

# Lectures

MVE055 / MSG810

Mathematical statistics and discrete mathematics

---

Moritz Schauer

Last updated October 20, 2021

GU & Chalmers University of Technology

# Teachers

---

Moritz Schauer: Instructor  
Room: H3029  
E-mail: [smoritz@chalmers.se](mailto:smoritz@chalmers.se)

Ruben Seyer: Teaching assistant  
E-mail: [rubense@chalmers.se](mailto:rubense@chalmers.se)

## Time table (1st week)

---

Lecture	Monday	13-16
Exercise	Tuesday	10-12
Lecture	Wednesday	10-12
Exercise	Thursday	10-12

## Student representatives

---

---

eli.adelhult@outlook.com	Eli Adelhult
jesper@acorneroftheweb.com	Jesper Führ
moltas.hultin@outlook.com	Moltas Hultin
alexandra.lindvall@telia.com	Alexandra Lindvall
riyatagra@gmail.com	Riya Tagra

---

## Course overview

<https://chalmers.instructure.com/courses/15306>

# Examination

---

“För godkänd på kursen krävs godkänd på de tre grupparbetana samt godkänd på skriftlig tentamen. Betyget på kursen baseras på betyget på tentan.”

Examination consists of two parts.

## Exam:

- Exam takes place on campus. **Will** look similar to the last exam.

## 3 group assignments:

- First assignment: “Skiplist”.
- Groups of up to four students.
- ↳ Find yourself a group on canvas "Assignment groups".
- One student hands in for the group on canvas.
- Required for passing but does not affect course grade.

# Course content

---

In **probability theory** we construct and analyse mathematical models for phenomena that exhibit uncertainty and variation.  
Highlight: Markov chains.

In **statistics** we observe data and we want to infer the probabilistic model or parameters of such a model: **inverse probability**.

**Generating functions** allow to solve recursive equations.

**The law of large number** describes what happens if you perform the same experiment a large number of times.

**Regression** to find linear relationships between inputs/explanatory variables and outputs/explained variables.

## Example: Probability vs statistics

---

What is the probability to throw 10 times heads in a row with a fair coin.

This is the 10th time you throw head in a row... is that coin fair!?



## Describing data

---

# Visual inspection

---

When analysing a data set, it is a good idea to first visualise it graphically.

**Example:**

Throwing a dice 20 times we obtained the following results:

1, 3, 3, 3, 1, 6, 6, 5, 1, 4, 6, 1, 4, 5, 1, 1, 2, 3, 6, 5.

# Frequency table and histogram

Everything starts with data and tables.

If the observations take values in a small set, then we can summarise the data in a frequency table showing how many outcomes we have for each possible outcome.

For our results

1, 3, 3, 3, 1, 6, 6, 5, 1, 4, 6, 1, 4, 5, 1, 1, 2, 3, 6, 5

we get

Outcome	1	2	3	4	5	6
Count	6	1	4	2	3	4
Proportion	0.30	0.05	0.20	0.10	0.15	0.20

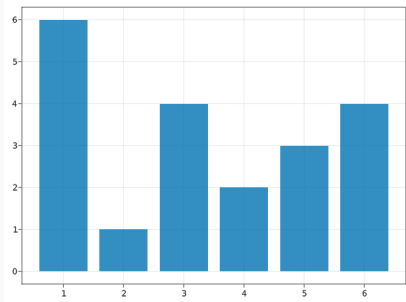
# Tricky denominators

ZIP	Neighborhood	Estimated Population	At Least 1 Dose	At Least 1 Dose (%)	Fully Vaccinated	Fully Vaccinated (%) ▼
10004	Financial District	2,972	3,718	100%	3,341	100%
10006	Financial District	3,382	4,087	100%	3,599	100%
10018	Hell's Kitchen/Midtown Manhattan	11,791	18,861	100%	15,089	100%
10036	Hell's Kitchen/Midtown Manhattan	27,242	35,718	100%	30,586	100%
10001, 10118	Chelsea/NoMad/West Chelsea	27,613	29,985	100%	25,988	94%
10019, 10020	Hell's Kitchen/Midtown Manhattan	43,522	45,518	100%	40,120	92%
11355	Flushing/Murray Hill/Queensboro Hill	78,853	76,759	97%	71,043	90%
10017	East Midtown/Murray Hill	15,613	15,705	100%	14,059	90%
10007	TriBeCa	6,991	6,512	93%	5,997	86%
10022	East Midtown	30,896	27,664	90%	25,509	83%

New York City Health Department, 2021-08-08.

# Bar chart

Using the frequency table we can draw a bar chart. For each value we draw a bar whose height is proportional to the number of observations for that value.



```
using StatsBase, GLMakie
x = [1, 3, 3, 3, 1, 6, ..., 5, 1, 1, 2, 3, 6, 5,]
barplot(counts(x, 1:6))
```

# Histogram

---

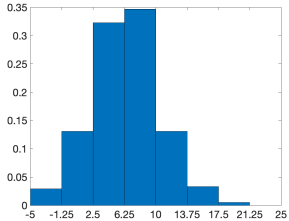
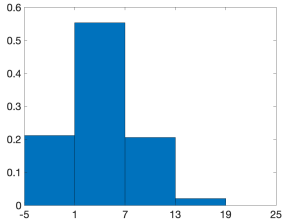
**Task:** Summarise 1000 real numbers which are the outcome of some experiment,

12.15, 17.33, 0.96, 13.44, 11.27, 4.76, 8.26, 11.37, 24.31, 21.07, ...

A bar chart doesn't make sense because the data does not have only a few different values. We can use a histogram:

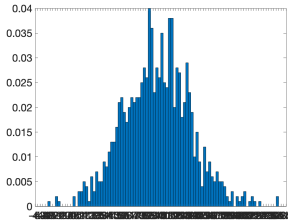
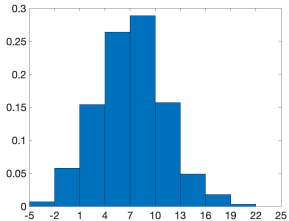
- Divide the data into a number of classes (intervals) and then calculate the number of observations in each class.
- Draw bars where the height is proportional to the number of observations in the class and the width equals the interval width.

# Histogram



4 classes

7 classes

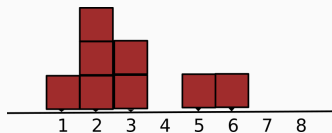


9 classes

200 classes

## Sample statistics for location

Case	1	2	3	4	5	6	7	8
Value	2	3	2	6	5	1	2	3



Weights on a bar



# Sample median

---

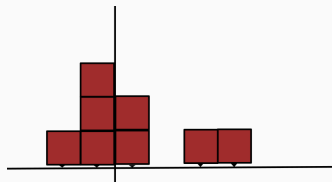
- To obtain the **sample median**, write the values in sorted order and take the middle one.

If there is an even number of values in the data set, take the average of the two middle most.

# Median

Median

Value	1	2	2	2	3	3	5	6
-------	---	---	---	---	---	---	---	---

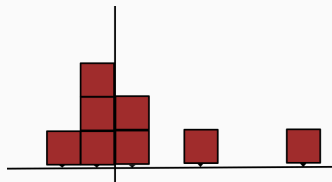


Median = 2.5

# Median

Median

Value	1	2	2	2	3	3	5	8
-------	---	---	---	---	---	---	---	---



Median = 2.5

## Sample mean

---

- The (sample) mean, denoted as  $\bar{x}$ , can be calculated as

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

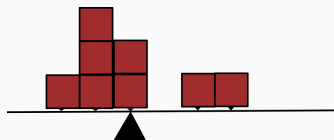
where  $x_1, x_2, \cdots, x_n$  are the  $n$  observed values.

In words: Sum the values of all cases in the data set and divide by the total number of values.

## Sample mean

Mean

Value	1	2	2	2	3	3	5	6
-------	---	---	---	---	---	---	---	---

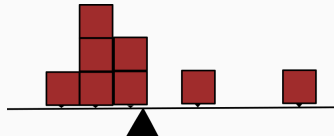


$$\text{Mean } \bar{x} = \frac{1 \cdot 1 + 3 \cdot 2 + 2 \cdot 3 + 1 \cdot 5 + 1 \cdot 6}{8} = 3$$

## Sample mean

Mean

Value	1	2	2	2	3	3	5	8
-------	---	---	---	---	---	---	---	---



$$\text{Mean } \bar{x} = \frac{1 \cdot 1 + 3 \cdot 2 + 2 \cdot 3 + 1 \cdot 5 + 1 \cdot 8}{8} = 3.25$$

## Sample statistics for variation/spread

**Sample variance:** The sample variance of a data set  $x_1, \dots, x_n$  is given by

$$s^2 = \frac{1}{n-1} \sum_{i=1}^n (x_i - \bar{x})^2 = \frac{1}{n-1} ((x_1 - \bar{x})^2 + \dots + (x_n - \bar{x})^2)$$

Sometimes convenient to use the formula

$$s^2 = \frac{1}{n-1} \left( \sum_{i=1}^n x_i^2 - n\bar{x}^2 \right) = \frac{1}{n-1} (x_1^2 + \dots + x_n^2 - n\bar{x}^2)$$

**Sample standard deviation  $s$ :** the square root  $\sqrt{s^2}$  of the sample variance.

## Example 1 (cont.)

For the dice throw example

1, 3, 3, 3, 1, 6, 6, 5, 1, 4, 6, 1, 4, 5, 1, 1, 2, 3, 6, 5

we obtain the mean

$$\bar{x} = (1 + 3 + 3 + \dots + 3 + 6 + 5)/20 = 67/20 = 3.35$$

Sorting the values and taking the central one we obtain the median 3.

The variance is

$$s^2 = ((1 - 3.35)^2 + (3 - 3.35)^2 + \dots + (5 - 3.35)^2)/19 = 3.8184$$

and the standard deviation is  $s = 1.9541$ .



## Sample spaces

---

# Outcomes

---

In probability theory we consider experiments which have non-deterministic, variable or random outcomes. For example

1. Roll a die and count the eyes.
2. Ask a person on the street which party they would vote for.
3. Throw a handful of coins and count the heads.
4. Examine a unit from a manufacturing process.
5. Measure the round-trip time (ping) of a connection.

The result of the experiment is called **outcome**  $\omega$  (*utfall*). The set of possible outcomes is called the **sample space**  $\Omega$  (*utfallsrummet*).

↳ Sets  $\Omega$  and elements  $\omega$ .

## Sample spaces

---

- $\Omega = \{1,2,3,4,5,6\}$ .
- $\Omega = \{V, S, MP, C, L, M, KD, S, Others, No\ answer\}$ .
- $\Omega = \{(head, head), (head, tail), (tail, head), (tail, tail)\}$  (for 2 coins).
- $\Omega = \{defect, intact\}$ .
- $\Omega = [0, \infty)$  (seconds).

# Events

We group outcomes into **events**.

An event  $A$  is a set of outcomes, that is, a subset of the sample space  $\Omega$ .

Example for events:

1.  $A = \{1,3,5\}$ , that is “my die shows an odd number”.
2.  $A = \{C, L, M, KD\}$ , a “vote for the ‘Alliansen’ ”.
3.  $A = \{(\text{head},\text{head}),(\text{tail},\text{tail})\}$ , “both coins show the same face”.
4.  $A = \{\text{defect}\}$ , the “unit is broken”.
5.  $A = \{x: x \geq 0.5\}$ , round-trip-time larger than 0.5s.

An event  $A$  occurs if any of the outcomes  $\omega \in A$  occurs in the experiment.

# Outcome and sample space

## Outcome and sample space

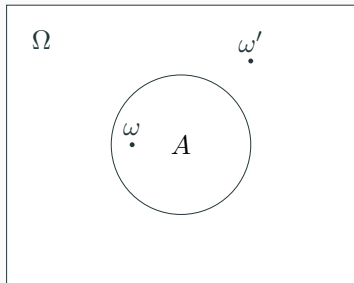
The *outcome*  $\omega$  is the result of a random experiment, and the set of all possible outcomes  $\Omega$  is called the *sample space*.

## Events

An event is a collection (a set of) different outcomes. The event  $A$ , as a set of outcomes, is therefore a subset of the sample space  $\Omega$ .

We like events because the probability of a single outcome might be too small or zero.

## Event, outcome and sample space



Event  $A$ , outcome  $\omega \in A$  and sample space  $\Omega$

And some other outcome  $\omega' \notin A$ .

# Overview: Intersection, union and complement

For events  $A$  and  $B$  we have defined:

## **Complement, $A^c$**

Set of all outcomes  $\omega$  not contained in  $A$ .  $A^c = \Omega \setminus A$ .

## **Union, $A \cup B$**

Set of all outcomes  $\omega$  in  $A$  or  $B$ .

## **Intersection, $A \cap B$**

Set of all outcomes  $\omega$  in  $A$  and  $B$ .

$A^c$ ,  $A \cup B$ ,  $A \cap B$  are also events.  $\emptyset$  and  $\Omega$  are also events, the impossible event and the sure event.

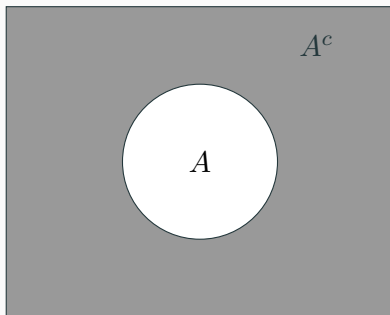
## **Mutually exclusive events**

If  $A \cap B = \emptyset$  then  $A$  and  $B$  are mutually exclusive events.

**Example:** The set  $\{2, 4, 6\}$  and the set  $\{1, 3, 5\}$  are disjoint.



# Complement

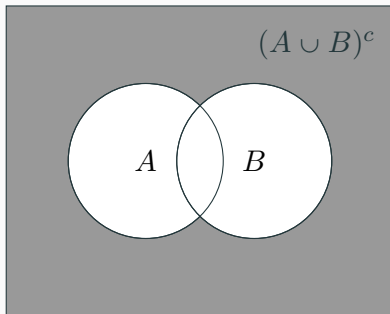


The **complement** of a  $A$  are all outcomes not in  $A$ .

$$A^c = \Omega \setminus A.$$

In the example with the die: Here  $A = \{1, 3, 5\}$ . So if the die shows a 2, then  $A^c = \{2, 4, 6\}$  happened.

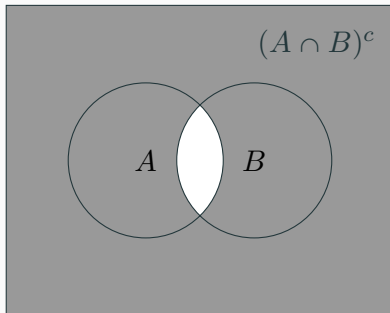
# Union



If we have events,  $A$  and  $B$  we can define  $A \cup B$ , the **union of  $A$  and  $B$** .

- $A \cup B$  occurs if  $A$  or  $B$  occur (or both).

# Intersection



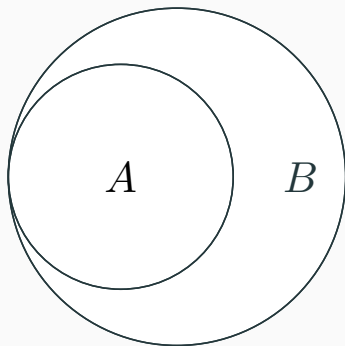
The **intersection**  $A \cap B$  are all elements both in  $A$  and  $B$ .

- So for  $A \cap B$  to occur, both  $A$  and  $B$  need to occur.

$A \cap B = \emptyset$  means that  $A$  and  $B$  exclude each other.

## Set inclusion

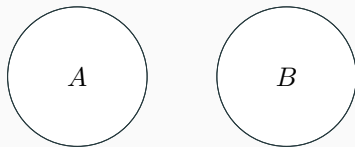
---



$$A \subset B.$$

## Disjoint sets

---



$$A \cap B = \emptyset.$$

## The empty set $\emptyset$

---

# Permutations and combinations

## Permutation

A specific order of a number of objects.

$(1, 3, 2, 5, 6, 4)$  is a permutation of the numbers 1 to 6.

## Combination

A selection of objects without regard for their order.

$\{1, 3, 5\}$  is a combination of 3 the of the numbers 1 to 6.

**Note**  $(1, 2) \neq (2, 1)$  but  $\{1, 2\} = \{2, 1\}$ .

# Permutations and combinations

## Multiplication principle

If there are  $a$  ways to make a choice and there are  $b$  ways to make a second choice, then there are  $ab$  ways to make a combined choice.

## Factorial

For  $n \in \mathbb{N}$  define  $n! = n \cdot (n - 1) \cdot (n - 2) \cdots 2 \cdot 1$  and  $0! = 1$ .  
 $n!$  is read “n-factorial”.

$$4! = 4 \cdot 3 \cdot 2 \cdot 1 = 24$$



# Calculate the number of combinations

## Number of combinations

The number of ways we can choose  $r$  objects out of a total of  $n$  distinct objects, ignoring their order, is given by

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

- ${}_nC_r$  is usually called binomial coefficient.

**Example:** Draw five cards from a poker set of 52 cards.

2 598 960 combinations are possible:

$$\binom{52}{5} = \frac{52!}{5!(52-5)!} = \frac{52 \cdot 51 \cdot 50 \cdot 49 \cdot 48}{5 \cdot 4 \cdot 3 \cdot 2 \cdot 1} = 2\,598\,960$$

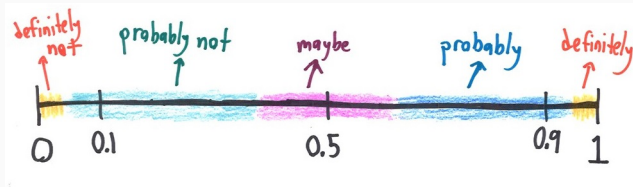
$${}_nP_r = \frac{n!}{(n-r)!}$$

# Probabilities

---

# Probabilities of events

- Probability is a numerical measure of how likely an **event** is to happen.



- Probability is a *proportion*, a number between 0 and 1.  
Notation

$P(\text{something that can happen}) = \text{a probability.}$

E.g.

$$P(\text{coin heads-up}) = \frac{1}{2}.$$

## Equally likely outcomes

What is probability? (How do we assign probability?)

- A classical and useful view considers equally likely outcomes.  
Then

$$P(A) = \frac{\text{number of outcomes for which } A \text{ occurs}}{\text{total number of outcomes}}$$

- Probability to throw an odd number with a fair die.

$$P(A) = \frac{|\{1, 3, 5\}|}{|\{1, 2, 3, 4, 5, 6\}|} = \frac{3}{6} = \frac{1}{2}$$

## Frequentist interpretation of probability

- Sometimes it is not reasonable to assume that all outcomes are equally likely.
- The **frequentist interpretation of probability**: Suppose we repeat a random experiment many times under identical conditions. As the number of repetitions  $n$  grows, we observe that the proportion  $n_A/n$  of times that an event  $A$  occur converges to a number. This number is the probability of  $A$ , or as formula

$$\frac{n_A}{n} \rightarrow P(A), \text{ where } n \rightarrow \infty$$

Example: With a fair die, we observe the proportion of times where  $A = \{\text{even number of eyes}\}$  occurs converge to  $\frac{1}{2}$ .

# Kolmogorov's axioms

Let  $\Omega$  be a sample space.

## Kolmogorov's axioms

A probability measure  $P$  is function  $A \mapsto P(A)$  assigning each event  $A \subset \Omega$  a probability, a positive number such that

1.  $0 \leq P(A) \leq 1$ .
2.  $P(\Omega) = 1$ .
3. For pairwise disjoint events  $A_1, A_2, \dots$

$$P\left(\bigcup_{i=1}^{\infty} A_i\right) = \sum_{i=1}^{\infty} P(A_i).$$

Especially for disjoint/mutually exclusive events  $A$  and  $B$ ,

$$P(A \cup B) = P(A) + P(B).$$

# Properties of probability distributions

The axioms determine all further properties of probabilities...

## Properties

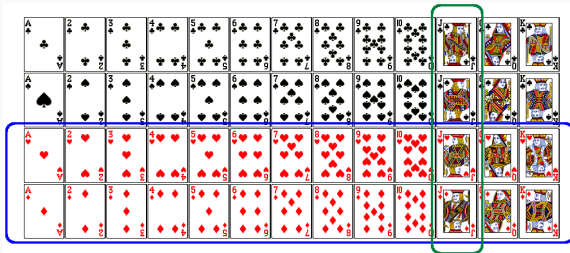
For the probability measure  $P$  it holds that:

1.  $P(\emptyset) = 0$ .
2.  $P(A^c) = 1 - P(A)$ .
3.  $P(A \cup B) = P(A) + P(B) - P(A \cap B)$ .

All these properties can be seen with the help of Venn diagrams.

# Probability of the union of non-disjoint events

What is the probability of drawing a jack or a red card from a well shuffled full deck (52 cards)?



$$\begin{aligned} P(\text{jack or red}) &= P(\text{jack}) + P(\text{red}) - P(\text{jack and red}) \\ &= \frac{4}{52} + \frac{26}{52} - \frac{2}{52} = \frac{28}{52} \end{aligned}$$




## General addition rule

---









$$P(A \cup B) = P(A) + P(B) - P(A \cap B)$$

## Combined experiment

Throw a coin ((1), ), and throw a 6 sided die. What is

$$P(\textcircled{1}, \textcircled{\cdot\cdot}) = \frac{1}{12}$$

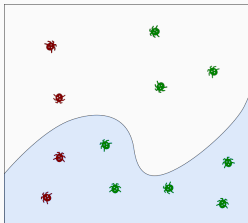
Use multiplication rule and the classical approach.

							
	$\frac{1}{12}$	$\frac{1}{12}$	$\frac{1}{12}$	$\frac{1}{12}$	$\frac{1}{12}$	$\frac{1}{12}$	$\frac{1}{2}$
	$\frac{1}{12}$	$\frac{1}{12}$	$\frac{1}{12}$	$\frac{1}{12}$	$\frac{1}{12}$	$\frac{1}{12}$	$\frac{1}{2}$
	$\frac{1}{6}$	$\frac{1}{6}$	$\frac{1}{6}$	$\frac{1}{6}$	$\frac{1}{6}$	$\frac{1}{6}$	1

The table also shows the *marginal probabilities*.

## Example with the bugs

Drawing a random bug out of the aquarium, with (g)reen and (r)ed bugs on (l)and and (w)ater.



	R	G			R	G	
L	2	3	5	L	$\frac{1}{6}$	$\frac{1}{4}$	$\frac{5}{12}$
W	2	5	7	W	$\frac{1}{6}$	$\frac{5}{12}$	$\frac{7}{12}$
	4	8	12		$\frac{1}{3}$	$\frac{2}{3}$	1

Frequency table and probability table.

## Flawed reasoning

Students at an elementary school are given a questionnaire that they are required to return after their parents have completed it.

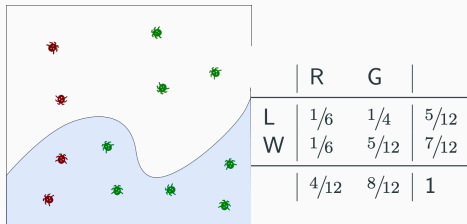
One of the questions asked is, “Do you find that your work schedule makes it difficult for you to spend time with your kids after school” Of the parents who replied, 85% said “no”.

Based on these results, the school officials conclude that a great majority of the parents have no difficulty spending time with their kids after school.

What went wrong?

# Conditional probability

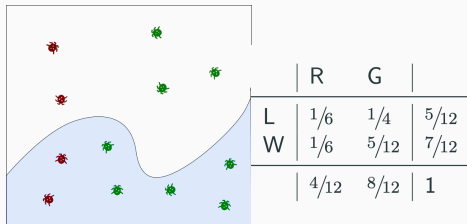
Drawing a random bug out of the aquarium, with (g)reen and (r)ed bugs on (l)and and (w)ater.



$$P(\text{is red}) = 1/3$$

# Conditional probability

Drawing a random bug out of the aquarium, with (g)reen and (r)ed bugs on (l)and and (w)ater.



We catch a red bug. What is the probability it is “dry”: 50%-50%

$$P(\text{lives on land} \mid \text{is red}) = \frac{P(\text{red and land})}{P(\text{is red})} = \frac{2/12}{4/12}$$

## Conditional probability

The *conditional probability* of the event of interest  $A$  given condition  $B$  is calculated as

$$P(A \mid B) = \frac{P(A \cap B)}{P(B)}$$

### Multiplication rule

If  $A$  and  $B$  represent two events, then

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

Note that this formula is simply the conditional probability formula, rearranged.

## Conditional distribution

If we know some event  $B$  occurs, the probability of  $A$  given the new information  $B$  can be calculated as follows:

### Conditional probability

Assume that  $P(B) > 0$ . The conditional probability of  $A$  given  $B$  is defined as

$$P(A | B) = \frac{P(A \cap B)}{P(B)}. \quad (3.1)$$

### Multiplication rule for probabilities

For events  $A$  and  $B$  it holds

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

The multiplication rule is useful to calculate probabilities of multiple events affecting each other.



# Bayes formula

---

## Bayes formula

For events  $A$  and  $B$

$$P(A | B) = \frac{P(B | A)P(A)}{P(B)}$$

Often it is useful to rewrite the denominator  $P(B)$

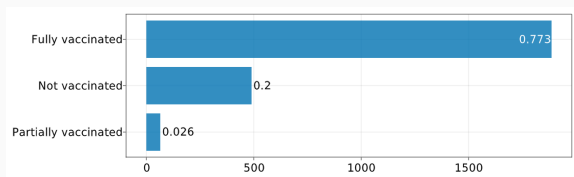
$$P(B) = P(B | A)P(A) + P(B | A^c)P(A^c)$$

# Base rate fallacy

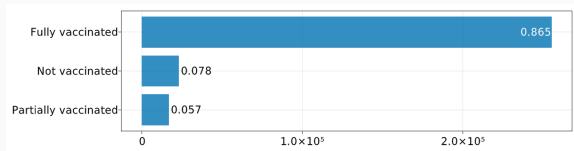
---

## Base rate fallacy

In the fourth wave (July 10 - August 16, 2021) about 2400 (or 0.825 %) people 16 or older in Iceland have been diagnosed with Covid-19:



But (young) adults in Iceland's population are highly vaccinated



$$P(\text{diagn} \mid \text{vacc}) = \frac{P(\text{vacc} \mid \text{diagn})P(\text{diagn})}{P(\text{vacc})} = \frac{0.773 \cdot 0.00825}{0.864} = 0.00738$$

## Base rate fallacy

---

$$P(\text{diagn} \mid \text{vacc}) = \frac{0.773 \cdot 0.00825}{0.864} = 0.00738$$

$$P(\text{diagn} \mid \text{not vacc}) = \frac{0.200 \cdot 0.00825}{0.0783} = 0.0211$$

$$P(\text{diagn} \mid \text{part. vacc}) = \frac{0.0262 \cdot 0.00825}{0.0570} = 0.00379 \text{ (sic!)}$$

## Independent events

Two events  $A$  and  $B$  are independent if knowing whether  $B$  occurred does not change the probability of  $A$

$$P(A \mid B) = P(A).$$

### Independent events

Two events  $A$  and  $B$  are independent if  $P(A \cap B) = P(A)P(B)$ .

Simple example: Throw a 6-sided die. Are  $A = \{5, 6\}$  and  $B = \{1, 3, 5\}$  dependent?

$$P(A)P(B) = \frac{2}{6} \frac{3}{6} = \frac{1}{6}, \quad P(A \cap B) = P(\{5\}) = \frac{1}{6}.$$

If I tell you  $A$  happened, that does not change probabilities of  $B$ :  
 $P(B \mid A) = P(B) = \frac{3}{6}$ .

# Random variables

---

# Random variables

---

## Random variables

A **random variable** is a numeric quantity whose value depends on the outcome of a random experiment.

Example:  $X$  is the number of eyes on a 6-sided die.

We denote random variables with capital letters, often  $X$  or  $Y$ .

Examples?

## Pair of dice

Throw a pair of dice, count the total number of eyes, call that random variable  $X$ . Consider the **event** that  $X = 7$ .

Event? What are the actual  $\omega$  making our event and sample space?  
You could take

$$A = \{\square\boxplus, \boxplus\square, \square\boxtimes, \boxtimes\square, \boxtimes\boxplus, \boxplus\boxtimes\}, \quad \Omega = \{\square\square, \dots, \boxplus\boxplus\}$$

$$P(X = 7) = P(A) = \frac{|A|}{|\Omega|} = \frac{6}{36}$$

Value $k$	2	3	4	5	6	7	8	9	10	11	12
Probability $P(X=k)$	$\frac{1}{36}$	$\frac{2}{36}$	$\frac{3}{36}$	$\frac{4}{36}$	$\frac{5}{36}$	$\frac{6}{36}$	$\frac{5}{36}$	$\frac{4}{36}$	$\frac{3}{36}$	$\frac{2}{36}$	$\frac{1}{36}$



## Pair of dice

The following holds for  $k \in \{2, \dots, 12\}$ :

$$P(X = k) = \frac{6 - |k - 7|}{36}$$

Check:

Value $k$	2	3	4	5	6	7	8	9	10	11	12	other
Probability $P(X=k)$	$\frac{1}{36}$	$\frac{2}{36}$	$\frac{3}{36}$	$\frac{4}{36}$	$\frac{5}{36}$	$\frac{6}{36}$	$\frac{5}{36}$	$\frac{4}{36}$	$\frac{3}{36}$	$\frac{2}{36}$	$\frac{1}{36}$	0

# Discrete random variables

## Discrete random variables

A random variable is called discrete if it is integer-valued or otherwise has only a finite or countable number of values.

Example:  $Y = X/2$  is discrete (but can take non-integers such as  $Y = 5.5$  as values.)

# Probability mass function

## Probability mass function

Define the probability mass function  $f$  of a discrete random variable  $X$  by

$$f(k) = P(X = k).$$

Also  $f(y) = 0$  for all real  $y$  such that  $P(X = y) = 0$ , okay?

Sometimes we write  $f_X$  to talk about  $X$ 's own probability mass function.

## Sum of two dice

---

$$f(k) = \begin{cases} \frac{6-|k-7|}{36} & \text{if } k \in \{2, 3, \dots, 12\} \\ 0 & \text{otherwise} \end{cases}$$

is the probability mass function for the random variable which counts the sum of two dice.

## Two coins

---

Flip two coins... count the number of heads. Call it  $X$ .

$$f(0) = \frac{1}{4}, f(1) = \frac{1}{2} \text{ and } f(2) = \frac{1}{4}$$

$$f(x) = 0 \text{ otherwise if } x \notin \{0, 1, 2\}.$$

Flip two coins... count the number of heads.  $f_X(0) = \frac{1}{4}$ ,  
 $f_X(1) = \frac{1}{2}$  and  $f_X(2) = \frac{1}{4}$ .

What is  $P(X \in \{1, 2\}) = P(1 \leq X \leq 2)$ ?

$$P(1 \leq X \leq 2) = f_X(1) + f_X(2) = \frac{3}{4}$$

Let  $Y = X/2$ . What is  $P(Y > 0)$ ?

$$P(Y > 0) = P(1 \leq X \leq 2) = f_X(1) + f_X(2) = \frac{3}{4}$$

## Rule

For integer valued  $X$

$$P(m \leq X \leq n) = \sum_{k=m}^n f(k)$$

for any integers  $m$  and  $n$ .

## Describing distributions

---

# Probability mass function

---

Not all functions are probability mass functions. Because they describe probability distributions, some conditions must hold.

$f(k)$  is a probability mass function if and only if

- $f(k) \geq 0$  for all  $k$ .
- $\sum_{\text{all } k} f(k) = 1$ .

If somebody gives you a probability mass function, there is a random variable for it.



# Distribution function

## Distribution function

Assume  $X$  is a discrete random variable. Its distribution function is given by

$$F(x) = P(X \leq x) = \sum_{k \leq x} f_X(k),$$

Flip two coins... count the number of heads. Call it  $X$ .

$f(0) = \frac{1}{4}$ ,  $f(1) = \frac{1}{2}$  and  $f(2) = \frac{1}{4}$ . Find  $F$ .

$$F(0) = f(0) = \frac{1}{4}$$

$$F(1) = f(0) + f(1) = \frac{1}{4} + \frac{1}{2}$$

$$F(2) = f(0) + f(1) + f(2) = 1$$

## Distribution function

---

What is the probability to throw  $k$  times heads in a row with a fair coin?

$$f(0) = \frac{1}{2}, \quad f(1) = \frac{1}{2} \cdot \frac{1}{2} = \frac{1}{4}, \quad f(2) = \frac{1}{8}, \quad f(k) = \left(\frac{1}{2}\right)^{k+1}$$

$$P(X > 0) = f(1) + f(2) + f(3) + \dots = 1 - P(X = 0) = 1 - f(0)$$

# Distribution function

---

For  $F(x)$  it holds

- $F(x)$  is increasing
- $F(x) \rightarrow 1$  for  $x \rightarrow \infty$ .
- $F(x) \rightarrow 0$  for  $x \rightarrow -\infty$ .

Also

- $P(a < X \leq b) = F(b) - F(a).$
- $P(X > a) = 1 - F(a).$
- For integer valued random variables:  
 $f(m) = F(m) - F(m - 1).$

# Expected value

We are often interested in the “average” outcome of a random variable.

## Expected value

The expected value of a random variable is defined as

$$E(X) = \sum_{\text{all } k} k f_X(k) \quad \text{if } X \text{ is discrete,}$$

## Recall: the average using fractions

*Data set:* grades of 24 students

5, 5, 6, 5, 6, 6, 6, 5, 5, 7, 6, 7, 5, 5, 5, 6, 6, 6, 5, 6, 5, 7, 6, 7

*Table:*

grade	$x_1 = 7$	$x_2 = 6$	$x_3 = 5$
fraction of students	$p_1 = 4/24$	$p_2 = 10/24$	$p_3 = 10/24$

*Average* One can write the average in different forms

$$\begin{aligned}\text{Average} &= \frac{5 + 5 + 6 + \cdots + 5 + 7 + 6 + 7}{24} \\ &= \frac{7 \cdot 4 + 6 \cdot 10 + 5 \cdot 10}{24} = 7 \cdot \frac{4}{24} + 6 \cdot \frac{10}{24} + 5 \cdot \frac{10}{24} = \sum_{i=1}^3 x_i \cdot p_i\end{aligned}$$

## Expected value

---

The expected value of a discrete random variable  $X$  with finitely many outcomes can also be written as

$$\begin{aligned}\mu = E(X) &= \sum_{\text{all } k} x_k \cdot \underbrace{P(X = x_k)}_{f(x_k)} \\ &= x_1 \cdot P(X = x_1) + x_2 \cdot P(X = x_2) + \cdots + x_n \cdot P(X = x_n)\end{aligned}$$

Here  $x_i$  are the  $n$  possible outcomes and  $P(X = x_i)$  are the probabilities of each outcome.

## Expected value

---

Flip two coins... count the number of heads.

$$f(0) = \frac{1}{4}, f(1) = \frac{1}{2} \text{ and } f(2) = \frac{1}{4}$$

$$E(X) = 0 \cdot \frac{1}{4} + 1 \cdot \frac{1}{2} + 2 \cdot \frac{1}{4} = 1$$



## Rules for computing expected values

---

For the expected value,

- $E(a) = a$ .
- $E(aX) = aE(X)$ .
- $E(aX + b) = aE(X) + b$ .
- $E(X + Y) = E(X) + E(Y)$ .

Here  $X$  and  $Y$  are any two random variables and  $a$  and  $b$  are constants.

If we transform the random variables by a function  $h$  we have:

**Theorem** ♡

$$E(h(X)) = \sum_{\text{all } k} h(k) f(k)$$

Coin example (with  $h(x) = x/2$ ):

$$\begin{aligned} E(X/2) &= \frac{0}{2} \cdot f_X(0) + \frac{1}{2} \cdot f_X(1) + \frac{2}{2} \cdot f_X(2) = \frac{1}{2} \\ &= (E(X))/2 \end{aligned}$$

## Common discrete distributions

---

## Bernoulli distribution

---

The **Bernoulli distribution** describes a random experiment that can either succeed (with probability  $p$ ) or fail (with probability  $1 - p$ .) Suppose we make a random experiment which succeeds with probability  $p$  and set

$$X = \begin{cases} 1, & \text{if the experiment succeeds} \\ 0, & \text{in case of failure.} \end{cases}$$

We have  $f(1) = p$  and  $f(0) = 1 - p$ .

Sometimes useful to write as  $f(k) = p^k(1 - p)^{1-k}$  for  $k \in \{0, 1\}$ .

# The binomial distribution

---

## Bernoulli distribution

A random variable  $X$  is Bernoulli distributed if it has probability mass function  $f(1) = p$  and  $f(0) = 1 - p$  and  $= 0$  otherwise. We write  $X \sim \text{Ber}(p)$ .

Examples?

# The binomial distribution

---

The **binomial distribution** describes the probability of having exactly  $k$  successes in  $n$  independent Bernoulli trials with probability of success  $p$ .

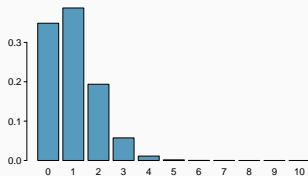
If  $X$  is binomial with parameters  $n$  and  $p$  we write:

$$X \sim \text{Bin}(n, p)$$

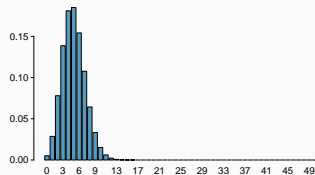
Ha, the sum of two coins with sides 0 and 1 is  $\text{Bin}(2, 0.5)$  distributed.

# The binomial distribution

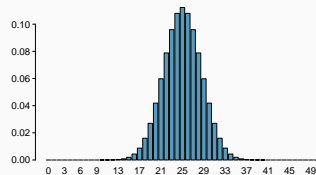
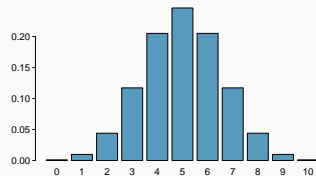
$n = 10$



$n = 50$



$p = 0.1$



$p = 0.5$

# The binomial distribution

The **binomial distribution** describes the probability of having exactly  $k$  successes in  $n$  independent Bernoulli trials with probability of success  $p$ .

If  $X$  is binomial with parameters  $n$  and  $p$  we write:

$$X \sim \text{Bin}(n, p)$$

## Binomial distribution

A random variable  $X$  is binomial distributed with parameters  $n, p$  if

$$P(X = k) = \binom{n}{k} p^k (1 - p)^{n-k} \qquad \binom{n}{k} = \frac{n!}{k!(n-k)!}$$



## Sum of binomial distributed random variables

### Sum of binomial distributed random variables.

If  $X_1 \sim \text{Bin}(n, p)$  and  $X_2 \sim \text{Bin}(m, p)$  are independent, then  $X_1 + X_2 \sim \text{Bin}(m + n, p)$ .

("Dropping  $m$  items, counting the broken ones, dropping  $n$  more items, counting the additional broken ones is the same as dropping  $m + n$  items...")

## Geometric distribution

---

The experiment consists of a series of independent Bernoulli trials with probability of success equal to  $p$ .

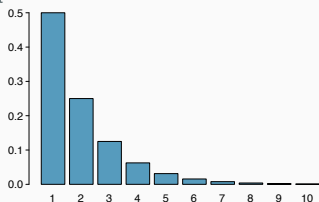
The random variable  $X$  denotes the number of trials needed to get the first success.

$p$  is called the parameter of  $X$ .

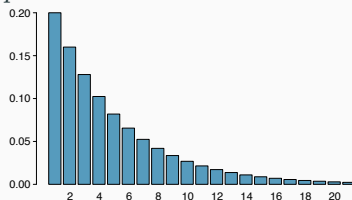
# The geometric distribution

The **geometric distribution** describes the probability distribution of the number of trials needed  $k$  to get the first success, for a single event succeeding with probability  $p$ . ( $k - 1$  failures and 1 success.)

$p = 0.5$



$p = 0.2$



# The geometric distribution

## Geometric distribution

A random variable  $X$  is geometrically distributed with parameters  $p$  if

$$P(X = k) = (1 - p)^{k-1}p, \quad k = 1, 2, \dots$$

We write  $X \sim \text{Geom}(p)$ .

# Continuous distributions

---

# Continuous distributions

## Continuous random variables

A continuous random variable can assume all values in one or several intervals of real numbers, and the probability of assuming a particular value is zero.

A continuous random variable  $X$  is described by its *probability density function (pdf)*  $f(x)$

$$P(a \leq X \leq b) = \int_a^b f(x)dx.$$

$$P(X = x) = 0$$

and

$$P(a \leq X \leq b) = P(a < X \leq b) = P(a \leq X < b) = P(a < X < b)$$

## Probability density function (pdf)

A function is a probability density function (pdf) if and only if

$$f(x) \geq 0 \quad \text{and} \quad \int_{-\infty}^{\infty} f(x)dx = 1.$$

## Example

Show that the function

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

is a pdf.

$$f(x) \geq 0 \quad \checkmark.$$

$$\begin{aligned} \int_{-\infty}^{+\infty} f(t) dt &= \int_{-\infty}^a 0 dt + \int_a^b \frac{1}{b-a} dt + \int_b^{\infty} 0 dt \\ &= \int_a^b \frac{1}{b-a} dt = \frac{b-a}{b-a} = 1 \quad \checkmark. \end{aligned}$$



## Cumulative distribution function

The cumulative distribution function  $F$  of a continuous distribution is

$$F(x) = P(X \leq x) = \int_{-\infty}^x f(t)dt.$$

$$P(a \leq X \leq b) = F(b) - F(a)$$

## Example

---

Find cumulative distribution function for  $X$  with pdf

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

$$F(x) = \int_{-\infty}^x f(t)dt = \begin{cases} 0 & x \leq a \\ \frac{x-a}{b-a} & x \in [a, b] \\ 1 & x \geq b. \end{cases}$$

## Expected value

The expected value is an “average” outcome of a random variable.

### Expected value

The expected value of a random variable is defined as

$$E(X) = \begin{cases} \int_{-\infty}^{\infty} xf(x)dx & \text{if } X \text{ is continuous,} \\ \sum_{\text{all } k} kf(k) & \text{if } X \text{ is discrete.} \end{cases}$$

## Rules for computing expected values

---

For the expected value,

- $E(a) = a$ .
- $E(aX) = aE(X)$ .
- $E(aX + b) = aE(X) + b$ .
- $E(X + Y) = E(X) + E(Y)$ .

Here  $X$  and  $Y$  are two random variables and  $a$  and  $b$  are constants.

The *same* rules:  $E$  is a linear operator on random variables.

# Uniform distribution

## Uniform distribution

The continuous distribution with pdf

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

is called the *uniform distribution*. Facts:  $EX = (a + b)/2$ .

$$EX = \int_{-\infty}^{\infty} xf(x)dx = \frac{1}{b-a} \int_a^b tdt = \frac{\frac{1}{2}b^2 - \frac{1}{2}a^2}{a-b} = (a+b)/2$$

If we transform the random variables by a function  $h$  we have:

## Theorem

$$E(h(X)) = \begin{cases} \sum_{\text{all } k} h(k)f(k), & \text{if } X \text{ is discrete,} \\ \dots \\ \int_{-\infty}^{\infty} h(x)f(x)dx, & \text{if } X \text{ is continuous.} \end{cases}$$

# Variance

---

## Variance and standard deviation

### Variance

The variance of a random variable is defined as

$$V(X) = E[(X - \mu)^2],$$

where  $\mu = E(X)$  is the expected value of  $X$ .

In words, this is the expected squared deviation of the mean. The variance can be calculated by

$$V(X) = \begin{cases} \sum_{\text{all } k} (k - \mu)^2 f(k), & \text{for discrete } X \\ \int_{-\infty}^{\infty} (x - \mu)^2 f(x) dx, & \text{for continuous } X. \end{cases}$$

Sometimes it is easiest to compute  $V(X) = E(X^2) - \mu^2$ .

The standard deviation of a random variable  $X$  is defined as  $\sigma = \sqrt{V(X)}$ .



## Rules for computing variance

---

For the variance

- $V(a) = 0$ .
- $V(aX) = a^2V(X)$ .
- $V(aX + b) = a^2V(X)$ .
- $V(X + Y) = V(X) + V(Y)$ , if  $X$  and  $Y$  are **independent**.

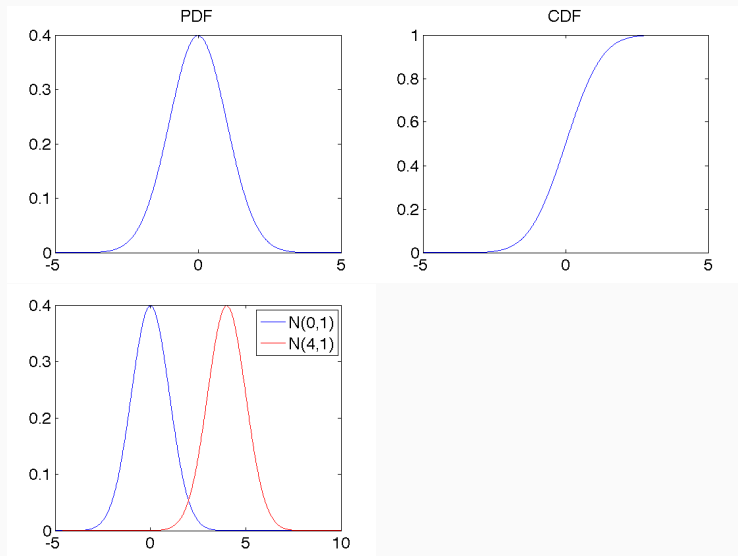
Here  $X$  and  $Y$  are two random variables and  $a$  and  $b$  are constants.

# Normal distributions

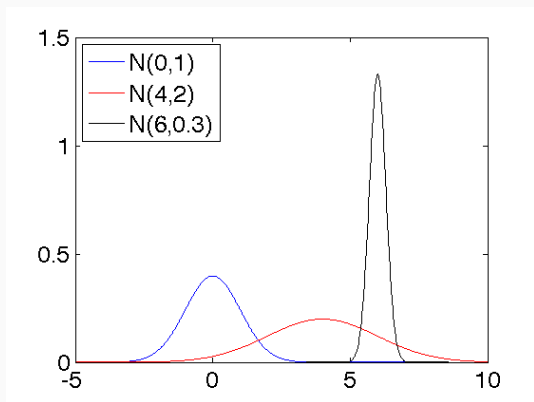
---

# Normal distribution

Density and distribution function of  $Z \sim N(0, 1)$  and  $N(4, 1)$



## pdf's for some other possible parameters



# Normal distribution

## Normal distribution $N(\mu, \sigma^2)$

A continuous  $X$  is normally distributed,  $N(\mu, \sigma^2)$ , with parameters  $\mu \in \mathbb{R}$  and  $\sigma > 0$ , if it has pdf

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

The distribution function is

$$F(x) = \int_{-\infty}^x = \dots \text{has no nice solution}$$

## Parameters

If  $X \sim N(\mu, \sigma^2)$  then  $E(X) = \mu$  and  $V(X) = \sigma^2$ .

## Normal distribution pdf

---

# Standard normal distribution

---

## Standard normal distribution

A continuous random variable  $Z$  is standard normally distributed if  $Z \sim N(0, 1)$ .  $E[Z] = 0$  and  $\text{Var}(Z) = 1^2$ .

We denote pdf and cdf by  $\varphi(x)$  and  $\Phi(x)$

## Theorem

If  $X \sim N(\mu, \sigma^2)$  then  $aX + b \sim N(a\mu + b, a^2\sigma^2)$ .

That means for  $X \sim N(\mu, \sigma^2)$  that

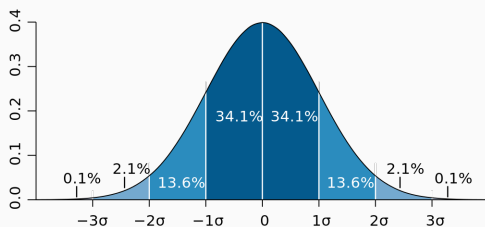
- $X = \mu + \sigma Z$  where  $Z \sim N(0, 1)$ .
- $Z = (X - \mu)/\sigma \sim N(0, 1)$ .

We use this to sample random variables, and to compute probabilities:

$$P(X < x) = P\left(\frac{X - \mu}{\sigma} < \frac{x - \mu}{\sigma}\right) = P\left(Z < \frac{x - \mu}{\sigma}\right) = \Phi\left(\frac{x - \mu}{\sigma}\right).$$



# Rule



Example: IQ values are normalized such that (approximately)

$$IQ \sim N(100, 15^2)$$

What is the probability that a random person scores 115 or more?  
Approx.  $13.6 + 2.1 + 0.1 = 15.8$ .

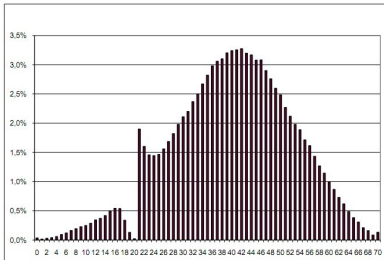
# Relict of the past: Normal distribution table

Table gives  $\Phi(z) = P(X \leq z)$  for  $Z \sim N(0, 1)$ .  
For negative values use that  $\Phi(-z) = 1 - \Phi(z)$ .

$z$	.00	.01	.02	.03	.04	.05	.06	.07	.08	.09
0.0 :	.5000	.5040	.5080	.5120	.5160	.5199	.5239	.5279	.5319	.5359
0.1 :	.5398	.5438	.5478	.5517	.5557	.5596	.5636	.5675	.5714	.5753
0.2 :	.5793	.5832	.5871	.5910	.5948	.5987	.6026	.6064	.6103	.6141
0.3 :	.6179	.6217	.6255	.6293	.6331	.6368	.6406	.6443	.6480	.6517
0.4 :	.6554	.6591	.6628	.6664	.6700	.6736	.6772	.6808	.6844	.6879
0.5 :	.6915	.6950	.6985	.7019	.7054	.7088	.7123	.7157	.7190	.7224
0.6 :	.7257	.7291	.7324	.7357	.7389	.7422	.7454	.7486	.7517	.7549
0.7 :	.7580	.7611	.7642	.7673	.7704	.7734	.7764	.7794	.7823	.7852
0.8 :	.7881	.7910	.7939	.7967	.7995	.8023	.8051	.8078	.8106	.8133
0.9 :	.8159	.8186	.8212	.8238	.8264	.8289	.8315	.8340	.8365	.8389
1.0 :	.8413	.8438	.8461	.8485	.8508	.8531	.8554	.8577	.8599	.8621
1.1 :	.8643	.8665	.8686	.8708	.8729	.8749	.8770	.8790	.8810	.8830
1.2 :	.8849	.8869	.8888	.8907	.8925	.8944	.8962	.8980	.8997	.9015
1.3 :	.9032	.9049	.9066	.9082	.9099	.9115	.9131	.9147	.9162	.9177
1.4 :	.9192	.9207	.9222	.9236	.9251	.9265	.9279	.9292	.9306	.9319
1.5 :	.9332	.9345	.9357	.9370	.9382	.9394	.9406	.9418	.9429	.9441
1.6 :	.9452	.9463	.9474	.9484	.9495	.9505	.9515	.9525	.9535	.9545
1.7 :	.9554	.9564	.9573	.9582	.9591	.9599	.9608	.9616	.9625	.9633
1.8 :	.9641	.9649	.9656	.9664	.9671	.9678	.9686	.9693	.9699	.9706
1.9 :	.9713	.9719	.9726	.9732	.9738	.9744	.9750	.9756	.9761	.9767
2.0 :	.9772	.9778	.9783	.9788	.9793	.9798	.9803	.9808	.9812	.9817
2.1 :	.9821	.9826	.9830	.9834	.9838	.9842	.9846	.9850	.9854	.9857
2.2 :	.9861	.9864	.9868	.9871	.9875	.9878	.9881	.9884	.9887	.9890
2.3 :	.9893	.9896	.9898	.9901	.9904	.9906	.9909	.9911	.9913	.9916
2.4 :	.9918	.9920	.9922	.9925	.9927	.9929	.9931	.9932	.9934	.9936
2.5 :	.9938	.9940	.9941	.9943	.9945	.9946	.9948	.9949	.9951	.9952

# High-school maturity exam in Poland

2.1. Poziom podstawowy



Wykres 1. Rozkład wyników na poziomie podstawowym

Histogram showing the distribution of scores for the obligatory Polish language test. "The dip and spike that occurs at around 21 points just happens to coincide with the cut-off score for passing the exam"

<http://freakonomics.com/2011/07/07/>

another-case-of-teacher-cheating-or-is-it-just-altruism/

## Moment generating function (m.g.f.)

---

Let  $X$  be a random variable

- The  $k^{\text{th}}$  moment for  $X$  is defined by  $E[X^k]$ .
- The moment generating function for  $X$  is defined by

$$m_X(t) = E[e^{tX}].$$

- Let  $m_X(t)$  be the m.g.f for  $X$ . Then

$$\left. \frac{d^k m_X(t)}{dt^k} \right|_{t=0} = E[X^k]$$

# Moment generating function for standard normal distribution

---

Let  $Z \sim N(0, 1)$ . Compute the mgf. Use  $h(x) = e^{tx}$  and transform:

$$\begin{aligned} m_X(t) &= \mathbb{E}[e^{tX}] = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} \underbrace{e^{tx}}_{h(x)} e^{-\frac{1}{2}x^2} dx \\ &= \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} e^{-\frac{1}{2}(x-t)^2} e^{\frac{1}{2}t^2} dx = e^{\frac{1}{2}t^2} \end{aligned}$$

## We have seen

---

- Probability mass functions  $fP(X = k) = \binom{n}{k} p^k (1-p)^{n-k}$ .
- Bernoulli – Bernoulli( $p$ ):  $X \in \{0, 1\}$
- Binomial – Bin( $n, p$ ):  $X \in \{0, 1, \dots, n\}$
- Geometric – Geom( $p$ ):  $X \in \{1, 2, 3, 4, \dots\}$
- Normal – N( $\mu, \sigma^2$ ):  $X \in (-\infty, \infty)$

What was the mean and the variance of  $X \sim \text{Bin}(n, p)$ ?

$$E(X) = np. \quad \text{Var}(X) = np(1 - p).$$

### Normal approximation of Binomial distribution

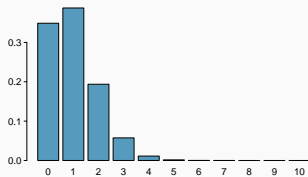
If  $X \sim \text{Bin}(n, p)$ ,  $X$  is approximately normally distributed with mean  $np$  and variance  $np(1 - p)$ ,

$$X \stackrel{\text{approx.}}{\sim} N(np, np(1 - p)),$$

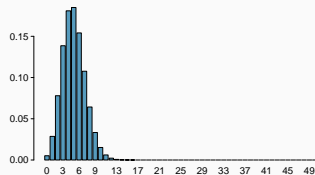
if both  $np > 5$  and  $n(1 - p) > 5$ .

# Normal approximation

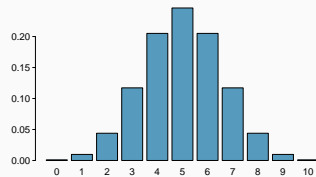
$n = 10$



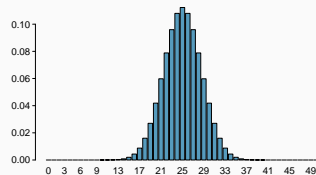
$n = 50$



$p = 0.1$



$p = 0.5$





## Discrete distributions today

---

- Poisson distribution –  $\text{Poisson}(\mu)$ : model the number of events that occur in a time interval, in a region or in some volume.
- Negative binomial distribution –  $\text{nBin}(r, p)$ : The number of trials  $X$  in a sequence of independent  $\text{Bernoulli}(p)$  trials before  $r$  successes occur
- Hypergeometric distribution –  $\text{Hyp}(N, n, r)$ : Draw sample of  $n$  objects without replacement out of  $N$ . The random variable  $X$  is the number of marked objects.

# Poisson distribution

---

The **Poisson distribution** is often used to model the number of events that occur in a time interval, in a region or in some volume.

(Named after Simeon Denis Poisson, 1781-1840.)

Some examples where this distribution fits well are

- The number of particles emitted per minute (hour, day) of a radioactive material.
- Call connections routed via a cell tower (GSM base station).

## Poisson distribution

$$X \sim \text{Poisson}(\mu)$$

A random variable  $X$  has Poisson distribution with parameter  $\mu$  if

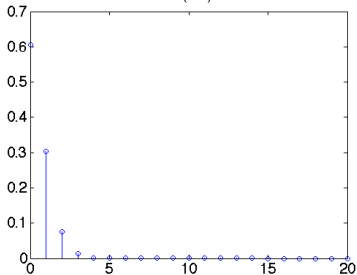
$$P(X = k) = \frac{e^{-\mu} \mu^k}{k!}, \quad k \in \{0, 1, 2, \dots\}.$$

## Sum of Poisson distributed random variables.

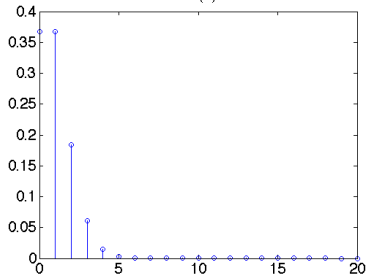
If  $X_1 \sim \text{Poisson}(\mu_1)$  and  $X_2 \sim \text{Poisson}(\mu_2)$  are independent, then  $X_1 + X_2 \sim \text{Poisson}(\mu_1 + \mu_2)$ .

# Poisson distribution

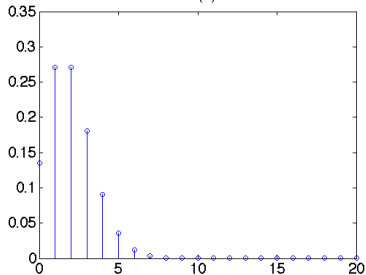
Po(0.5)



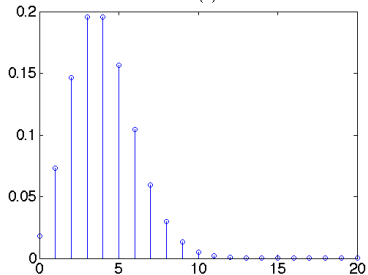
Po(1)



Po(2)



Po(4)





Number of chewing gums on a tile is approximately Poisson.

## Example

---

Let  $X$  be the number of typos on a printed page with a mean of 3 typos per page. Assume the typos occur independently of each other.

1. What is the probability that a randomly selected page has at least one typo on it?

$$P(X \geq 1) = 1 - P(X = 0) = 1 - f(0) = 1 - e^{-3}$$

2. What is the probability that three randomly selected pages have more than eight typos on it?

In this case  $\lambda = 9$  since we have in average 9 typos on three printed pages.

$$P(X > 8) = 1 - P(X \leq 8) \approx 1 - 0.456 \text{ by table II page 692}$$

# Poisson distribution as limit of a Binomial distribution

The Poisson distribution appears as limit of the Binomial distribution if  $n$  becomes large and  $p$  goes to 0:

## Theorem

Let  $n \rightarrow \infty$ ,  $p \rightarrow 0$ , and also  $np \rightarrow \mu$ . Then for fix  $k \geq 0$

$$\binom{n}{k} p^k (1-p)^{n-k} \rightarrow \frac{\mu^k e^{-\mu}}{k!} \quad (9.1)$$

Connection to the previous example:

- There is a large number  $n$  of atoms in the material and the probability that an atom decays in a unit of time  $p$  is very small.

# Negative binomial distribution

---

The number of trials  $X$  in a sequence of independent Bernoulli( $p$ ) trials before  $r$  successes occur has the **negative binomial distribution**.



## Negative binomial distribution

$$X \sim \text{nBin}(r, p)$$

The random variable  $X$  has a negative binomial distribution with parameter  $r$  and  $p$  if

$$P(X = k) = \binom{k-1}{r-1} p^r (1-p)^{k-r}, \quad k = r, r+1, \dots$$

Motivation: Probability of  $r$  successes in  $k$  trials:  $(1-p)^{k-r} p^r$ . The last attempt succeeds. The binomial coefficient gives the number of ways we assign the remaining  $r-1$  successes to the remaining  $k-1$  trials.

# Hypergeometric distribution

---

- Suppose we have  $N$  objects of which  $r$  are “marked”.
- Draw sample of  $n$  objects without replacement. The random variable  $X$  is the number of marked objects. Then  $X$  has hypergeometric distribution with parameters  $N, n, r$ .

## Hypergeometric distribution

$$X \sim \text{Hyp}(N, n, r)$$

The random variable  $X$  has hypergeometric distribution with parameters  $N$ ,  $n$  and  $r$  if

$$P(X = k) = \frac{\binom{r}{k} \binom{N-r}{n-k}}{\binom{N}{n}} \quad \max(0, n + r - N) \leq k \leq \min(n, r)$$

If  $n = 1$  then  $\text{Hyp}(N, 1, r) = \text{Bernoulli}(r/N)$ . If  $N$  and  $r$  are large compared to  $n$  we have  $\text{Hyp}(N, n, r) \approx \text{Bin}(n, r/N)$ .



## Continuous distributions today (all positive)

---

- Exponential distribution –  $\text{Exp}(\lambda)$ : Time between calls/visitors/people knocking on your door. (Poisson: How many ticks. Exponential: time between ticks.)
- Gamma distribution –  $\Gamma(\alpha, \beta)$ : Flexible distribution for positive random variables.
- $\chi^2$ -distribution –  $\chi^2(n)$ : Distribution for sum of squares of  $n$  independent  $N(0, 1)$  random variables.

## Exponential distribution

$$X \sim \text{Exp}(\lambda)$$

The density function of an **exponential distribution** with rate  $\lambda$  or is given by

$$f(x) = \lambda e^{-\lambda x}, \quad x \geq 0$$

or equivalently  $f(x) = \frac{1}{\beta} e^{-x/\beta}$  where  $\beta = \frac{1}{\lambda}$  is the scale.

$$E[X] = \beta \text{ and } \text{Var}(X) = \beta^2$$

The cumulative distribution function is given by

$$F(x) = 1 - e^{-\lambda x}.$$

# Exponential distribution

---

Assume objects arrive after exponentially distributed interarrival times.

$\lambda$  - how many arrivals per time unit.

$\beta$  - expected waiting time

## Gamma distribution

$$X \sim \text{Gamma}(\alpha, \beta)$$

A random variable  $X$  with density function

$$f(x) = \frac{1}{\Gamma(\alpha)\beta^\alpha} x^{\alpha-1} e^{-x/\beta}, \quad x > 0$$

for  $\beta > 0$  and  $\alpha > 0$  has a **Gamma distribution** with parameters shape  $\alpha$  and scale  $\beta$ , or .

$$E[X] = \alpha\beta \text{ and } \text{Var}(X) = \alpha\beta^2.$$

If  $X$  follows a Gamma distribution with parameters  $\alpha$  and  $\beta$ , then the m.g.f is given by  $m_X(t) = (1 - \beta t)^{-\alpha}$ .



## $\chi^2$ -distribution

$$X \sim \chi^2(n)$$

The Gamma distribution with parameters  $\beta = 2$  and  $\alpha = \frac{n}{2}$  is called  $\chi^2$ -distribution with  $n$  degrees of freedom.

$$E[X] = n \text{ and } \text{Var}(X) = 2n.$$

## Sum of squares

If  $Z_1, \dots, Z_n$  have standard normal distributions and are independent, then  $Z_1^2 + \dots + Z_n^2$  follow a  $\chi^2$ -distribution with  $n$  degrees of freedom.

## Moment generating function (m.g.f.)

---

Let  $X$  be a random variable

- The  $k^{\text{th}}$  moment for  $X$  is defined by  $E(X^k)$ .
- The moment generating function for  $X$  is defined by

$$m_X(t) = E(e^{tX}).$$

- Let  $m_X(t)$  be the m.g.f for  $X$ . Then

$$\left. \frac{d^k m_X(t)}{dt^k} \right|_{t=0} = E(X^k)$$

# Moment generating function for standard normal distribution

---

Let  $Z \sim N(0, 1)$ . Compute the mgf. Use  $h(x) = e^{tx}$  and transform:

$$\begin{aligned} m_X(t) &= \mathbb{E}[e^{tX}] = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} \underbrace{e^{tx}}_{h(x)} e^{-\frac{1}{2}x^2} dx \\ &= \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} e^{-\frac{1}{2}(x-t)^2} e^{\frac{1}{2}t^2} dx = e^{\frac{1}{2}t^2} \end{aligned}$$

# Bivariate distributions

## Definition

Informal: A two-dimensional or bivariate random variable  $(X, Y)$  produces a pair of random numbers.

For discrete random variables we have the joint density (probability mass function)

$$f_{X,Y}(i, j) = P(X = i, Y = j) = P(X = i \text{ and } Y = j).$$

Here  $f_{X,Y}(i, j) \geq 0$  and  $\sum_{\text{all } i, j} f_{X,Y}(i, j) = 1$ .

## Example

Let  $X$  and  $Y$  be the number of girls, respectively boys in a randomly chosen Swedish family. The joint density function  $f_{X,Y}(x,y)$  is given in the table below.

$Y$	0	1	2	3	4
$X$					
0	0.38	0.16	0.04	0.01	0.01
1	0.17	0.08	0.02		
2	0.05	0.02	0.01		
3	0.02	0.01			
4	0.02				

$$\sum_{\text{all } x,y} f_{X,Y}(x,y) = 1$$

$$P(X = 0 \text{ and } Y = 1) = f_{X,Y}(0,1) = 0.16$$

$$P(X = 2) = f_{X,Y}(2,0) + f_{X,Y}(2,1) + f_{X,Y}(2,2) = 0.08$$

## Expected values ♡

$$E(h(X, Y)) = \sum_{\text{all } i, j} h(i, j) f_{X,Y}(i, j).$$

For example:

$$E(X + Y) = \sum_{\text{all } i, j} (i + j) f_{X,Y}(i, j)$$

with  $h(i, j) = i + j$ .

## Expected number of children

$X$  and  $Y$  be the number of girls, respectively boys in a randomly chosen Swedish family.

$E(X + Y)$  is the expected number of girls + boys = children. So  $h(i, j) = i + j$ .

$Y$	0	1	2	3	4
$X$					
0	0.38	0.16	0.04	0.01	0.01
1	0.17	0.08	0.02		
2	0.05	0.02	0.01		
3	0.02	0.01			
4	0.02				

$$E(X + Y) = (0 + 0) \cdot 0.38 + (1 + 0) \cdot 0.17 + \dots = 1.08$$

## Marginal distributions

Given a pair of discrete random variables  $(X, Y)$  with joint density  $f_{X,Y}$  density for  $X$  and  $Y$  are given by

$$f_X(i) = \sum_{\text{all } j} f_{X,Y}(i, j)$$
$$f_Y(j) = \sum_{\text{all } i} f_{X,Y}(i, j).$$

and called **marginal densities** (marginal p.m.f.'s.)



	Y	0	1	2	3	4	$f_X$
X							
0		0.38	0.16	0.04	0.01	0.01	0.60
1		0.17	0.08	0.02			0.27
2		0.05	0.02	0.01			0.08
3		0.02	0.01				0.03
4		0.02					0.02
$f_Y$		0.64	0.27	0.07	0.01	0.01	1

## Continuous bivariate random variables

---

For a pair of continuous random variables: a function  $f_{X,Y}(x,y)$  with properties

1.  $f_{X,Y}(x,y) \geq 0$ ,

2.  $\int \int f_{X,Y}(x,y) dx dy = 1$ , and

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



## Marginal distributions

For a bivariate continuous random variable  $(X, Y)$ , the probability density functions for  $X$  and  $Y$  are given by

$$f_X(x) = \int f_{X,Y}(x, y) dy$$

$$f_Y(y) = \int f_{X,Y}(x, y) dx$$

## Expected value

For a two-dimensional random variable  $(X, Y)$ , the expected values of  $X$  and  $Y$  are given by

$$E(X) = \begin{cases} \sum_{\text{all } i, j} i f_{X,Y}(i, j), & \text{for } X \text{ discrete,} \\ \int \int x f_{X,Y}(x, y) dx dy, & \text{for } X \text{ continuous,} \end{cases}$$

and

$$E(Y) = \begin{cases} \sum_{\text{all } i, j} j f_{X,Y}(i, j), & \text{for } Y \text{ discrete,} \\ \int \int y f_{X,Y}(x, y) dx dy, & \text{for } Y \text{ continuous.} \end{cases}$$

## Conditional distribution

The *conditional distribution of  $X$  given  $Y = y$*  is defined by its density

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

provided that  $f_Y(y) > 0$ .

### Independent random variables

Two random variables  $X$  and  $Y$  are called independent if their bivariate density can be written as product of the marginal densities:

$$f_{X,Y}(u, v) = f_X(u)f_Y(v).$$

There is no “samvariation”, knowing  $X$  does not explain  $Y$ , etc.



## Covariance

Covariance between random variables  $X$  and  $Y$  is defined as  $\text{Cov}(X, Y) = E[(X - \mu_X)(Y - \mu_Y)]$ , where  $\mu_X = E(X)$  and  $\mu_Y = E(Y)$ .

- According to the definition,

$$\text{Cov}(X, Y) = \begin{cases} \sum_{\text{all } i, j} (i - \mu_X)(j - \mu_Y)f_{X,Y}(i, j), & \text{discrete} \\ \int \int (x - \mu_X)(y - \mu_Y)f_{X,Y}(x, y)dx dy, & \text{cont.} \end{cases}$$

- Note that  $\text{Cov}(X, X) = V(X)$ .
- **If**  $X$  and  $Y$  are independent, **then**  $\text{Cov}(X, Y) = 0$  and  $E(XY) = E(X)E(Y)$ .
- Unit??



## Rules for covariance

---

$\text{Cov}(X, Y)$  can be calculated as

$$\text{Cov}(X, Y) = E(XY) - E(X)E(Y).$$

For two random variables  $X$  and  $Y$ , and two numbers  $a$  and  $b$  we have

$$V(aX + bY) = a^2V(X) + b^2V(Y) + 2ab \text{ Cov}(X, Y).$$

Examples:

$$V(2X) = V(X + X) = V(X) + V(X) + 2 \text{ Cov}(X, X) = 4V(X)$$

$$V(X + Y) = V(X) + V(Y) \text{ when } X \text{ and } Y \text{ are independent}$$

---

(“Fun” thing to do: look up the law of cosines.)

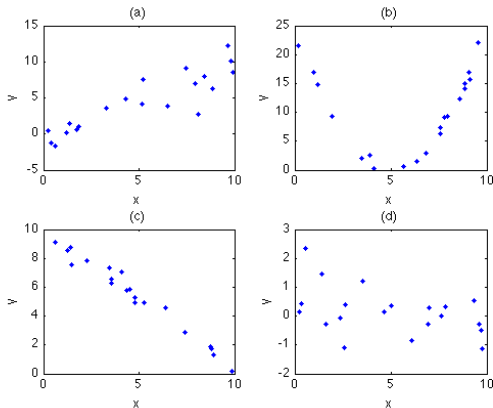
## Correlation

The correlation coefficient is defined as

$$\rho(X, Y) = \frac{\text{Cov}(X, Y)}{\sqrt{V(X)V(Y)}}.$$

- A measure of linear relationship (linjär samvariation) of  $X$  and  $Y$ .
- It holds  $-1 \leq \rho \leq 1$ .
- $X$  and  $Y$  are called **uncorrelated** if  $\rho(X, Y) = 0$  (there is no “linjär samvariation”).
- Unit??

# Visualisation



Assume 2d measurements  $(x_i, y_i)$ . A scatter plot is a two-dimensional plot in which each  $(x_i, y_i)$  measurement is represented as a point in the  $x$ - $y$ -plane.

## Descriptive statistic for bivariate data

The sample covariance is defined as,

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

and sample correlation coefficient is defined as

$$r_{xy} = \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^n (x_i - \bar{x})^2} \sqrt{\sum_{i=1}^n (y_i - \bar{y})^2}} = \frac{c_{xy}}{s_x s_y}$$

The sample correlation is a measure of linear dependence.

In the picture  $r_{xy} = 0.8067$  i (a),  $r_{xy} = 0.2912$  i (b),  
 $r_{xy} = -0.9884$  i (c), och  $r_{xy} = 0.3640$  i (d).

We have the following relationship between dependence and correlation:

- If  $X$  and  $Y$  are independent, then they are also uncorrelated.
- (Thus if  $X$  and  $Y$  are uncorrelated, they do not need to be independent.)

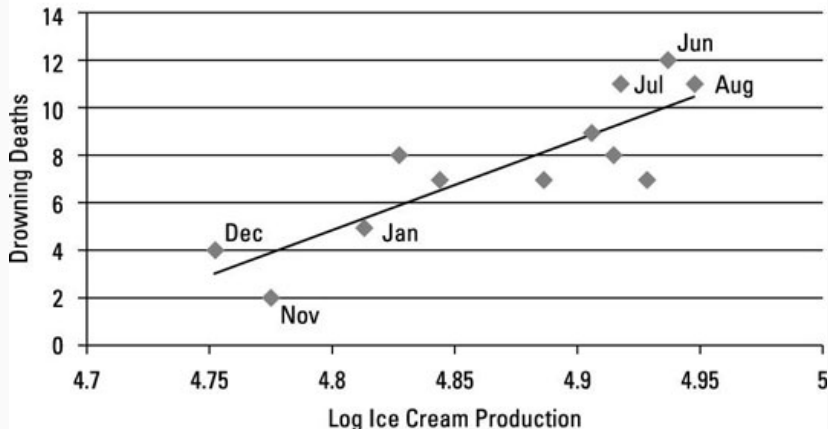
This is natural because two random variables are independent if there is no “samvariation” at all, while they are not correlated if there is no “*linjär* samvariation”.

## Correlation, dependence and causality

---

- Correlation does not say anything about causality!\*
- Sometimes correlation is present but can be explained by a third variable which was not measured.
- Month with high ice cream sales tend to have more drowning accidents. Time to ban ice cream? In this example, an important variable which perhaps was not measured is the sunshine. Such variables are sometimes called **confounding variables**.

Ice Cream and Drowning Scatter, 2006

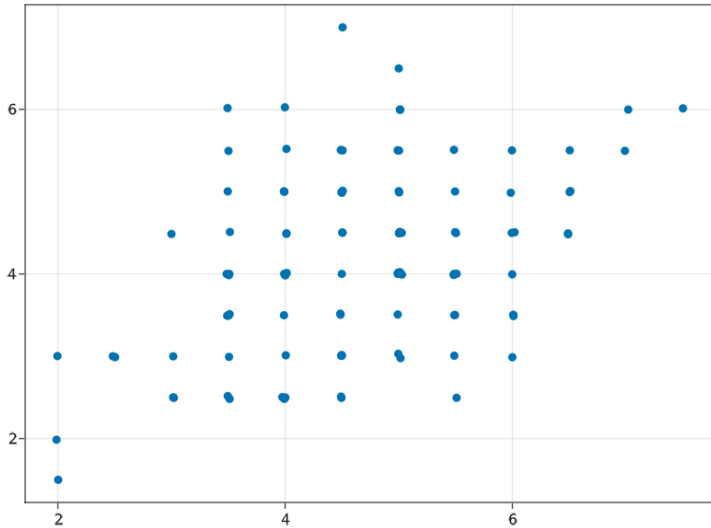


<https://twitter.com/dannagal/status/1244082688899919872>,  
October 20, 2021

- Correlation can also be introduced by selection effects.
- Exam with two questions, one difficult, one easy. A student achieves  $X$  out of 10 points on the easy question,  $Y$  out of 10 points on the difficult question (random).
- Say  $X$  and  $Y$  slightly positively correlated. *But* only students with  $X + Y \geq 10$  pass. Say I tell you the student has passed.
- Passing students performance on easy questions may now be negatively correlated with performance on the difficult question.



# Exam points

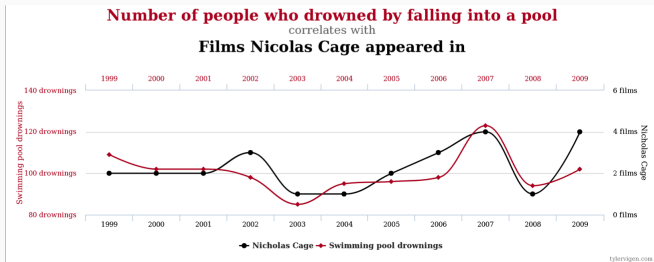


# Causality

---

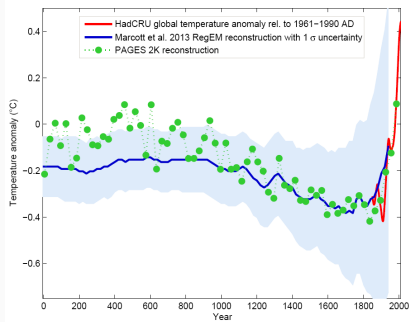
- If we want to know/predict what will change if we perform an action we need insight into causality.
- Will the number of drowning accidents change if we ban ice?
- There are many causal statements in the news!
  - “Do not skip breakfast if you want to reduce the risk of coronary heart disease”
- Be careful...
  - Candidate for a confounding variable: stress.
- We need to understand the science to answer causal questions!  
We will come back to this later.

# Cherry picking



<http://www.tylervigen.com/spurious-correlations>

# Thinking statistics: Global warming



Two millennia of mean surface temperatures according to different reconstructions from climate proxies with the instrumental temperature record overlaid in red.

Stefan Rahmstorf: Paleoclimate: The End of the Holocene.

<http://www.realclimate.org/index.php/archives/2013/09/paleoclimate-the-end-of-the-holocene/>.

Web. 3 Feb. 2019.

## Markov chains

---

The weather in the land Oz is R (rainy), S (sunny) or C (cloudy).  
Weather of the last 30 days:

*RRRRRRRCCSCCSRCSSRRRRRCRSCSCCRCRS...*

What do you expect for the weather of tomorrow? There have been no two nice days in a row and after sun we have 2 times rain, 3 times clouds.

	<i>R</i>	<i>S</i>	<i>C</i>
<i>R</i>	4/7	1/7	2/7
<i>S</i>	2/5	0	3/5
<i>C</i>	3/10	4/10	3/10

## Markov chains

---

There, they never have two nice days in a row and if it was C (cloudy) yesterday, there is a 0.25 probability of R (rain) today.

For each day, the weather of the next day is random and we represent the probabilities by a matrix

	$R$	$S$	$C$
$R$	0.5	0.25	0.25
$S$	0.5	0	0.5
$C$	0.25	0.25	0.5

Each *row* contains the probability for next days weather depending on current weather.

## Markov chain

A Markov chain consists of:

A set of states:  $\{s_1, \dots, s_n\}$ .

A matrix of transition probabilities

$$\mathbf{P} = \begin{pmatrix} p_{11} & \dots & p_{1n} \\ \vdots & & \\ p_{n1} & \dots & p_{nn} \end{pmatrix}$$

containing the probability  $p_{ij}$  to move from state  $s_i$  to state  $s_j$

— — — —

sv: övergångssannolikhet, övergångsmatrisen

## Markov property

The transition probability does only depend on the current state:

$$p_{ij} = P(\text{next state is } s_j \mid \text{current state is } s_i \text{ and the state before ....})$$



# Transition probabilities

---

Transition probabilities are conditional probabilities:

$$p_{ij} = \text{P}(\text{next state is } s_j \mid \text{current state is } s_i)$$

That means **rows** sum to 1:  $\sum_{\text{all } j} p_{ij} = 1.$

## What is the weather in three days

---

The probability that the Markov chain, starting in states  $s_i$ , will be in state  $s_j$  after  $n$  steps is given by the  $ij$ 'th entry of

$$\mathbf{P}^n = \mathbf{P} \cdot \dots \cdot \mathbf{P}$$

( $n$ -fold matrix product.)

## Example

Suppose we want to compute the probability that, given that it is rainy today, the weather will be cloudy in two days.

	$R$	$S$	$C$
$R$	0.5	0.25	0.25
$S$	0.5	0	0.5
$C$	0.25	0.25	0.5

$$\begin{aligned} p_{13}^{(2)} &= p_{11}p_{13} + p_{12}p_{23} + p_{13}p_{33} \\ &= 0.5(0.25) + 0.25(0.5) + 0.25(0.5) = 0.375 \end{aligned}$$

$$\mathbf{P}^2 = \begin{pmatrix} 0.4375 & 0.1875 & \mathbf{0.375} \\ 0.375 & 0.25 & 0.375 \\ 0.375 & 0.1875 & 0.4375 \end{pmatrix}$$

# Probability vectors

A **probability vector** is a row vector that gives the probabilities of being at each state at a certain step.

The probability vector which represents the initial state of a Markov chain is starting vector and is denoted by  $\mathbf{u}^{(0)}$  or simply  $\mathbf{u}$ . The probability vector at step  $k$  is denoted by  $\mathbf{u}^{(k)}$ .

## 1 step

If  $\mathbf{u}_k$  is the probability vector at step  $k$ , then the vector

$$\mathbf{u}^{(k+1)} = \mathbf{u}^{(k)} \mathbf{P}$$

is the probability vector at step  $k + 1$ .

## n steps

If  $\mathbf{u}$  is the starting vector of a Markov Chain, then the probability vector at step  $n$  is given by

$$\mathbf{u}^{(n)} = \mathbf{u} \mathbf{P}^n.$$

## Example

---

In the previous example, if the initial probability vector is  $\mathbf{u} = (1/3, 2/3, 0)$ , then the probability vector on day 2 will be

$$\begin{aligned}\mathbf{u}^{(2)} = \mathbf{u}\mathbf{P}^2 &= \begin{pmatrix} 1/3 & 2/3 & 0 \end{pmatrix} \begin{pmatrix} 0.4375 & 0.1875 & 0.375 \\ 0.375 & 0.25 & 0.375 \\ 0.375 & 0.1875 & 0.4375 \end{pmatrix} \\ &= \begin{pmatrix} 0.3958 & 0.2292 & 0.3750 \end{pmatrix}\end{aligned}$$

This means that on day 2, there is a 39.58% chance of rain, 22.92% chance that the weather will be nice and 37.5% chance that it will be cloudy.

A Markov chain is said to be regular if there exists  $n$  such that all the elements of the matrix  $\mathbf{P}^n$  are nonzero. The Markov chain of the previous example is regular since

$$\mathbf{P}^2 = \begin{pmatrix} 0.4375 & 0.1875 & 0.375 \\ 0.375 & 0.25 & 0.375 \\ 0.375 & 0.1875 & 0.4375 \end{pmatrix}$$

(all the values are strictly positive)

## Stationary distribution

---

If the Markov chain is regular then,  $\mathbf{P}^n \rightarrow \mathbf{Q}$  where

$$\mathbf{Q} = \begin{pmatrix} q_1 & q_2 & \dots & q_n \\ q_1 & q_2 & \dots & q_n \\ \vdots & \vdots & \ddots & \vdots \\ q_1 & q_2 & \dots & q_n \end{pmatrix}$$

$q_j$  is the probability to be at state  $s_j$  on the long run.

$$q\mathbf{P} = q$$

## Absorbing states

---

A state is said to be absorbing if it is impossible to leave it, that is  $p_{ii} = 1$ .

A Markov chain is called absorbing if it has at least one absorbing state, and if from every state it is possible to go to an absorbing state.

In an absorbing Markov chain, a state that is not absorbing is called transient.

Example:

$$\begin{pmatrix} 1 & 0 & 0 & 0 \\ 0 & \frac{1}{2} & \frac{1}{2} & 0 \\ 0 & 0 & 1 & 0 \\ \frac{1}{4} & \frac{1}{2} & 0 & \frac{1}{4} \end{pmatrix}$$



The transition matrix of an absorbing Markov chain with  $r$  absorbing states and  $t$  transient states can be written as

$$P = \begin{pmatrix} \mathbf{Q} & \mathbf{R} \\ 0 & \mathbf{I}_r \end{pmatrix}$$

where  $\mathbf{I}_r$  is the identity matrix, 0 is the zero matrix (all elements are zeros),  $\mathbf{Q}$  is a  $t \times t$  -matrix and  $\mathbf{R}$  is a  $t \times r$  nonzero matrix.

This form is called the canonical form.  $\mathbf{P}^n = \begin{pmatrix} \mathbf{Q}^n & \star \\ 0 & \mathbf{I}_r \end{pmatrix}$  where  $\star$  is a  $t \times r$  matrix.  $\mathbf{Q}^n$  gives the probability for being in each of the transient states after  $n$  steps for each possible transient starting state.

## Samples and point estimators

---

**Example:** (5.27, 4.07, 5.48, 3.38) are measurements of the weight of  $n = 4$  randomly (independent) selected cats.

The weight of a cat is modelled as normal random variable  $X_1, X_2, X_3, X_4$  each  $N(\mu, (1.2)^2)$ -distributed with unknown parameter  $\mu$ . Here  $N(\mu, (1.2)^2)$  is a model for the population of *all cats*.

(5.27, 4.07, 5.48, 3.38) is a sample of  $X_1, X_2, X_3, X_4$ .

### Definition: Sample

A **sample**  $(x_1, \dots, x_n)$  of size  $n$  is made of  $n$  independent observations (realisations) of a random variable. Or – the same – of random variables  $X_1, \dots, X_n$  where all  $X_i$  are independent and equally distributed (thus have the same distribution).

**Example:** (5.27, 4.07, 5.48, 3.38) are measurements of the weight of  $n = 4$  randomly (independent) selected cats.

The weight of a cat is modelled as normal random variable  $X_1, X_2, X_3, X_4$  each  $N(\mu, (1.2)^2)$ -distributed with unknown parameter  $\mu$ . Here  $N(\mu, (1.2)^2)$  is a model for the population of *all cats*.

(5.27, 4.07, 5.48, 3.38) is a sample of  $X_1, X_2, X_3, X_4$ .

## Definition: Anti-Example

---

(5.27, 5.27, 5.27, 5.27, 5.27) is perhaps not a sample

(lack of independence because some genius just weighted the same cat over and over).

Like in the “cat”-example we can often say what kind of distribution is appropriate for  $X$  but we do not know the right parameters.

Many statistical problems can be reduced to the following question: Given the observations  $x_1, \dots, x_n$ , what can we say about the parameters in the distribution of  $X_i$  (assuming each  $X_i$  is drawn independently from the same distribution)?

**Definition: i.i.d.**

We write  $X_1, X_2, \dots, X_n \stackrel{\text{i.i.d.}}{\sim} D$  if  $X_1, X_2, \dots, X_n$  are independently and identically distributed with distribution  $D$ .

## The sample mean as estimator

$\bar{X}^{(n)} = \frac{1}{n} \sum_{i=1}^n X_i$  is the **sample mean**.

**Example:** Let (5.27, 4.07, 5.48, 3.38) our sample.

$\bar{x}^{(4)} = (5.27 + 4.07 + 5.48 + 3.38)/4 = 4.55$  is a realisation  $\bar{X}^{(n)}$ .

We model  $\bar{X}^{(n)}$  itself as random variable with its own expectation, variance and realization etc. Now with  $\mu = E(X_1) = E(X_2) = \dots$  and  $\sigma^2 = \text{Var}(X_1) = \text{Var}(X_2) = \dots$

$$E\left(\frac{1}{n} \sum_{i=1}^n X_i\right) = \frac{1}{n} \sum_{i=1}^n EX_i \stackrel{(*)}{=} \mu$$

$$\text{Var}\left(\frac{1}{n} \sum_{i=1}^n X_i\right) \stackrel{i.i.d}{=} \frac{1}{n^2} \sum_{i=1}^n \text{Var}(X_i) = \frac{\sigma^2}{n}$$

Ah! Smaller uncertainty, 4.55 is perhaps closer to  $\mu$  than most the values in our sample which vary from  $\mu$  by  $\sigma$ .



# The sample mean as random variable

---

## Expectation and variance of the sample average

$$E(\bar{X}^{(n)}) = \mu \text{ and } \text{Var}(\bar{X}^{(n)}) = \sigma^2/n.$$

Quiz: How fast goes uncertainty down if  $n$  increases?

## Standard error of the mean

$\frac{\sigma}{\sqrt{n}}$  is called **standard error of the mean**.

## Point estimate and standard error

---

**Example:** Take (5.27, 4.07, 5.48, 3.38) our sample. Model  $X_1, \dots, X_n \stackrel{i.i.d}{\sim} N(\mu, \sigma^2)$  with  $n = 4$  and  $\sigma = 1.2$  and  $\mu$  unknown.

$\bar{x}^{(4)} = (5.27 + 4.07 + 5.48 + 3.38)/4 = 4.55$  is an estimate for  $\mu$

The standard error associated with  $\bar{x}^{(4)}$  is  $\sigma/\sqrt{n} = 1.2/\sqrt{4} = 0.6$ .

Our estimate

$$\mu \approx 4.55 \pm 0.6$$

## The sample mean as random variable: Gaussian case

---

### Average of Gaussian distributed random variables.

Let  $X_1, \dots, X_n$  an independent sample of a  $N(\mu, \sigma^2)$  r.v. Then  $\bar{X}^{(n)}$  is  $N(\mu, \sigma^2/n)$ -distributed.

## Estimation

An estimator for a parameter  $\theta$  is a function  $\hat{\theta}(X_1, \dots, X_n)$  mapping the observations into the parameter space  $\Theta$ .

**Example:**  $\bar{X}^{(n)}$  is an estimator for  $\mu = EX_1 = EX_2 = \dots$

$\hat{\theta}$  can refer both to a random variable and to actual observed values.

- $\hat{\theta}(X_1, \dots, X_n)$  is a random variable with a certain distribution (random in  $\rightarrow$  random out).
- $\hat{\theta}(x_1, \dots, x_n)$  is a number calculated from data. This is called the point estimate of the parameter.

Two important qualities of estimators:

- *unbiased*:  $E(\hat{\theta}(X_1, \dots, X_n)) = \theta$ .
- Small variance in large samples:  $V(\hat{\theta}(X_1, \dots, X_n))$  small if  $n$  large.

If the expected value of the estimator is the true value (the estimator is unbiased), that means that the estimated values center on average around the true value if we make several repeated samples of size  $n$ .

- For a given sample, the value need not be close to the true value.
- The standard deviation of an unbiased estimate gives an indication of how far it may be from the actual value.
- Often the **standard error of the estimate** is reported, which is the standard deviation of the estimate.

## Sample mean and sample variance

Consider an i.i.d sample  $(X_1, \dots, X_n)$  and assume that  $E(X_i) = \mu$  and  $V(X_i) = \sigma^2$ .

The **sample mean**  $\hat{\mu} = \bar{X}^{(n)}$  is an unbiased estimator of  $\mu$ , that is  $E(\hat{\mu}) = \mu$ . It has standard error  $\sqrt{V(\hat{\mu})} = \frac{\sigma}{\sqrt{n}}$ .

An unbiased estimator for the variance  $\sigma^2$  is the **sample variance**

$$S^2 = \frac{1}{n-1} \sum_{i=1}^n (X_i - \bar{X})^2.$$

Sample variance can also be computed as

$$S^2 = \frac{n \sum_{i=1}^n X_i^2 - (\sum_{i=1}^n X_i)^2}{n(n-1)}$$



## Percentiles and quantiles

The  $p^{th}$  percentile  $P$  is the value of  $X$  such that  $p\%$  or less of the observations are less than  $P$  and  $(100 - p)\%$  or less are greater than  $P$ .  $p^{th}$  percentiles are  $p\%$ -quantiles.

In particular,  $P_{25}$  is the  $25^{th}$  percentile or the first quartile denoted also by  $Q_1$ .  $P_{50}$  is the  $50^{th}$  percentile or the second quartile  $Q_2$ , which is also the median, and  $P_{75}$  is the  $75^{th}$  percentile or the third quartile  $Q_3$ .

Note that  $Q_1 = \frac{n+1}{4}$  th ordered observation,  $Q_2 = \frac{2(n+1)}{4} = \frac{n+1}{2}$  th ordered observation, and  $Q_3 = \frac{3(n+1)}{4}$  th ordered observation.

## Example

---

Given the following set of data :

18, 1, 20, 15, 12, 15, 14, 7, 11, 9, 6, 4

Order the numbers from the lowest to the highest

1, 4, 6, 7, 9, 11, 12, 14, 15, 15, 18, 20

$$\bar{x}^{(12)} = \frac{1+4+\dots+18+20}{12} = 11.$$

$$\text{Median: } Me = \frac{11+12}{2} = 11.5.$$

## Example

---

Given the following set of data :

18, 1, 20, 15, 12, 15, 14, 7, 11, 9, 6, 4

Variance

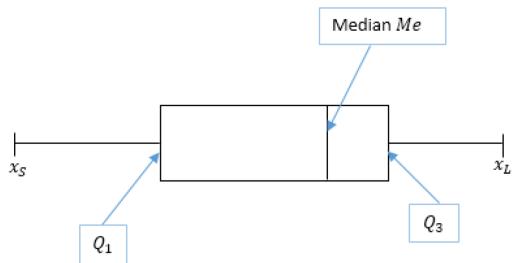
$$s^2 = \frac{(20 - 11)^2 + (18 - 11)^2 + \cdots + (-7)^2 + (-10)^2}{12 - 1} \approx 33.3$$

Order the numbers from the lowest to the highest

1, 4, 6, 7, 9, 11, 12, 14, 15, 15, 18, 20

$$Q_1 = 6.25, Q_3 = 15.$$

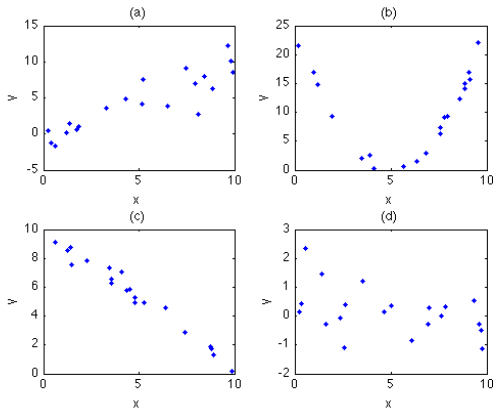
# Boxplot



## Bivariate samples

---

# Visualisation



Assume 2d measurements  $(x_i, y_i)$ . A scatter plot is a two-dimensional plot in which each  $(x_i, y_i)$  measurement is represented as a point in the  $x$ - $y$ -plane.

The *sample* covariance is defined as,

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

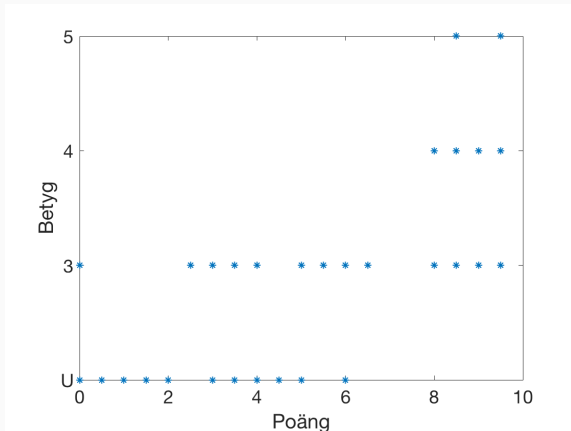
and is an unbiased estimator of the covariance  $\text{Cov}(X, Y)$ .

The sample correlation coefficient is defined as

$$r_{xy} = \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^n (x_i - \bar{x})^2} \sqrt{\sum_{i=1}^n (y_i - \bar{y})^2}} = \frac{c_{xy}}{s_x s_y}$$

The sample correlation is an empirical measure of linear dependence.

## Example: Course results 2017



Exam grade ( $Y$ ) versus points in exam question 5 ( $X$ ).

Correlation:  $r_{xy} = 0.7261$



## Sum of Gaussian r.v.

Let  $X \sim N(\mu_X, \sigma_X^2)$  and  $Y \sim N(\mu_Y, \sigma_Y^2)$  with  $X$  and  $Y$  independent. Then

$$X + Y \sim N(\mu_X + \mu_Y, \sigma_X^2 + \sigma_Y^2)$$

Note: A normal random variable with mean  $\mu$  and variance  $\sigma^2$  has moment generating function  $m(t) = \exp(t\mu + t^2\sigma^2/2)$ . So if you tell me your moment generating function, I tell you if you are normally distributed and if, what your parameters are. We can prove the theorem by computing and identifying the m.g.f of  $X + Y$  (next slide)

## Proof with m.g.f.

---

So we now  $m_X(t) = \mathbb{E} \exp(tX) = \exp(t\mu_X + t^2\sigma_X^2/2)$  and  $m_Y(t) = \mathbb{E} \exp(tY) = \exp(t\mu_Y + t^2\sigma_Y^2/2)$ .

We compute and identify  $m_{X+Y}$

$$\begin{aligned} m_{X+Y}(t) &= \mathbb{E} \exp(t(X + Y)) = \mathbb{E} (\exp(tX) \exp(tY)) \\ &\stackrel{indep}{=} \mathbb{E} (\exp(tX)) \mathbb{E} (\exp(tY)) \\ &= m_X(t) m_Y(t) = \exp(t\mu_X + t^2\sigma_X^2/2) \exp(t\mu_Y + t^2\sigma_Y^2/2) \\ &= \exp(t(\mu_X + \mu_Y) + t^2(\sigma_X^2 + \sigma_Y^2)/2) \end{aligned}$$

which is m.g.f of  $N(\mu_X + \mu_Y, \sigma_X^2 + \sigma_Y^2)$  so  $X + Y$  must be  $N(\mu_X + \mu_Y, \sigma_X^2 + \sigma_Y^2)$  distributed.

## Central limit theorem/CLT

---

## Recall

---

If  $X \sim N(\mu, \sigma^2)$ , then

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

If  $X_1, \dots, X_n \sim N(\mu, \sigma^2)$  independent, then

$$\bar{X}^{(n)} \sim N(\mu, \sigma^2/n).$$

then

$$\frac{\bar{X}^{(n)} - \mu}{\sigma/\sqrt{n}} \sim N(0, 1)$$

## Normal approximation of Binomial distribution

If  $X_1 \dots X_n \sim \text{Ber}(p)$ . Then  $X = \sum X_i \sim \text{Bin}(n, p)$ .

$X$  is approximately normally distributed

$$X \stackrel{\text{approx.}}{\sim} N(np, np(1-p)),$$

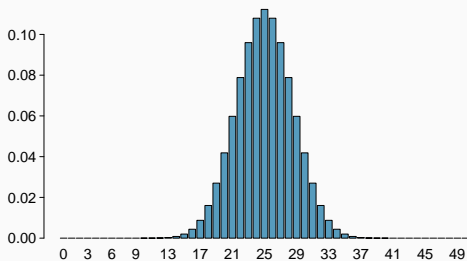
Thus **again** for  $\bar{X}^{(n)} = \frac{1}{n} \sum X_i$ ,

$$\bar{X}^{(n)} \stackrel{\text{approx.}}{\sim} N(p, p(1-p)/n),$$

or

$$\frac{\bar{X}^{(n)} - p}{\sqrt{p(1-p)/n}} \stackrel{\text{approx.}}{\sim} N(0, 1)$$

## Normal approximation



$$n = 50, p = 0.5$$

# Central limit theorem

## Central limit theorem (CLT)

If  $X_1, \dots, X_n$  are independent and equally distributed random variables with expected value  $\mu$  and variance  $\sigma^2 < \infty$ , then

$$P\left(\frac{\bar{X}^{(n)} - \mu}{\sigma/\sqrt{n}} \leq x\right) \rightarrow F(x), \quad \text{for } n \rightarrow \infty.$$

where  $F$  is the distribution function of  $N(0, 1)$ .

This means,

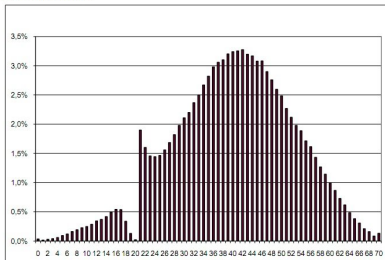
- $\bar{X} = n^{-1} \sum_{i=1}^n X_i$  is approximatively  $N(\mu, \text{SE}^2)$ -distributed, where  $\text{SE} = \sigma/\sqrt{n}$  is the **standard error**

for large  $n$ .

How large is large? Depends on the distribution of the  $X_i$ 's.

# High-school maturity exam in Poland

2.1. Poziom podstawowy



Wykres 1. Rozkład wyników na poziomie podstawowym

Histogram showing the distribution of scores for the obligatory Polish language test. "The dip and spike that occurs at around 21 points just happens to coincide with the cut-off score for passing the exam"

<http://freakonomics.com/2011/07/07/>

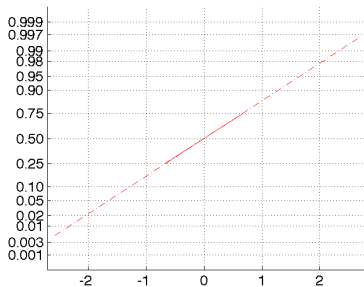
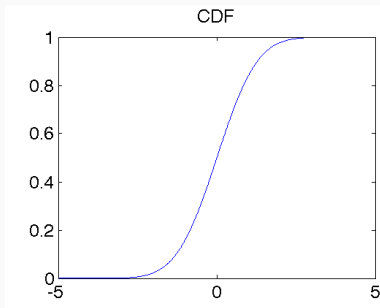
another-case-of-teacher-cheating-or-is-it-just-altruism/



## Normal probability plot

---

## Normal probability plot



The standard normal distribution function (cdf) is

$$F(x) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^x e^{-y^2/2} dy$$

It is possible to transform the scaling on the y-axis so that  $F$  becomes a straight line in the plot.

## Normal probability plot

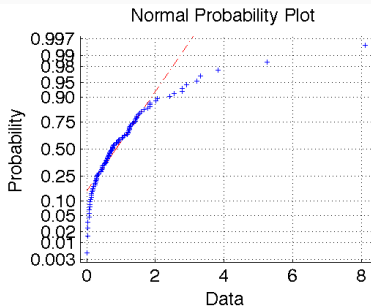
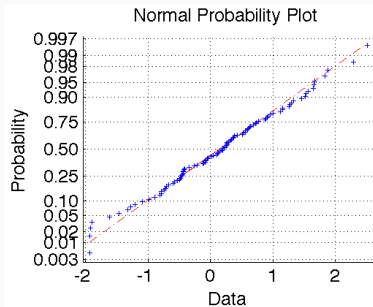
Suppose we have the data  $x_1, \dots, x_n$  and want to see if a normal distribution is a reasonable model for the data. We can use the normal probability plot for this.

First we compute the *empirical distribution function*

$$F^*(x) = \underbrace{\frac{1}{n} \sum_{i=1}^n \mathbb{I}(x_i \leq x)}_{\text{proportion of values smaller than } x}$$

We plot the points  $F^*(x_j)$  in the normal probability diagram, and if the data is normally distributed, these points should lie along a straight line.

## Normal probability plot



**Example:** left normally distributed data and right exponentially distributed data in normal probability diagram. In Matlab: `normplot`.

## Confidence interval

---

## Confidence interval

If  $X_1, \dots, X_n$  i.i.d random variables with distribution depending on a parameter  $\theta$ , with  $\theta_0$  being the unknown value. A  $100(1 - \alpha)\%$  confidence interval for  $\theta$  with confidence level  $1 - \alpha$  is an interval  $I_\theta = [A, B]$  computed from the data such that

$$P(A \leq \theta_0 \leq B) = 1 - \alpha.$$

## Confidence interval for parameter $\mu$ of a normal distribution

Let  $X_1, \dots, X_n$  be independent  $N(\mu, \sigma^2)$ .

**Known variance  $\sigma^2$**

$$I_\mu = (A, B) = \left( \bar{X}^{(n)} - 1.96 \frac{\sigma}{\sqrt{n}}, \bar{X}^{(n)} + 1.96 \frac{\sigma}{\sqrt{n}} \right)$$

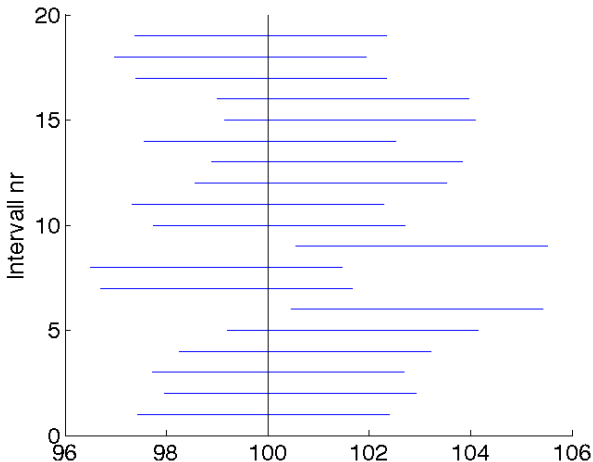
is a confidence interval for  $\mu$  with confidence level 95%.

Here 1.96 is the  $0.975 = (100 - 2.5)\%$  quantile of  $Z \sim N(0, 1)$ :

$$P(-1.96 < Z < 1.96) = 0.95.$$

$$P\left(-1.96 < \frac{\bar{X}^{(n)} - \mu}{\sigma/\sqrt{n}} < 1.96\right) = 0.95.$$

$$P(A \leq \mu \leq B) = 0.95$$



20 confidence intervals for  $\mu$ , that where each constructed from 20 different samples of 10  $N(100, 16)$ -observations.



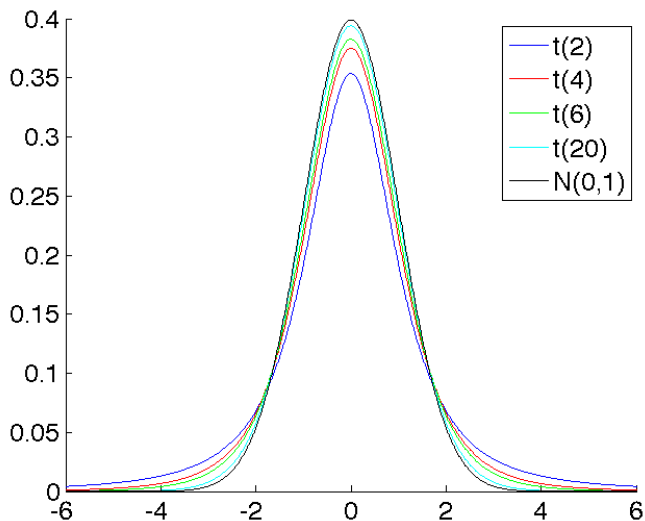
- $[A, B]$  is a random interval, because  $A$  and  $B$  are random variables (transformations of the random variables  $X_1, \dots, X_n$ ).
- Interpretation. Let  $\mathbf{x}_1 = (x_{11}, \dots, x_{n1})$ ,  $\mathbf{x}_2 = (x_{12}, \dots, x_{n2})$ ,  $\dots$  be repeated measurements of  $X_1, \dots, X_n$ . If we make the confidence interval for  $\theta$  based on every  $\mathbf{x}_i$ , then  $100(1 - \alpha)\%$  of these intervals cover the true value  $\theta_0$ .

## Table 2: Quantiles of the normal distribution

Table gives  $P(X > \lambda_\alpha) = \alpha$  for  $X \sim N(0, 1)$

$\alpha$	.1	.05	.025	.01	.005	.001	...	.00001
$\lambda_\alpha$	1.2816	1.6449	1.9600	2.3263	2.5758	3.0902	...	4.2649

## $t(n)$ -distribution



**Table 3: Quantiles of the  $t$ -distribution**

Table gives  $P(X > t_\alpha(f)) = \alpha$  for  $X \sim t(f)$ .

$\alpha$	.1	.05	.025	.01	.001
$t_\alpha(1)$	3.0777	6.3138	12.706	31.820	318.31
$t_\alpha(2)$	1.8856	2.9200	4.3027	6.9646	22.327
$t_\alpha(3)$	1.6377	2.3534	3.1824	4.5407	10.215
$t_\alpha(4)$	1.5332	2.1318	2.7764	3.7469	7.1732
$t_\alpha(5)$	1.4759	2.0150	2.5706	3.3649	5.8934
$t_\alpha(6)$	1.4398	1.9432	2.4469	3.1427	5.2076
$t_\alpha(7)$	1.4149	1.8946	2.3646	2.9980	4.7853
$t_\alpha(8)$	1.3968	1.8595	2.3060	2.8965	4.5008
$t_\alpha(9)$	1.3830	1.8331	2.2622	2.8214	4.2968
$t_\alpha(10)$	1.3722	1.8125	2.2281	2.7638	4.1437
$t_\alpha(15)$	1.3406	1.7531	2.1314	2.6025	3.7328
$t_\alpha(20)$	1.3253	1.7247	2.0860	2.5280	3.5518
$t_\alpha(30)$	1.3104	1.6973	2.0423	2.4573	3.3852
$t_\alpha(40)$	1.3031	1.6839	2.0211	2.4233	3.3069
$t_\alpha(60)$	1.2958	1.6706	2.0003	2.3901	3.2317
$t_\alpha(\infty)$	1.2816	1.6449	1.9600	2.3263	3.0902

## Confidence interval for $\mu$ of a normal distribution

Let  $X_1, \dots, X_n$  be independent  $N(\mu, \sigma^2)$ .

**Known variance  $\sigma^2$**

$$I_\mu = \left( \bar{X} - z_{\alpha/2} \frac{\sigma}{\sqrt{n}}, \bar{X} + z_{\alpha/2} \frac{\sigma}{\sqrt{n}} \right)$$

is a confidence interval for  $\mu$  with confidence level  $1 - \alpha$ .

**Unknown variance  $\sigma^2$**

$$I_\mu = \left( \bar{X} - t_{\alpha/2}(n-1) \frac{s}{\sqrt{n}}, \bar{X} + t_{\alpha/2}(n-1) \frac{s}{\sqrt{n}} \right)$$

is a confidence interval for  $\mu$  with confidence level  $1 - \alpha$ . Here  $s^2$  is the sample variance and  $t_{\alpha/2}(n-1)$  are the  $(1 - \alpha/2)$ -quantiles of the  $t(n-1)$ -distribution.

## Quiz

---

$x_1, \dots, x_n$  are a sample of i.i.d observations with distribution depending on a parameter  $\theta$ .

Winnie computes a 95 % confidence interval for  $\theta$ .

Piglet computes a 90 % confidence interval for  $\theta$  using the same data.

Which interval is smallest? Piglet's 90 % confidence interval.

## Confidence interval for $\mu$ from central limit theorem

- By the CLT the sample mean  $\bar{X}^{(n)}$  is approximatively  $N(\mu, \sigma^2/n)$ -distributed for large  $n$ .
- If we have a sample with known variance  $\sigma^2$ ,

$$I_\mu = \left( \bar{X} - z_{\alpha/2} \frac{\sigma}{\sqrt{n}}, \bar{X} + z_{\alpha/2} \frac{\sigma}{\sqrt{n}} \right)$$

is a confidence interval for the mean  $\mu$  with confidence level  $1 - \alpha$ .

- If  $\sigma$  is not known we can estimate it by  $S$ . For the estimate to be good, it is important that  $n$  is large and the distribution for  $X_i$  is not too heavy tailed.
- Since  $n$  is big, we use  $t_{\alpha/2}(n-1) \approx z_{\alpha/2}$ , so if  $\sigma$  is unknown, we use

$$I_\mu = \left( \bar{X} - z_{\alpha/2} \frac{s}{\sqrt{n}}, \bar{X} + z_{\alpha/2} \frac{s}{\sqrt{n}} \right).$$

## Confidence interval for $\sigma^2$ for the normal distribution

### Confidence interval for $\sigma$

If  $X_1, \dots, X_n$  are independent  $N(\mu, \sigma^2)$  then a confidence interval with confidence level  $1 - \alpha$  for  $\sigma$  is

$$I_\sigma = \left( \sqrt{\frac{(n-1)s^2}{\chi_{\alpha/2}^2(n-1)}}, \sqrt{\frac{(n-1)s^2}{\chi_{1-\alpha/2}^2(n-1)}} \right).$$

Here  $\chi_{\alpha/2}^2(n-1)$  are the  $(1 - \alpha/2)$ -quantiles of the  $\chi^2(n-1)$  distribution.

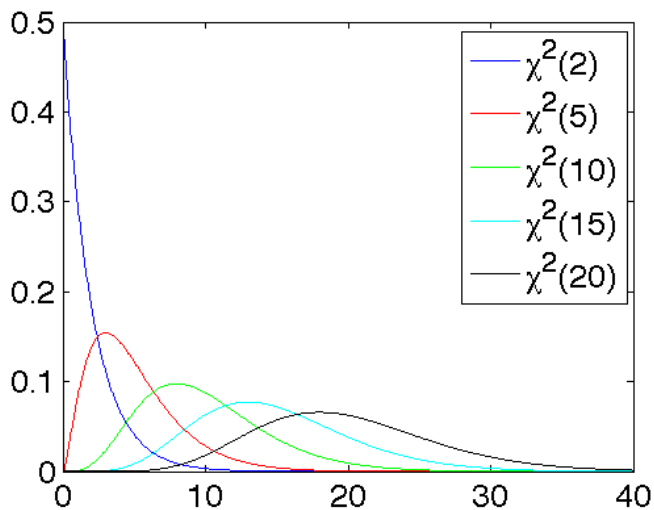
If  $Z_i$  are independent  $N(0, 1)$ , it holds

$$\sum_{i=1}^n Z_i^2$$

is  $\chi^2(n)$ -distributed



## $\chi^2(n)$ -distribution



## Confidence interval for $\sigma^2$ for the normal distribution

### Confidence interval for $\sigma$

If  $X_1, \dots, X_n$  are independent  $N(\mu, \sigma^2)$  then a confidence interval with confidence level  $1 - \alpha$  for  $\sigma$  is

$$I_\sigma = \left( \sqrt{\frac{(n-1)s^2}{\chi_{\alpha/2}^2(n-1)}}, \sqrt{\frac{(n-1)s^2}{\chi_{1-\alpha/2}^2(n-1)}} \right).$$

Important: In contrast to the confidence interval for the expected value, the confidence interval for the variance is very sensitive to deviations from the normal distribution.

# Summary

---

For a confidence interval

- for the expected value  $\mu$ 
  - of the normal distribution: Slide: confidence interval for  $\mu$  of a normal distribution
    - Known  $\sigma$  or large  $n$ : use confidence interval based on normal quantiles.
    - Small  $n$  and unknown  $\sigma$ : use quantiles based on  $t$ -distribution.
  - of a general distribution
    - Large  $n$ : use confidence interval based on normal quantiles (valid approximation by CLT). Slide: Confidence interval for  $\mu$  from central limit theorem.
- for the variance  $\sigma^2$ 
  - of the normal distribution: Slide: Confidence interval for  $\sigma^2$  for the normal distribution.

# Hypothesis tests

---

# Hypothesis tests

---

An important problem in statistics is to test whether a theory or a *research hypothesis* is right or wrong.

Examples of such problems include:

- Does a new drug have any effect?  $\text{Mean effect} > 0$
- Do smokers die sooner than non-smokers?  $\text{Mean life time difference} < 0$
- Does the measuring device have a systematic error?  $\text{Mean measurement error} \neq 0$

Answers the statistical analysis could give are

1. that the research hypothesis is supported by the data (and possibly a quantification of the degree of support),
2. that the data doesn't support the hypothesis,
3. a decision rule.

## Example

---

The length of a certain lumber from a national home building store is supposed to be 2.5 m.

A builder wants to check whether the lumber cut by the lumber mill has a mean length different smaller than 2.5 m.

A statistical formulation of this problem is that we want to test the **null hypothesis**

$$H_0: \text{mean length} = 2.5 \text{ m}$$

against the **alternative/research hypothesis**

$$H_1: \text{mean length} < 2.5 \text{ m}$$

$H_1$  is actionable knowledge. If  $H_1$  is true she needs to write an angry letter.

## Example

---

- You have new laboratory equipment to measure the chlorine content in water and want to check it. You mix water with true chlorine content 60 (you can do that very precisely), and take 6 measurements.
- Results of the measurement are  $\bar{x} = 59.62$  and  $s^2 = 4.6920$ .
- Assume that the measurements are samples of a random variable  $X \sim N(\mu, \sigma^2)$ .
- The question now is whether we can claim that the new equipment has systematic measurement error,  $\mu \neq 60$ .



## Setup

A statistical formulation of this problem is that we want to test the **null hypothesis**

$$H_0: \mu = 60$$

against the **alternative hypothesis** or **research hypothesis**

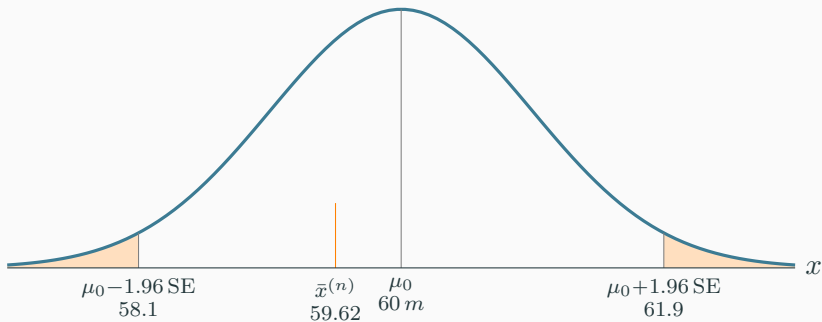
$$H_1: \mu \neq 60.$$

If the test we perform finds that there is a systematic error,  $H_0$  is rejected in favour of  $H_1$ .

Is  $H_1$  actionable knowledge?

### Choosing the alternative $H_1$

Choose  $H_1$  such if someone would tell you it is true, you can do something useful with that knowledge!



$$\text{SE} \approx \frac{\sqrt{4.6920}}{\sqrt{5}}$$

The **outcome** of a hypothesis test can be:

- Reject  $H_0$  (accept  $H_1$ .)
  - Action!
- Do not reject  $H_0$ 
  - Could be lack of data, or  $H_0$  being correct. The question of  $H_0$  or  $H_1$  is truly left open. Meh. Should still report it though.

## Decision errors

		Decision	
		fail to reject $H_0$	reject $H_0$
Truth	$H_0$ true	✓	Type 1 Error
	$H_1$ true	Type 2 Error	✓

- A **Type 1 Error** is rejecting the null hypothesis when  $H_0$  is true. We want to avoid that, control the probability for this error.
- A Type 2 Error is failing to reject the null hypothesis when  $H_1$  is true.

## Burden of proof

---

If we again think of a hypothesis test as a criminal trial then it makes sense to frame the verdict in terms of the null and alternative hypotheses:

$H_0$  : Defendant is innocent

$H_1$  : Defendant is guilty

Which type of error is being committed in the following circumstances?

- Declaring the defendant innocent when they are actually guilty

Type 2 error

- Declaring the defendant guilty when they are actually innocent

Type 1 error

Which error do you think is the worse error to make?

# Statistical reasoning

*Classical logic:* If the null hypothesis is correct, then **these data can not occur**.

These data have occurred.

Therefore, the null hypothesis is **false**.

*Tweak the language, so that it becomes **probabilistic**...      Statistical reasoning:*

If the null hypothesis is correct, then **these data are highly unlikely**.

These data have occurred.

Therefore, the null hypothesis is **unlikely**.

## Definition

In statistical hypothesis testing, a **result has statistical significance** when it is very unlikely to have occurred under the null hypothesis. So significance corresponds to "statistical evidence against the null".

The **significance level**  $\alpha$  is the (tolerated) probability of making a type I error:

If you want to take a decision in the case the test fails to reject  $H_0$ , you should compute the type II error probability first. This is typically difficult.

Therefore we should avoid far reaching decisions if our tests fail to reject  $H_0$ .

## Tests from confidence intervals

---

**Data** (samples from a distribution with unknown parameter  $\mu$ ).

**Hypothesis** about parameter. Here  $H_0 : \mu = \mu_0$  and  $H_1 : \mu \neq \mu_0$ .

**Significance level**  $\alpha$ , e.g  $\alpha = 5\%$ .

**Decision rule:** Compute a  $(1 - \alpha)(= 95\%)$ -confidence interval  $[A, B]$  for the parameter  $\mu$ . If the  $\mu_0 \notin [A, B]$ , reject  $H_0$ .

**Type 1 error:** This rule has type 1 error of 5 %, so this is a valid test for level  $\alpha = 5\%$ .



## Tests with test statistics

**Data** (samples with unknown population parameter  $\mu$ ).

**Hypothesis** about parameter. Here  $H_0 : \mu = \mu_0$  and  $H_1 : \mu \begin{matrix} \neq \\ \geq \\ < \end{matrix} \mu_0$ .

**Significance level**  $\alpha$ , e.g  $\alpha = 5\%$ .

**Test statistic**  $T$ : Typically,  $T$  comes from an estimator for our parameter with known distribution under  $H_0$ .

$$T = \frac{\bar{X} - \mu_0}{\sigma/\sqrt{n}} \quad (\text{example})$$

**Decision rule:** Reject  $H_0$  if the  $p$ -value is less than the significance level  $\alpha$ .

or: Reject  $H_0$  if the  $T_{obs}$  is in the critical region/rejection region (see next slide).

**Type I error:** The type I error for this test is  $\leq \alpha$ .

## Critical region

The **critical region**  $C_\alpha$  of a test are those values of the test statistic  $T$  for which  $H_0$  can be rejected while obeying significance level  $\alpha$ . Typically represented by one or two critical values.

We compute rejection region for the data. We reject  $H_0$  if  $T_{obs}$  is in the rejection region.

## Example: critical region for mean of normal population

---

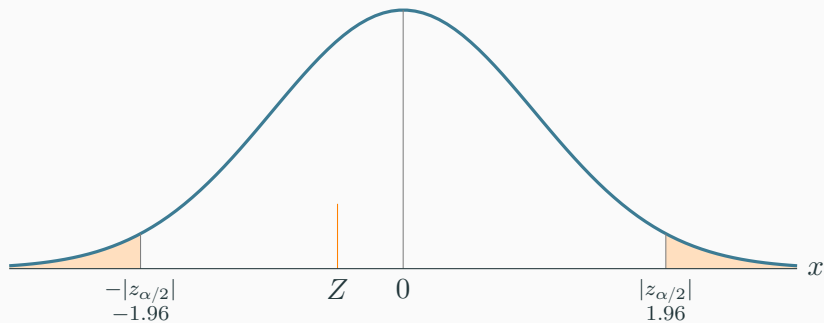
We want to use a quantity  $T$  that we know the distribution of under  $H_0$ , so that we can calculate the critical region.

In case of the normal distribution with known variance

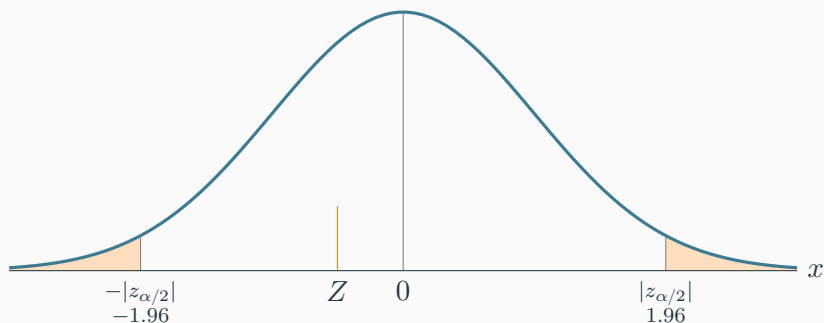
$$(T =) Z = \frac{\bar{X} - \mu_0}{\sigma/\sqrt{n}}$$

we know that  $Z$  under  $H_0$  is  $N(0, 1)$ -distributed and

Reject  $H_0$  at level  $\alpha$  if  $|Z| > z_{\alpha/2}$ .



Rejection region for  $\alpha = 0.05$ .



Rejection region for  $\alpha = 0.05$  (on the  $x$ -axis below the yellow area).

Rule: Reject  $H_0$  (yeah) if  $Z$  is in the rejection region.

## Example: $p$ -value for mean of normal population

---

### $p$ -value

The  $p$ -value is the probability **under the null hypothesis  $H_0$**  to obtain a test statistic  $T$  with more evidence for the alternative (more “extreme”) than the one we observed,  $t_{obs}$ .

## Example: $p$ -value for normal distribution (two-sided)

Again we want to use a quantity  $T$  that we know the distribution of under  $H_0$ , so that we can calculate the  $p$ -value.

In case of the normal distribution with known variance

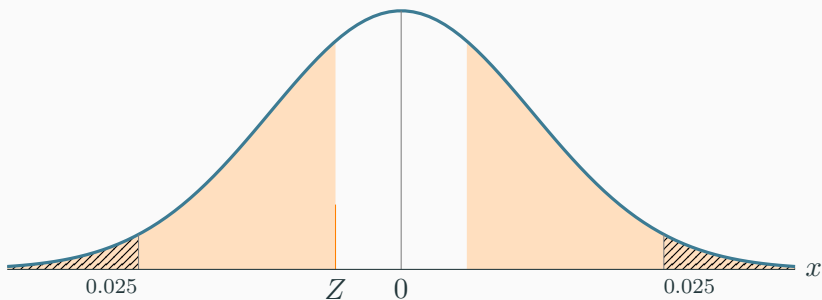
$$T = \frac{\bar{X} - \mu_0}{\sigma/\sqrt{n}}$$

we know that  $T$  under  $H_0$  is  $N(0, 1)$ -distributed and

$$p = P(|T| \geq |T_{obs}|) = 2 \cdot P(T \geq |T_{obs}|) = 2(1 - \Phi(|T_{obs}|)).$$

We compute  $p$  for the data. We reject  $H_0$  if  $p \leq \alpha$

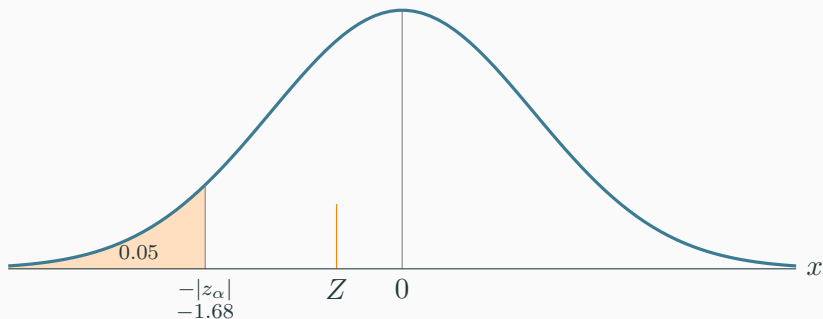
We compute rejection region for the data. We reject  $H_0$  if  $T_{obs}$  is in the rejection region.



Yellow area:  $p$ -value, dashed area:  $\alpha = 0.05$ .

Rule: Reject  $H_0$  if  $p \leq \alpha$





One-sided rejection region for  $\alpha = 0.05$ .

Rule: Reject  $H_0$  if  $Z$  is inside the rejection region.

## Example: $p$ -value for normal distribution (one-sided)

Again we want to use a quantity  $T$  that we know the distribution of under  $H_0$ , so that we can calculate the  $p$ -value.

In case of the normal distribution with known variance

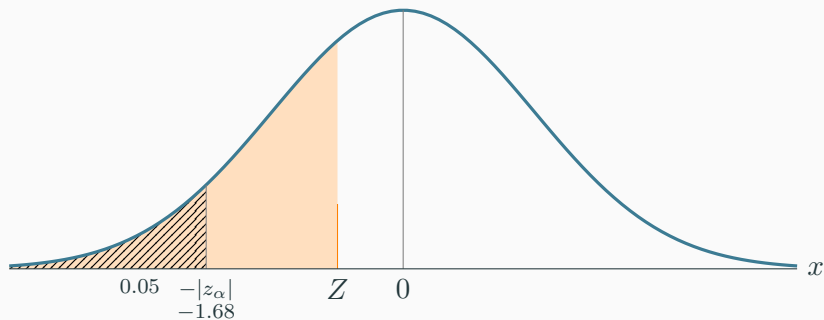
$$T = \frac{\bar{X} - \mu_0}{\sigma/\sqrt{n}}$$

we know that  $T$  under  $H_0$  is  $N(0, 1)$ -distributed.

- 1.) Check if  $T$  is on the right side to give evidence in favour of  $H_1$ .
- 2.)  $p = P(T \text{ more extreme than } T_{obs})$  on the right side  
 $1 - \Phi(|T_{obs}|).$

We compute  $p$  for the data. We reject  $H_0$  if  $p \leq \alpha$

We compute rejection region for the data. We reject  $H_0$  if  $T_{obs}$  is in the orange rejection region.



Yellow area:  $p$  value, dashed area:  $\alpha = 0.05$ .

Rule: Reject  $H_0$  if  $p \leq \alpha$ .

## How many observations are needed?

---

A test detects a deviation of  $\mu - \mu_0$  more easily if:

- If the significance level  $\alpha$  is not very small.
- The number of observations  $n$  is large.
- The population variance relatively  $\sigma^2$  is small.

## Estimating proportions

---

# Estimating proportions

---

## Example

Suppose we want to estimate the proportion  $p$  of people who own tablets in a certain city. 250 randomly selected people are surveyed, 98 of them reported owning tablets. An estimate for the population proportion is given by  $\hat{p} = \frac{98}{250} = 0.392$ .

In general we want to study a particular trait in a population too large to sample completely. We ask about the proportion of the population with this trait.

## Estimating a proportion

---

- We choose a random sample  $X_1, \dots, X_n$  from the population.

- 

$$X_i = \begin{cases} 1 & \text{ith member of the sample has the trait} \\ 0 & \text{otherwise} \end{cases}$$

- The **point estimator** is based on the

$$\hat{p} = \frac{\sum_{i=1}^n X_i}{n} \quad (\text{proportion in the sample}) \quad .$$

## Bernouli random variables

---

Why do we write  $\hat{p} = \frac{\sum_{i=1}^n X_i}{n}$  as sum of random variables.

$P(X_i = 1) = p$ ,  $P(X_i = 0) = 1 - p$ .  $X_i$  are Bernoulli random variables with parameter  $p$ !

We know a lot about them. E.g.

$$E(X_i) = 0 \cdot (1 - p) + 1 \cdot p = p$$

$n\hat{p}$  is the sum of Bernoulli random variables, hence  $\text{Bin}(n, p)$  distributed. So ...



## Unbiasedness

The expectation of  $\hat{p}$ :

$$E(\hat{p}) = \frac{1}{n} \sum_{i=1}^n E(X_i) = \frac{1}{n} (\underbrace{p + p + \cdots + p}_{n \text{ times}}) = p$$

$$E(\hat{p}) = p$$

$\hat{p}$  is an unbiased estimator for the proportion  $p$ .

## Variance

The variance of  $\hat{p}$  tells us how good as estimator  $\hat{p}$  is.

$$\text{Var}(X_i) = E(X_i^2) - E(X_i)^2 = p - p^2 = p(1 - p)$$

$$\Rightarrow \text{Var}(\hat{p}) = \frac{\sum \text{Var}(X_i)}{n^2} = \frac{np(1-p)}{n^2} = \frac{p(1-p)}{n}$$

### Standard error

The variance of  $\hat{p}$ :

$$\text{Var}(\hat{p}) = \frac{p(1 - p)}{n}$$

The standard error is

$$\text{SE} = \sqrt{\text{Var}(\hat{p})} \approx \frac{\sqrt{\hat{p}(1 - \hat{p})}}{\sqrt{n}}$$

How many more observations do I need to reduce the standard error by a factor 2? 4 times as much

## Example (ctd.)

---

Recall  $\hat{p} = \frac{98}{250} = 0.392$ .

The standard error the estimated proportion of people who own a tablet is

$$SE = \frac{\sqrt{\hat{p}(1 - \hat{p})}}{\sqrt{n}} = \frac{\sqrt{0.392(1 - 0.392)}}{\sqrt{250}} = \sqrt{\frac{0.392(0.608)}{250}}$$

## Confidence interval on $\hat{p}$ .

**Normal approximation** When we take  $n$  large enough, by the central limit theorem,  $\hat{p}$  is approximately normally distributed with mean  $p$  and variance  $p(1 - p)/n$ .

### Confidence interval

A  $100(1 - \alpha)\%$  confidence interval is defined by

$$(\hat{p} - z_{\alpha/2}\text{SE}, \hat{p} + z_{\alpha/2}\text{SE})$$

where  $\text{SE} = \sqrt{\hat{p}(1 - \hat{p})/n}$  and  $P(-z_{\alpha/2} \leq Z \leq z_{\alpha/2}) = 1 - \alpha$  for  $Z \sim N(0, 1)$

E.g. for a 95 % CI  $z_{\alpha/2} = 1.96$ .

## Example (ctd.)

A 95 % C.I. on the proportion of people who own a tablet is given by  $(\hat{p} - z_{\alpha/2}\text{SE}, \hat{p} + z_{\alpha/2}\text{SE})$  where  $\hat{p} = \frac{38}{250}$ ,  $z_{\alpha/2} = 1.96$ ,  $\text{SE}^2 = \frac{0.392(0.608)}{250}$ .

$$\left( 0.392 - 1.96\sqrt{\frac{0.392(0.608)}{250}}, 0.392 + 1.96\sqrt{\frac{0.392(0.608)}{250}} \right)$$

$$= (0.3315, 0.4525).$$

“We are 95% confident that proportion of people owning a tablet is somewhere in the interval (0.3315, 0.4525).”

# Hypothesis test for hypothesis about proportion

We can test hypotheses about the a population proportion:

$$H_0 : p = p_0 \quad \text{and} \quad H_1 : p \begin{matrix} \neq \\ > \\ < \end{matrix} p_0$$

Our test statistic is the  $z$ -value

$$z = \frac{\hat{p} - p_0}{\sqrt{p_0(1 - p_0)/n}}$$

where  $p_0$  is the null value, the value of  $p$  used in the null hypotheses.

The corresponding r.v.  $Z$  is approximately standard normal distributed for large  $n$ .

## Minimum sample size

$n$  is considered large enough if  $np_0 > 5$  and  $n(1 - p_0) > 5$  (both).

## Example

### Example

Newborn babies are more likely to be boys than girls. A random sample found 13 173 boys were born among 25 468 newborn children. The sample proportion of boys was 0.5172. Is this sample evidence that the birth of boys is more common than the birth of girls in the entire population? Let  $\alpha = 0.05$ .

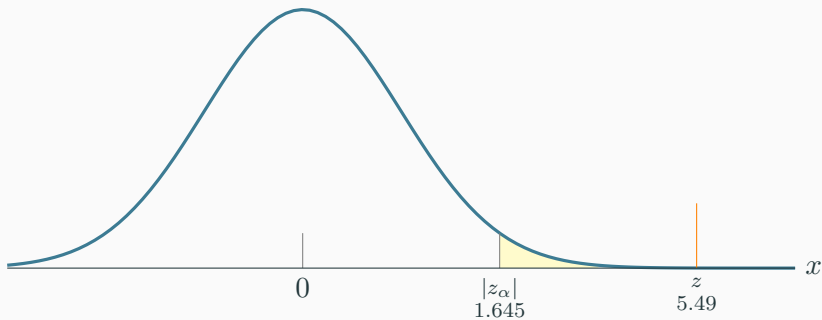
Test

$$H_0: p = 0.5 \quad \text{and} \quad H_1: p > 0.5.$$

at significance level  $\alpha = 0.05$ .

Since  $n$  is large,  $z = \frac{\hat{p}-0.5}{\sqrt{0.5(0.5)/25468}}$  is approximately normally distributed. The critical point is  $z_{0.95} = 1.645$  and  $z = \frac{0.5172-0.5}{\sqrt{0.5(0.5)/25468}} = 5.49$  which is in the rejection region.

Therefore  $H_0$  is rejected and hence the sample gives evidence that the proportion of boys is higher than that of girls.



Rejection region for  $\alpha = 0.05$  (on the  $x$ -axis below the yellow area).



## Comparing two proportions

---

Suppose we have two populations and we want to **compare** the proportions in the populations that have a certain trait. Denote the unknown proportions  $p_1$  and  $p_2$ .

### Example

We are interested in comparing the proportion of researchers who use a certain computer program in their research in two different fields: pure mathematics and probability and statistics.

**Populations:** Researchers in the pure math field and researchers in the probability and statistics field. **Trait of interest:** Usage of the computer program.

## Point estimator and SE for the difference between two proportions

---

Suppose that  $p_1$  is the true proportion of population 1 and  $p_2$  is that of population 2.

- From each population we take a random sample of sizes  $n_1$ ,  $n_2$  such that the samples are independent from each other.
- For each sample we compute the point estimate:  $\hat{p}_1$  and  $\hat{p}_2$ .
- A point estimator for  $p_1 - p_2$  is  $\hat{p}_1 - \hat{p}_2$ .
- For large samples,  $\hat{p}_1 - \hat{p}_2$  is approximately normal with mean  $p_1 - p_2$  and variance  $p_1(1 - p_1)/n_1 + p_2(1 - p_2)/n_2$  where  $n_1$  and  $n_2$  are the sample sizes from population 1 and 2 respectively.

## Confidence interval

A  $100(1 - \alpha)\%$  C.I. on  $p_1 - p_2$  is given by

$$(\hat{p} - z_{\alpha/2}\text{SE}, \hat{p} + z_{\alpha/2}\text{SE}) =$$

$$\hat{p}_1 - \hat{p}_2 \pm z_{\alpha/2} \sqrt{\hat{p}_1 (1 - \hat{p}_1) / n + \hat{p}_2 (1 - \hat{p}_2) / n_2}$$

## Example

---

We take a sample of size 375 from population 1 and 375 from population 2. The number of researchers that use a computer program we get from population 1 is 195 and that of researchers from population 2 is 232.

Then  $\hat{p}_1 = \frac{195}{375} = 0.52$  and  $\hat{p}_2 = \frac{232}{375} = 0.619$  A point estimate for the difference  $p_1 - p_2$  is  $0.52 - 0.619 = -0.099$ . The standard deviation is

$$\sqrt{0.52(0.48)/375 + 0.619(0.381)/375} = 0.036$$

## Example (ctd.)

---

A 95% confidence interval for  $p_1 - p_2$  is

$$(0.52 - 0.619 - 1.96(0.036), 0.52 - 0.619 + 1.96(0.036)) \\ (-0.17, -0.028)$$

Since the interval does not contain 0 and is negative-valued, we can say with 95% level of confidence that the proportion of researchers from population 2 is higher than that of population 1.

## Comparisons

---

# Comparisons

---

A common situation is that you want to make comparisons between different samples. Examples of when this may be of interest include

- We want to compare performances of two designs.
- We want to investigate the effect of a new drug.

Today we will examine two types of comparisons

- Independent samples (measurements of two populations)
- Paired samples (samples are pairs of related measurements)

## Paired samples

---



## Paired samples:

---

A common situation is that the measurements are made in pairs. For example when you take different measurements of the same subjects, e.g. strength of the right arm and strength of the left arm.

We set up a model which has  $n$  pairs of observations

$$X_1, Y_1, \quad X_2, Y_2, \quad \dots, \quad X_n, Y_n.$$

For each measurement, we form the difference, which is assumed to be normally distributed:

$$D_i = X_i - Y_i \stackrel{iid}{\sim} \mathcal{N}(\mu_{\text{diff}}, \sigma^2)$$

Summary: We test whether  $H_0: \mu_{\text{diff}} = 0$  against an alternative. This is done as usual for normally distributed measurements with known or unknown variance.

## Independent samples

---

# Independent samples

---

Assume we have two independent samples from different populations:

- $n_1$  observations  $X_1, X_2, \dots, X_{n_1}$  from  $N(\mu_1, \sigma_1^2)$ .
- Also  $n_2$  observations  $Y_1, Y_2, \dots, Y_{n_2}$  from  $N(\mu_2, \sigma_2^2)$ .

Summary: Build test/CI for  $H :: \mu_1 - \mu_2$ . We'll start with estimator  $\bar{D} = \bar{X} - \bar{Y}$  of  $\mu_1 - \mu_2$ .

## Paired or not

---

1. Compare pre-class (beginning of semester) and post-class (end of semester) scores of students. Paired.
2. Assess gender-related salary gap by comparing salaries of 10 randomly sampled men and 12 women. Independent.
3. Compare artery thicknesses at the beginning of a study and after 2 years of taking Vitamin E for the same group of patients. Paired.
4. Measure the strength of the left arm vs right arm of each subject. Paired.
5. Assess gender-related salary gap by comparing salaries of 10 randomly sampled men and 10 women. Still independent.

## Example for samples

You would like to know whether a new wheat variety yields a higher harvest than the existing variety. You select six fields that differ in fertility and climate, and divide each field into two parts in which each variety is grown.

Field nr	1	2	3	4	5	6
Harvest sort 1, kg/ha	7529	8913	6534	6503	6896	8023
Harvest sort 2, kg/ha	7239	8726	6129	6351	6644	7711
Difference $D_i$	290	187	405	152	252	312

We test  $H_0 : \mu_{\text{diff}} = 0$  against  $H_1 : \mu_{\text{diff}} \neq 0$  at level  $\alpha = 0.05$ . We have  $\bar{D} = 266.3$  and  $s_D = 91$  and look up  $t_{0.025}(5) = 2.57$

$$I_{\mu_{\text{diff}}} = (\bar{D} \pm t_{0.025}(5) \cdot s_D / \sqrt{6}) = (171, 362)$$

As  $0 \notin I_{\mu_{\text{diff}}}$  we reject  $H_0$ .

## Independent samples

---

Assume we have two independent samples

- $n_1$  observations  $X_1, X_2, \dots, X_{n_1}$  from  $N(\mu_1, \sigma_1^2)$ .
- Also  $n_2$  observations  $Y_1, Y_2, \dots, Y_{n_2}$  from  $N(\mu_2, \sigma_2^2)$ .

We want to test whether  $\mu_1$  and  $\mu_2$  differ ( $H_0: \mu_1 = \mu_2$ ).

Introduce  $\mu_{\text{diff}} = \mu_1 - \mu_2$  with estimator  $\bar{D} = \bar{X} - \bar{Y}$ . Test

$$H_0: \mu_{\text{diff}} = 0,$$

$$H_1: \mu_{\text{diff}} \neq 0 \quad (\text{or against } H_1: \mu_{\text{diff}} > 0, \text{ or } \dots)$$

But what is the standard error??

## 3 cases

---

We distinguish between 3 cases:

**Case 1:**  $\sigma_1$  and  $\sigma_2$  are known.

**Case 2:**  $\sigma_1 = \sigma_2 = \sigma$  where  $\sigma$  is unknown.

**Case 3:**  $\sigma_1$  and  $\sigma_2$  are unknown and not necessarily the same.

If the case is not known, we may first have to test whether  $\sigma_1 = \sigma_2$  with the

**Preliminary test:**

$$H_0 : \frac{\sigma_1}{\sigma_2} = 1$$

$$H_1 : \frac{\sigma_1}{\sigma_2} \neq 1$$

## Case 1: Known $\sigma_1$ and $\sigma_2$

If  $\sigma_1$  and  $\sigma_2$  are known it holds that

$$\text{SE} = \text{SE}(\bar{X}_1 - \bar{X}_2) = \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}.$$

In a hypothesis test we use that under  $H_0$

$$Z = \frac{\bar{X}_1 - \bar{X}_2}{\text{SE}} \sim \text{N}(0, 1)$$

with p-value  $p = 2(1 - \Phi(|Z_{\text{obs}}|))$ .

A confidence interval for  $\mu_{\text{diff}} = \mu_1 - \mu_2$  is given by

$$I_{\mu_{\text{diff}}} = (\hat{\mu}_{\text{diff}} \pm z_{\alpha/2} \text{SE}) = \left( \bar{x}_1 - \bar{x}_2 \pm z_{\alpha/2} \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}} \right)$$



## Case 2: $\sigma_1 = \sigma_2 = \sigma$ where $\sigma$ unknown

### Pooled estimate of variance

For 2 normally distributed samples  $N(\mu_j, \sigma^2)$ ,  $j = 1, 2$  an unbiased estimate of  $\sigma^2$  is

$$s_p^2 = \frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{(n_1 - 1) + (n_2 - 1)}. \quad \text{Step 1!}$$

With

$$\text{SE} = \sqrt{s_p^2 \left( \frac{1}{n_1} + \frac{1}{n_2} \right)} \quad \text{Step 2!}$$

one has under  $H_0$  that

$$T = \frac{\bar{X}_1 - \bar{X}_2}{\text{SE}} \sim t(n_1 + n_2 - 2)$$

Confidence interval:  $I_{\mu_{\text{diff}}} = (\bar{x}_1 - \bar{x}_2 \pm t_{\alpha/2}(n_1 + n_2 - 2) \text{SE})$ .

### Case 3: $\sigma_1 \neq \sigma_2$ unknown

#### Theorem

For two normally distributed samples

$$T = \frac{\bar{X}_1 - \bar{X}_2}{\sqrt{s_1^2/n_1 + s_2^2/n_2}}$$

is approximately  $t(df)$ -distributed where

$$df = \frac{(s_1^2/n_1 + s_2^2/n_2)^2}{\frac{(s_1^2/n_1)^2}{n_1-1} + \frac{(s_2^2/n_2)^2}{n_2-1}}$$

We can now create confidence intervals and perform hypothesis tests in the same way as before:

$$I_{\mu_{\text{diff}}} = \left( \hat{\mu}_{\text{diff}} \pm t_{\alpha/2}(f) \sqrt{s_1^2/n_1 + s_2^2/n_2} \right).$$

### Example (Exercise 10.14)

To decide whether or not to purchase a new hand-held laser scanner for use in inventorying stock, tests are conducted on the scanner currently in use and on the new scanner. There data are obtained on the number of 7-inch bar codes that can be scanned per second:

new	old
$n_1 = 61$	$n_2 = 61$
$\bar{x}_1 = 40$	$\bar{x}_2 = 29$
$s_1^2 = 24.9$	$s_2^2 = 22.7$

1. Find the pooled variance.
2. Find a 90% CI on  $\mu_1 - \mu_2$ .
3. Does the new laser appear to read more bar codes per second on the average?

1. Find the pooled variance.

$$s_2^p = \frac{60(24.9) + 60(22.7)}{120} = 23.8$$

2. Find a 90% CI on  $\mu_1 - \mu_2$ .

$t$ -distribution with  $df = 120$ .  $t_{\alpha/2} = t_{0.05} = 1.658$  (note that the table does not give the values for degrees of freedom greater than 100 , use then an approximation). A 90% CI is therefore

$$(40 - 29 \pm 1.658\sqrt{23.8(1/61 + 1/61)}) = (9.54, 12.45)$$

3. Does the new laser appear to read more bar codes per second on the average?

Yes, since the interval does not contain 0 and is positive-valued.

## Preliminary test: Comparison of variance

---

Denote with  $F_\alpha(df_1, df_2)$  the  $\alpha$ -quantile of the  $F$ -distribution. A confidence interval for  $\sigma_1^2/\sigma_2^2$  is

$$I_{\sigma_1^2/\sigma_2^2} = \left[ \frac{s_1^2/s_2^2}{F_{\alpha/2}(n_1 - 1, n_2 - 1)}, \frac{s_1^2/s_2^2}{F_{1-\alpha/2}(n_1 - 1, n_2 - 1)} \right]$$

Use for a hypothesis test  $H_0: \sigma_1^2/\sigma_2^2 = 1$  (same as  $H_0: \sigma_1^2 = \sigma_2^2$ )

.

# Linear regression

---

# What is linear regression

---

Regression is a technique used for estimating the relationship between variables.

Often we want to predict a variable  $Y$  (the dependent variable) in terms of another variable  $x$  (the independent variable) (or more generally understand the relationship between  $Y$  and  $x$ ).

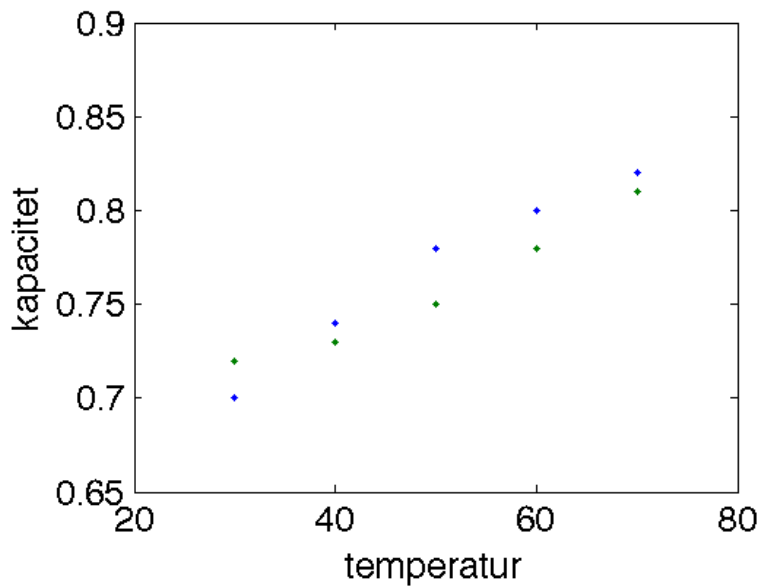
## Example

We want to investigate how the specific heat capacity of a substance (the ability of the substance to store heat energy) depends on temperature.

For each of the five temperatures, two heat capacity measurements are made with the following results:

Temperature ( $^{\circ}\text{C}$ )	30	40	50	60	70
Heat capacity	0.70	0.74	0.78	0.80	0.82
	0.72	0.73	0.75	0.78	0.81





## Model description

---

We have measured a response variable  $Y$  for fixed values of an explanatory variable  $x$  that can be controlled without errors.

We use a linear model for  $(Y_i, x_i), i = 1, \dots, n$ :

$$Y_i = \beta_0 + \beta_1 x_i + \varepsilon_i \quad (20.1)$$

- $\varepsilon_i$  are independent  $N(0, \sigma^2)$  random variables describing measurement errors.
- $\beta_0$  is the intercept parameter.
- $\beta_1$  is the slope parameter.

Another way of writing the model is

$$Y_i \sim \text{N}(\beta_0 + \beta_1 x_i, \sigma^2).$$

The expected value of  $Y$  is determined by the linear relationship with  $x$ , and the variance of measurement error  $\sigma^2$  describes the variation of the individual observations around the expected value  $\beta_0 + \beta_1 x$ . **Assumption:**  $Y_i$  are independent..

Given a sample (visualized by a scatterplot)

$$(Y_1, x_1), (Y_2, x_2), \dots, (Y_n, x_n)$$

we want to estimate the line with parameters  $\beta_0$  and  $\beta_1$  as well as  $\sigma^2$ , the variation of the  $Y_i$ -values from the regression line  $\beta_0 + \beta_1 x$  at  $x_i$ .

With the estimated parameters, we can predict  $Y$  for a given value of  $x$ .

## Least squares estimator

---

$\beta_0$  and  $\beta_1$  are estimated by the method of least-squares which is done by minimizing

$$\text{SSE} = \sum_{i=1}^n \epsilon_i^2 = \sum_{i=1}^n (y_i - \beta_0 - \beta_1 x_i)^2$$

Let  $b_0$  and  $b_1$  values of  $\beta_0$  and  $\beta_1$  respectively minimizing the SSE. Then,

$$b_1 = \frac{n \sum_{i=1}^n x_i y_i - \left( \sum_{i=1}^n x_i \right) \left( \sum_{i=1}^n y_i \right)}{n \sum_{i=1}^n x_i^2 - \left( \sum_{i=1}^n x_i \right)^2}$$

and

$$b_0 = \bar{y} - b_1 \bar{x}$$

## Least squares estimator

---

An estimator for the variance parameter  $\sigma^2$  is  $s^2 = \frac{Q_0}{n-2}$  where

$$Q_0 = \sum_{i=1}^n (y_i - b_0 - b_1 x_i)^2$$

( $b_0$  and  $b_1$  your estimates).

## Different way of computing the estimate

The LS-estimators for  $\beta_0$  and  $\beta_1$  are

$$b_1 = S_{xy}/S_{xx} \text{ and } b_0 = \bar{y} - b_1\bar{x}$$

where

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

$$S_{yy} = \sum_{i=1}^n (y_i - \bar{y})^2 = \sum_{i=1}^n y_i^2 - n\bar{y}^2$$

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

An estimator for the variance parameter  $\sigma^2$  is  $s^2 = \frac{Q_0}{n-2}$  where

$$Q_0 = \sum_{i=1}^n (y_i - b_0 - b_1 x_i)^2 = S_{yy} - b_1 S_{xy} = S_{yy} - \frac{S_{xy}^2}{S_{xx}}$$

## Estimators for the example

We estimate parameters of the regression line in the example. We have  $\bar{x} = 50$ ,  $\bar{y} = 0.763$  and

$$S_{xx} = \sum_{i=1}^{10} x_i^2 - 10\bar{x}^2 = 27000 - 10 \cdot 50^2 = 2000$$

$$S_{yy} = \sum_{i=1}^{10} y_i^2 - 10\bar{y}^2 = 5.8367 - 10 \cdot 0.763^2 = 0.01501$$

$$S_{xy} = \sum_{i=1}^{10} x_i y_i - 10\bar{x}\bar{y} = 386.8 - 10 \cdot 50 \cdot 0.763 = 5.3$$

and therefor the estimate

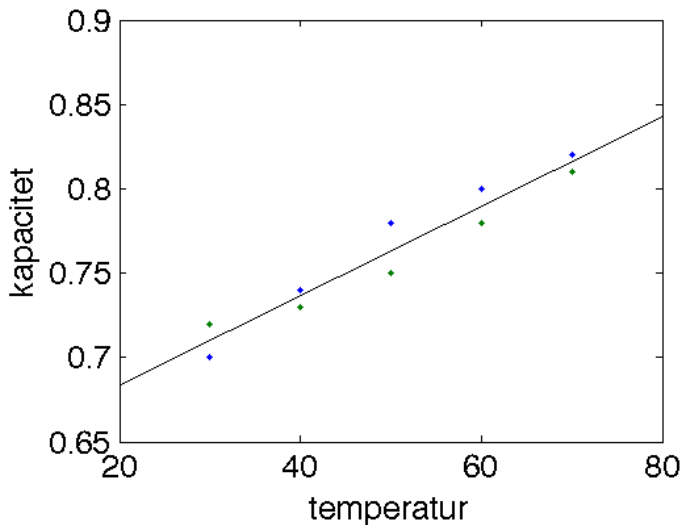
$$b_1 = S_{xy}/S_{xx} = 5.3/2000 = 0.00265$$

$$b_0 = \bar{y} - b_1\bar{x} = 0.6305$$

$$s^2 = \frac{1}{n-2} \left( S_{yy} - \frac{S_{xy}^2}{S_{xx}} \right) = 0.00012, \quad s = \sqrt{0.00012} = 0.011$$



The estimated regression line is  $b_0 + b_1x$



### Shortcut

The book often uses

$$\frac{S_{xy}}{S_{xx}} = \frac{nS_{xy}}{nS_{xx}} = \frac{n \sum_{i=1}^n x_i y_i - (\sum_{i=1}^n x_i)(\sum_{i=1}^n y_i)}{n \sum_{i=1}^n x_i^2 - (\sum_{i=1}^n x_i)^2}$$

## Example 2

---

Let  $X$  denote the number of lines of executable SAS code, and let  $Y$  denote the execution time in seconds. The following is a summary information:

$$n = 10 \quad \sum_{i=1}^{10} x_i = 16.75 \quad \sum_{i=1}^{10} y_i = 170$$

$$\sum_{i=1}^{10} x_i^2 = 28.64 \quad \sum_{i=1}^{10} y_i^2 = 2898 \quad \sum_{i=1}^{10} x_i y_i = 285.625$$

Estimate the line of regression.

## Example 2

---

Using the **shortcut**...

$$b_1 = \frac{10(285.625) - (16.75)(170)}{10(28.64) - (16.75)^2} = 1.498$$

$$b_0 = \frac{170}{10} - 1.498 \frac{16.75}{10} = 14.491$$

Estimated model:

$$Y_i = 1.498x_i + 14.491 + \epsilon_i$$

Our estimator for  $\beta_1$  is  $B_1 = \hat{\beta}_1$  (the random quantity w. value  $b_1$ ).

### Properties of the estimator for the slope

We have  $E(\bar{Y}) = \beta_0 + \beta_1 \bar{x}$  and  $V(\bar{Y}) = \frac{\sigma^2}{n}$ . The book shows using  $\sum_{i=1}^n (x_i - \bar{x}) = 0$  and the rules of expectation and variance

$$E(B_1) = \beta_1 \qquad V(B_1) = \frac{\sigma^2}{S_{xx}} = \frac{\sigma^2}{\sum_{i=1}^n (x_i - \bar{x})^2}$$

So we see that  $B_1$  is an unbiased estimator for  $\beta_1$ .

Our estimator for  $\beta_0$  is  $B_0 = \hat{\beta}_0$  (the random quantity with value  $b_0$ .)  $\hat{\mu}_0(x_0) = B_0 + B_1x_0$  is an estimator for  $E(\beta_0 + \beta_1x_0)$  (=  $EY$  if  $Y = \beta_0 + \beta_1x_0 + \epsilon$ )

### Properties of estimators for intercept and prediction of $Y$

With  $\hat{\mu}_Y(x_0) = B_0 + B_1x_0$  also

$$E(\hat{\mu}_Y(x_0)) = \beta_0 + \beta_1x_0$$

with

$$V(\hat{\mu}_Y(x_0)) = \sigma^2 \left[ \frac{1}{n} + \frac{(x_0 - \bar{x})^2}{S_{xx}} \right]$$

With  $x_0 = 0$  we see that  $B_0$  is unbiased.

## Distribution of the estimators

### Theorem

For normally distributed  $\varepsilon_i$  it holds that  $\bar{Y}$ ,  $B_0$ ,  $B_1$  and  $\hat{\mu}_Y(x_0) = B_0 + B_1x_0$  are also normally distributed.

Because the estimator is a sum of  $Y_i$ , by the CLT this also holds approximately if the distribution of the  $\varepsilon_i$  deviates from the normal distribution.

### Theorem

If  $\varepsilon_i$  is normally distributed it holds that

$$\frac{(n-2)S^2}{\sigma^2} \sim \chi^2(n-2)$$

further  $S^2$  is independent of  $\bar{Y}$ ,  $B_0$ ,  $B_1$  and  $\hat{\mu}_Y(x_0)$ .

## Confidence interval and test

---

Let  $\theta$  one of  $\beta_0$ ,  $\beta_1$  or  $\mu_Y(x_0) = \beta_0 + \beta_1 x_0$ .

We know that these estimates are normally distributed and have determined the variance of the estimates.

If  $\text{SE}(\hat{\theta})$  denotes the standard error of the estimator, the statistic

$$T = \frac{\hat{\theta} - \theta}{\text{SE}(\hat{\theta})} \sim t(n - 2)$$

is often used for tests and a confidence interval is,

$$I_{\theta} = (\hat{\theta} \pm t_{\alpha/2}(n - 2) \text{SE}(\hat{\theta}))$$



## Example

Consider the previous example and suppose we want to see if there is a relation between  $X$  and  $Y$  with a significance level  $\alpha = 5\%$ . There is a relation between  $X$  and  $Y$  if and only if  $\beta_1 \neq 0$ , which is our alternative hypothesis. Let  $H_0 : \beta_1 = 0$ . We have a two tailed test.

$b_1 = 1.498$ ,  $S_{xx} = \left( n \sum_{i=1}^n x_i^2 - (\sum_{i=1}^n x_i)^2 \right) / n = 0.584$ ,  $S_{yy} = 8$  and  $S_{xy} = 0.875$ .

Therefore  $SSE = 8 - 1.498(0.875) = 6.69$  and

$$s^2 = SSE/8 = 0.84$$

The test statistic is

$$T = \frac{b_1 - 0}{\sqrt{S^2/S_{XX}}} = \frac{1.498}{\sqrt{0.84/0.584}} = 1.25$$

$t_{0.025} = 2.306$ . Hence, we do not reject the hypothesis.

## Example

---

A 95% C.I. on  $\beta_0$  in our previous example is given by

$$\begin{aligned} &14.491 \pm 2.306\sqrt{0.84(28.64)/5.84} \\ &(14.491 - 4.68, 14.491 + 4.68) \\ &(9.81, 19.181) \end{aligned}$$

We are 95% sure that the true regression line crosses the  $y$ -axis between the points  $y = 9.81$  and  $y = 19.81$ .

## Confidence interval

- Confidence interval for  $\beta_0$ :

$$I_{\beta_0} = \left( \hat{\beta}_0 \pm t_{\alpha/2}(n-2)s\sqrt{\frac{1}{n} + \frac{\bar{x}^2}{S_{xx}}} \right)$$

- Confidence interval for  $\beta_1$ :

$$I_{\beta_1} = \left( \hat{\beta}_1 \pm t_{\alpha/2}(n-2)\frac{s}{\sqrt{S_{xx}}} \right)$$

- Confidence interval for  $\mu_Y(x_0) = \beta_0 + \beta_1 x_0$ :

$$I_{\mu_Y(x_0)} = \left( \hat{\beta}_0 + \hat{\beta}_1 x_0 \pm t_{\alpha/2}(n-2)s\sqrt{\frac{1}{n} + \frac{(x_0 - \bar{x})^2}{S_{xx}}} \right)$$

## Prediction interval

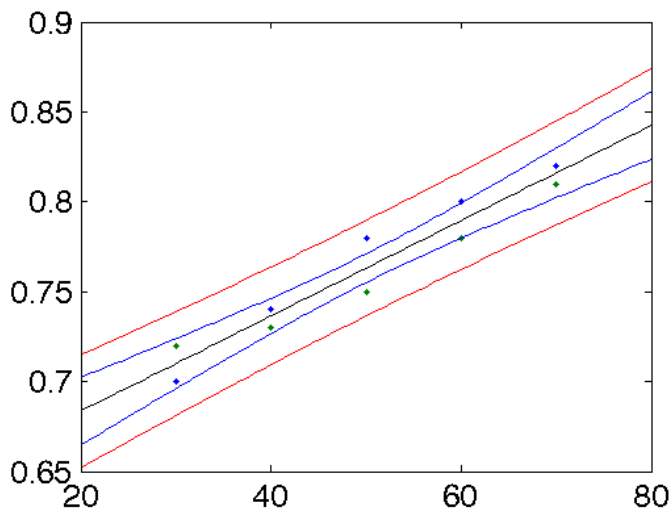
- Sometimes you want to know where a future observation will be for a certain value of  $x$ , for this use a **prediction interval**:
- The difference between a prediction interval  $I_{Y(x_0)}$  and a confidence interval  $I_{\mu_Y(x_0)}$  is that  $I_{\mu_Y(x_0)}$  indicates where the expected value (the line!) is likely, while  $I_{Y(x_0)}$  indicates where a future observation is likely.
- Since observations scatter around the regression line, the prediction interval must be wider than the confidence interval, and it can be shown that

$$\hat{Y}(x_0) \sim N \left( \beta_0 + \beta_1 x_0, \sigma^2 \left( 1 + \frac{1}{n} + \frac{(x_0 - \bar{x})^2}{S_{xx}} \right) \right).$$

The prediction interval is

$$I_{Y(x_0)} = \left[ \hat{\beta}_0 + \hat{\beta}_1 x_0 \pm t_{\alpha/2}(n-2) s \sqrt{1 + \frac{1}{n} + \frac{(x_0 - \bar{x})^2}{S_{xx}}} \right]$$

## Confidence interval and prediction interval



# Model validation

---

## Model validation

---

A very important part of a regression analysis is the validation of the model. This means that we must ensure that it is appropriate to use a simple regression model. The most common method for this is the calculation of residuals.

$$e_i = y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i$$

For the regression to be valid the residuals

- must be distributed approximately normally with expected value 0,
- do not reveal any special structure as a function of  $x$ .
- Have about the same variation for all different values of  $x$ . For example, the variance for large values of  $x$  should not increase.

Check this visually by drawing the residuals as a function of  $x$  and using normal distribution plots.

## Example

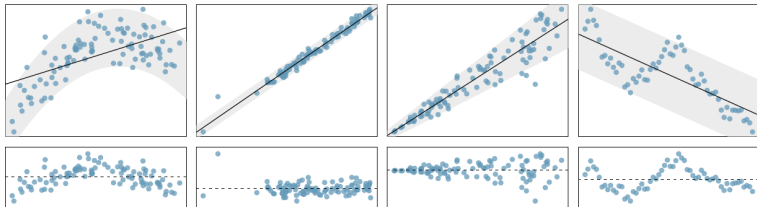


Figure 8.12: Four examples showing when the methods in this chapter are insuf-



## Example

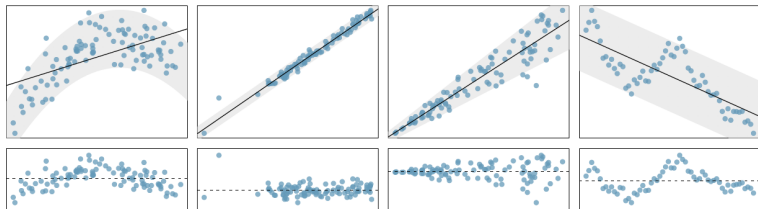


Figure 8.12: Four examples showing when the methods in this chapter are insufficient to apply to the data. First panel: linearity fails. Second panel: there are outliers, most especially one point that is very far away from the line. Third panel: the variability of the errors is related to the value of  $x$ . Fourth panel: a time series data set is shown, where successive observations are highly correlated.