep it.

Here: $Z = -1.828$, $K = (-\infty, -1.645]$;
 $Z \in K$, hence reject H_0

p-values:

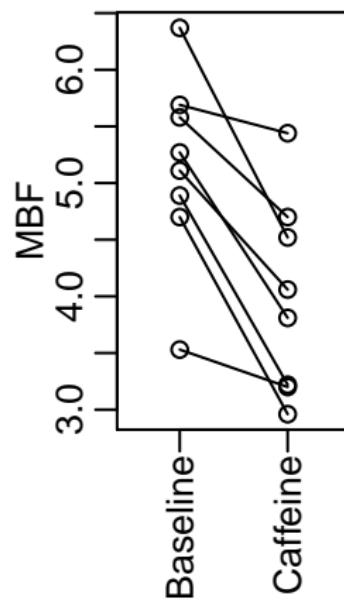
- for binomial distribution: $p = 0.053$
- for normal approximation: $p = 0.034$



Example: effect of caffeine on blood flow

- Study: does drinking coffee affect blood flow during exercise?
- Doctors measured myocardial blood flow (MBF) of 8 subjects during bicycle exercise before (Y_i) and after caffeine consumption (Z_i) ($i = 1, \dots, 8$)
- Question: is there a systematic difference of the blood flow before and after caffeine consumption?

(Source: Namdar et al. (2006))



Paired-samples (or one-sample) t test

Consider differences $X_i = Z_i - Y_i$, $i = 1, 2, \dots, n = 8$.

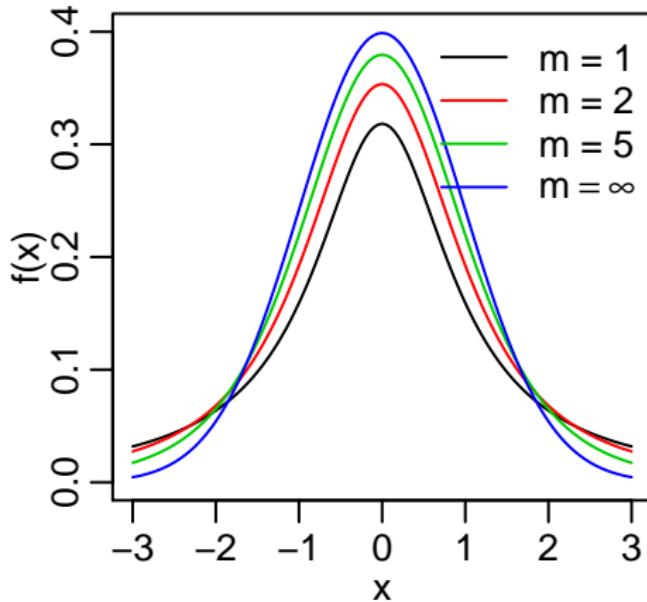
- ① **Model:** $X_1, \dots, X_n \stackrel{\text{i.i.d}}{\sim} \mathcal{N}(\mu, \sigma^2)$, with unknown σ^2
- ② **Null hypothesis:** $H_0 : \mu = \mu_0 = 0$
Alternative hypothesis: $H_A : \mu \neq \mu_0$
- ③ **Test statistic:** $T = \frac{\sqrt{n}(\bar{X} - \mu_0)}{s_x} = \frac{\text{obs. mean} - \text{exp. mean}}{\text{standard error}}$

Distribution of T under H_0 : Student's t distribution with $n - 1$ degrees of freedom

- ④ Choose **significance level**: e.g. $\alpha = 5\%$

Student's t distribution

- Notation: $T \sim t_m$
- Distribution characterized by "degrees of freedom", m
- Resembles the standard normal distribution more and more for growing m
- α quantile denoted by $t_{m,\alpha}$
- Because of symmetry:
 $t_{m,\alpha} = -t_{m,1-\alpha}$
- R function to calculate quantiles: `qt`



Paired-samples (or one-sample) t test: continued

⑤ Range of rejection

$$\begin{aligned}K &= \left(-\infty, -t_{n-1,1-\frac{\alpha}{2}}\right] \cup \left[t_{n-1,1-\frac{\alpha}{2}}, \infty\right) \\&= (-\infty, -2.365] \cup [2.365, \infty)\end{aligned}$$

Paired-samples (or one-sample) t test: continued

⑤ Range of rejection

$$\begin{aligned}K &= \left(-\infty, -t_{n-1,1-\frac{\alpha}{2}}\right] \cup \left[t_{n-1,1-\frac{\alpha}{2}}, \infty\right) \\&= \left(-\infty, -2.365\right] \cup \left[2.365, \infty\right)\end{aligned}$$

- ⑥ **Test decision:** reject H_0 if $T \in K$, otherwise keep it.
Here: $T = -5.188 \in K$, hence reject H_0

Paired-samples (or one-sample) t test: continued

⑤ Range of rejection

$$\begin{aligned}K &= \left(-\infty, -t_{n-1,1-\frac{\alpha}{2}}\right] \cup \left[t_{n-1,1-\frac{\alpha}{2}}, \infty\right) \\&= \left(-\infty, -2.365\right] \cup \left[2.365, \infty\right)\end{aligned}$$

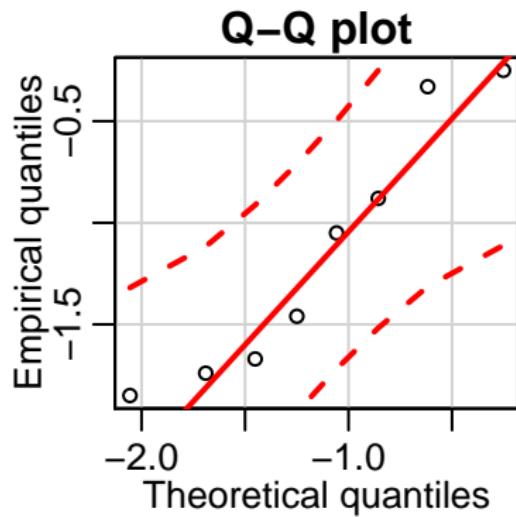
⑥ **Test decision:** reject H_0 if $T \in K$, otherwise keep it.

Here: $T = -5.188 \in K$, hence reject H_0

p-value: $p = 0.00127$

Testing assumptions of t test

Remember: model $X_1, \dots, X_n \stackrel{\text{i.i.d}}{\sim} \mathcal{N}(\mu, \sigma^2)$, with unknown σ^2



Confidence interval

Definition (confidence interval)

The **confidence interval** / for the parameter μ with confidence level $1 - \alpha$ is the set of all parameter values which are compatible with the data set in the sense of hypothesis testing.

Formally:

$$I = \{\mu_0 \mid \text{null hypothesis } H_0 : \mu = \mu_0 \text{ is not rejected}\}$$

Here: formally for t test; defined in analogue way for any other test and parameter.

Confidence interval for μ

Form of confidence interval depends on alternative hypothesis:

- $H_A : \mu \neq \mu_0 \Rightarrow I = \left[\bar{x} - t_{n-1,1-\alpha/2} \frac{s_x}{\sqrt{n}}, \bar{x} + t_{n-1,1-\alpha/2} \frac{s_x}{\sqrt{n}} \right] = [-1.680, -0.628]$
- $H_A : \mu < \mu_0 \Rightarrow I = \left(-\infty, \bar{x} + t_{n-1,1-\alpha} \frac{s_x}{\sqrt{n}} \right] = (-\infty, -0.732]$
- $H_A : \mu > \mu_0 \Rightarrow I = \left[\bar{x} - t_{n-1,1-\alpha} \frac{s_x}{\sqrt{n}}, \infty \right) = [-1.575, \infty)$

R function t.test |

```
> bloodflow <- read.table("data/bloodFlow.csv", header = TRUE, sep =  
",")  
> t.test(bloodflow$Caffeine, bloodflow$Baseline, paired = TRUE, alternative  
= "two.sided", conf.level = 0.95)
```

Paired t-test

data: bloodflow\$Caffeine and bloodflow\$Baseline

Baseline t = -5.1878, df = 7, p-value = 0.00127

alternative hypothesis: true difference in means is not equal to 0

95 percent confidence interval:

-1.6796357 -0.6278643

sample estimates:

mean of the differences -1.15375

R function t.test II

```
> t.test(bloodflow$Caffeine, bloodflow$Baseline, paired = TRUE,  
alternative = "less", conf.level = 0.95)
```

Paired t-test

data: bloodflow\$Caffeine and bloodflow\$Baseline

t = -5.1878, df = 7, p-value = 0.000635

alternative hypothesis: true difference in means is less than 0

95 percent confidence interval:

-Inf -0.7324011

sample estimates:

mean of the differences

-1.15375

Sign test

- bloodflow data set: consider differences $X_i = Z_i - Y_i$,
 $i = 1, 2, \dots, n = 8$; aim: check whether mean is significantly different from (or below) zero.
- Assumptions for t test: normality X_i 's
- What could be done if they were *not* met?

Sign test

- bloodflow data set: consider differences $X_i = Z_i - Y_i$,
 $i = 1, 2, \dots, n = 8$; aim: check whether mean is significantly different from (or below) zero.
- Assumptions for t test: normality X_i 's
- What could be done if they were *not* met?
- Alternative: **sign test**

Sign test

Consider differences $X_i = Z_i - Y_i$, $i = 1, 2, \dots, n = 8$.

- ① **Model:** X_1, \dots, X_n i.i.d. from *arbitrary distribution with median m*
- ② **Null hypothesis:** $H_0 : m = m_0 = 0$
Alternative hypothesis: $H_A : m \neq m_0$
- ③ **Test statistic:** $V = \#\{i | X_i > m_0\}$: number of values larger than m_0 .

Sign test

Consider differences $X_i = Z_i - Y_i$, $i = 1, 2, \dots, n = 8$.

- ① **Model:** X_1, \dots, X_n i.i.d. from *arbitrary distribution with median m*
- ② **Null hypothesis:** $H_0 : m = m_0 = 0$
Alternative hypothesis: $H_A : m \neq m_0$
- ③ **Test statistic:** $V = \#\{i | X_i > m_0\}$: number of values larger than m_0 .

Distribution of V under H_0 : $V \sim \text{Bin}(n, 0.5)$

- ④ Choose **significance level**: e.g. $\alpha = 5\%$

Sign test

- ⑤ **Range of rejection:** $K = [0, c] \cup [n - c, n]$ such that

$$P_{H_0}[V \in K] \lesssim \alpha.$$

c determined by binomial distribution: $P_{H_0}[V \in K] = 2P_{H_0}[V \leq c]$.

Values for $c = 0, 1, 2, 3$:

```
> 2*pbinom(0:3, n, 0.5)
```

```
[1] 0.0078125 0.0703125 0.2890625 0.7265625
```

Hence take $c = 0$ (small data set!)

Sign test

- ⑤ **Range of rejection:** $K = [0, c] \cup [n - c, n]$ such that
 $P_{H_0}[V \in K] \lesssim \alpha$.
 c determined by binomial distribution: $P_{H_0}[V \in K] = 2P_{H_0}[V \leq c]$.
Values for $c = 0, 1, 2, 3$:
`> 2*pbinom(0:3, n, 0.5)`
`[1] 0.0078125 0.0703125 0.2890625 0.7265625`
Hence take $c = 0$ (small data set!)
- ⑥ **Test decision:** reject H_0 if $V \in K$, otherwise keep it.
Here: $V = 0 \in K$, hence reject H_0

Sign test

- ⑤ **Range of rejection:** $K = [0, c] \cup [n - c, n]$ such that

$$P_{H_0}[V \in K] \lesssim \alpha.$$

c determined by binomial distribution: $P_{H_0}[V \in K] = 2P_{H_0}[V \leq c]$.

Values for $c = 0, 1, 2, 3$:

```
> 2*pbinom(0:3, n, 0.5)
```

```
[1] 0.0078125 0.0703125 0.2890625 0.7265625
```

Hence take $c = 0$ (small data set!)

- ⑥ **Test decision:** reject H_0 if $V \in K$, otherwise keep it.

Here: $V = 0 \in K$, hence reject H_0

p-value: $p = 0.00781$

Sign test in R: two-sided

Calculate V and use `binom.test`:

```
> V <- sum(bloodflow$Caffeine > bloodflow$Baseline)  
> binom.test(V, n, p = 0.5, alternative = "two.sided", conf.level = 0.95)
```

Exact binomial test

data: V and n

number of successes = 0, number of trials = 8, p-value = 0.007812
alternative hypothesis: true probability of success is not equal to 0.5

95 percent confidence interval:

0.0000000 0.3694166

sample estimates:

probability of success

0

Sign test in R: one-sided

```
> V <- sum(bloodflow$Caffeine > bloodflow$Baseline)
> binom.test(V, n, p = 0.5, alternative = "less", conf.level = 0.95)
Exact binomial test
data: V and n number of successes = 0, number of trials = 8, p-value =
0.003906
alternative hypothesis: true probability of success is less than 0.5
95 percent confidence interval:
0.000000 0.312344
sample estimates:
probability of success
0
```

Wilcoxon signed-rank test

Consider again the bloodflow data set with the differences $X_i = Z_i - Y_i$, $i = 1, 2, \dots, n = 8$.

- ① **Model:** X_1, \dots, X_n i.i.d. from *arbitrary* distribution median m ; distribution of differences X_i symmetric around m
- ② **Null hypothesis:** $H_0 : m = 0$
Alternative hypothesis: $H_A : m \neq 0$
- ③ **Test statistic:**
 - ▶ Order differences by absolute value $|X_i|$
 - ▶ Assign rank R_i to absolute differences; smallest one has rank 1, largest one rank n .
 - ▶ Test statistic: $W = \sum_{i=1}^n \text{sign}(X_i)R_i$

Distribution of W under H_0 : for $n \geq 10$, $W \approx \mathcal{N}(0.5, \sigma^2)$ with

$$\sigma^2 = \frac{n(n+1)(2n+1)}{6}$$

For small n , complicated distribution; use R function `wilcox.test`

Wilcoxon signed-rank test: bloodflow data

i	X_i	$ X_i $
1	-1.85	1.85
2	-0.25	0.25
3	-0.88	0.88
4	-1.46	1.46
5	-1.05	1.05
6	-1.67	1.67
7	-1.74	1.74
8	-0.33	0.33

Wilcoxon signed-rank test: bloodflow data

i	X_i	$ X_i $
1	-1.85	1.85
2	-0.25	0.25
3	-0.88	0.88
4	-1.46	1.46
		sort by $ X_i $:
5	-1.05	1.05
6	-1.67	1.67
7	-1.74	1.74
8	-0.33	0.33

Wilcoxon signed-rank test: bloodflow data

i	X_i	$ X_i $
1	-1.85	1.85
2	-0.25	0.25
3	-0.88	0.88
4	-1.46	1.46
5	-1.05	1.05
6	-1.67	1.67
7	-1.74	1.74
8	-0.33	0.33

sort by $|X_i|$:

i	X_i	$ X_i $	R_i	signed rank
2	-0.25	0.25	1	-1
8	-0.33	0.33	2	-2
3	-0.88	0.88	3	-3
5	-1.05	1.05	4	-4
4	-1.46	1.46	5	-5
6	-1.67	1.67	6	-6
7	-1.74	1.74	7	-7
1	-1.85	1.85	8	-8

Wilcoxon signed-rank test: bloodflow data

i	X_i	$ X_i $
1	-1.85	1.85
2	-0.25	0.25
3	-0.88	0.88
4	-1.46	1.46
5	-1.05	1.05
6	-1.67	1.67
7	-1.74	1.74
8	-0.33	0.33

sort by $|X_i|$:

i	X_i	$ X_i $	R_i	signed rank
2	-0.25	0.25	1	-1
8	-0.33	0.33	2	-2
3	-0.88	0.88	3	-3
5	-1.05	1.05	4	-4
4	-1.46	1.46	5	-5
6	-1.67	1.67	6	-6
7	-1.74	1.74	7	-7
1	-1.85	1.85	8	-8

Test statistic W : sum of last column. $W = -36$

Wilcoxon signed-rank test: continued

- ④ Choose **significance level**: e.g. $\alpha = 5\%$
- ⑤ **Range of rejection**: $K = (-\infty, 0.5 - c] \cup [0.5 + c, \infty)$ such that $P_{H_0}[W \in K] \lesssim \alpha$.

Exact value of c complicated to get for $n < 10$; use R instead.

- ⑥ **Test decision**: reject H_0 if $W \in K$, otherwise keep it.

Wilcoxon signed-rank test in R

```
> wilcox.test(bloodflow$Caffeine, bloodflow$Baseline, paired = TRUE,  
exact = TRUE, alternative = "two.sided", conf.level = 0.95)  
Wilcoxon signed rank test  
data: bloodflow$Caffeine and bloodflow$Baseline V = 0, p-value =  
0.007813  
alternative hypothesis: true location shift is not equal to 0
```

Wilcoxon signed-rank test in R: one-sided

```
> wilcox.test(bloodflow$Caffeine, bloodflow$Baseline, paired = TRUE,  
exact = TRUE, alternative = "less", conf.level = 0.95)  
Wilcoxon signed rank test  
data: bloodflow$Caffeine and bloodflow$Baseline  
V = 0, p-value = 0.003906  
alternative hypothesis: true location shift is less than 0
```

Overview: parametric and non-parametric tests for bloodflow data

Test	p-value, 2-sided	p-value, 1-sided
t test	0.00127	0.00063
sign test	0.00781	0.00391
Wilcoxon signed-rank test	0.00781	0.00391

Comparison: parametric and non-parametric one-sample tests

Parametric tests

- Assumptions about distribution family of data
- Examples: z test, t test
- Limited applicability
- More power when assumptions met

Nonparametric test

- No assumptions about distribution family of data
- Examples: sign test, Wilcoxon signed-rank test
- Wider applicability
- Less power when parametric assumptions would apply

Comparison: parametric and non-parametric one-sample tests

Parametric tests

- Assumptions about distribution family of data
- Examples: z test, t test
- Limited applicability
- More power when assumptions met

Nonparametric test

- No assumptions about distribution family of data
- Examples: sign test, Wilcoxon signed-rank test
- Wider applicability
- Less power when parametric assumptions would apply

Choice of test

Use **parametric** tests whenever this is appropriate; use **non-parametric** tests only when you need to.

Which problems arise when sample is small?

Part X

Hypothesis Testing for Two Samples

Learning objectives

- Choose between a one-sample and a two-sample test
- Conduct a two-sample t-test
- Explain and conduct a randomization test

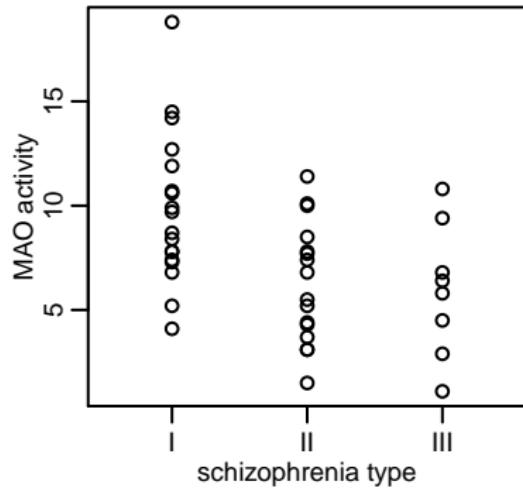
Suggested literature

This lecture is partly based on the following source:

- Samuels et al. (2012), Chapters 7.1, 7.2, 7.3, 7.6

Example: Monoamine oxidase and schizophrenia

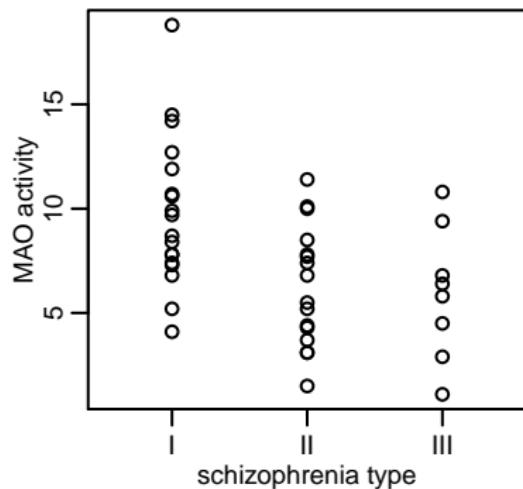
- Monoamine oxidase (MAO): enzyme thought to play a role in regulation of behavior
- Study: measured levels of MAO activity in 42 patients with different three types of schizophrenia



(Source: Potkin et al. (1978))

Example: Monoamine oxidase and schizophrenia

- Monoamine oxidase (MAO): enzyme thought to play a role in regulation of behavior
- Study: measured levels of MAO activity in 42 patients with different three types of schizophrenia
- Are different types of schizophrenia associated with different levels of MAO activity?

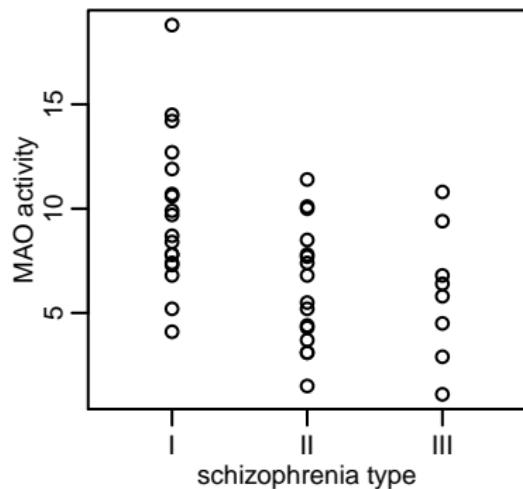


(Source: Potkin et al. (1978))

Example: Monoamine oxidase and schizophrenia

- Monoamine oxidase (MAO): enzyme thought to play a role in regulation of behavior
- Study: measured levels of MAO activity in 42 patients with different three types of schizophrenia
- Are different types of schizophrenia associated with different levels of MAO activity?
- What's the difference to the bloodflow data set in Part IX?

(Source: Potkin et al. (1978))



MAO data set: t-test

First possibility to test difference of MAO activity: 2-sample, unpaired, 2-sided **t-test**; $\{X_i\}_i, \{Y_i\}_i$: MAO level for type I and II patients, resp.

① Model:

$$X_1, \dots, X_n \stackrel{\text{i.i.d.}}{\sim} \mathcal{N}(\mu_X, \sigma^2)$$

$$Y_1, \dots, Y_m \stackrel{\text{i.i.d.}}{\sim} \mathcal{N}(\mu_Y, \sigma^2)$$

② Null hypothesis: $H_0: \mu_X = \mu_Y$

Alternative hypothesis: $H_A: \mu_X \neq \mu_Y$

③ Test statistic: $T = \frac{\bar{X} - \bar{Y}}{s_{\text{pool}} \sqrt{1/n + 1/m}} = 3.1151$, where

$$s_{\text{pool}}^2 = \frac{1}{n+m-2} \left((n-1)s_x^2 + (m-1)s_y^2 \right)$$

Distribution of T under H_0 : $T \sim t_{n+m-2}$

MAO data set: t-test

- ④ Choose **significance level**: e.g.
 $\alpha = 5\%$

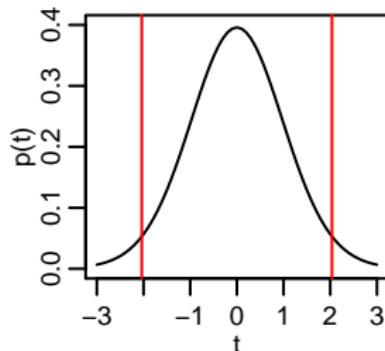
- ⑤ **Range of rejection**:

$$K = (-\infty, -t_{n+m-2, 1-\alpha/2}] \cup [t_{n+m-2, 1-\alpha/2}, \infty)$$

$t_{k,\alpha}$: α -quantile of t distribution with k degrees of freedom (df)

Here: $df = n + m - 2 = 32$;

$$t_{n+m-2, 1-\alpha/2} = t_{32, 0.975} = 2.0369$$



Calculate quantiles in R: >
`qt(0.975, n+m-2)`
[1] 2.036933

Note: borders of range of rejection

- One-sided test: + or - $(1 - \alpha)$ -quantile of distribution
- Two-sided test: + and - $(1 - \alpha/2)$ -quantile of distribution

MAO data set: t-test

⑥ **Test decision:** reject H_0 if $T \in K$, otherwise keep it.

Here: $T = 3.1151$, $K = (-\infty, -2.0369] \cup [2.0369, \infty)$; $X \in K$, hence
reject H_0

```
> 2*(1 - pt(T, n + m - 2))
```

```
[1] 0.003863469
```

MAO data set: t-test

⑥ **Test decision:** reject H_0 if $T \in K$, otherwise keep it.

Here: $T = 3.1151$, $K = (-\infty, -2.0369] \cup [2.0369, \infty)$; $X \in K$, hence
reject H_0

p-value: lowest significance level α for which H_0 is rejected

Here: $p = 2 * (1 - F(T))$, where F is the CDF of the t distribution with
 $n + m - 2$ degrees of freedom.

```
> 2*(1 - pt(T, n + m - 2))
```

```
[1] 0.003863469
```

Quick way of doing a t-test: R

```
> t.test(x, y, alternative = "two.sided", paired = FALSE, conf.level = 0.95)
```

Welch Two Sample t-test

data: x and y

t = 3.1578, df = 31.647, p-value = 0.003483

alternative hypothesis: true difference in means is not equal to 0

95 percent confidence interval:

1.249945 5.798666

sample estimates:

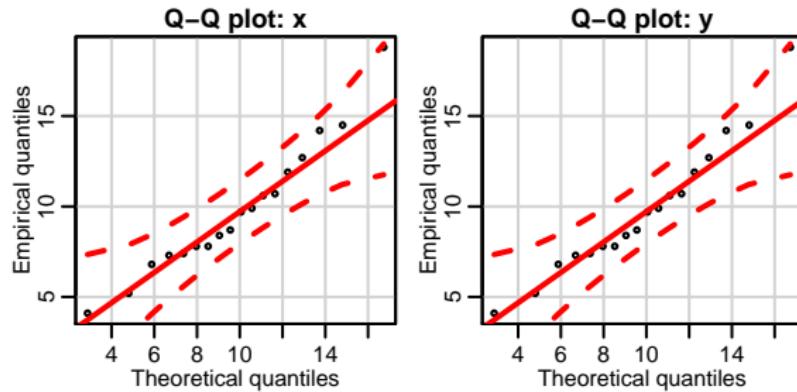
mean of x mean of y

9.805556 6.281250

t-test: review

Commonly used, standard test. Nevertheless: never forget about assumptions!

Checking normality assumptions for X and Y :



MAO data: permutation test

Permutation test: nonparametric test (i.e., test that does not assume the data to come from a parametric distribution family)

① Model:

$$X_1, \dots, X_n \stackrel{\text{i.i.d.}}{\sim} F_X(\cdot)$$

$$Y_1, \dots, Y_m \stackrel{\text{i.i.d.}}{\sim} F_Y(\cdot)$$

② Null hypothesis: $H_0 : F_X = F_Y$

Alternative hypothesis: $H_A : F_X \neq F_Y$

MAO data: permutation test

③ **Test statistic:** $D = \bar{X} - \bar{Y} = 3.5243$

Distribution of D under H_0 : approximated by resampling, not known in advance (*nonparametric* test)!

Resampling:

- Choose number of repetitions N
- Randomly assign n values of $\{X_i\} \cup \{Y_i\}$ to “type I”, and the remaining m values to “type II”; calculate D as if random assignment was true.
- Repeat previous step N times

MAO data: permutation test

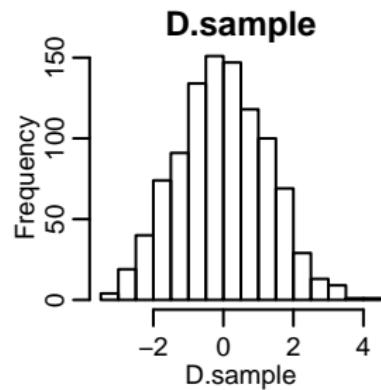
Resampling in R:

```
> set.seed(12)
> xy.all <- c(x, y)
> mean.diff <- function(group1)
mean(xy.all[group1]) - mean(xy.all[-group1])
> (D <- mean.diff(1:n))
[1] 3.524306
> N <- 1000
> D.sample <-
replicate(N,mean.diff(sample.int(n+m, size
= n)))
```

MAO data: permutation test

Resampling in R:

```
> set.seed(12)
> xy.all <- c(x, y)
> mean.diff <- function(group1)
mean(xy.all[group1]) - mean(xy.all[-group1])
> (D <- mean.diff(1:n))
[1] 3.524306
> N <- 1000
> D.sample <-
replicate(N,mean.diff(sample.int(n+m, size
= n)))
```



MAO data: permutation test

- ④ Choose significance level: e.g.

$$\alpha = 5\%$$

- ⑤ Range of rejection:

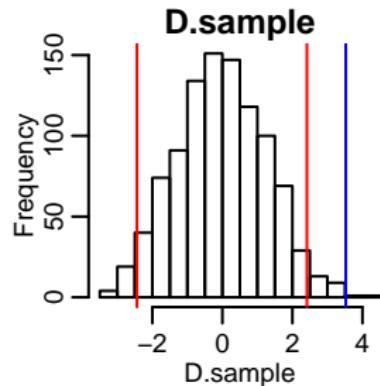
$$K = (-\infty, c_l] \cup [c_u, \infty),$$

c_l : empirical $\alpha/2$ -quantile of resampling distribution,

c_u : empirical $1 - \alpha/2$ -quantile of resampling distribution

Here: $c_l = -2.4381$, $c_u = 2.4155$;

actual mean difference: $D = 3.5243$



Calculating borders of rejection region in R:

```
> quantile(D.sample, c(0.025, 0.975))
```

2.5% 97.5%

-2.438090 2.415469

MAO data: permutation test

- ⑥ **Test decision:** reject H_0 if $D \in K$, otherwise keep it.

Here: $D = 3.5243$, $K = (-\infty, -2.4381] \cup [2.4155, \infty)$; $D \in K$, hence reject H_0

p-value: fraction of permutations for which the *absolute* difference of the means is larger than $|D|$

```
> sum(abs(D.sample) >= abs(D))/N
```

```
[1] 0.002
```

Slightly better version also considers true value as belonging to random sample (less prone to underestimate p-value):

```
> (sum(abs(D.sample) >= abs(D)) + 1)/(N + 1)
```

```
[1] 0.002997003
```

Permutation test in R

Permutation test is implemented in R package `perm`, in function `permTS`:

```
> library(perm)
```

```
> permTS(x, y, alternative = "two.sided", paired = FALSE, conf.level = 0.95, method = "exact.mc")
```

Exact Permutation Test Estimated by Monte Carlo

data: x and y

p-value = 0.008

alternative hypothesis: true mean x - mean y is not equal to 0

sample estimates:

mean x - mean y

3.524306

p-value estimated from 999 Monte Carlo replications

99 percent confidence interval on p-value:

0.0006769666 0.0218893569

Note: different p-values due to randomness; `permTS` indicates confidence interval on p-value for same reason

Permutation test: review

- No parametric assumption, also works for non-normal samples. **Pro:** wider applicability
- Accuracy of p -value increases with N . **Contra:** computational costs
- Caveat: many different ways of resampling; results may depend on method you use

Alternative to permutation test: try to transform data to make normality assumption hold. In biology, a lot of log-normal data \rightsquigarrow log-transformation gives normally distributed data

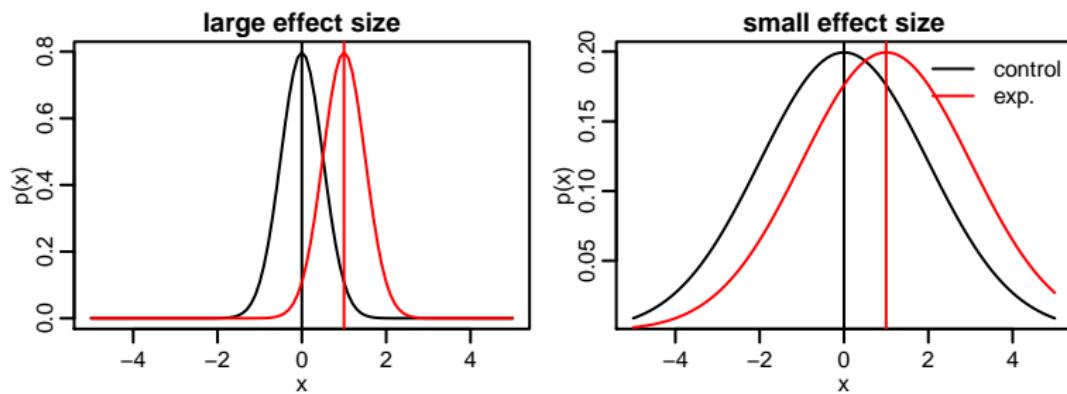
t-test: tweaking for significance?

- Recall null hypothesis of t-test, $\mu = \mu_0$: statement about *mean*, not about *full distribution*
- Null hypothesis is never *exactly* true. With enough data, we can always reject null hypothesis, get arbitrarily high p-values
- Some journals even forbid publication of p-values...

Effect size

Two samples: experimental group $\{X_i\}_i$, control group $\{Y_i\}_i$

$$\text{effect size} = \frac{\bar{X} - \bar{Y}}{s_{\text{pool}}}$$



MAO data: effect size

In the MAO data set, we have

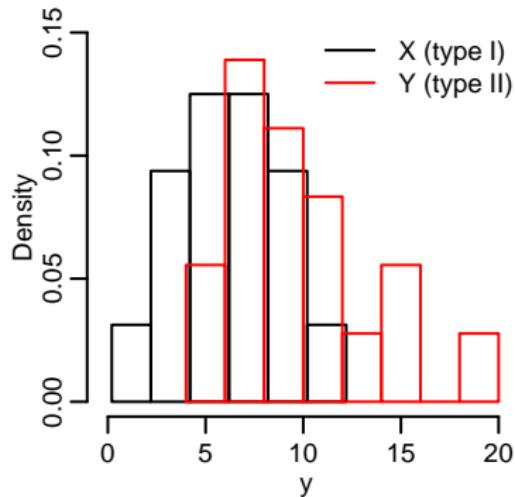
$$\bar{X} = 9.806$$

$$\bar{Y} = 6.281$$

$$s_{\text{pool}} = 3.293,$$

hence an effect size of

$$\frac{9.806 - 6.281}{3.293} = 1.07$$



Part XI

Multiple Testing

Learning objectives

- Know the difference between false positive rate (FPR), false discovery rate (FDR) and family-wise error rate (FWER)
- Correct p-values (Bonferroni, Holm, and Benjamini-Hochberg correction) in order to account for multiple hypothesis tests
- Choose an appropriate correction of p-values in a given experiment

Example: finding differentially expressed genes

- Hereditary breast cancer: often due to mutations in gene BRCA1 or BRCA2
- Histopathological traits in cancer tissue often characteristic for mutation in one or the other gene
- Question: can we also find genes (beside BRCA1 and BRCA2) that are **differentially expressed** in the two cancer types, i.e. whose expression levels in both cancer types are significantly different?

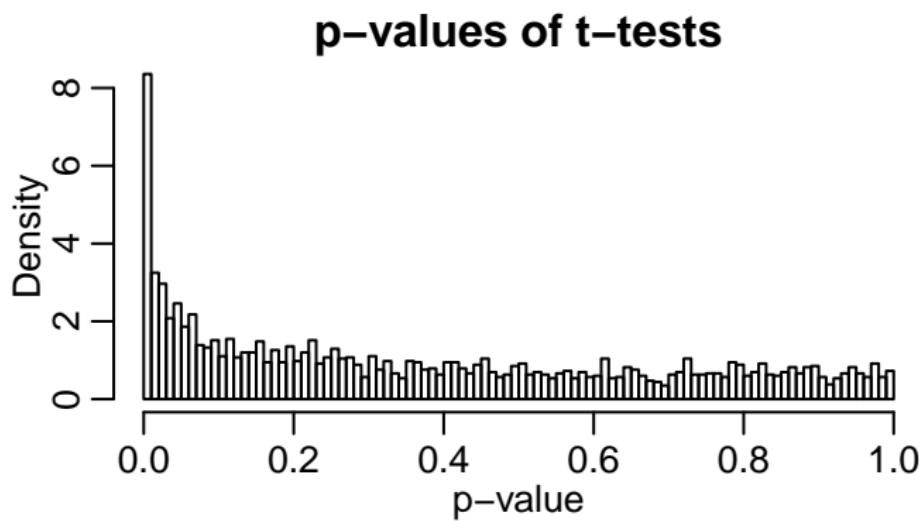
Example: finding differentially expressed genes

- Hereditary breast cancer: often due to mutations in gene BRCA1 or BRCA2
- Histopathological traits in cancer tissue often characteristic for mutation in one or the other gene
- Question: can we also find genes (beside BRCA1 and BRCA2) that are **differentially expressed** in the two cancer types, i.e. whose expression levels in both cancer types are significantly different?

Study of Hedenfalk et al. (2001):

- Measured RNA levels of 3170 genes in 7 samples of cancer tissue of BRCA1 mutation carriers and in 7 samples of cancer tissue of BRCA2 mutation carriers
- How would you investigate the question: **which genes are differentially expressed in the two cancer types?**

p-values in the study of Hedenfalk et al. (2001)



Which distribution would one expect if the null hypothesis (no differential expression) was true for all genes?

Distribution of p-values under null hypothesis

- If the null hypothesis is true, the p-value of a hypothesis test is **uniformly distributed** on the interval $[0, 1]$: $p \sim \mathcal{U}([0, 1])$
- This is true for every statistical hypothesis test (see derivation)
- Performing many tests simultaneously: interesting p-values = small p-values exceeding uniform distribution

Distribution of p-values under null hypothesis

- If the null hypothesis is true, the p-value of a hypothesis test is **uniformly distributed** on the interval $[0, 1]$: $p \sim \mathcal{U}([0, 1])$
- This is true for every statistical hypothesis test (see derivation)
- Performing many tests simultaneously: interesting p-values = small p-values exceeding uniform distribution
- Study of Hedenfalk et al. (2001): 265 p-values < 0.01 ; how many would you expect if H_0 was true for all 3170 genes?

Biological implication of Hedenfalk et al. (2001)

- Lab work: further investigations with differentially expressed genes
- Analyzing biological effect of a gene: great effort
- Therefore: keep number of false positives low

Biological implication of Hedenfalk et al. (2001)

- Lab work: further investigations with differentially expressed genes
- Analyzing biological effect of a gene: great effort
- Therefore: keep number of false positives low

Numbers of true/false positives/negatives:

		Decision		Total
		H_0	H_A	
Truth	H_0	true negatives: U	type I errors: V	m_0
	H_A	type II errors: T	true positives: S	$m - m_0$
Total		$m - R$	R	m

Which quantities are random, which are fix? Which are known, which are unknown?

False positive rate (FPR)

False positive rate: rate of false positives *among cases with true null hypothesis* (= genes without differential expression):

Definition (False positive rate)

$$\text{False positive rate: } \text{FPR} = E \left[\frac{\text{FP}}{\text{FP} + \text{TN}} \right] = E \left[\frac{V}{m_0} \right]$$

False positive rate (FPR)

False positive rate: rate of false positives *among cases with true null hypothesis* (= genes without differential expression):

Definition (False positive rate)

$$\text{False positive rate: } \text{FPR} = E \left[\frac{\text{FP}}{\text{FP} + \text{TN}} \right] = E \left[\frac{V}{m_0} \right]$$

FPR controlled by **significance level** α :

- $\alpha = E \left[\frac{V}{m_0} \right] = \text{FPR}$
- α also called **comparison-wise type I error rate**
- Test procedure that guarantees a FPR of (at most) α :
 - ▶ for each test case (e.g. gene), calculate p-value
 - ▶ reject null hypotheses whose p-value is smaller than α ; accept others

Controlling FPR for data of Hedenfalk et al. (2001)

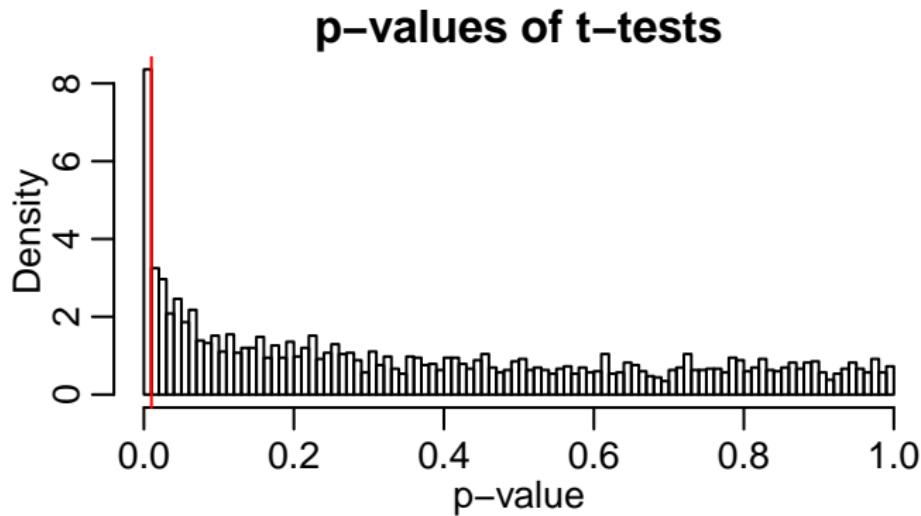
Raw data:

Gene	Expression level in BRCA1	Expression level in BRCA2	p-value of t-test
catenin	(7 samples)	(7 samples)	0.0121
phosphofructokinase, platelet	(7 samples)	(7 samples)	0.9949
ADP-ribosylation factor 3	(7 samples)	(7 samples)	0.0418
uroporphyrinogen III synthase	(7 samples)	(7 samples)	0.8458
ribosomal protein L26	(7 samples)	(7 samples)	0.2519
G protein	(7 samples)	(7 samples)	0.6587
:	:	:	:

Controlling FPR at $\alpha = 1\%$

- rejecting all null hypothesis with p-values smaller than 0.01, i.e.
- declaring all genes with a p-value smaller than 0.01 as differentially expressed

Controlling FPR for data of Hedenfalk et al. (2001)



- Reject null hypotheses to the left of the red threshold line
- Problem: rejecting many null hypotheses by chance

Family-wise error rate (FWER)

Family-wise error rate: probability of having one or more false positives in the whole study:

Definition (Family-wise error rate)

Family-wise error rate: $\text{FWER} = P[\text{1 or more type I errors}] = P[V \geq 1]$

Family-wise error rate (FWER)

Family-wise error rate: probability of having one or more false positives in the whole study:

Definition (Family-wise error rate)

Family-wise error rate: $\text{FWER} = P[\text{1 or more type I errors}] = P[V \geq 1]$

- FWER controlled by **experiment-wise type I error rate** $\bar{\alpha}$
- Test procedure that guarantees a FWER of (at most) $\bar{\alpha}$:
 - ▶ for each test case (e.g. gene), calculate p-value
 - ▶ **adjust p-value**
 - ▶ reject null hypotheses whose **adjusted** p-value is smaller than $\bar{\alpha}$; accept others

Controlling FWER: Bonferroni method

Controlling FWER at level of $\bar{\alpha} = 10\%$ over all $m = 3170$ genes:

- for each test case (gene), calculate p-value; this yields m p-values P_1, P_2, \dots, P_m
(as before, use any appropriate test; for the data of Hedenfalk et al. (2001) e.g. a t-test)
- **adjust** p-value: $P_{\text{adj},i} = \min\{m \cdot P_i, 1\} = \min\{3170P_i, 1\}$
- reject null hypotheses whose **adjusted** p-value is smaller than $\bar{\alpha}$; accept others

Procedure guarantees FWER $\leq \bar{\alpha}$ (see blackboard)

Method published by Bonferroni (1936); Dunn (1961)

Bonferroni method for data of Hedenfalk et al. (2001)

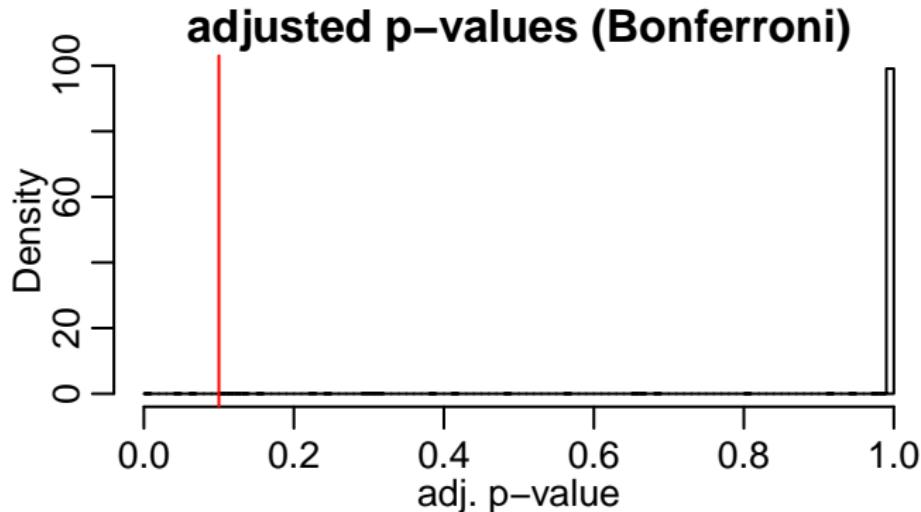
Raw data:

Gene	Expression level		p-value	
	in BRCA1	in BRCA2	raw	adjusted
catenin	(7 samples)	(7 samples)	0.0121	1
phosphofructokinase, platelet	(7 samples)	(7 samples)	0.9949	1
ADP-ribosylation factor 3	(7 samples)	(7 samples)	0.0418	1
uroporphyrinogen III synthase	(7 samples)	(7 samples)	0.8458	1
ribosomal protein L26	(7 samples)	(7 samples)	0.2519	1
G protein	(7 samples)	(7 samples)	0.6587	1
:	:	:	:	:

Controlling FWER at $\bar{\alpha} = 10\%$

- rejecting all null hypothesis with **adjusted** p-values smaller than 0.1, i.e.
- declaring all genes with an **adjusted** p-value smaller than 0.1 as differentially expressed

Bonferroni method for data of Hedenfalk et al. (2001)



- Reject null hypotheses to the left of the red threshold line
- Smallest adjusted p-values: 0.010, 0.050, 0.070, 0.110, 0.120
- Problem: very conservative ("restrictive") procedure, low power

Controlling FWER: Holm method

Controlling FWER at level of $\bar{\alpha} = 10\%$ over all $m = 3170$ genes:

- for each test case (gene), calculate p-value; this yields m p-values P_1, P_2, \dots, P_m
- order p-values: $P_{(1)} \leq P_{(2)} \leq P_{(3)} \leq \dots \leq P_{(m)}$
- **adjust** p-values:
 - ▶ $P_{\text{adj},(1)} = \min\{m \cdot P_{(1)}, 1\}$
 - ▶ $P_{\text{adj},(2)} = \min\{(m - 1) \cdot P_{(2)}, 1\}$; if value below $P_{\text{adj},(1)}$, replace it by $P_{\text{adj},(1)}$
 - ▶ $P_{\text{adj},(3)} = \min\{(m - 2) \cdot P_{(3)}, 1\}$; if value below $P_{\text{adj},(2)}$, replace it by $P_{\text{adj},(2)}$
 - ▶ $P_{\text{adj},(i)} = \min\{(m - i + 1) \cdot P_{(i)}, 1\}$; if value below $P_{\text{adj},(i-1)}$, replace it by $P_{\text{adj},(i-1)}$
- reject null hypotheses whose **adjusted** p-value is smaller than $\bar{\alpha}$;
accept others

Procedure guarantees FWER $\leq \bar{\alpha}$

Method published by Holm (1979)

Holm method for data of Hedenfalk et al. (2001)

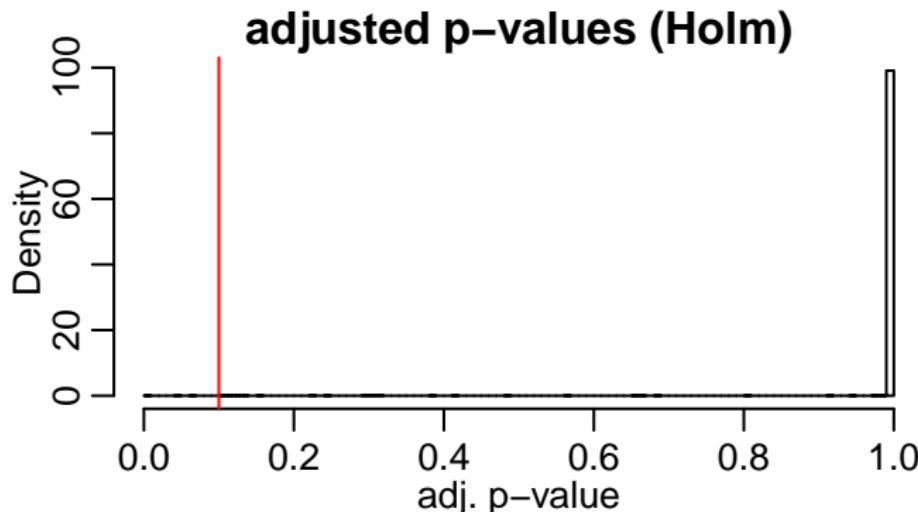
Raw data:

Gene	Expression level		p-value	
	in BRCA1	in BRCA2	raw	adjusted
catenin	(7 samples)	(7 samples)	0.0121	1
phosphofructokinase, platelet	(7 samples)	(7 samples)	0.9949	1
ADP-ribosylation factor 3	(7 samples)	(7 samples)	0.0418	1
uroporphyrinogen III synthase	(7 samples)	(7 samples)	0.8458	1
ribosomal protein L26	(7 samples)	(7 samples)	0.2519	1
G protein	(7 samples)	(7 samples)	0.6587	1
:	:	:	:	:

Controlling FWER at $\bar{\alpha} = 10\%$

- rejecting all null hypothesis with **adjusted** p-values smaller than 0.1, i.e.
- declaring all genes with an **adjusted** p-value smaller than 0.1 as differentially expressed

Holm method for data of Hedenfalk et al. (2001)



- Reject null hypotheses to the left of the red threshold line
- Smallest adjusted p-values: 0.010, 0.050, 0.070, 0.110, 0.120
- Always (slightly) higher power than Bonferroni method

Interpretation of adjusted p-values

Recall from Part IX:

Definition (p-value)

The **p-value** is the smallest significance level α for which we reject a null hypothesis for the given data set.

First definition can be generalized for **adjusted** p-values:

Definition (Adjusted p-value)

The **adjusted p-value** of a certain null hypothesis is the smallest experiment-wise type I error rate $\bar{\alpha}$ for which we reject this hypothesis for the given data set.

Interpretation of adjusted p-values

Recall from Part IX:

Definition (p-value)

The **p-value** is the smallest significance level α for which we reject a null hypothesis for the given data set.

Definition (p-value: alternative definition)

The p-value is the probability under the null hypothesis to find the actual outcome or a more extreme one.

First definition can be generalized for **adjusted** p-values:

Definition (Adjusted p-value)

The **adjusted p-value** of a certain null hypothesis is the smallest experiment-wise type I error rate $\bar{\alpha}$ for which we reject this hypothesis for the given data set.

Adjusted p-values

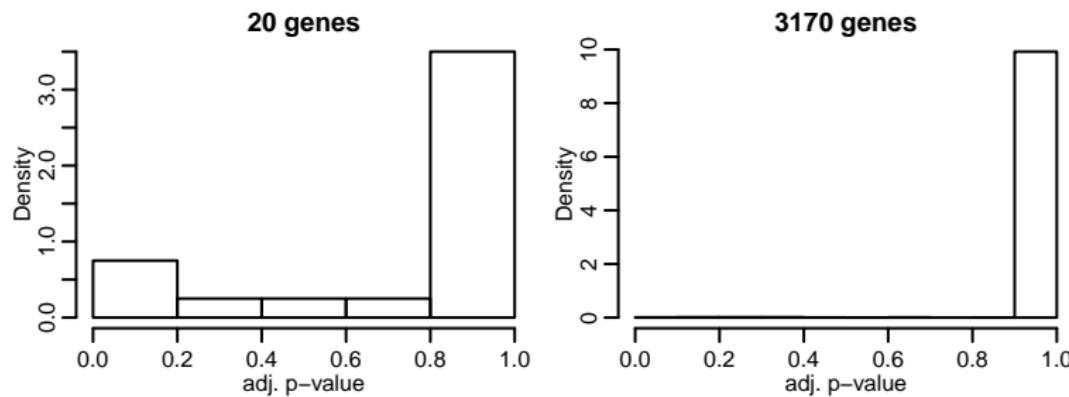
Definition (Adjusted p-value)

The **adjusted p-value** of a certain null hypothesis is the smallest experiment-wise type I error rate $\bar{\alpha}$ for which we reject this hypothesis for the given data set.

- We assign an individual adjusted p-value to a **single** hypothesis
- This adjusted p-value depends on data for **all** test cases (e.g. genes) in the experiment:
 - ▶ the **raw** p-value is calculated from data that belongs to **one** test case
 - ▶ the raw p-value is corrected to get an **adjusted** p-value with other p-values or data that belongs to **all** test cases
- Interpretation of an adjusted p-value in the study of Hedenfalk et al. (2001): which is the smallest experiment-wise type I rate $\bar{\alpha}$ for which I declare *this* gene differentially expressed?

Controlling FWER

Adjusted p-values (Holm correction) when considering the first 20 and all (3170) genes:



Are the most significant genes in the data of Hedenfalk et al. (2001) by chance the first 20 ones? Or do adjusted p-values depend on the number of tests? And why?

False discovery rate (FDR)

False discovery rate: rate of false positives *among cases declared significant* (= genes with *declared* differential expression):

Definition (False positive rate)

$$\text{False discovery rate: } \text{FDR} = E \left[\frac{\text{FP}}{\text{FP} + \text{TP}} \right] = E \left[\frac{V}{R} \right]$$

Aim: make sure that FDR is not larger than \bar{q} (\bar{q} : any value chosen between 0 and 1, similar meaning as $\bar{\alpha}$:

- for each test case (e.g. gene), calculate p-value
- **adjust** p-values to get corresponding **q-values**
- reject null hypotheses whose q-value is smaller than \bar{q} ; accept others

Controlling FDR: Benjamini-Hochberg method

Controlling FDR at level $\bar{q} = 10\%$:

- for each test case (gene), calculate p-value; this yields m p-values P_1, P_2, \dots, P_m
- order p-values: $P_{(1)} \leq P_{(2)} \leq P_{(3)} \leq \dots \leq P_{(m)}$
- **adjust** p-values to get q-values:
 - ▶ $Q_{(1)} = \min\{m \cdot P_{(1)}, 1\}$
 - ▶ $Q_{(2)} = \min\{\frac{m}{2} \cdot P_{(2)}, 1\}$; if value below $Q_{(1)}$, replace it by $Q_{(1)}$
 - ▶ $Q_{(3)} = \min\{\frac{m}{3} \cdot P_{(3)}, 1\}$; if value below $Q_{(2)}$, replace it by $Q_{(2)}$
 - ▶ $Q_{(i)} = \min\{\frac{m}{i} \cdot P_{(i)}, 1\}$; if value below $Q_{(i-1)}$, replace it by $Q_{(i-1)}$
- reject null hypotheses whose q-value is smaller than \bar{q} ; accept others

Procedure guarantees $FDR \leq \bar{q}$

Method published by Benjamini and Hochberg (1995)

Benjamini-Hochberg method for data of Hedenfalk et al. (2001)

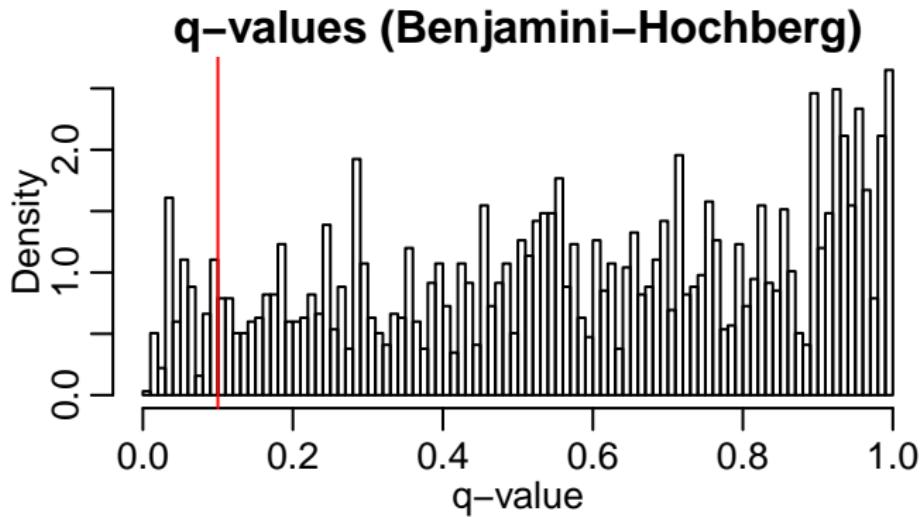
Raw data:

Gene	Expression level in BRCA1	Expression level in BRCA2	p-value	q-value
catenin	(7 samples)	(7 samples)	0.0121	0.1316
phosphofructokinase, platelet	(7 samples)	(7 samples)	0.9949	0.9971
ADP-ribosylation factor 3	(7 samples)	(7 samples)	0.0418	0.9441
uroporphyrinogen III synthase	(7 samples)	(7 samples)	0.8458	0.9441
ribosomal protein L26	(7 samples)	(7 samples)	0.2519	0.5549
G protein	(7 samples)	(7 samples)	0.6587	0.8557
:	:	:	:	:

Controlling FDR at $\bar{q} = 10\%$

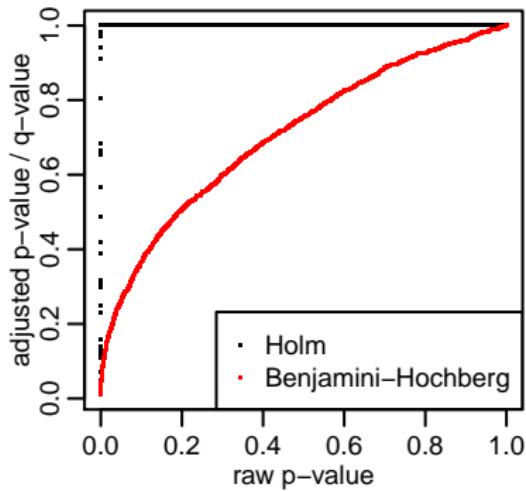
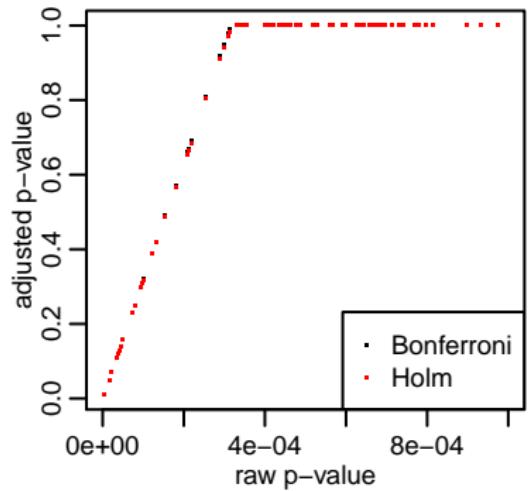
- rejecting all null hypothesis with q-values smaller than 0.1, i.e.
- declaring all genes with a q-value smaller than 0.1 as differentially expressed

Benjamini-Hochberg method for data of Hedenfalk et al. (2001)



- Reject null hypotheses to the left of the red threshold line
- Smallest q-values: 0.010, 0.019, 0.019, 0.019, 0.019
- Higher power than methods controlling FWER

Comparison: raw and adjusted p-values, q-values



Overview: multiple testing corrections

Controlling FWER

- Control probability of a type I error in the whole experiment
- Useful when
 - ▶ conclusion from any inference is likely to be erroneous when individual inference is erroneous (e.g.: testing different treatments against standard, choosing best one)
 - ▶ number of tests not higher than 20 to 50
- Conservative acceptance of null hypotheses

Controlling FDR

- Control rate of false positives among cases declared significant
- Useful when
 - ▶ outcome from different inferences interesting by itself (e.g., finding causes for a disease; finding differentially expressed genes)
 - ▶ number of test cases higher than 500
- Less conservative acceptance of null hypotheses

Part XII

Simple Linear Regression

Learning objectives

- Know the concept of linear regression: model, interpretation of coefficients, estimation of coefficients
- Perform a linear regression in R and interpreting its output
- Check model assumptions after performing a linear regression

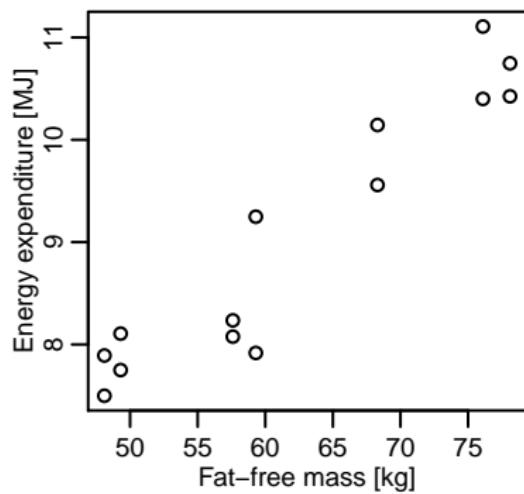
Suggested literature

This lecture is partly based on the following source:

- Samuels et al. (2012), Chapters 12.3, 12.4, 12.5

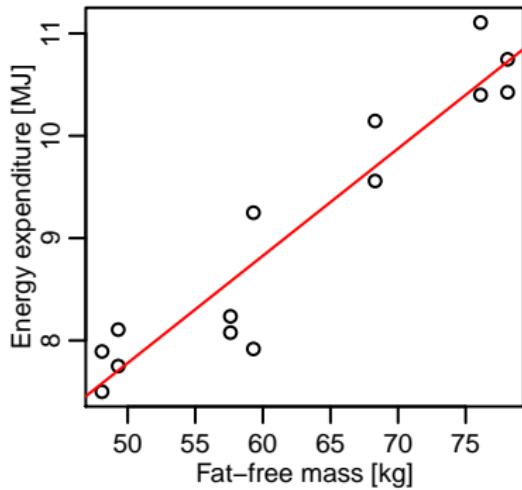
Example: body mass and energy expenditure

- Measurements of fat-free mass and energy expenditure in 24 h of 7 subjects (Webb, 1981)



Example: body mass and energy expenditure

- Measurements of fat-free mass and energy expenditure in 24 h of 7 subjects (Webb, 1981)
- Daily energy expenditure seems to grow linearly with fat-free mass
- Which measure (that you know...) measures how well fat-free mass explains energy expenditure?



Simple linear regression

Model for simple linear regression:

$$Y_i = \beta_0 + \beta_1 x_i + E_i, \quad i = 1, \dots, n,$$

$$E_1, \dots, E_n \stackrel{\text{i.i.d.}}{\sim} \mathcal{N}(0, \sigma^2)$$

Symb.	Name	Meaning	Example
Y_i	response variable	variable we want to predict	energy expenditure
x_i	explanatory variable, covariate	variable we know, which is easy to measure	fat-free mass
E_i	error or noise variables	deviation from a perfect straight line	

Error variables account for

- (unexplainable) individual deviation from a mean
- unmeasured variables having an influence on response variable: e.g. activity during the day

Simple linear regression: naming

Model for simple linear regression:

$$Y_i = \beta_0 + \beta_1 x_i + E_i, \quad i = 1, \dots, n,$$

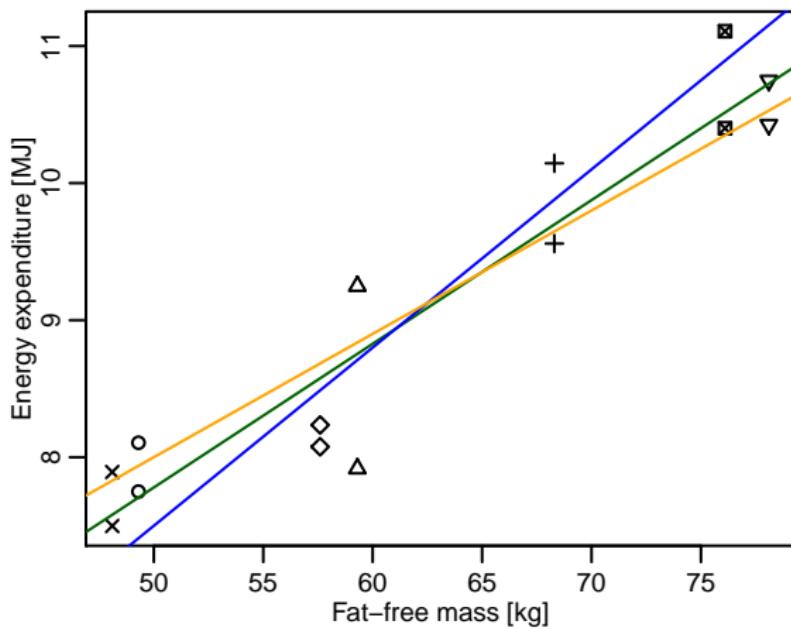
$$E_1, \dots, E_n \stackrel{\text{i.i.d.}}{\sim} \mathcal{N}(0, \sigma^2)$$

What the name of the model means:

- “simple”: only one explanatory variable (otherwise: “multiple linear regression model”, see later)
- “linear”: response variable is a linear function of the coefficients β_0, β_1

Estimating the line

How do we find a line that fits the data well? Which of the lines shown is the best one?



Estimating the line: overview

- Parameters to be estimated: β_0 , β_1 and σ^2
- Estimate β_0 and β_1 by minimizing **residual sum of squares**
- Estimate σ^2 by estimating the variance of the **residuals**

$$R_i = Y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i$$

Residual sum of squares

- **Residual:** difference between measured and predicted value of response variable
- Measured value: Y_i (measured energy expenditure)
- Predicted value for given parameters β_0 and β_1 : $\beta_0 + \beta_1 x_i$ (energy expenditure predicted based on fat-free mass of subject)

Definition (Residuals)

The i -th residual ($i = 1, \dots, n$) is defined as $R_i = Y_i - \beta_0 - \beta_1 x_i$.

The **residual sum of squares** is defined as $\text{RSS} = \sum_{i=1}^n R_i^2$.

Minimizing the residual sum of squares

- RSS is quality measure for fitted line: when line fits data well, residuals are small, and hence RSS is small
- This motivates estimator for coefficients: choose β_0 and β_1 such that RSS is minimal
- Minimizers $\hat{\beta}_0$ and $\hat{\beta}_1$ of RSS are **unbiased** estimators for the true coefficients β_0 and β_1

Estimating the error variance

- Error variables E_i not directly observable: we only know values of x_i and Y_i
- Use trick to estimate variance σ^2 of error variables: approximate error variables by residuals
- $R_i = Y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i$
- Estimate $\hat{\sigma}^2 = \frac{1}{n-2} \sum_{i=1}^n R_i^2$

Estimating the error variance

- Error variables E_i not directly observable: we only know values of x_i and Y_i
- Use trick to estimate variance σ^2 of error variables: approximate error variables by residuals
- $R_i = Y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i$
- Estimate $\hat{\sigma}^2 = \frac{1}{n-2} \sum_{i=1}^n R_i^2$
- Why denominator $n-2$ and not $n-1$ as before?
- Rule of thumb: use factor $1/(n - \text{number of est. parameters})$
- Normally, we estimate variance and mean (1 additional parameter); here, we estimate variance, β_0 and β_1 (2 additional parameters)

Linear regression in R

```
> energymass <- read.table("data/energy.csv", sep = ",", header = TRUE)
> ## Convert energy to MJ...
> energymass$energy <- 4.1868e - 3*energymass$energy
> energy.fit <- lm(energy ~ mass, data = energymass)
> summary(energy.fit)
```

Call:

```
lm(formula = energy ~ mass, data = energymass)
```

Residuals:

Min 1Q Median 3Q Max

-0.83689 -0.25948 -0.02941 0.37778 0.59247

Coefficients:

Estimate	Std. Error	t value	Pr(> t)
----------	------------	---------	----------

(Intercept)	2.53831	0.65519	3.874	0.00221 **
-------------	---------	---------	-------	------------

mass	0.10482	0.01033	10.143	3.07e-07 ***
------	---------	---------	--------	--------------

—

Signif. codes:

0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.433 on 12 degrees of freedom

Multiple R-squared: 0.8955, Adjusted R-squared: 0.8868

F-statistic: 102.9 on 1 and 12 DF, p-value: 3.073e-07

Lots of output...

Call:

```
lm(formula = energy ~ mass, data = energymass)
```

Residuals:

Min 1Q Median 3Q Max

-0.83689 -0.25948 -0.02941 0.37778 0.59247

Coefficients:

Estimate Std. Error t value Pr(>|t|)

(Intercept) 2.53831 0.65519 3.874 0.00221 **

mass 0.10482 0.01033 10.143 3.07e-07 ***

—

Signif. codes:

0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.433 on 12 degrees of freedom

Multiple R-squared: 0.8955, Adjusted R-squared:

0.8868

F-statistic: 102.9 on 1 and 12 DF, p-value:

3.073e-07

- **Coefficients:**

$$\hat{\beta}_0 = 2.53831,$$

$$\hat{\beta}_1 = 0.10482$$

- **Standard deviation of error variables:**

$$\hat{\sigma} = 0.433$$

- **Subjects in the study:**

$n = \text{degrees of freedom}$
+ number of estimated parameters;

$$n = 12 + 2 = 14$$

Interpretation of output I

Answer the following questions based on the model output on the previous slides:

- When you weigh 4 kg more than your brother, how much more energy are you likely to expend per day?
- What's the estimated amount of energy a person with a fat-free mass of 65 kg expends per day?
- Indicate a 95% confidence interval for the daily energy expenditure of a person of 65 kg.

Significance of whole model: F test

p-value of "F-statistic": p-value to null hypothesis $\beta_0 = \beta_1 = 0$

Call:

```
lm(formula = energy ~ mass, data = energymass)
```

Residuals:

Min 1Q Median 3Q Max

-0.83689 -0.25948 -0.02941 0.37778 0.59247

Coefficients:

Estimate Std. Error t value Pr(>|t|)

(Intercept) 2.53831 0.65519 3.874 0.00221 **

mass 0.10482 0.01033 10.143 3.07e-07 ***

Signif. codes:

0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.433 on 12 degrees of freedom

Multiple R-squared: 0.8955, Adjusted R-squared: 0.8868

F-statistic: 102.9 on 1 and 12 DF, p-value: 3.073e-07

Significance of explanatory variable: t-test

Has fat-free mass an influence on energy expenditure? Let's test it:

① **Model:** $Y_i = \beta_0 + \beta_1 x_i + E_i, E_1, \dots, E_n \stackrel{\text{i.i.d.}}{\sim} \mathcal{N}(0, \sigma^2)$

② **Null hypothesis:** $H_0 : \beta_1 = 0$

Alternative hypothesis: $H_A : \beta_1 \neq 0$

③ **Test statistic:** $T = \frac{\text{calc. coeff.} - \text{exp. coeff.}}{\text{standard error}} = \frac{\hat{\beta}_1 - 0}{\widehat{\text{se}}(\hat{\beta}_1)}$

Distribution of T under H_0 : $T \sim t_{n-2}$ (Student's t distribution with $n - 2$ degrees of freedom)

④ Choose **significance level**: e.g. $\alpha = 5\%$

⑤ **Range of rejection:** $K = (-\infty, -t_{n-2, 1-\frac{\alpha}{2}}] \cup [t_{n-2, 1-\frac{\alpha}{2}}, \infty)$

⑥ **Test decision:** reject if $T \in K$, i.e. if $|T| > t_{n-2, 1-\frac{\alpha}{2}}$

Significance of explanatory variable: t-test

p-value of t-test already in R output; no need to perform test by hand:

Call:

```
lm(formula = energy ~ mass, data = energymass)
```

Residuals:

Min 1Q Median 3Q Max

-0.83689 -0.25948 -0.02941 0.37778 0.59247

Coefficients:

Estimate Std. Error t value Pr(>|t|)

(Intercept) 2.53831 0.65519 3.874 0.00221 **

mass 0.10482 0.01033 10.143 3.07e-07 ***

—
Signif. codes:

0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.433 on 12 degrees of freedom

Multiple R-squared: 0.8955, Adjusted R-squared: 0.8868

F-statistic: 102.9 on 1 and 12 DF, p-value: 3.073e-07

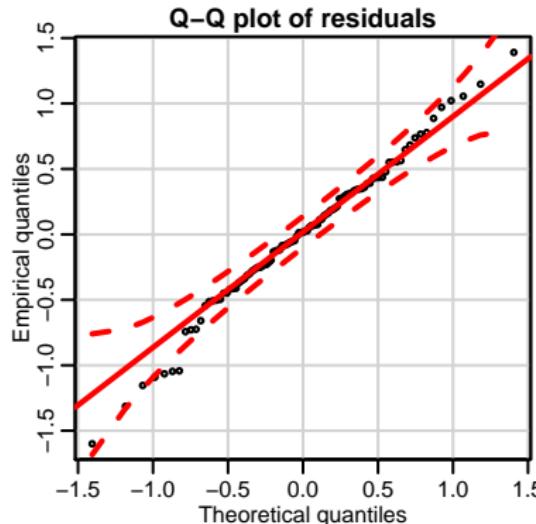
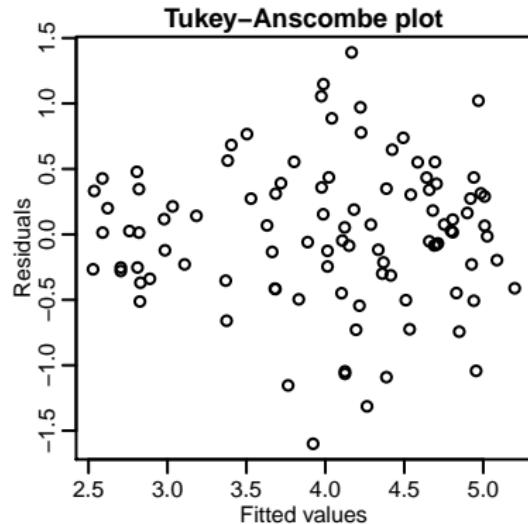
- Confidence interval for β_1 :

$$I = \left[\hat{\beta}_1 - \text{se}(\hat{\beta}_1) t_{n-2,1-\frac{\alpha}{2}}, \hat{\beta}_1 + \text{se}(\hat{\beta}_1) t_{n-2,1-\frac{\alpha}{2}} \right]$$

Residual analysis: checking model assumptions

Checking model assumptions based on two plots:

- **linearity and i.i.d. assumption** on errors with **Tukey-Anscombe plot**: residuals vs. fitted values
- **normality assumption** on errors with **Q-Q plot**

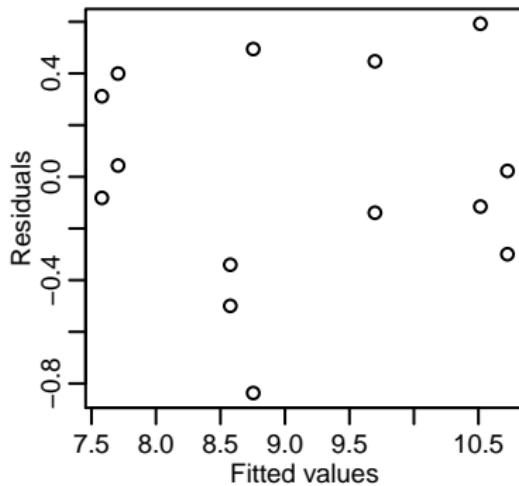


Tukey-Anscombe plot

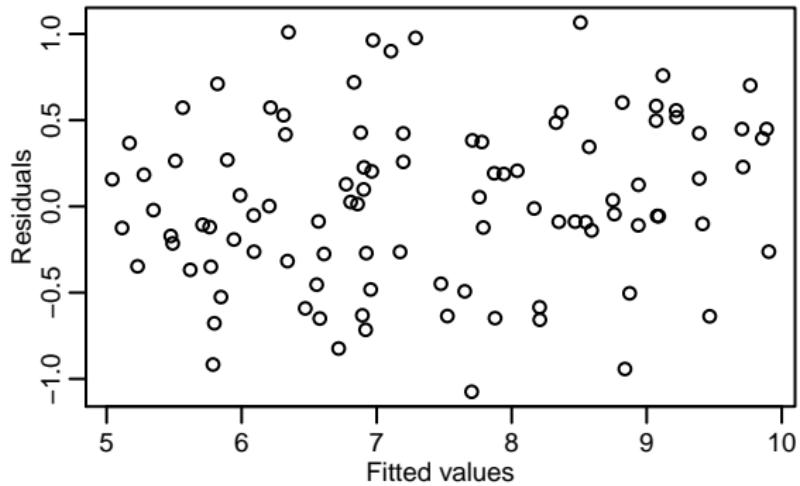
Checking linearity and i.i.d. assumption on error variables:

- plot residuals R_i vs. fitted response variables $\hat{Y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i$ (**Tukey-Anscombe plot**)
- if i.i.d. assumption holds, points in Tukey-Anscombe plot should fluctuate randomly around the horizontal axis, without visible pattern
- Tukey-Anscombe plot in R:

```
> plot(fitted(energy.fit), resid(energy.fit), xlab =  
"Fitted values", ylab = "Residuals")
```

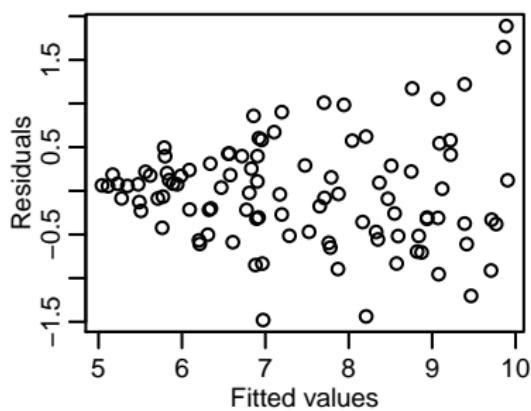


Good Tukey-Anscombe plot



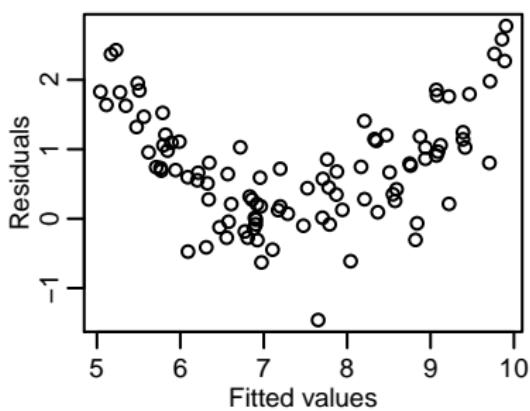
Model violations visible in Tukey-Anscombe plot

Errors of different variance:



Cone; may be corrected via
log-transform $Y_i \mapsto \log(Y_i)$

Non-linear trend:

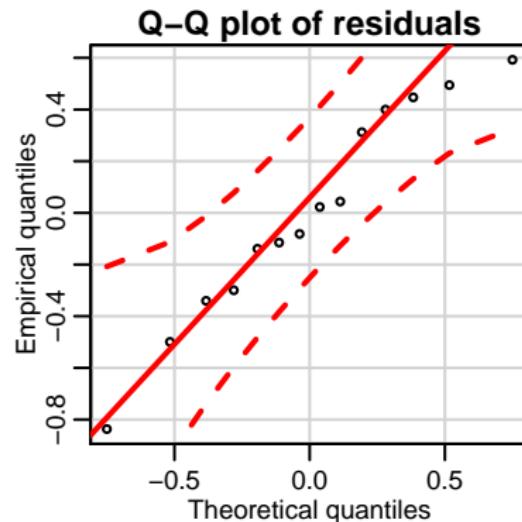


Quadratic trend; may be corrected by
adding x_i^2 as explanatory variable

Q-Q plots for residuals

Good old Q-Q plot to check assumption on distribution . . .

```
> library(car)  
> qqPlot(resid(energy.fit), dist = "norm",  
mean = mean(resid(energy.fit)), sd = sd(resid(energy.fit)),  
xlab = "Theoretical quantiles", ylab = "Empirical quantiles",  
main = "Q-Q plot of residuals")
```



```
main = "Q-Q plot of residuals")
```

Coefficient of determination R^2

- **Coefficient of determination R^2** indicates how well the data points fit a line ("goodness of fit")
- Definition: coefficient of determination = squared empirical correlation between measured (Y_i) and predicted (\hat{Y}_i) response variables
- Recall from Part V: $R^2 = \left(\frac{s_{\hat{y}y}}{s_{\hat{y}}s_y} \right)^2$

Coefficient of determination in R output

Call:

```
lm(formula = energy ~ mass, data = energymass)
```

Residuals:

Min 1Q Median 3Q Max

-0.83689 -0.25948 -0.02941 0.37778 0.59247

Coefficients:

Estimate	Std. Error	t value	Pr(> t)
----------	------------	---------	----------

(Intercept) 2.53831 0.65519 3.874 0.00221 **

mass 0.10482 0.01033 10.143 3.07e-07 ***

—

Signif. codes:

0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.433 on 12 degrees of freedom

Multiple R-squared: 0.8955, Adjusted R-squared: 0.8868

F-statistic: 102.9 on 1 and 12 DF, p-value: 3.073e-07

Manual calculation:

```
> cor(energymass$energy, fitted(energy.fit))^2
```

```
[1] 0.8955353
```

Connection to empirical correlation

- In the linear regression model, we assume the covariate to be fix, not random
- However, if we are not interested in the uncertainty about it, we could model it as random
- This would lead to the model $Y = \beta_0 + \beta_1 X + \varepsilon$
- In this model, we have $\text{Cov}(X, Y) = \beta_1 \cdot \text{Var}(X)$: the regression coefficient β_1 is connected to the covariance between X and Y
- Indeed, the estimator for β_1 is

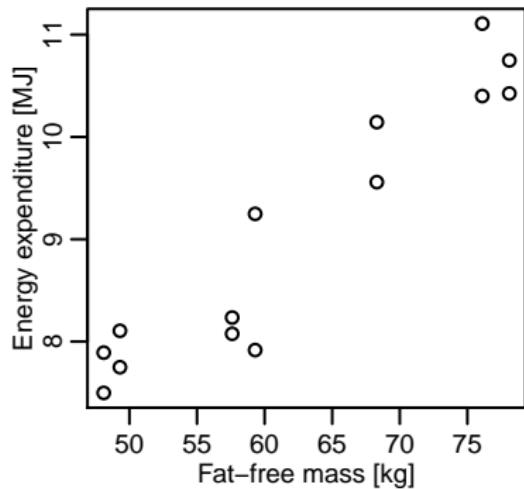
$$\hat{\beta}_1 = \frac{s_{xy}}{s_x^2} ,$$

where s_x denotes the empirical standard deviation of x_1, \dots, x_n and s_{xy} stands for

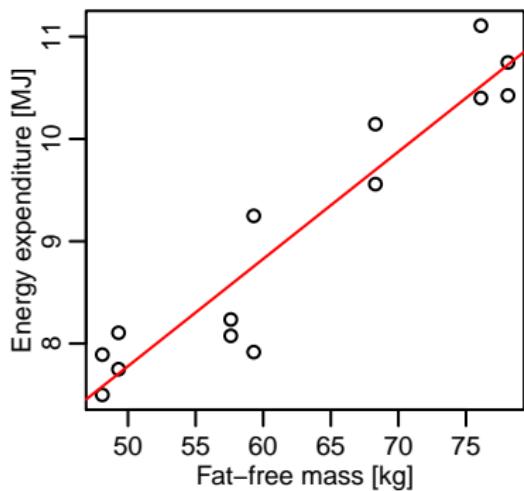
$$\frac{1}{n-1} \sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})$$

(see also Part V of the lecture)

Simple linear regression: summary



Simple linear regression: summary



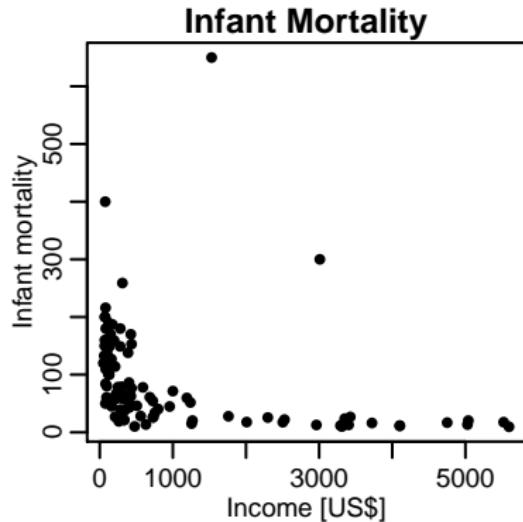
- Simple linear regression model:

$$Y_i = \beta_0 + \beta_1 x_i + E_i, \quad i = 1, \dots, n,$$
$$E_1, \dots, E_n \stackrel{\text{i.i.d.}}{\sim} \mathcal{N}(0, \sigma^2)$$

- Parameters to be estimated: β_0 , β_1 and σ^2
- Estimate β_0 and β_1 by minimizing **residual sum of squares** $\text{RSS} = \sum_{i=1}^n R_i^2$.
- Estimate σ^2 by estimating the variance of the **residuals**
$$R_i = Y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i$$

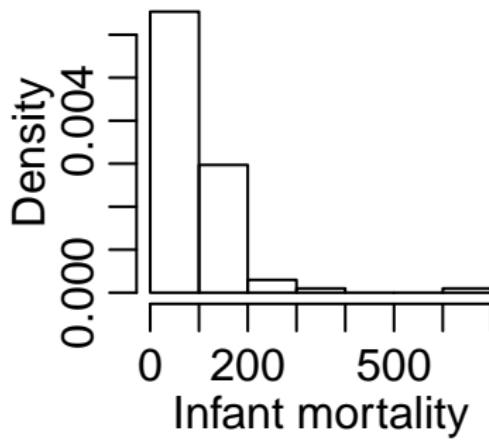
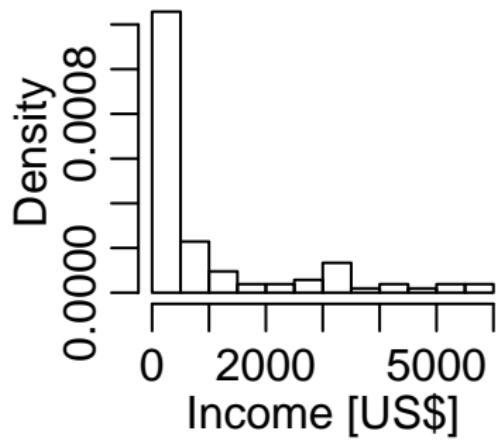
Example: income and child mortality

- Right: plot of child mortality (# children dying before age of 5 per 1000 newborn babies) versus per-capita income in US\$
- Famous data set first published in the *New York Times* in 1975
- Special features of the data set?



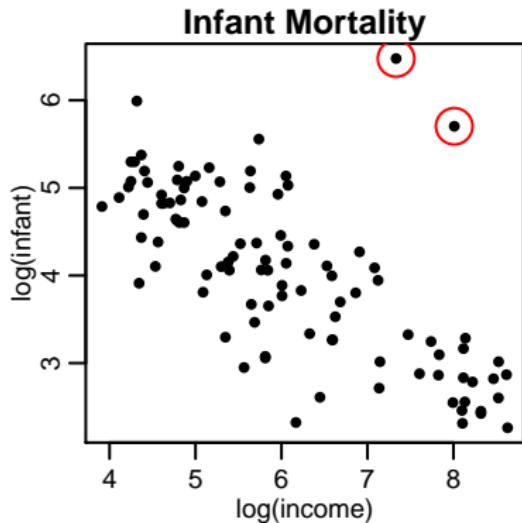
Child mortality data set

Histograms for child mortality data set:



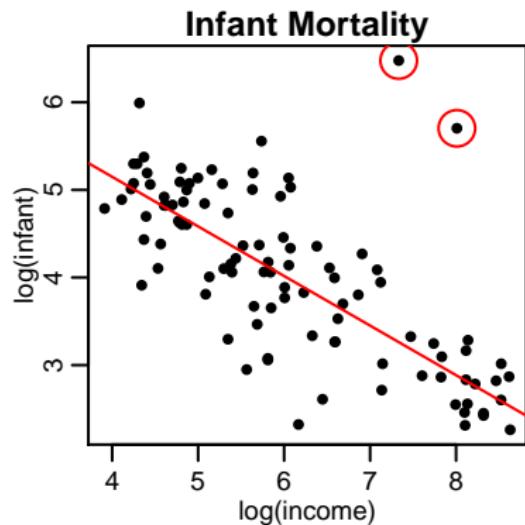
Fitting linear model to transformed data

- Notation: X_i : per-capita income of i -th country; Y_i : infant mortality of i -th country
- Log-transform both variables: $X_i = \log(X_i)$, $Y_i = \log(Y_i)$ (reason for doing that...?)
- Remove outliers (Afghanistan and Zambia)



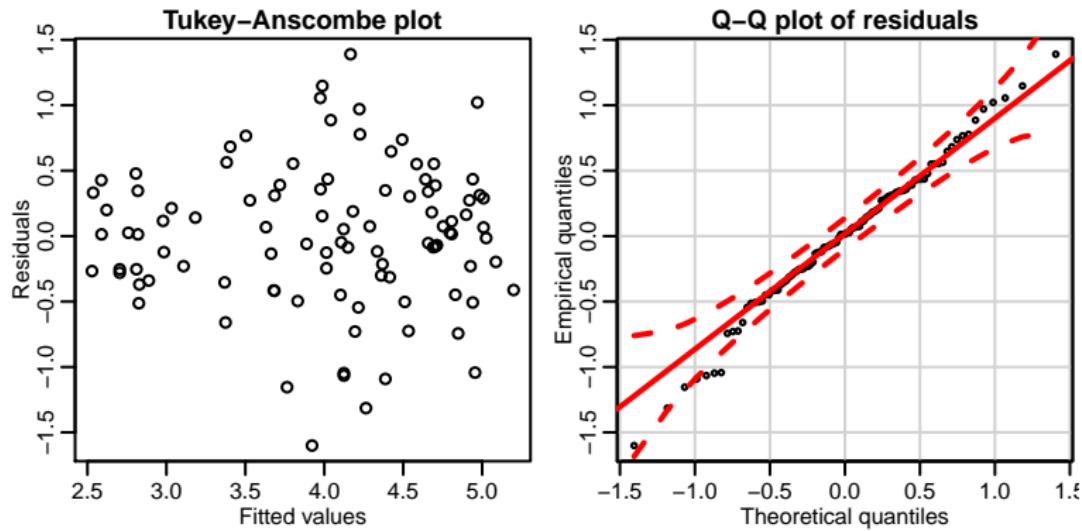
Fitting linear model to transformed data

- Notation: X_i : per-capita income of i -th country; Y_i : infant mortality of i -th country
- Log-transform both variables: $X_i = \log(X_i)$, $Y_i = \log(Y_i)$ (reason for doing that...?)
- Remove outliers (Afghanistan and Zambia)



Residual analysis for transformed variables

Tukey-Anscombe plot and Q-Q plot of residuals:



Back-transform fitted model

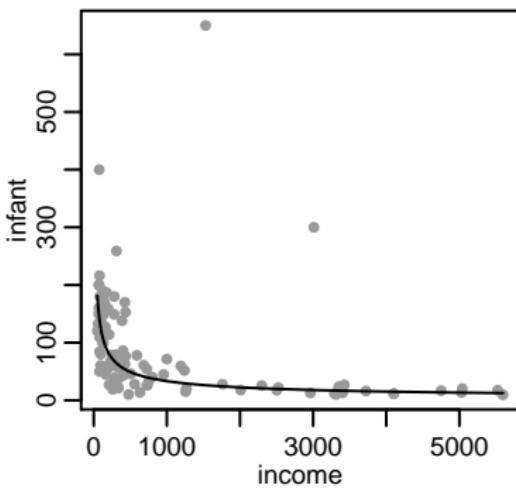
In R (fitted object stored as `im.fit`):

```
> plot(infant ~ income, data = Leinhardt,  
  pch = 20, col = gray(0.6))  
> pred <- exp(predict(im.fit, newdata =  
  data.frame(income = Leinhardt$income)))  
> ord <- order(Leinhardt$income)  
> lines(Leinhardt$income[ord], pred[ord])
```

- Fitted model of the form

$$\log(Y_i) = \hat{\beta}_0 + \hat{\beta}_1 \log(X_i) + E_i$$

- Hence model gives us predictions for $\log(Y)$, not Y
- Can back-transform to initial scale by exponential function



References

Die Welt in Zahlen 2011. 2011.

Yoav Benjamini and Yosef Hochberg. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the royal statistical society. Series B (Methodological)*, pages 289–300, 1995.

Carlo E Bonferroni. *Teoria statistica delle classi e calcolo delle probabilita*. Libreria internazionale Seeber, 1936.

Olive Jean Dunn. Multiple comparisons among means. *Journal of the American Statistical Association*, 56(293): 52–64, 1961.

Ingrid Hedenfalk, David Duggan, Yidong Chen, Michael Radmacher, Michael Bittner, Richard Simon, Paul Meltzer, Barry Gusterson, Manel Esteller, Mark Raffeld, et al. Gene-expression profiles in hereditary breast cancer. *New England Journal of Medicine*, 344(8):539–548, 2001.

Sture Holm. A simple sequentially rejective multiple test procedure. *Scandinavian journal of statistics*, pages 65–70, 1979.

Gunther Maier and Peter Weiss. *Modelle diskreter Entscheidungen*. Springer Science & Business Media, 2013.

Takeo Mizutani and Tomotari Mitsuoka. Effect of intestinal bacteria on incidence of liver tumors in gnotobiotic c3h/he male mice. *Journal of the National Cancer Institute*, 63(6):1365–1370, 1979.

Mehdi Namdar, Pascal Koepfli, Renate Grathwohl, Patrick T Siegrist, Michael Klainguti, Tiziano Schepis, Raphael Delaloye, Christophe A Wyss, Samuel P Fleischmann, Oliver Gaemperli, et al. Caffeine decreases exercise-induced myocardial flow reserve. *Journal of the American College of Cardiology*, 47(2):405–410, 2006.

Colin A Nurse. Interactions between dissociated rat sympathetic neurons and skeletal muscle cells developing in cell culture: II. synaptic mechanisms. *Developmental biology*, 88(1):71–79, 1981.

Steven G Potkin, H Eleanor Cannon, Dennis L Murphy, and Richard Jed Wyatt. Are paranoid schizophrenics biologically different from other schizophrenics? *New England Journal of Medicine*, 298(2):61–66, 1978.

Myra L Samuels, Jeffrey A Witmer, and Andrew Schaffner. *Statistics for the life sciences*. Pearson education, 2012.

Werner A Stahel. Ausblick. In *Statistische Datenanalyse*, pages 348–353. Springer, 2002.

Richard W Van Norman. *Experimental biology*. Prentice Hall, 1971.

Paul Webb. Energy expenditure and fat-free mass in men and women. *The American journal of clinical nutrition*, 34(9):1816–1826, 1981.