

# Hypothesis Formulation and Testing

2IAB1

Week 6

2023-2024

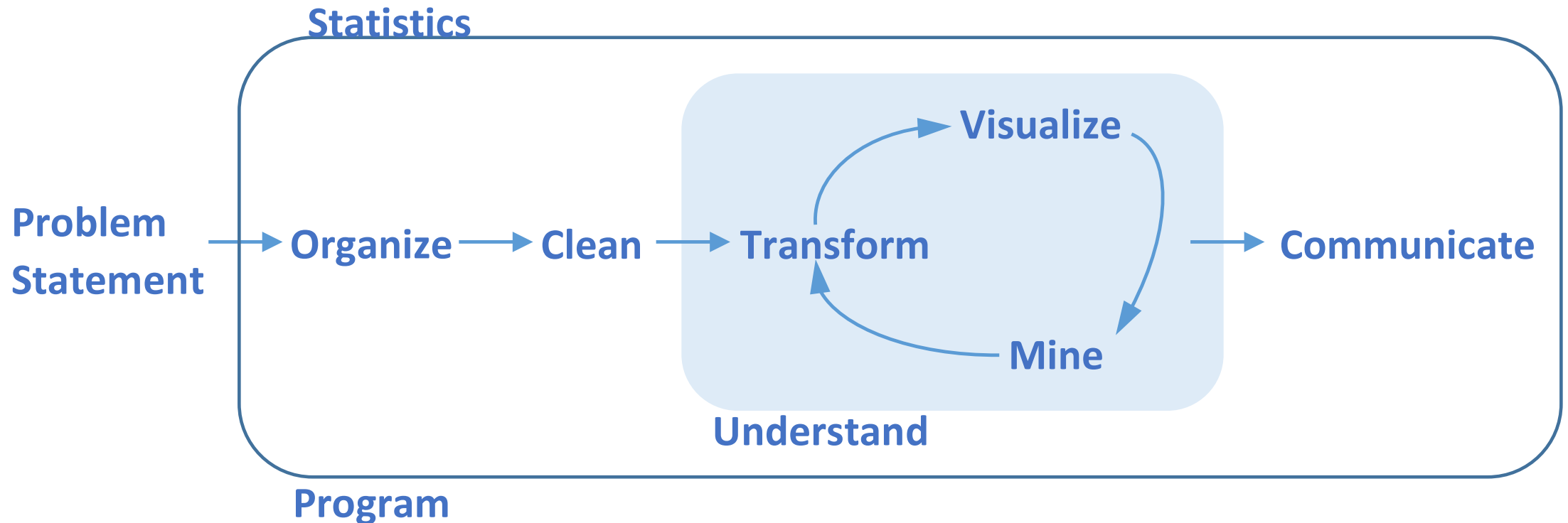


# Today: the last module of new material

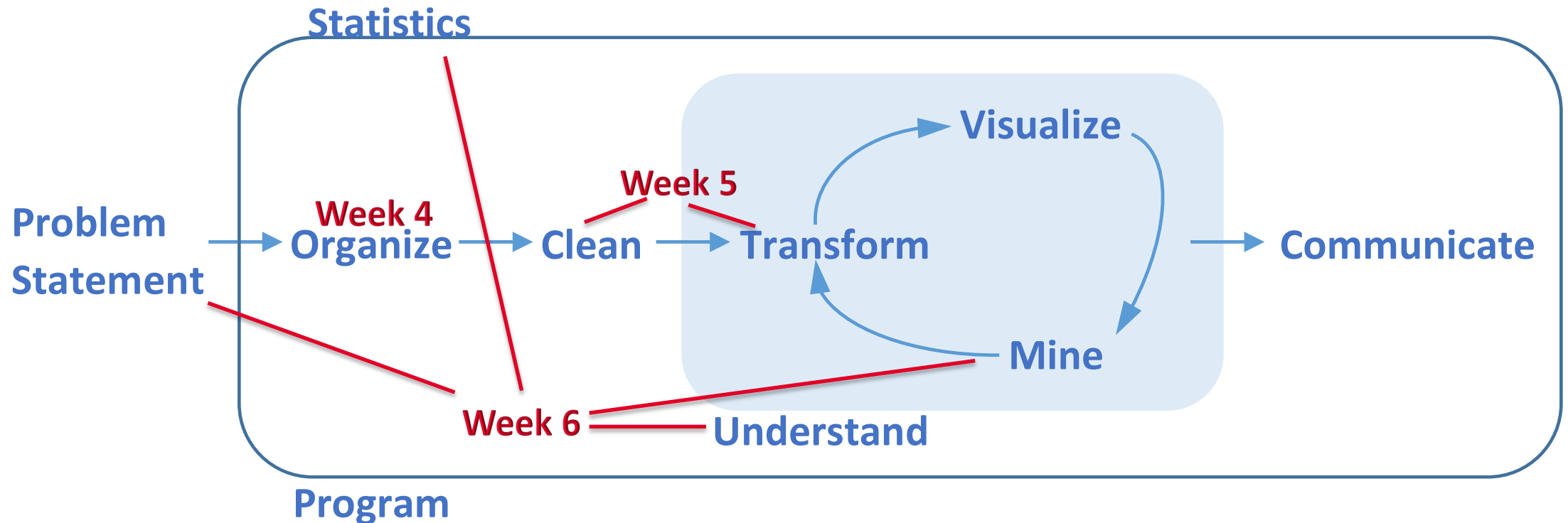
## What's next (in the New Year)?

- Week 7: focus on GA2
  - During the lecture:
    - discussion of pitfalls in GA2 and
    - recommendations for pitches (by *Career & Professional Skills* coaches)
  - During lab sessions and at home: GA2
- Week 8: GA2 pitches and learning for the exam
  - During the lecture: preparation to the exam
  - During lab sessions: the GA2 grand finale – a pitch session (compulsory presence!!!)
  - At home: exam practice with exercises from old exams (see the Canvas homepage)

# Where Do We Stand?



# Lectures 4-6 → Assignment 2



# Learning objectives

---

- learn to express uncertainty about summary statistics through confidence intervals
- learn to translate problem statements into statistical hypotheses
- learn to perform and interpret hypothesis tests
- learn to check assumptions on statistical and data mining methods

# Discrete probability distributions:

## Discrete uniform distribution and binomial distribution

# Discrete uniform distribution

- **discrete uniform distribution** – all outcomes have the same **probability**:

$$P(X = k) = \frac{1}{n}, \quad k = 1, \dots, n$$

- $X$  is a random variable denoting the outcome
- The number of possible outcomes is finite or countable

- **Example:** the number of eyes when throwing a *fair* dice



$$P(X = 1) = \dots = P(X = 6) = \frac{1}{6}$$

- **Definition:** A **probability** is a number in the interval  $[0, 1]$  that indicates how likely a certain outcome is (the higher, the more likely the outcome).
- A **probability mass function (pmf)**  $P(X = x)$  describes the probability distribution for a **discrete random variable**  $X$  assigning a probability to a every possible **outcome**  $x$ .
- **Key property of probabilities:** The sum of probabilities over all possible outcomes is equal to 1.

# Probabilities – fair coin example

H – head  
T – tail



Modelling successes (yes/no data) is like modelling flipping coins

- a coin might be **biased**, i.e. **not fair**: heads and tails might be **not** equally likely

How likely do we get **HHT** if we flip a fair coin? (HHT = Head, Head, Tail)

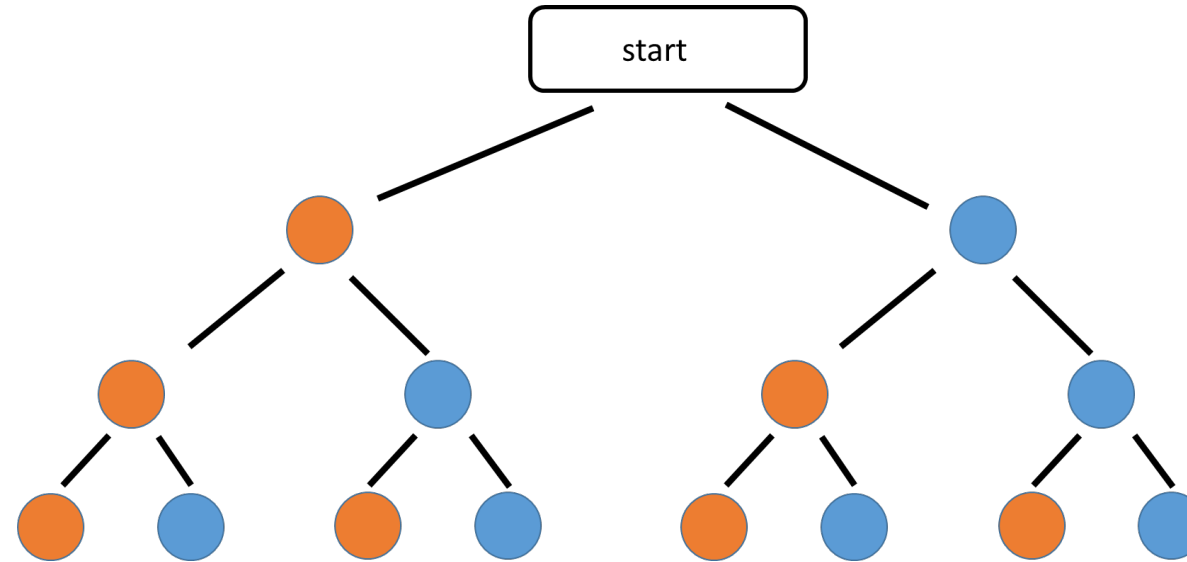
We indicate **H** with an **orange disc** and **T** with a **blue disc**.

This boils down to counting sequences of two possible items (**H** and **T**).

*You can do such a calculation for any other sequence in the same way.*



# Possible sequences – equal probabilities



The number of possible sequences  $2 \times 2 \times 2 = 8$ : HHH, HHT, HTH, HTT, THH, THT, TTH, TTT.

Thus, the **probability** of HHT =  $1/8$ . Notation  $P(HHT) = \frac{1}{8}$ .

What is the probability of getting heads 1 time? Options: HTT, THT, TTH, so probability =  $3/8$

What is the probability of 0 or 1 heads? TTT, HHT, THH, HTH, so probability =  $4/8$

# Calculating probabilities



**Exercise 1:** What is the probability to get *only heads* in a sequence of length 10?

There are  $2 \cdot 2 \cdot 2 \dots \cdot 2 = 2^{10}$  different sequences.

So the probability of only heads  $P(10 \text{ heads}) = \frac{1}{2^{10}} = \frac{1}{1024}$

**Note:** 10 heads is the same as 0 tails.

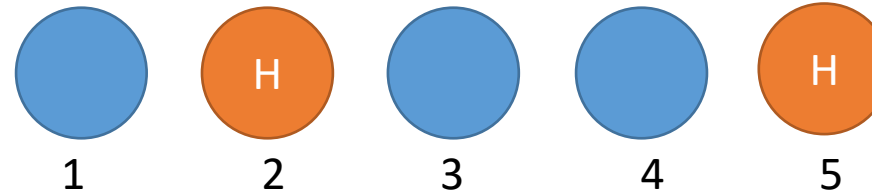
**Exercise 2:** What is the probability of having at least one head in a sequence of length 10?

**Use the key property:** *Probabilities of all possible disjoint outcomes add up to 1.*

$P(0 \text{ tails}) + P(\text{at least one head}) = 1$  (no other possibilities)

So,  $P(\text{at least one head}) = 1 - P(0 \text{ tails}) = 1 - \frac{1}{1024} = \frac{1023}{1024}$ .

# Binomial coefficients



**How many options are there to get 2 heads in 5 tries?**

Choose 2 positions for heads.

5 possibilities for the first head and 4 for the second one:

(1,2); (1,3); (1,4); (1,5); (2,1);... ; (3,1);... ; (4,1); ... ; (4,5) –  $5 \times 4 = 20$  options,

but double counts: heads on positions 1,2 = heads on positions 2,1 !

So there are  $\frac{5 \times 4}{2} = 10$  options

# General properties binomial coefficients

**Definition:** Factorial:  $n! = n \times (n - 1) \times \cdots \times 2 \times 1$ . Extra:  $0! = 1$

**Definition: Binomial coefficient**  $\binom{n}{k} \stackrel{\text{def}}{=} \frac{n!}{k!(n-k)!}$  is the number of sequences consisting of  $k$  items of one type and  $(n - k)$  items of another type

$\binom{n}{k}$  is pronounced as “ **$n$  choose  $k$** ”

**Note:** there is symmetry between the two types (5 coins choose 3 heads = 5 coins choose 2 tails),  
so we have  $\binom{n}{k} = \binom{n}{n-k}$

**Side remark:** the binomial coefficient also appears in formulas for expressions like

$$(a + b)^5 \text{ (“binomial formula of Newton”)}$$

# Application of binomial coefficients



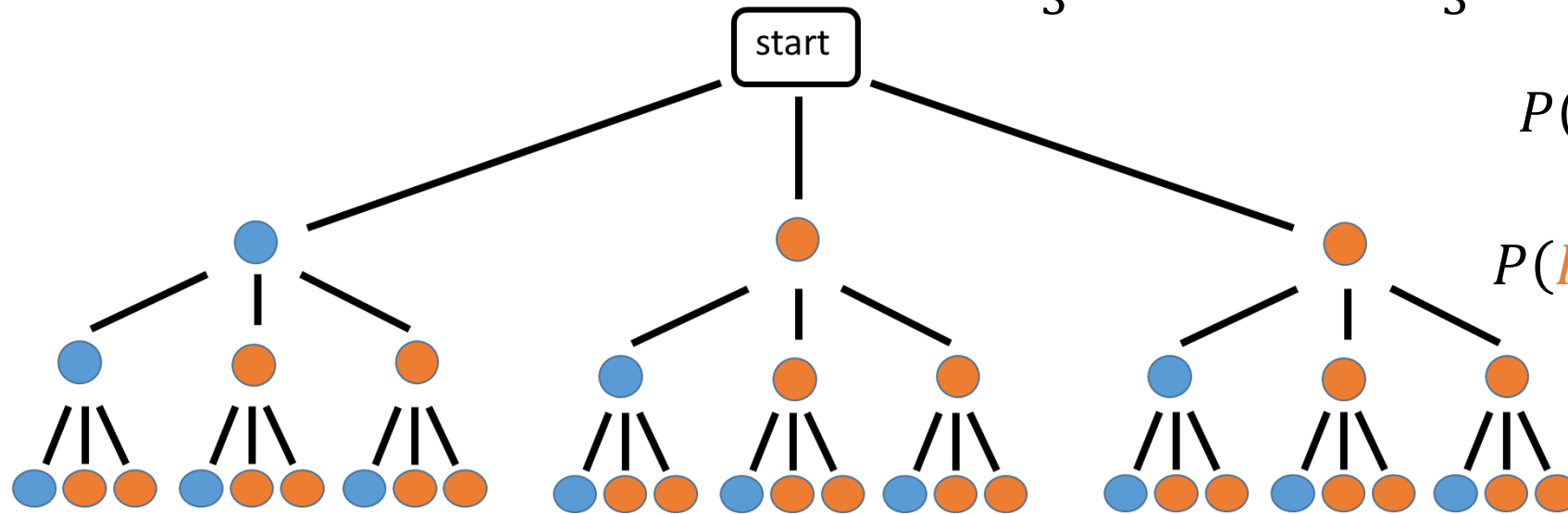
Probability of 2 heads in 5 coin flips?

- $\binom{5}{2} = \frac{5!}{2!3!} = \frac{120}{2 \cdot 6} = 10$  sequences with exactly 2 H in 5 coin flips.
- The number of possible H/T sequences of length 5:  $2 \times 2 \times 2 \times 2 \times 2 = 2^5 = 32$
- Probability:  $P(2\text{H in 5 tries}) = \frac{\binom{5}{2}}{2^5} = \frac{10}{32}$

# Probabilities – biased coin

## (unequal probabilities for head and tail)

Example: a biased coin with  $P(H) = \frac{2}{3}$  and  $P(T) = \frac{1}{3}$  and 3 coin flips.



$$P(HHT) = \frac{2}{3} \times \frac{2}{3} \times \frac{1}{3} = \frac{4}{27}$$

$$P(HHT) = P(HTH) = P(THH)$$

There are  $\binom{3}{2} = 3$  **sequences** with 2 heads and 1 tail, the probability of each of them equals  $\frac{4}{27}$

$$P(2 \text{ heads}) = 3 \times \frac{4}{27} = \frac{12}{27}.$$

# General setting: binomial distribution

Let  $P(H) = p$ , so  $P(T) = 1 - p$ , and let  $X$  be a **random variable** denoting the **number of heads in  $n$  coin flips**.

Then the probability to get  $k$  heads in  $n$  coin flips is:

$$P(X = k) = \binom{n}{k} p^k (1 - p)^{n-k}, \quad k = 0, 1, \dots, n$$

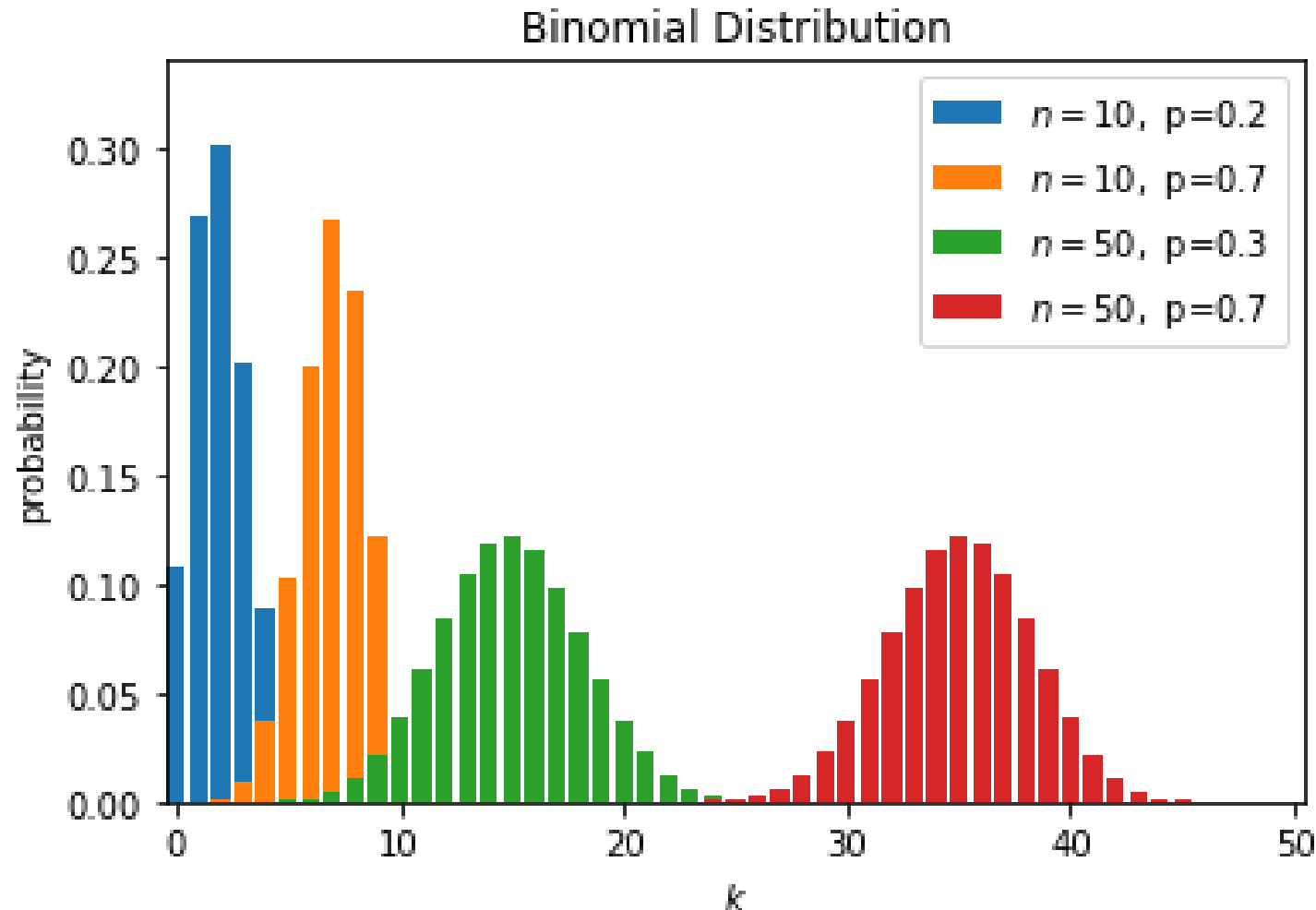
**The same formula holds in the general setting:**  $n$  *independent* observations, each with success probability  $p$

- i.e. failure probability =  $1 - p$
- “successes” and “failures” instead of heads and coins, e.g. a train coming on time or not

This mathematical object is known as the **binomial probability distribution** (shortened to binomial distribution).

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

# Examples binomial probability distributions





# Cumulative distribution function

The **probability mass function (pmf)**  $P(X = k) = \binom{n}{k} p^k (1 - p)^{n-k}$  gives probabilities for **exactly**  $k$  successes.

The **cumulative distribution function**  $F$  (often abbreviated as **distribution function**) gives **cumulative** probabilities, for **at most**  $l$  successes:

$$F(\ell) = P(X \leq \ell) = \sum_{k=0}^{\ell} \binom{n}{k} p^k (1 - p)^{n-k}$$

Note: For the binomial distribution,  $P(X = k) = P(X \leq k) - P(X \leq k - 1)$

# Some other probability distributions

The binomial distribution is just an example of a probability distribution.  
Other probability distributions include (***no exam questions about them***):

- **geometric distribution** (how many tries you need before the first success)

$$P(X = k) = (1 - p)^{k-1}p,$$



- **Poisson distribution**  $P(X = k) = e^{-\lambda} \frac{\lambda^k}{k!}, k = 0, 1, 2, \dots$

(for modelling counts of “rare” events, e.g. the number of calls per hour in a call centre)



# Expectation (mean): averaging probabilities

The analogue of the *sample mean* for a probability distribution is called the **expectation (mean)**.

For a random variable  $X$  with a finite or a countable set of outcomes:

$$E(X) = \sum_k kP(X = k)$$

For a dice:

$$E(X) = \frac{1}{6} \cdot 1 + \frac{1}{6} \cdot 2 + \frac{1}{6} \cdot 3 + \frac{1}{6} \cdot 4 + \frac{1}{6} \cdot 5 + \frac{1}{6} \cdot 6 = 3.5$$

For a **binomial distribution** one can derive an explicit formula:

$$E(X) = np$$

# Variance

Sample variance: a number that indicates the spread of a data set.

**Variance:** the analogue of sample variance for a probability distribution:

$$\text{Var}(X) = \sum_k (k - E(X))^2 P(X = k) = E(X - E(X))^2$$

**For a dice:**  $\text{Var}(X) = ((1 - 3.5)^2 + (2 - 3.5)^2 + (3 - 3.5)^2 + (4 - 3.5)^2 + (5 - 3.5)^2 + (6 - 3.5)^2) \cdot \frac{1}{6} = 2\frac{11}{12}$

**For the binomial distribution:**  $\text{Var}(X) = np(1 - p)$

# Summary discrete probability distributions

- Probabilities are numbers between 0 and 1
- The **probabilities** of all possible **disjoint** outcomes **sum up to 1** and this may be used to simplify computations of complicated outcomes.
- The **binomial** distribution is a discrete probability distribution that **counts the number of successes** in a finite sequence of **independent** observations
- Probability distributions with discrete outcomes can be described by a probability mass function or equivalently, through a (cumulative) distribution function
- The expectation or expected value of a distribution indicates the “average outcome” (weights of outcomes are the probabilities)
- The variance of a probability distribution is an indication of the spread of the outcomes (again weighted with the probabilities)

# Continuous probability distributions

## Normal distribution

# Continuous data

Continuous probability distributions are meant for continuous data.

*Reminder:*

- **numerical data** - data that has intrinsic numerical value
  - **continuous** data – data that can attain any value on a given measurement scale
    - interval data (continuous data for which only differences have meaning, there is no fixed “zero point”). Examples: temperature in Celsius, pH, clock time, IQ scores, birth year, longitude.
    - ratio data (continuous data for which both differences and ratios make sense; it has a fixed “zero point”). Examples: movie budget, temperature in Kelvin, distance, time duration.
  - **discrete** data – data that can only attain certain values (e.g., integers). Examples: the number of days with sunshine in a certain year, the number of traffic incidents.

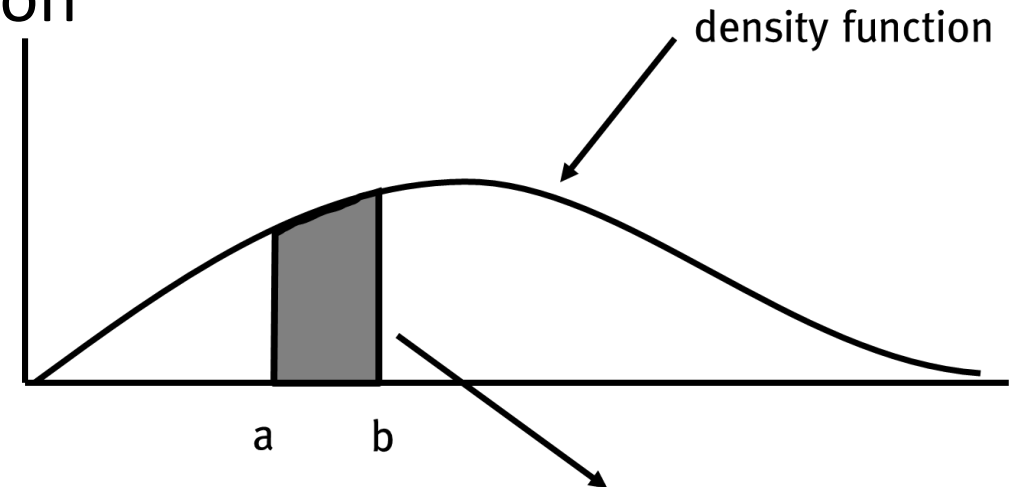
# Continuous distribution – density function

Discrete distributions: probability mass function

Continuous distributions: **density function**

The value of the density function  $f$  has no direct interpretation!

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



**The area under the density curve shows probability!**

area denotes probability that observation falls between a and b

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

$$\int_{-\infty}^{+\infty} f(x) dx = 1$$

$$\text{Expectation: } E(X) = \int x f(x) dx$$

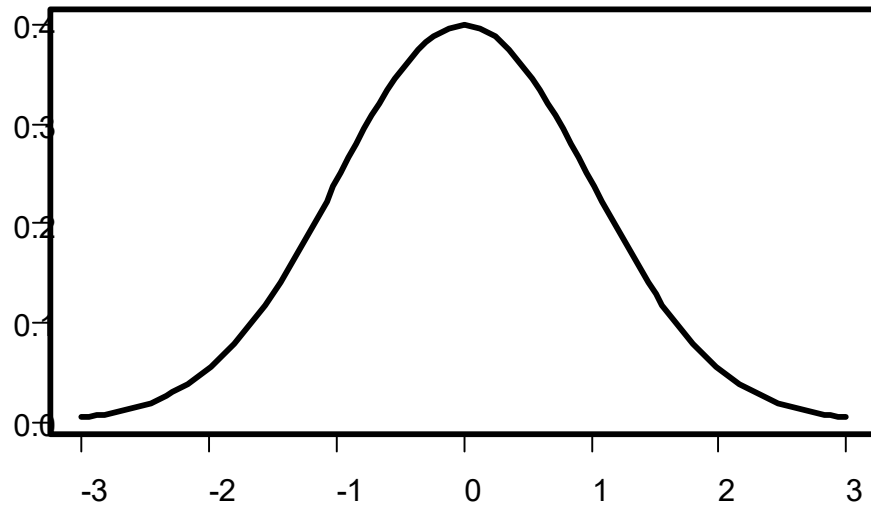
$$\text{Variance: } Var(X) = \int (x - E(X))^2 f(x) dx$$

$$\text{Standard deviation: } \sigma = \sqrt{Var(X)}$$

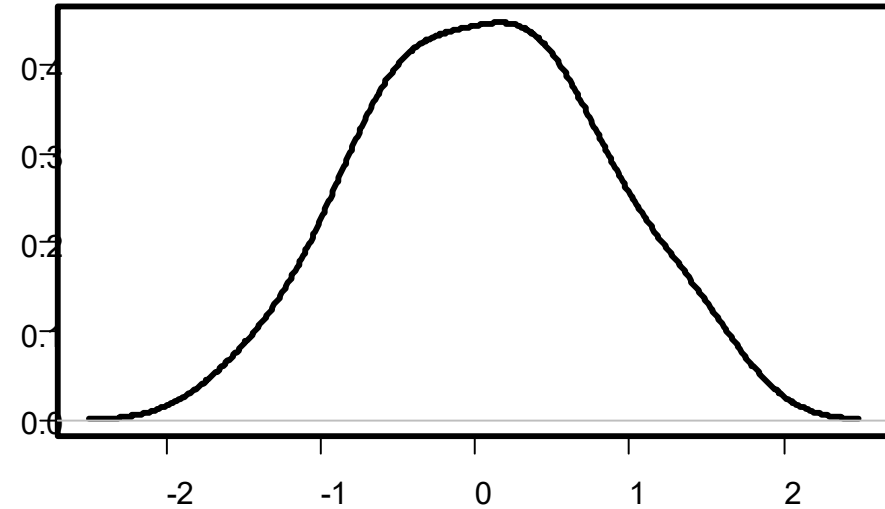


# Kernel density plots

In week 1 you learned about (kernel) density plots.



Density of probability distribution



Kernel density plot:  
Estimate of the density based on data

# Example: computing probabilities

Let  $X$  have the probability density function  $f$  plotted in the picture.

- a) What is its maximal value?
- b) Compute  $P(0.25 \leq X \leq 1)$ .

a)  $X$  takes values between -1 and 1.

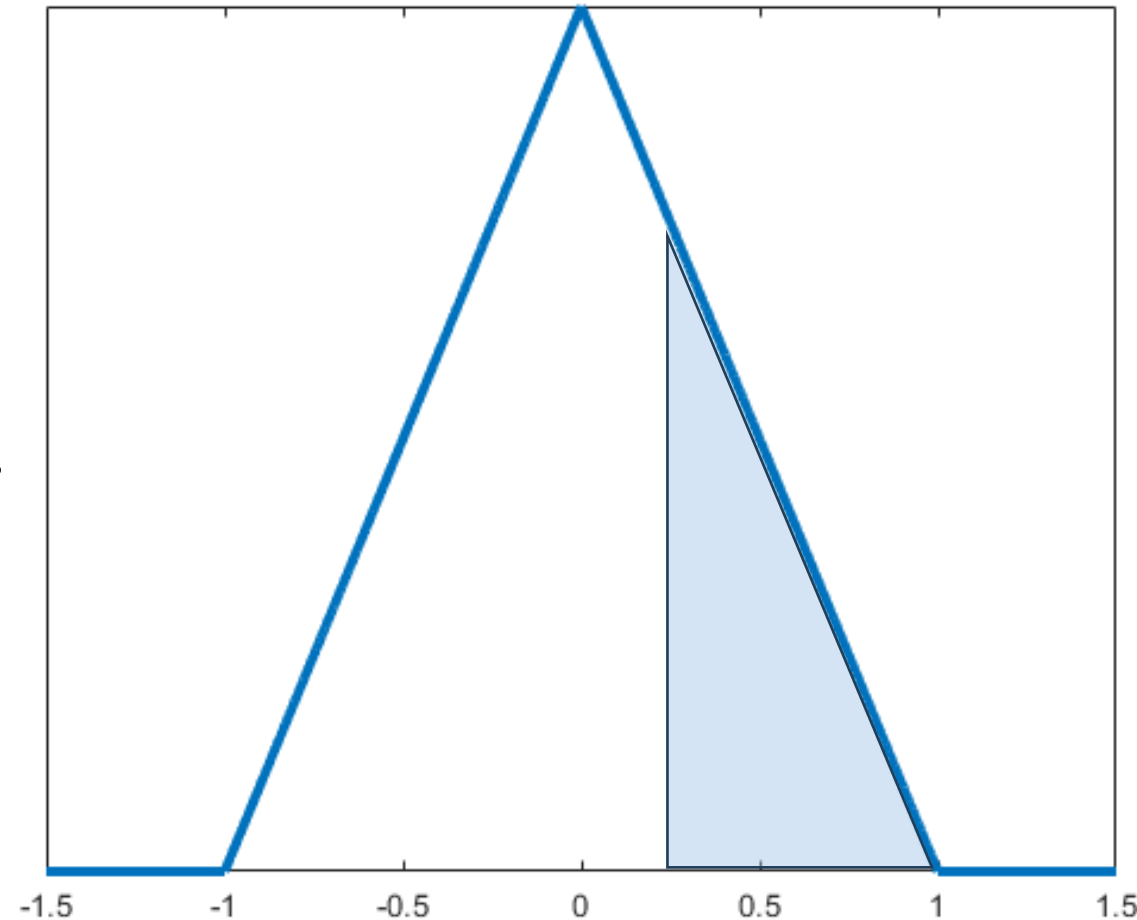
The area under the curve (the total probability) equals 1.

The area of the triangle is  $\frac{1}{2} \cdot 2 \cdot f(0)$ .

So  $f(0) = 1$  – the maximal value

b)  $f(0.25) = 0.75$  (line  $y = 1 - x$ )

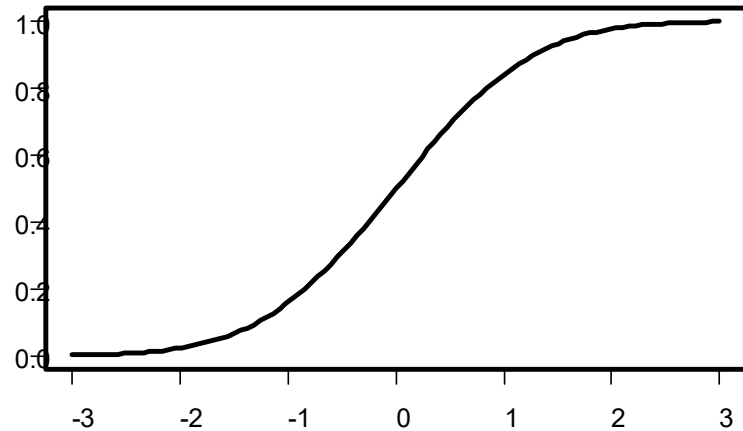
$$P(0.25 \leq X \leq 1) = \frac{1}{2} \cdot (1 - 0.25) \cdot f(0.25) = \frac{9}{32}$$



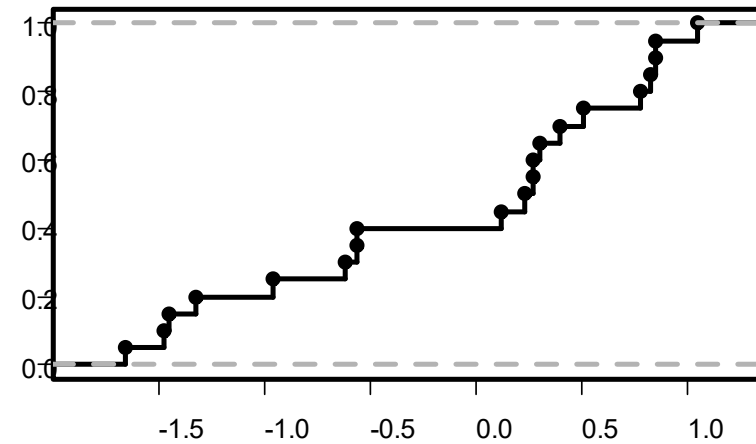
# Cumulative distribution function vs Empirical cumulative distribution function

The (cumulative) distribution function  $F$  is the function mapping  $x$  to  $P(X \leq x)$ :

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



Cumulative distribution function



ECDF: empirical cumulative distribution function - estimate of cumulative distribution function based on data

# Continuous versus discrete distributions

- summation  $\leftrightarrow$  integration
- taking differences  $\leftrightarrow$  differentiation

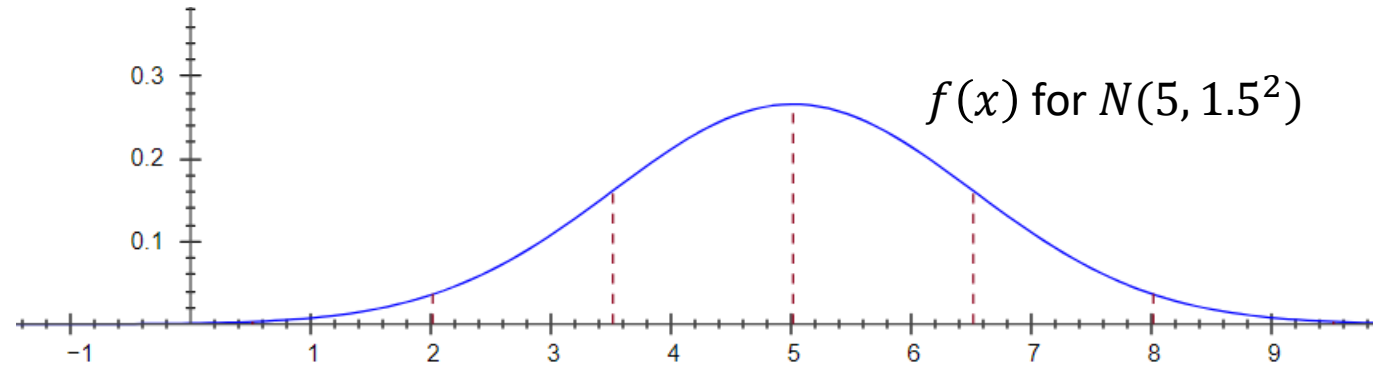
Discrete (integer-valued)	Continuous
$F(x) = \sum_{k \leq x} P(X = k)$	$F(x) = \int_{-\infty}^x f(u) du$
$P(X = k) = F(k) - F(k - 1)$	$f(x) = F'(x)$
$\sum_k P(k) = 1$	$\int_{-\infty}^{\infty} f(x) dx = 1$

# Normal distribution

- Notation:  $N(\mu, \sigma^2)$
- a **continuous** probability distribution over real numbers (probability 0 for individual values)
- **symmetric** around the **mean  $\mu$**
- The distribution is **fully specified by** the values of the **mean  $\mu$**  and **standard deviation  $\sigma$** .

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

Other names for normal distribution:  
**Gaussian distribution, bell curve distribution.**

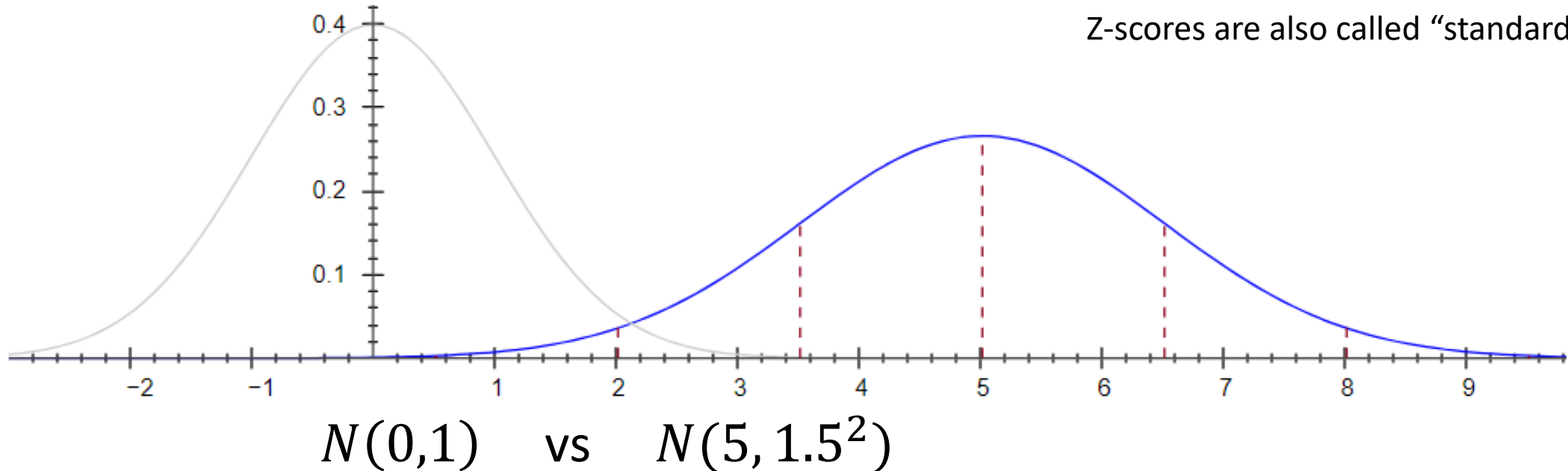


Software or lookup tables are needed to compute probabilities.

# Standard normal distribution

- Special case  $\mu = 0, \sigma^2 = 1$  (and hence  $\sigma = 1$ ) is called “standard normal distribution”
- If  $X \sim N(\mu, \sigma^2)$ , then  $\frac{X-\mu}{\sigma} \sim N(0, 1)$ , i.e. **z-scores of  $X$**  have the **standard** normal distribution
- **z-score**  $Z = \frac{X-\mu}{\sigma}$  shows how many standard deviations  $X$  is away from  $\mu$

Z-scores are also called “standard scores”

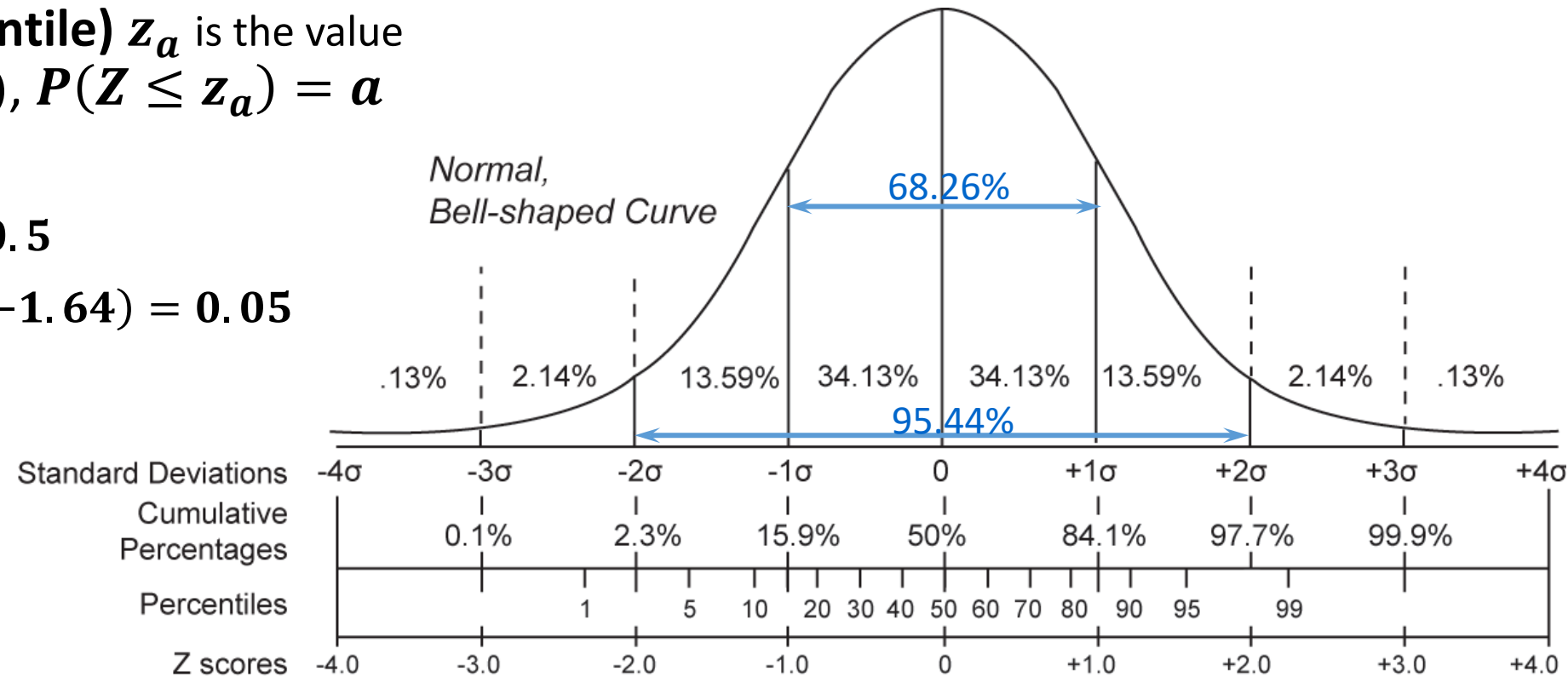


# Standard normal distribution: quantiles

- **normal quantile (percentile)  $z_a$**  is the value such that for  $Z \sim N(0, 1)$ ,  $P(Z \leq z_a) = a$

**Frequently used values:**

- $z_{0.5} \approx 0$ , i.e.  $P(Z \leq 0) = 0.5$
- $z_{0.05} \approx -1.64$ , i.e.  $P(Z \leq -1.64) = 0.05$
- $z_{0.023} \approx -2$
- $z_{0.0013} \approx -3$



**Important:**

A **z-score** is computed for a given value  $x$  of a random variable  $X$ , e.g. for  $X \sim (183, 9.7^2)$  and  $x = 173.3$ ,  $z\text{-score} = \frac{173.3 - 183}{9.7} = -1.0$ .

A **normal quantile** is looked up for a given probability  $p$ , e.g. for  $p = 0.05$ , the normal quantile  $z_{0.05} = -1.64$

# Normal distribution: usage examples

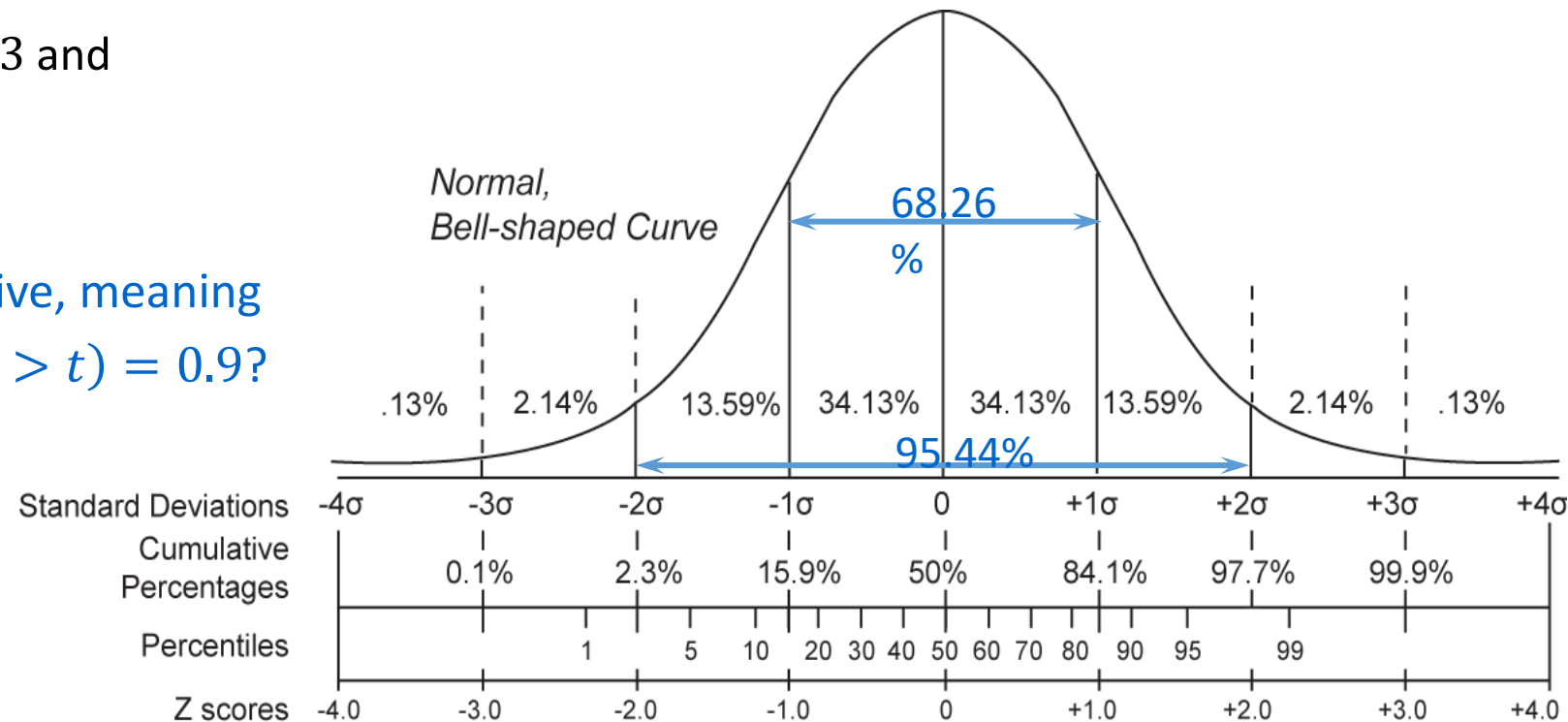
Let  $X \sim N(15, 25)$ , with  $X$  being time left till the start of an exam when a student enters the exam room.

- What is the probability that a student comes to the exam too late?

- “Too late” means  $X < 0$ !
- Compute the z-score of 0:  $\frac{0-15}{5} = -3$  and
- use Python or tables to find:  
 $P(X \leq 0) \approx 0.0013$

- By which time are students likely to arrive, meaning for which value  $t$  is the probability  $P(X > t) = 0.9$ ?

- $P(X > t) = 1 - P(X \leq t)$ ,
- so  $P(X \leq t) = 0.1$
- $z_{0.1} \approx -1.28$ , so  $\frac{t-15}{5} = -1.28$
- $t \approx 8.6$  minutes before the exam





# (Almost) normally distributed variables

---

- Birth weight
- Measurement errors
- User response time
- Network latency
- ...

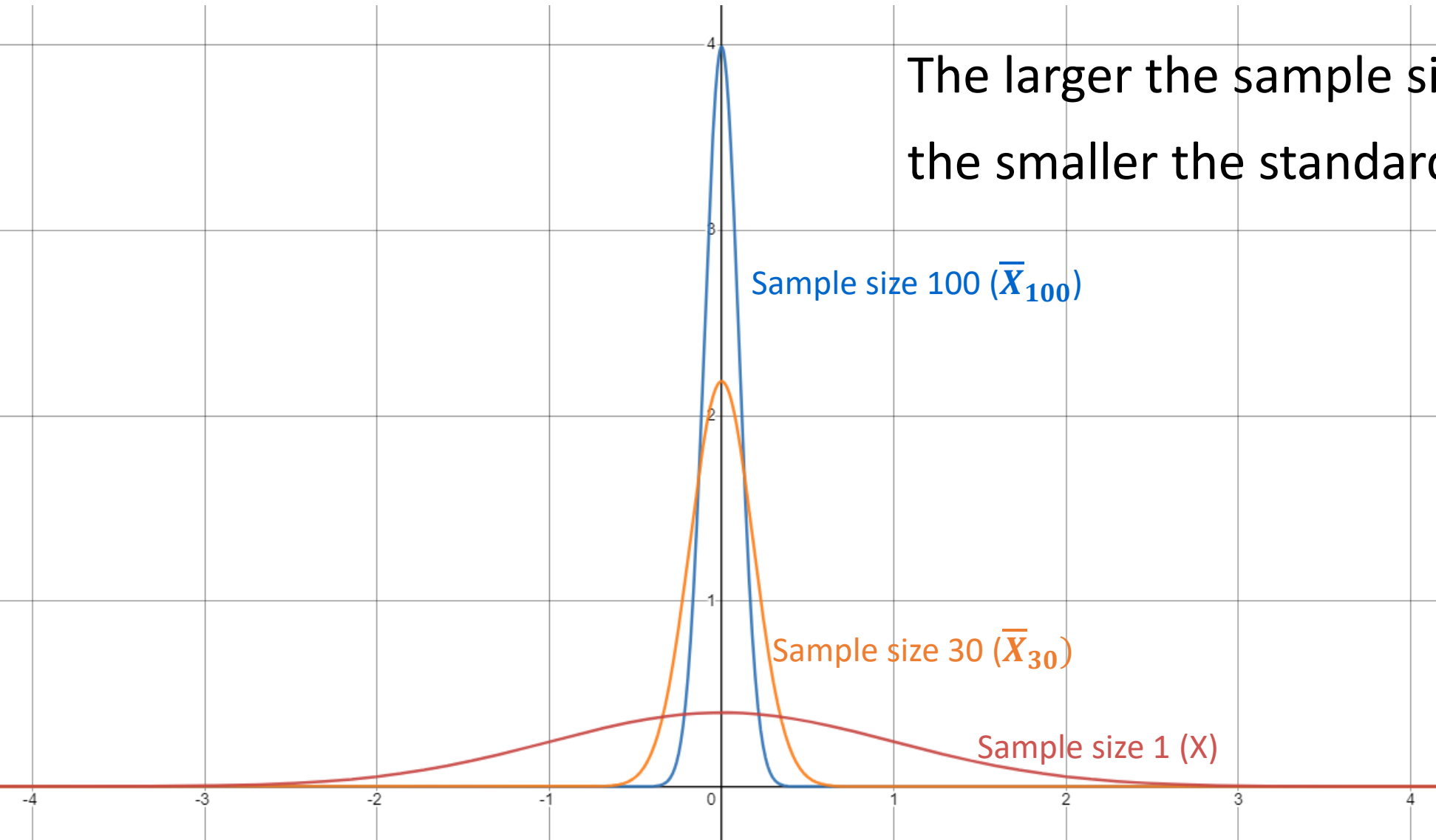
# Summary: normal distribution as an example of continuous distributions

- The normal distribution is fully specified by its mean and variance
- The mean and variance of the normal distribution can be estimated by the sample mean and sample variance, respectively.
- Continuous distributions are described by the density function and probabilities by areas under the density function.
- Kernel density plots estimate the density of a probability distribution
- ECDF estimates the cumulative distribution function

# Estimations

$$N(0, 1), N\left(0, \frac{1}{30}\right), N\left(0, \frac{1}{100}\right)$$

The larger the sample size ,  
the smaller the standard deviation of  $\bar{X}$



# Central limit theorem

- Also called the **law of large numbers**

Let  $\bar{X} = \frac{X_1 + X_2 + \dots + X_n}{n}$ , where  $X_i$  are independent random variables with expectation  $E(X)$  and standard deviation  $\sigma_X$  (so  $Var(X) = \sigma_X^2$ ), with a **large  $n$** .

Then the distribution of  $\bar{X}$  can be approximated by  $N(E(X), \frac{\sigma_X^2}{n})$ .

**Note:** although  $X$  is not normally distributed, the sample mean  $\bar{X}$  is! (  $\lim_{n \rightarrow \infty}$  )

# Toss a coin many times!

and count the number of H.

Notation:  $\hat{p} = \frac{X}{n}$  – estimate of  $p$

# Estimation – general setup

**Estimate of the success probability  $p$  of the binomial distribution:**

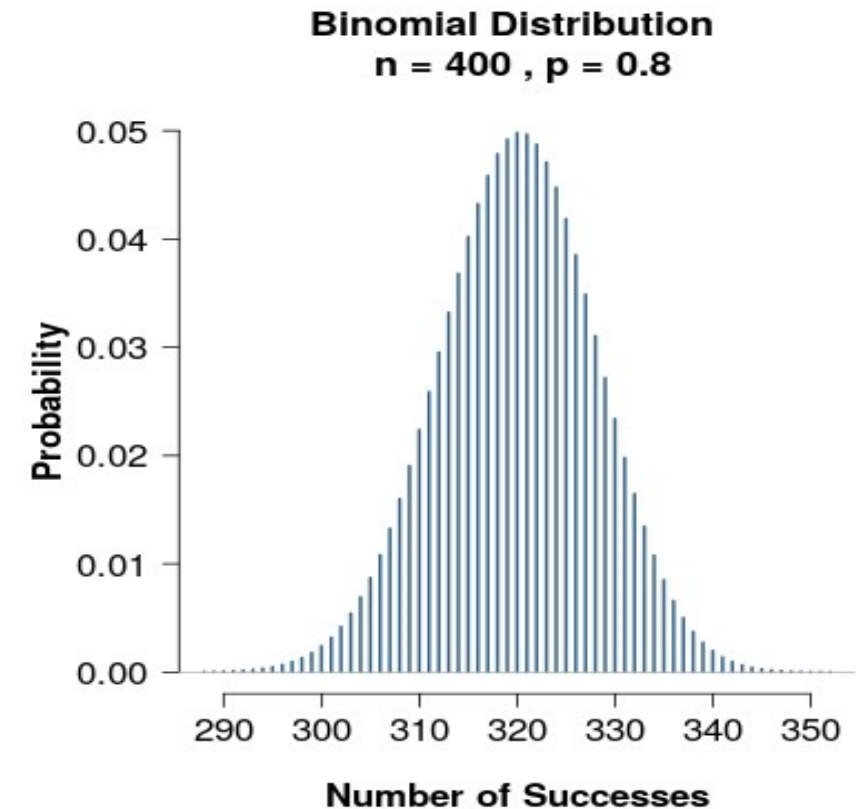
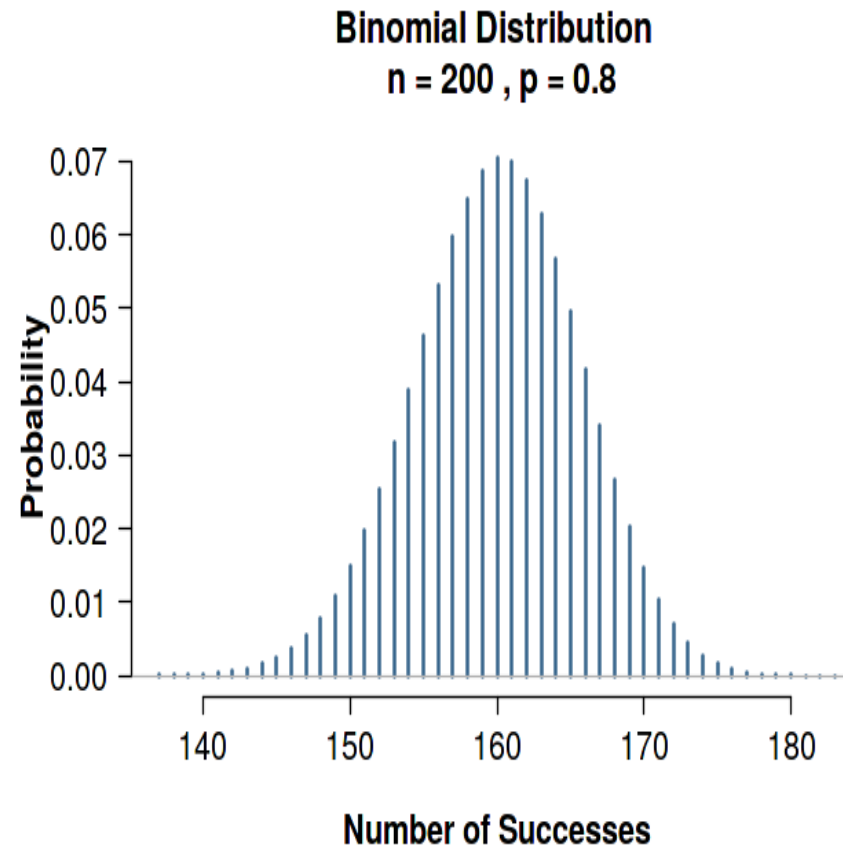
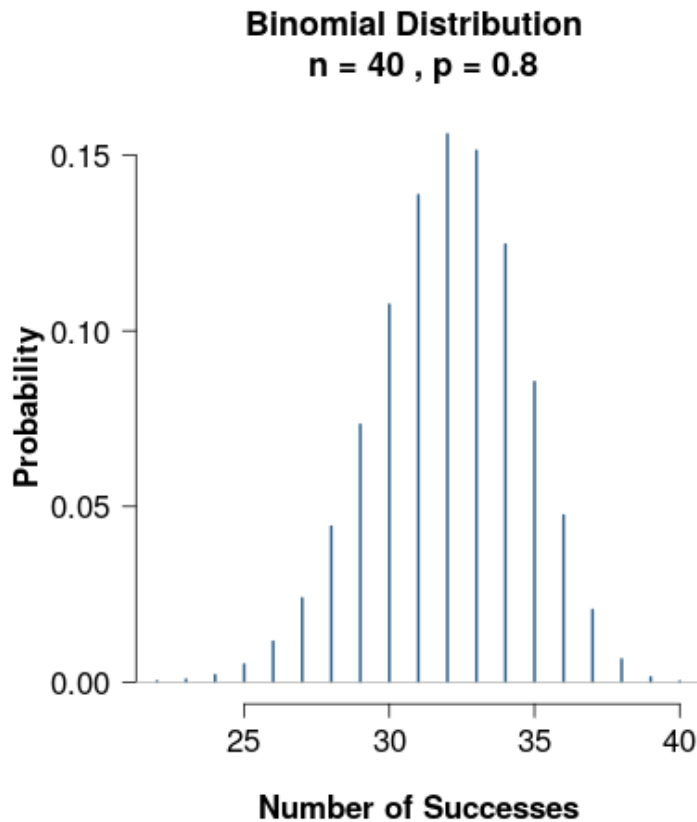
$$\hat{p} = \frac{X}{n}$$

$$E(X) = np, \text{ so } E(\hat{p}) = p$$

$$Var(X) = np(1 - p), \text{ so } Var(\hat{p}) = \frac{p(1-p)}{n} \quad (\text{reminder: } Var(X) = E(X - E(X))^2)$$

# Binomial distribution – many observations

What happens if we get more observations?





# Also here: Central Limit Theorem

- $\hat{p}$  is a random variable
- $\hat{p}$  has the normal distribution if your sample is large!
- $\hat{p} \sim N(p, \frac{p(1-p)}{n})$
- The standard deviation of  $\hat{p}$  gets smaller for larger samples!
  - $Var(\hat{p}) = \frac{p(1-p)}{n}, \quad \sigma_{\hat{p}} = \sqrt{\frac{p(1-p)}{n}}$

# Summary: estimations

---

- Use the normal distribution to estimate the mean based on large samples
- Compute point estimates (single numbers) for proportions by computing the share of successful cases in the total number
- For large sample sizes, the binomial distribution is well approximated by the normal distribution

# Confidence Intervals

# Confidence interval

**Confidence interval** is an **interval containing the true unknown value that we wish to estimate** (e.g., a mean or a proportion) at a given **confidence level** (e.g. 95%)

## **Probability interpretation:**

When the estimate computation procedure is repeated many times for different samples of the same size coming from the same population, the confidence interval at a confidence level 95% computed based on a sample will contain the true value of the parameter in 95% of the cases, and it won't in 5% of the cases.

In this case:

**significance level**  $\alpha = 0.05$

**confidence level**  $1 - \alpha = 0.95$

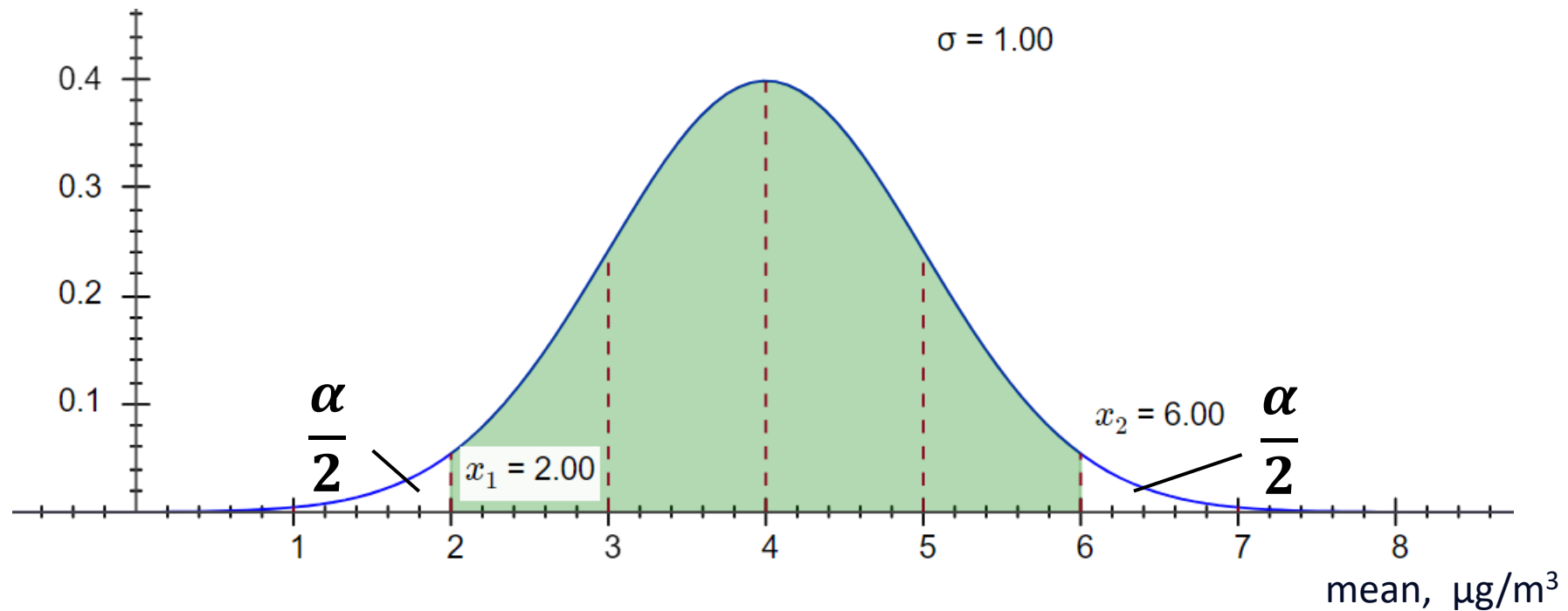
Most popular choices of  $\alpha$ : 0.01, 0.05, 0.1.

The context of the problem defines the choice!

You might need  $\alpha = 0.0001$ !

# Two-sided confidence intervals

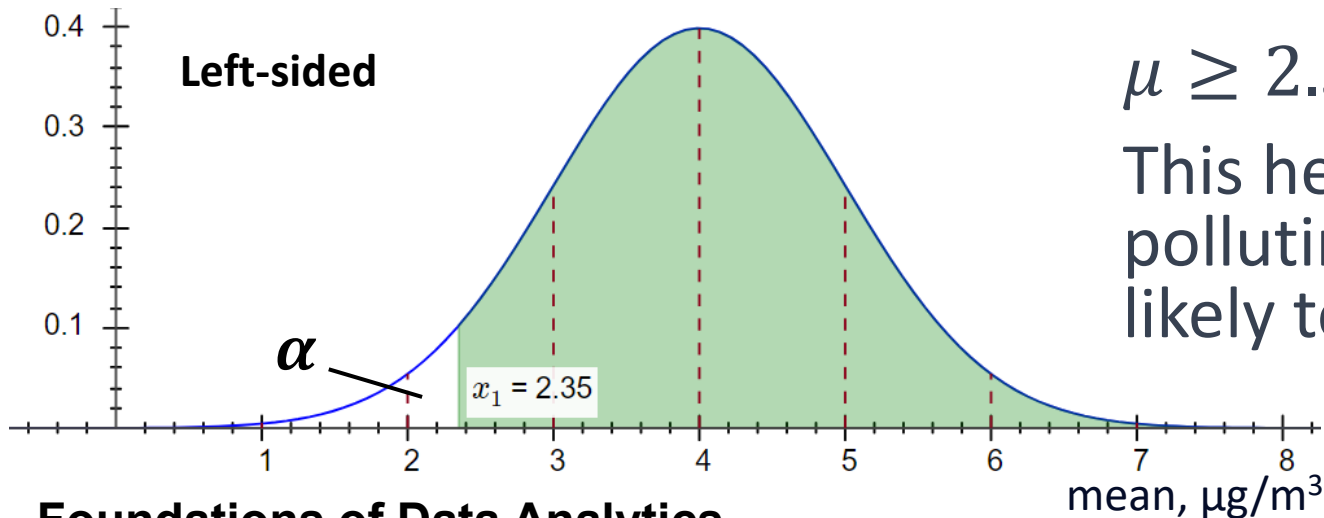
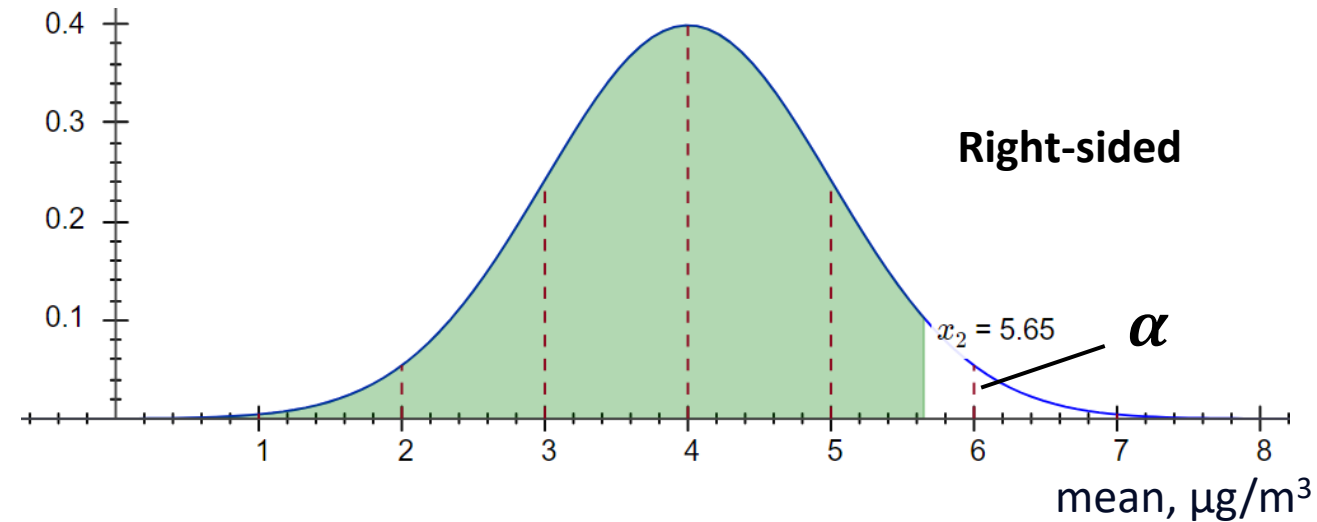
- What is the confidence interval for the concentration of  $PM_{2.5}$  at a confidence level of 95%? (based on a sample)



# One-sided confidence intervals

$\mu \leq 5.68$  at a confidence level of 95%.

This helps when assessing health risks: the concentration is not likely to exceed that threshold.



$\mu \geq 2.35$  at a confidence level of 95%.

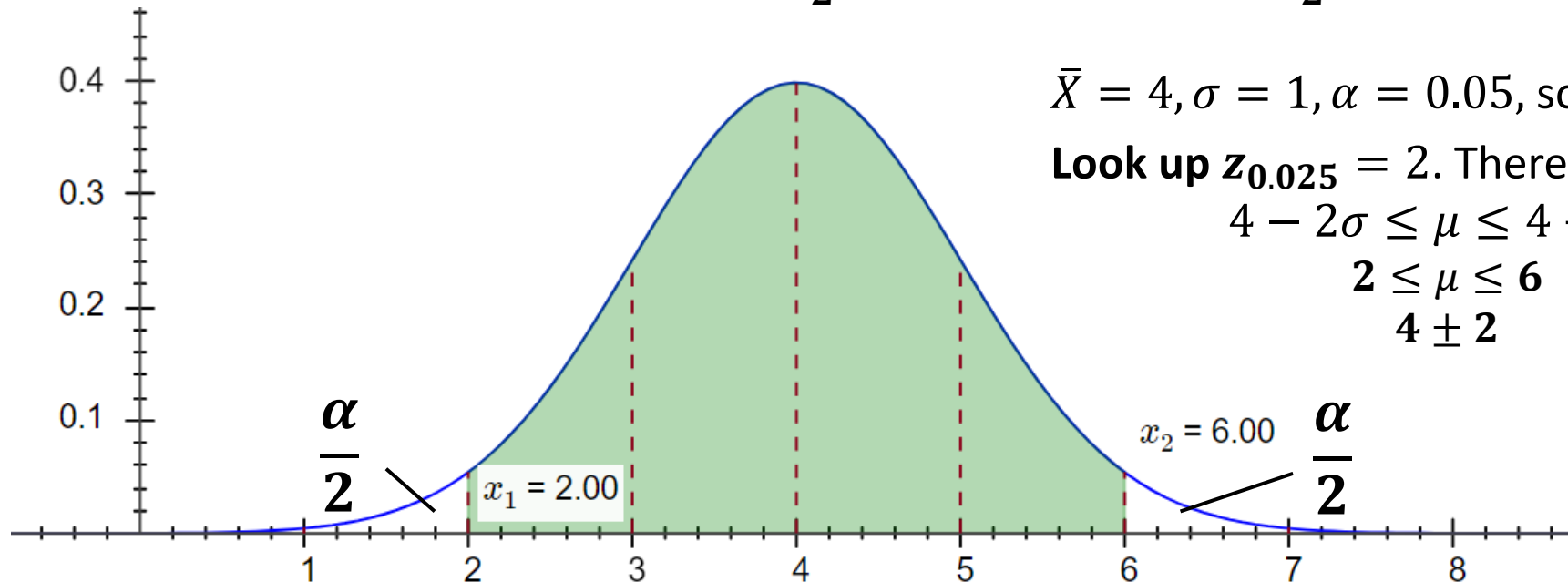
This helps to make a clear case against a polluting company: the pollution level is not likely to be below 2.35

# Use the normal distribution!

**Problem:** find a confidence interval for  $\mu_X$  for a given  $\alpha$ .

You have a sample of size  $n$  and know the standard deviation  $\sigma_X$ .

- Compute the confidence interval:  $\bar{X} - z_{\frac{\alpha}{2}} \frac{\sigma}{\sqrt{n}} \leq \mu \leq \bar{X} + z_{\frac{\alpha}{2}} \frac{\sigma}{\sqrt{n}}$



$\bar{X} = 4, \sigma = 1, \alpha = 0.05$ , so  $\frac{\alpha}{2} = 0.025$

Look up  $z_{0.025} = 2$ . Therefore:

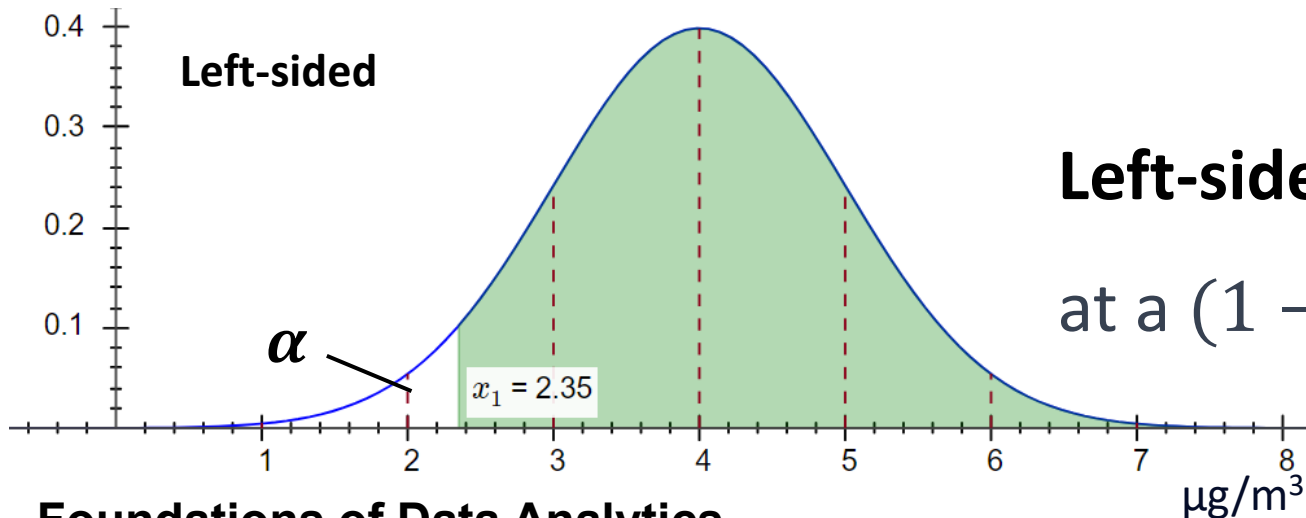
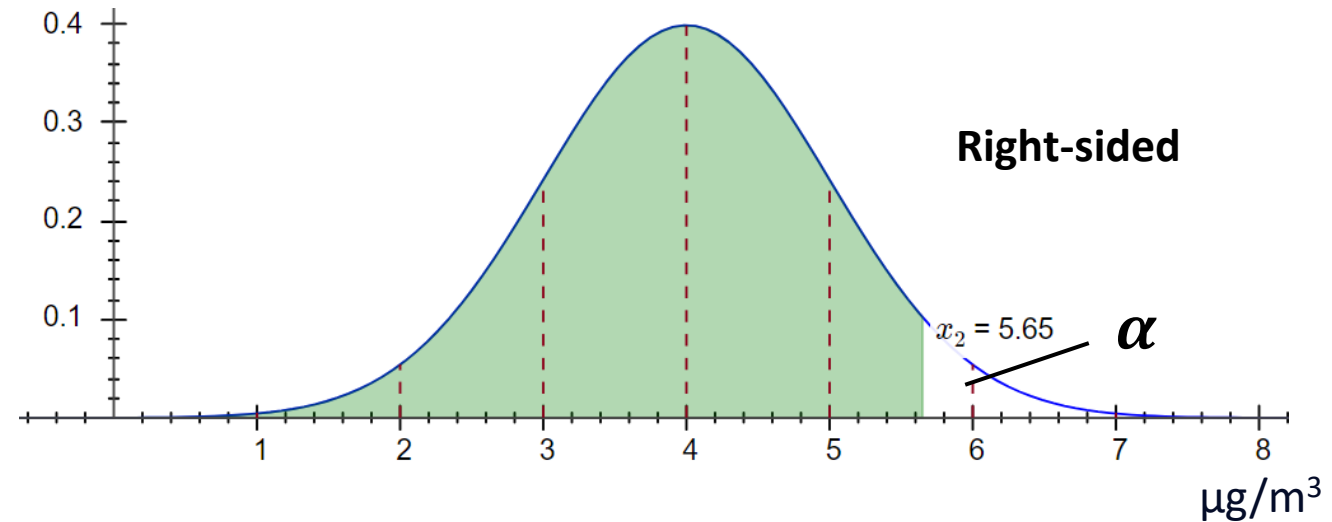
$$4 - 2\sigma \leq \mu \leq 4 + 2\sigma$$

$$2 \leq \mu \leq 6$$

$$4 \pm 2$$

# One-sided confidence intervals

**Right-sided:**  $\mu \leq \bar{X} + z_{\alpha} \frac{\sigma}{\sqrt{n}}$   
at a  $(1 - \alpha)$  confidence level



**Left-sided:**  $\bar{X} - z_{\alpha} \frac{\sigma}{\sqrt{n}} \leq \mu$   
at a  $(1 - \alpha)$  confidence level



# Derivation of confidence interval (optional)

We use that  $\bar{X} \sim N\left(\mu, \frac{\sigma^2}{n}\right)$  and thus  $\frac{\bar{X}-\mu}{\sigma/\sqrt{n}} \sim N(0,1)$ . Hence,

$$\begin{aligned} 1 - \alpha &= P\left(-z_{\alpha/2} \leq \frac{\bar{X}-\mu}{\frac{\sigma}{\sqrt{n}}} \leq z_{\alpha/2}\right) = P\left(-z_{\alpha/2} \frac{\sigma}{\sqrt{n}} \leq \bar{X} - \mu \leq z_{\alpha/2} \frac{\sigma}{\sqrt{n}}\right) \\ &= P\left(\mu - z_{\alpha/2} \frac{\sigma}{\sqrt{n}} \leq \bar{X} \leq \mu + z_{\alpha/2} \frac{\sigma}{\sqrt{n}}\right) \end{aligned}$$

Since  $\mu - z_{\alpha/2} \frac{\sigma}{\sqrt{n}} \leq \bar{X} \leq \mu + z_{\alpha/2} \frac{\sigma}{\sqrt{n}} \Leftrightarrow \bar{X} - z_{\alpha/2} \frac{\sigma}{\sqrt{n}} \leq \mu \leq \bar{X} + z_{\alpha/2} \frac{\sigma}{\sqrt{n}}$ ,

we have a formula for the confidence interval.

# Confidence intervals – some remarks

## Confidence interval example

$\bar{x} = 1.4, \sigma = 0.3, n = 25, \alpha = 0.05$ :  $1.4 \pm 1.96 \frac{0.3}{\sqrt{25}} = 1.4 \pm 0.117$  (2-sided 95% confidence interval)

*Changing one of the parameters:*

**Larger standard deviation → wider confidence interval**, less certainty about estimate:

$\bar{x} = 1.4, \sigma = 0.4, n = 25, \alpha = 0.05$ :  $1.4 \pm 1.96 \frac{0.4}{\sqrt{25}} = 1.4 \pm 0.16$

$$\bar{x} \pm z_{\alpha/2} \frac{\sigma}{\sqrt{n}}$$

**Larger sample → tighter interval**, more certainty about estimate:

$\bar{x} = 1.4, \sigma = 0.3, n = 100, \alpha = 0.05$ :  $1.4 \pm 1.96 \frac{0.3}{\sqrt{100}} = 1.4 \pm 0.06$

**Lower confidence → tighter interval** because less guaranteed coverage

$\bar{x} = 1.4, \sigma = 0.3, n = 25, \alpha = 0.10$ :  $1.4 \pm 1.28 \frac{0.3}{\sqrt{25}} = 1.4 \pm 0.08$

# Confidence interval for the mean: unknown $\sigma$

If  $\sigma$  is not known, use the **sample standard deviation  $s$**  and a similar formula based on the **Student  $t$ -distribution with  $n - 1$  degrees of freedom**, where  $n$  is the sample size:

$$\bar{x} \pm t_{n-1;\alpha/2} \frac{s}{\sqrt{n}}$$

The **quantile  $t_{n-1;\alpha/2}$**  of the  $t$ -distribution depends on the confidence level **and on the sample size!**

Use software or tables to find  $t_{n-1;\alpha/2}$ .

**Example:**  $\bar{x} = 1.4$ ,  $s = 0.3$ ,  $n = 25$ ,  $\alpha = 0.05$ ;

$t_{24;0.025} = 2.06$  ; the boundaries of the confidence interval are  $1.4 \pm 2.06 \cdot \frac{0.3}{\sqrt{25}} = 1.4 \pm 0.12$

Note: the interval for unknown  $\sigma$  is slightly wider than for a known  $\sigma$ .

Also: the higher  $n$ , the smaller  $t_{n-1;\alpha/2}$  (so narrower interval)

# Confidence interval – proportion

We use the Central Limit Theorem (i.e., the fact that the binomial distribution can be approximated by a normal distribution):  $\hat{p} \sim N(p, \frac{p(1-p)}{n})$

$$\text{Confidence interval: } \hat{p} - z_{\alpha/2} \sqrt{\frac{\hat{p}(1-\hat{p})}{n}} \leq p \leq \hat{p} + z_{\alpha/2} \sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$$

*Note:* the width of the confidence interval depends on the confidence level, the sample size  $n$ , *and also* on the estimated success probability.

# Confidence intervals – comparing distributions

Confidence intervals also exist for:

- **the difference of two means** (normal distributions, several cases depending on assumptions about the standard deviations)

$$\overline{x}_1 - \overline{x}_2 \pm z_{\alpha/2} \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}$$

Example:

comparing mean concentrations of *NO* in Amsterdam and in Eindhoven

- **the difference of two proportions** (binomial distributions).

$$\widehat{p}_1 - \widehat{p}_2 \pm z_{\frac{\alpha}{2}} \sqrt{\frac{\widehat{p}_1(1-\widehat{p}_1)}{n_1} + \frac{\widehat{p}_2(1-\widehat{p}_2)}{n_2}}$$

Example:

Comparing percentages of days on which *NO* concentration is above the threshold in Amsterdam and in Eindhoven

# How to compute confidence intervals

- select right type of interval
  - one-sided or two-sided
  - means or proportions
  - one or two samples
  - for means: can variances assumed to be known or equal (comparison)
- select confidence level (standard choice is 95%)
- compute confidence interval
  - use formulas
  - use Python functions from the statsmodels library

# Summary confidence intervals

- To get an impression about how accurate we know the proportion, we need an interval
- Probability interpretation: when a 95% confidence interval for a certain parameter is applied to many data sets, it will contain in 95% of the cases the true value of the parameter
- To obtain an interval, we need to perform calculations with the probability distribution of the point estimates.
- Use explicit formula and use `scipy` in Python for quantile `norm.ppf` or `t.ppf(0.05,24)`
- Use ready made confidence intervals from the `statsmodels` library in Python
- There are confidence intervals for single parameters but also for differences of parameters.
- No need to learn intervals by heart, just know the different types and read the manual
- Confidence intervals indicate how certain we are about a value that we estimated from data: wider intervals means more uncertainty

# Hypothesis Testing



# Hypotheses and Court



Standard legal procedure:

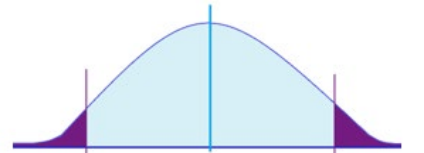
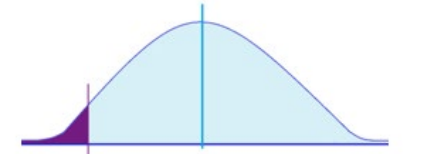
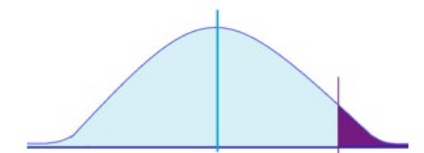
1. suspect assumed to be not guilty by court (**null hypothesis**), prosecution believes the suspect is guilty (**alternative hypothesis**)
2. prosecution brings **evidence** for guilt (data)
3. in case of **insufficient evidence** → acquittal (null hypothesis **not** rejected)
4. in case of **sufficient evidence** → conviction (null hypothesis **rejected**)

Note the asymmetry in the procedure!

# Null and alternative hypotheses: $H_0$ and $H_a$

Manufacturer: weight  $X \sim N(1, 0.002)$  – the weight of a sugar pack is  $\approx 1$  kg

**Null hypothesis:**  $H_0: \mu = 1$

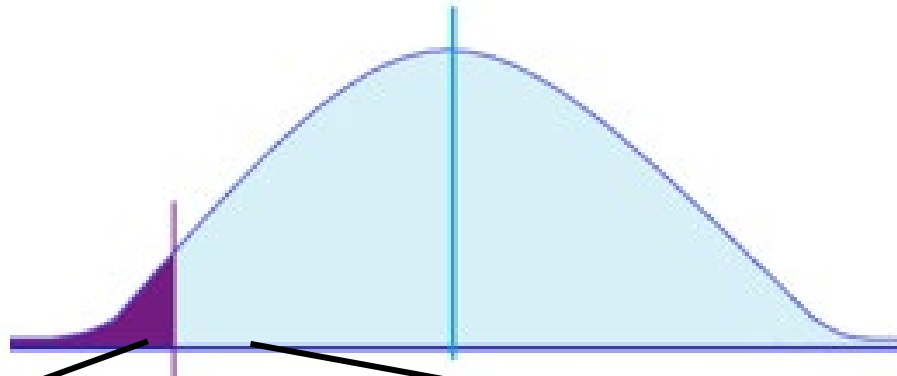
Suspicion	Alternative hypothesis	To accept $H_a$ , $\bar{X}$ should be extreme enough	
Factory: a machine malfunctions	$H_a: \mu \neq 1$	Two-sided	
Consumer: too little sugar	$H_a: \mu < 1$	Left-tailed one-sided	
Transport company: too much sugar	$H_a: \mu > 1$	Right-tailed one-sided	



# Choice of null and alternative hypothesis

Common practice: **choose what you think/hope/afraid to be true as  $H_a$** .  
You **cannot** accept the null hypothesis! You can only reject or not reject it!

Compute a *test statistic*

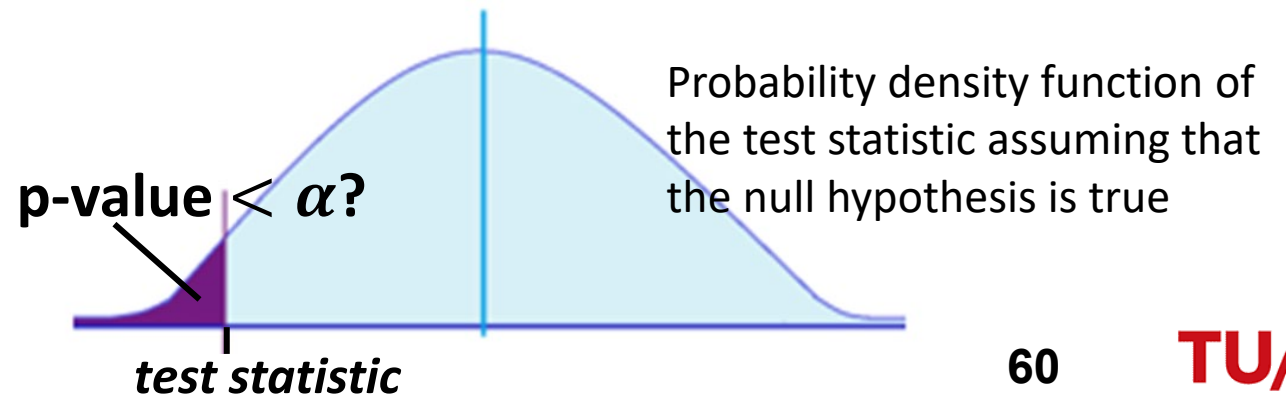


If the *test statistic* is sufficiently extreme,  
 **$H_0$  is rejected** in favour of  $H_a$

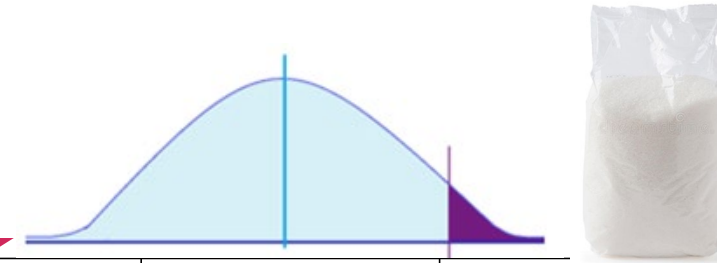
If the *test statistic* is **NOT** sufficiently extreme,  
 **$H_0$  is NOT rejected**

# Decision procedure based on p-values

- Choose the **significance level  $\alpha$**  (often 0.05 or 0.01, but other values can be used as well!)
- Compute a **test statistic** for your sample, e.g.:
  - **Z-score**  $\frac{\bar{X}-\mu}{s/\sqrt{n}}$  (when the population standard deviation is known) – **standard normal distribution**
  - **T-Score**  $\frac{\bar{X}-\mu}{s/\sqrt{n}}$  (when the population standard deviation is **not** known) – **Student  $t$ -distribution**
  - The **number  $k$  of successes** in a set of  $n$  independent observations
  - **Proportion estimate  $\hat{p}$**
- Compute the **p-value** – the **probability of observing a more extreme test statistic in the direction of the alternative hypothesis than the one you computed**, assuming the null hypothesis were true.
  - i.e. there were 3 successes and  $P(k \leq 3) = 0.04$
- **If p-value  $< \alpha$ , reject  $H_0$**   
**If not, then we do not reject  $H_0$ .**



# Example: using p-values



**Manufacturer: “More than 90% of our sugar bags contain  $\geq 1$  kg!”**

- i.e. they assume the binomial distribution with  $p = 0.9$ 
  - success: bag weight  $\geq 1$  kg!, failure: bag weight  $< 1$  kg
- $H_0: p = 0.9, H_a: p > 0.9$
- Chosen **significance level**:  $\alpha = 0.05$
- They went for  $n = 20$  (the number of tested bags)
- **Test statistics: the number  $k$  of successes**
  - $k = 19$  in their sample
- **p-value**  $P(k \geq 19) = 1 - P(k \leq 18) = 1 - 0.6083 = 0.3917$
- $0.3917 > 0.05$ , hence  $H_0$  cannot be rejected!
- Even if they got  $k = 20$ ,  $P(k \geq 20) = P(k = 20) = 0.1216$ ,  $0.1216 \geq 0.05$ , and that is not enough to reject  $H_0$ !

Number $k$ of successes	$P(X = k)$	$P(X \leq k)$
0	0.0000	0.0000
1	0.0000	0.0000
2	0.0000	0.0000
3	0.0000	0.0000
4	0.0000	0.0000
5	0.0000	0.0000
6	0.0000	0.0000
7	0.0000	0.0000
8	0.0000	0.0000
9	0.0000	0.0000
10	0.0000	0.0000
11	0.0001	0.0001
12	0.0004	0.0004
13	0.0020	0.0024
14	0.0089	0.0113
15	0.0319	0.0432
16	0.0898	0.1330
17	0.1901	0.3231
18	0.2852	0.6083
19	0.2702	0.8784
20	0.1216	1.0000

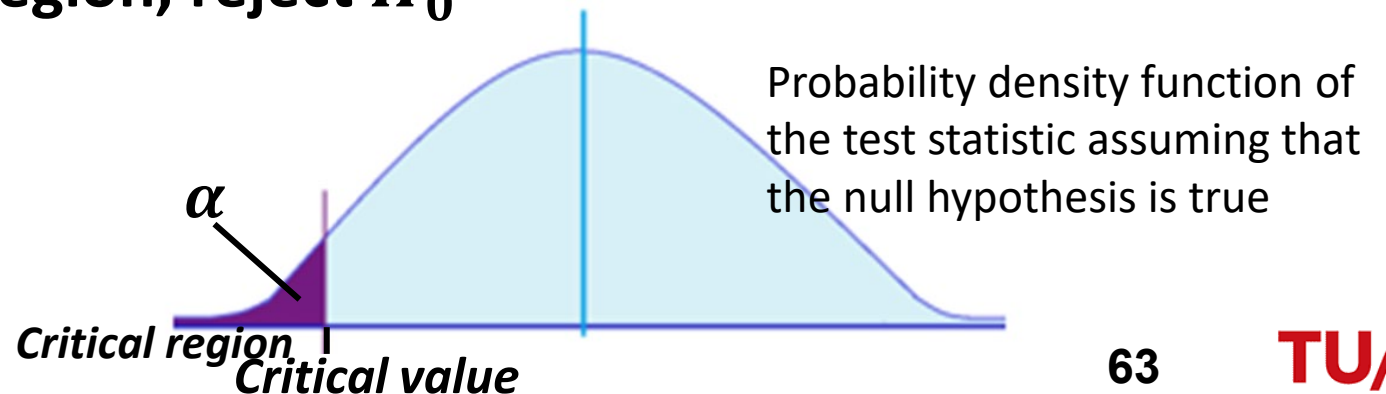
# Test size to conclude effect

**Manufacturer: “More than 90% of our sugar bags contain  $\geq 1$  kg!”**

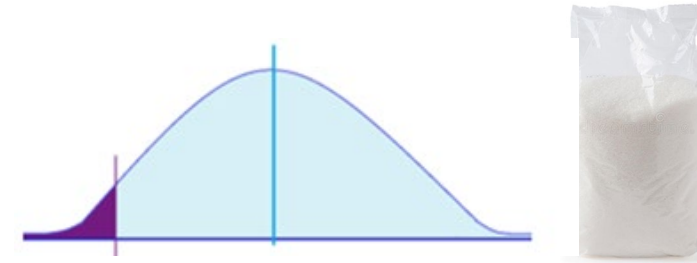
- i.e. they assume the binomial distribution with  $p = 0.9$ 
  - success: bag weight  $\geq 1$  kg!, failure: bag weight  $< 1$  kg
- $H_0: p = 0.9, H_a: p > 0.9$
- Chosen **significance level**:  $\alpha = 0.05$
- How many bags all with weight  $\geq 1$  kg would we need to have in the sample to reject  $H_0$ ?
- $n$  bags,  $0.9^n < 0.05$ 
  - $n > \log_{0.9} 0.05$
  - $n > 28.4$
- i.e. we can reject  $H_0$  if we have a sample of size = 29 (or more bags) all weighting more than 1kg

# Decision procedure based on critical values

- Choose the **significance level  $\alpha$**
- **Choose** an appropriate **test statistic**
- **Compute** the **critical values** based on  $\alpha$  and the hypotheses
  - Critical values are the boundaries for the tail(s) of the distribution beyond which the null hypothesis will be rejected
- **Compute** the **test statistic** from the sample and **compare it to the critical values**
- **If the test statistic is in the critical region, reject  $H_0$**   
**If not, then we do not reject  $H_0$ .**



# Example: using critical values



Manufacturer: “The weight of a bag  $\sim N(1010, 25)$ ”

Consumer organisation: “Not true! There is less sugar than the manufacturer claims!”

- $H_0: \mu = 1010, H_a: \mu < 1010$ , known  $\sigma = 5$
- Chosen **significance level:  $\alpha = 0.01$**  and sample size  $n = 100$ 
  - *weight* can be approximated by  $N(\mu, \frac{\sigma^2}{n})$ , i.e. by  $N(1010, 0.25)$
- **Compute the critical value:**
  - $z_{0.01} \approx -2.23$
  - Critical value =  $1010 - 2.23 \cdot \sqrt{0.25} = 1008.885$ 
    - If you want to round it, go in the direction of the extreme, e.g. 1008.8 !

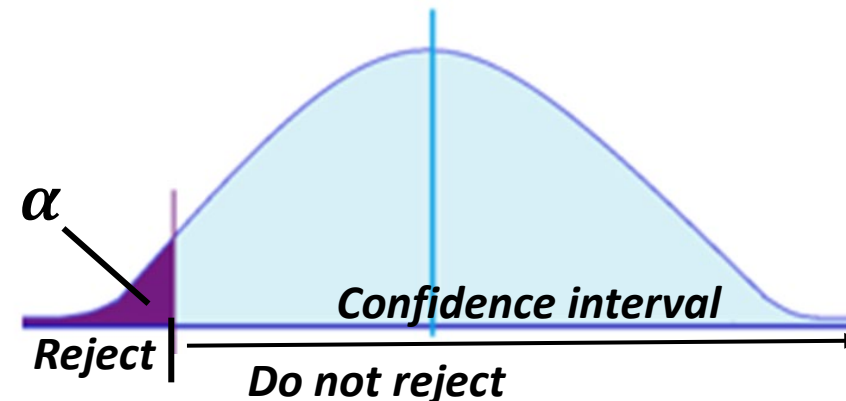
If the sample mean is below the critical value, e.g.  $1008.75 < 1008.885$ ,  $H_0$  can be rejected, otherwise we cannot reject  $H_0$



# Decision procedure based on confidence intervals

Consider e.g.  $H_0: \mu = 1000$  ;  $H_a: \mu < 1000$

- Choose the **significance level  $\alpha$**
- **Compute the confidence interval at a confidence level  $1 - \alpha$  from your sample**
- **Compute the test statistic from the sample and compare it to the critical values**
- **If  $\mu = 5$  does not belong to the confidence interval, reject  $H_0$**   
If 5 is in the confidence interval, then we **do not reject  $H_0$** .

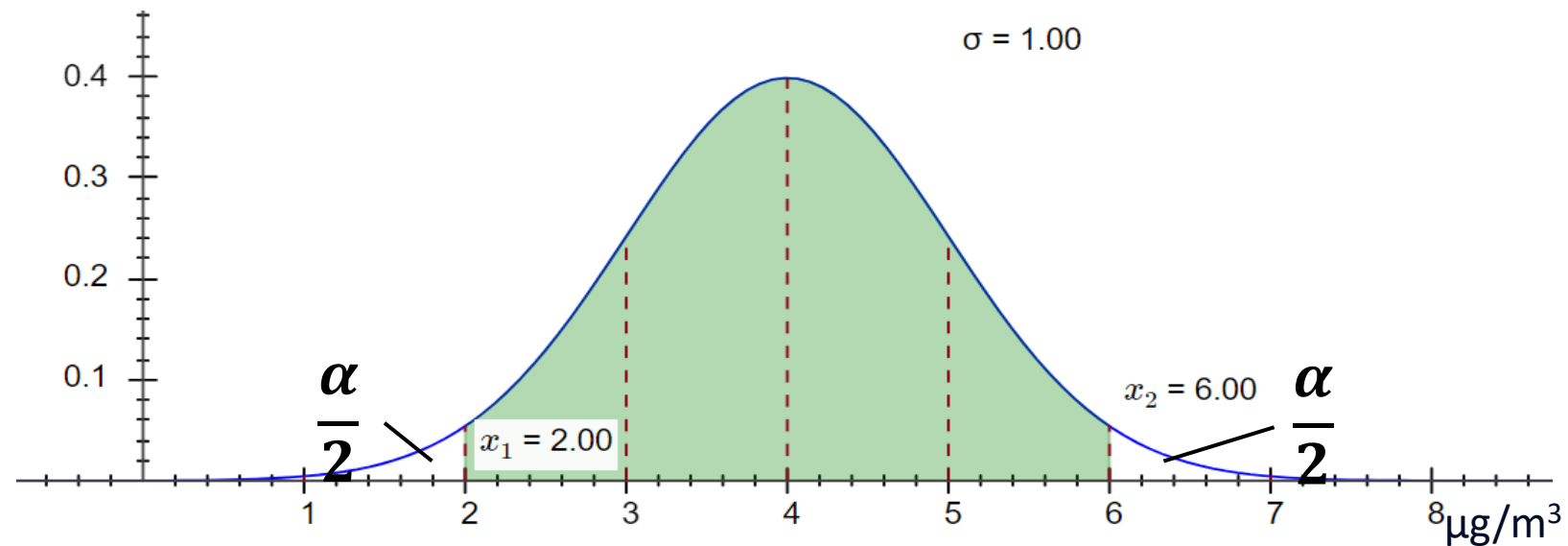


# Confidence intervals and hypothesis testing

Example:  $\alpha = 0.05$ . We computed 95%-confidence interval for  $\mu$  from our sample: (2;6)

**Case 1:**  $H_0: \mu = 5$  ;  $H_a: \mu \neq 5$ .  $5 \in (2; 6) \Rightarrow$  do **not** reject  $H_0$  with  $\alpha = 0.05$

**Case 2:**  $H_0: \mu = 7$  ;  $H_a: \mu \neq 7$ .  $7 \notin (2; 6) \Rightarrow$  **reject**  $H_0$  with  $\alpha = 0.05$

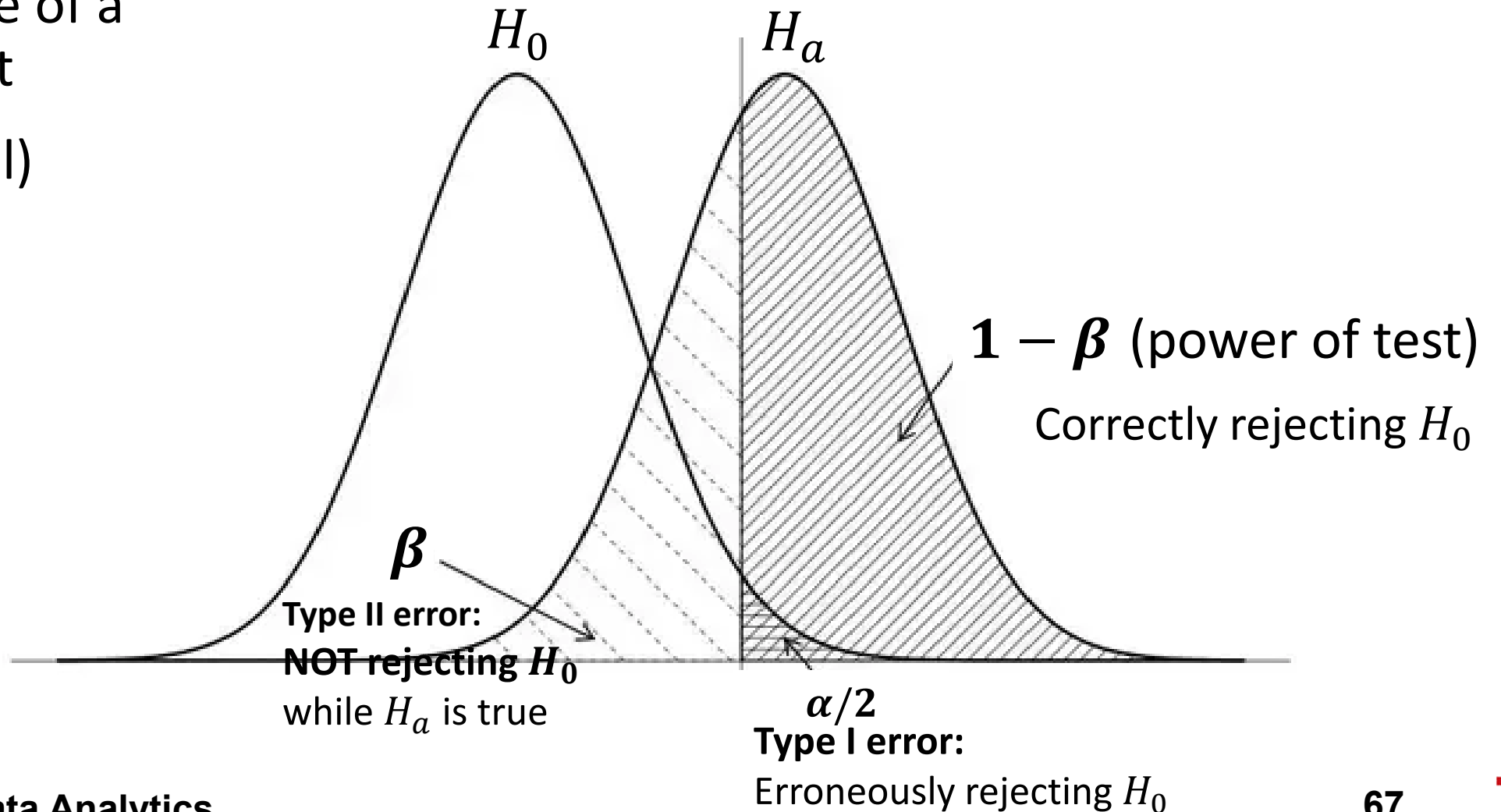


Confidence interval for the concentration of  $PM_{2.5}$  at a confidence level of 95%

# Error types

on an example of a  
two-sided test

( $\frac{\alpha}{2}$  for each tail)



# Errors in hypothesis testing

Hypothesis testing relies on a sample and we can therefore *never* be *sure* that we make a correct decision.

**Type I error:** rejecting  $H_0$  while  $H_0$  is true, **probability  $\alpha$**  (“significance”))

**Type II error:** NOT rejecting  $H_0$ , while  $H_a$  is true, probability  $\beta$

The **power of a test** is the **probability of correctly rejecting  $H_0$**  (so  $1 - \beta$ ).

# Choosing hypothesis tests

- Is it **one data set (sample)**, or are there **two data sets (sample)**?
- Is it about **proportions or means**?
- Is the alternative hypothesis **two-sided or one-sided** (use context to decide)?
- Are there any **assumptions** (e.g., **known or equal variances**)?

**Warning:** names as z-test and t-test are often used (also in Python).

There are multiple z-tests and t-tests.

The names refer to the distribution of the test statistic and may help you in finding the right test if there are options

Test for mean: variance/stddev known (z-test) and variance/stddev unknown (t-test)

# Overview two-sided tests

Examples of hypotheses:

- **one-sample mean**  $H_0: \mu = 10 ; H_a: \mu \neq 10$
- **two-sample means** (independent groups)

$$H_0: \mu_1 = \mu_2 ; H_a: \mu_1 \neq \mu_2$$

or more generally

$$H_0: \mu_1 = \mu_2 + 5 ; H_a: \mu_1 \neq \mu_2 + 5$$

- **one-sample proportion**  $H_0: p = 0.3 ; H_a: p \neq 0.3$
- **two-sample proportions**  $H_0: p_1 = p_2 ; H_a: p_1 \neq p_2$

# Summary: Hypothesis Testing

- Probabilities are useful to quantify evidence for the null hypothesis
- Hypotheses are rejected when the outcome of the experiment is too unlikely (too low probability) if the null hypothesis would be true
- If there are few observations (“small sample size”), then there may not be enough evidence to reject a null hypothesis
- Hypothesis testing is similar to legal procedures in courts
- There are many hypothesis tests (not only for means and proportions, but e.g. for variances), but the basic principles stay the same.
- Choice between one or two-sided alternatives is dictated by context
- Hypothesis testing may be performed using one of the following methods:
  - p-values
  - confidence intervals
  - critical regions (not treated in this course)

# Assumptions – Hypothesis Tests

Previous topic: Statistical Formulation of Hypothesis Tests





"All models are wrong, but some are useful."

(c. 1976)

"Remember that all models are wrong; the practical question is how wrong do they have to be to not be useful."

George E. P. Box (1919-2013)

# Conclusion validity

---

To which extent are conclusions **influenced by assumptions made in the data analysis?**

# Assumptions behind hypothesis tests and confidence intervals

- observations in your data set are **independent** from each other
- observation in your data set **come from the same probability distribution**
- tests and confidence intervals for **means: normality or “large sample size”**
- tests and confidence intervals for **proportions:  $np > 5, n(1 - p) > 5$**   
(conditions to ensure approximate normality)

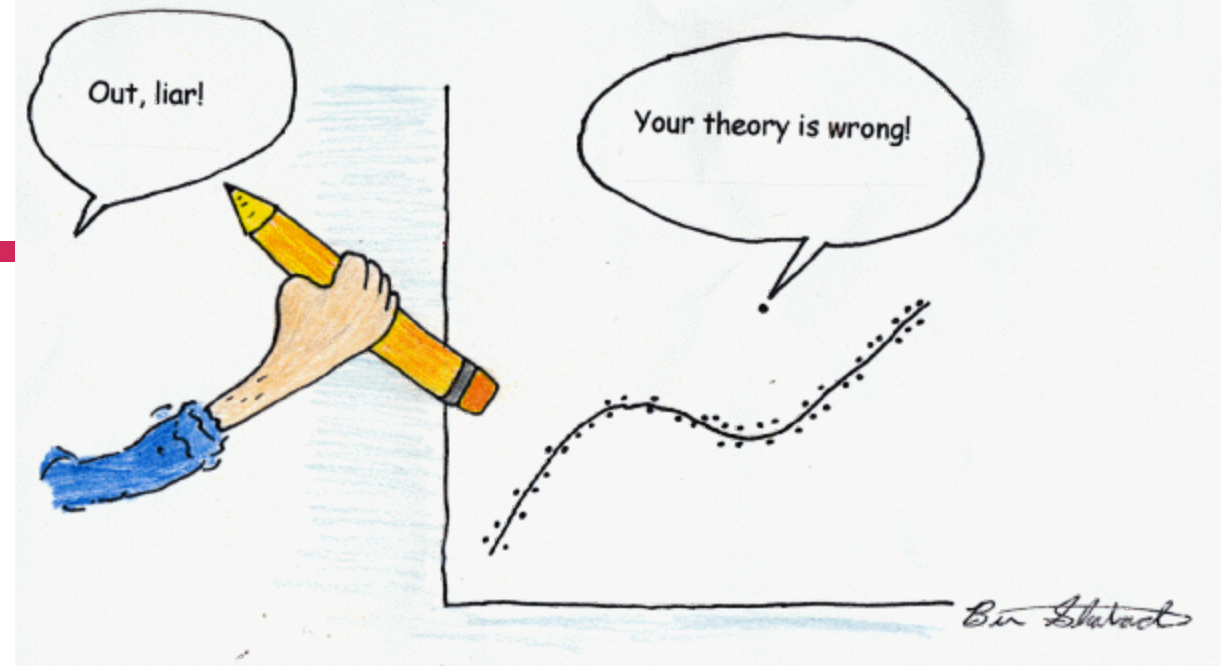
**Software may not check for you whether these assumptions are met.**

If not, you may get ***wrong answers!***

It is your responsibility to **check these assumptions.**

# Outliers

- distinguish between
  - wrong/incorrect data
  - data that do not fit a statistical model
- Outliers could distort analysis results
- Simple graphical tool: Box-and-Whisker plot
- **Rule of thumb:** data points more than 2 or 3 standard deviations away from mean are suspect



Outliers should not just be deleted unless there is a good contextual reason!

# Normality testing

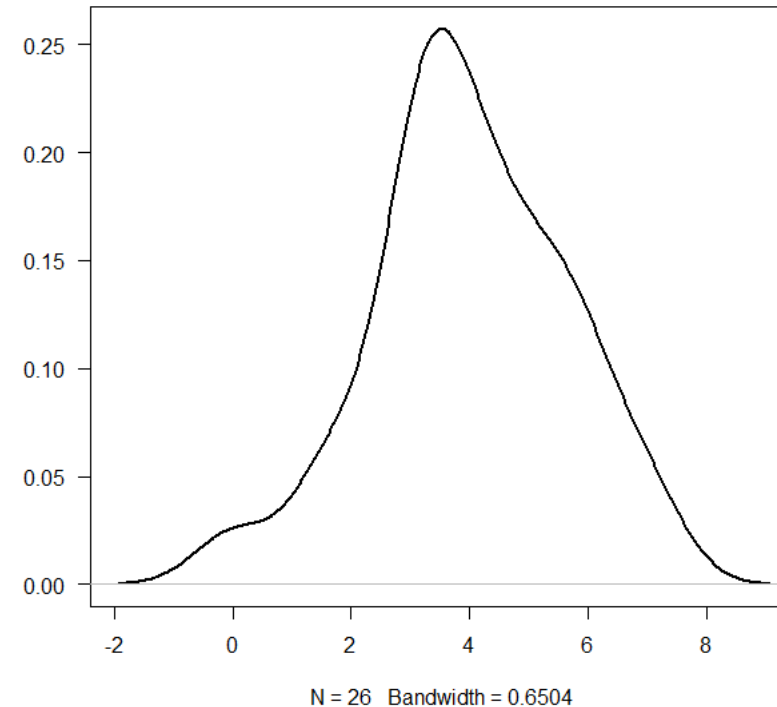
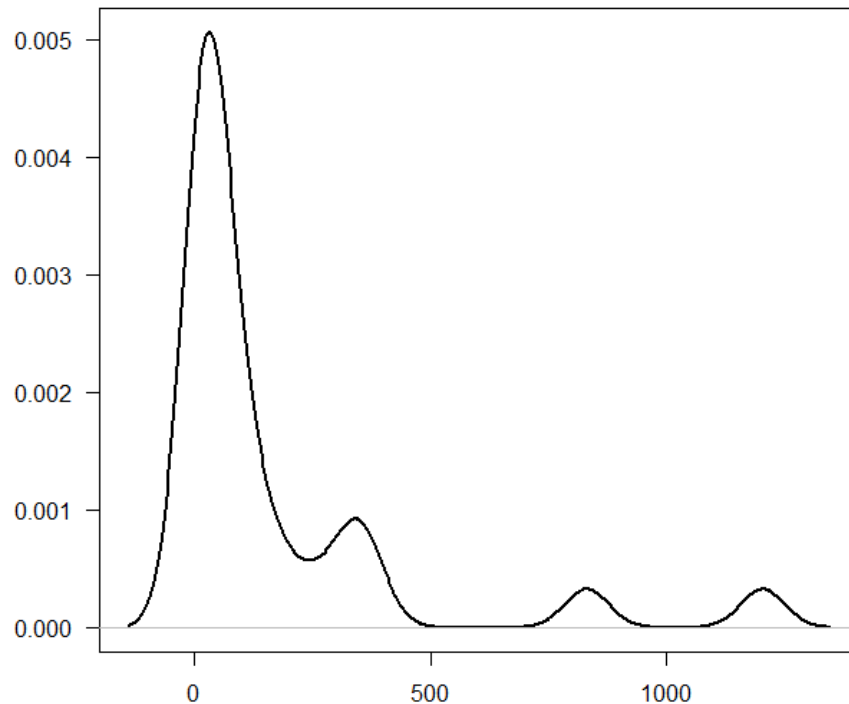
How:

- graphical (gives insight why normality may not be appropriate) :
  - kernel density plot (good for global assessment of shape)
  - normal probability plot (good for detecting whether there are problems)
- goodness-of-fit test (gives objective decision criterion):
  - Anderson-Darling test

Note: **caution when  $n < 20$**  (a single outlier may distort everything)

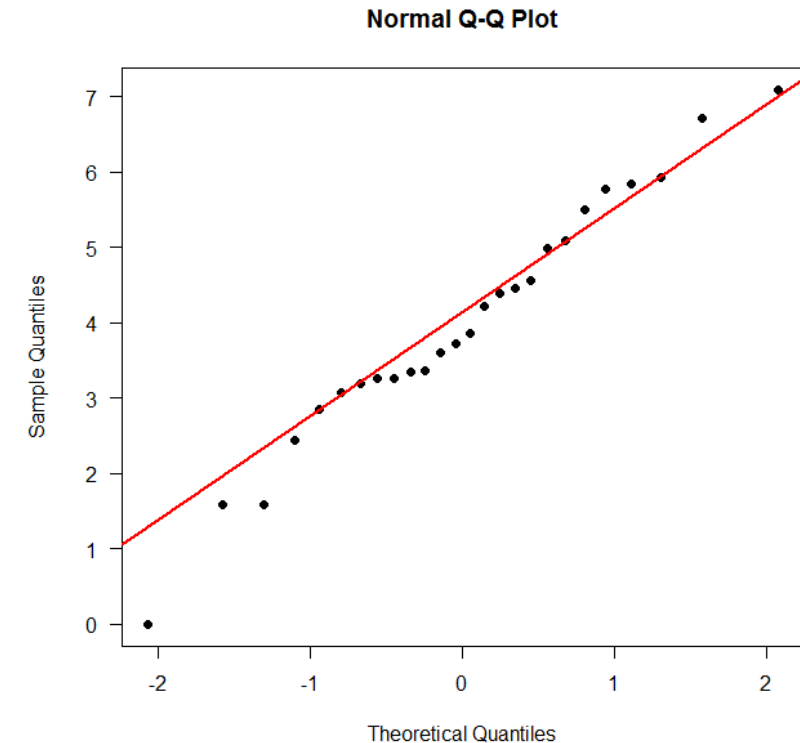
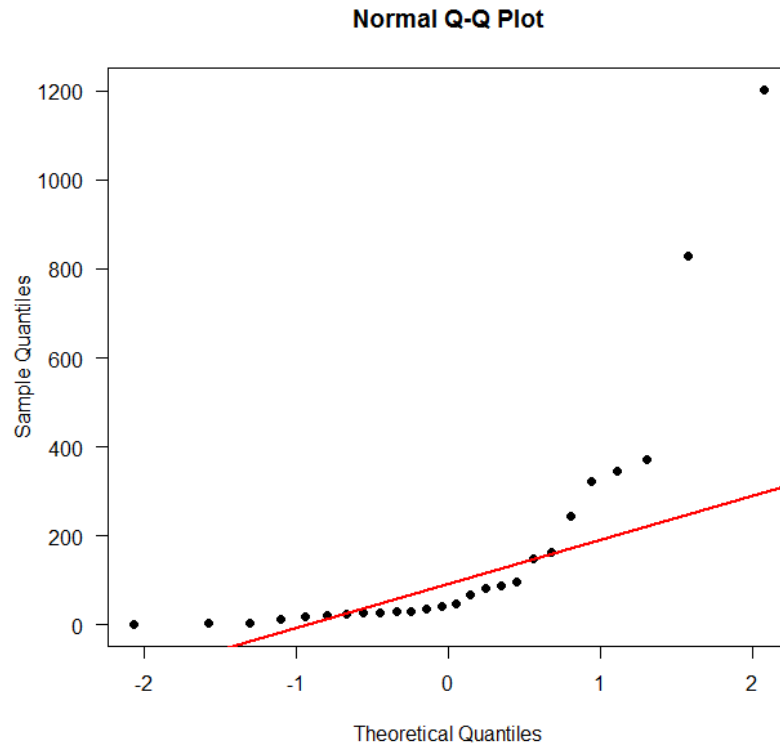
# Normality testing - kernel density plot

- good for “seeing” global shape, symmetry, bell shaped
- not good enough for tails



# Normality testing – normal probability plot

- similar to the ECDF (an improved cumulative histogram)
- trick: transform y-axis so that ECDF becomes straight line



# Normality testing : Anderson-Darling test

This is a statistical test to be used together with the graphical methods. A statistical test gives an objective answer, the graphical methods give insight why the data are distributed like a normal distribution.

$H_0$ : data comes from normal distribution

$H_a$ : data does not come from normal distribution

This test requires software to perform

Interpretation of  $p$ -value: if  $p$  is small, then reject  $H_0$

Practical choice: small threshold (e.g. 0.01 instead 0.05), since the t-test is robust against moderate deviations of normality.



# What to do when normality fails

---

Means and proportions are constructed using sums.

If the sample size is “large”, then such sums may be approximately normally distributed by the **Central Limit Theorem**. In such cases confidence intervals and  $p$ -values can be trusted.

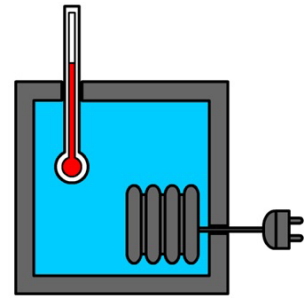
There is no general rule what “large” is (it depends on the data; sometimes 50 is mentioned as a rule of thumb).

# Summary Assumptions - Hypothesis Tests

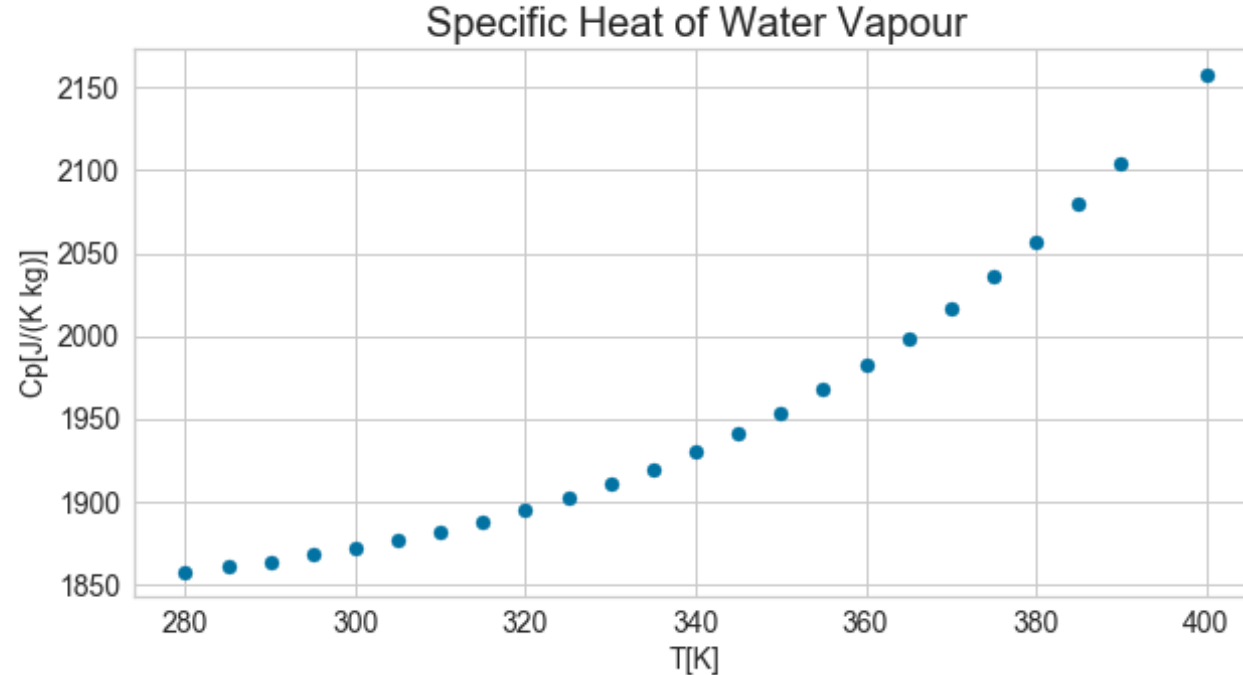
- Hypothesis tests come with assumptions; it is your responsibility to check them – if the assumptions are not met, you may get wrong answers
- Check for outliers using rule of thumb and/or box and whisker plots
- Check for normality should consist of
  - graphical inspection: kernel density plot, normal probability plot
  - goodness-of-test, e.g. Anderson-Darling test
- For large data sets, the Central Limit Theorem may cause the test statistic to be approximately distributed, even if the data is not normal

# Assumptions - Linear Regression

# Linear regression

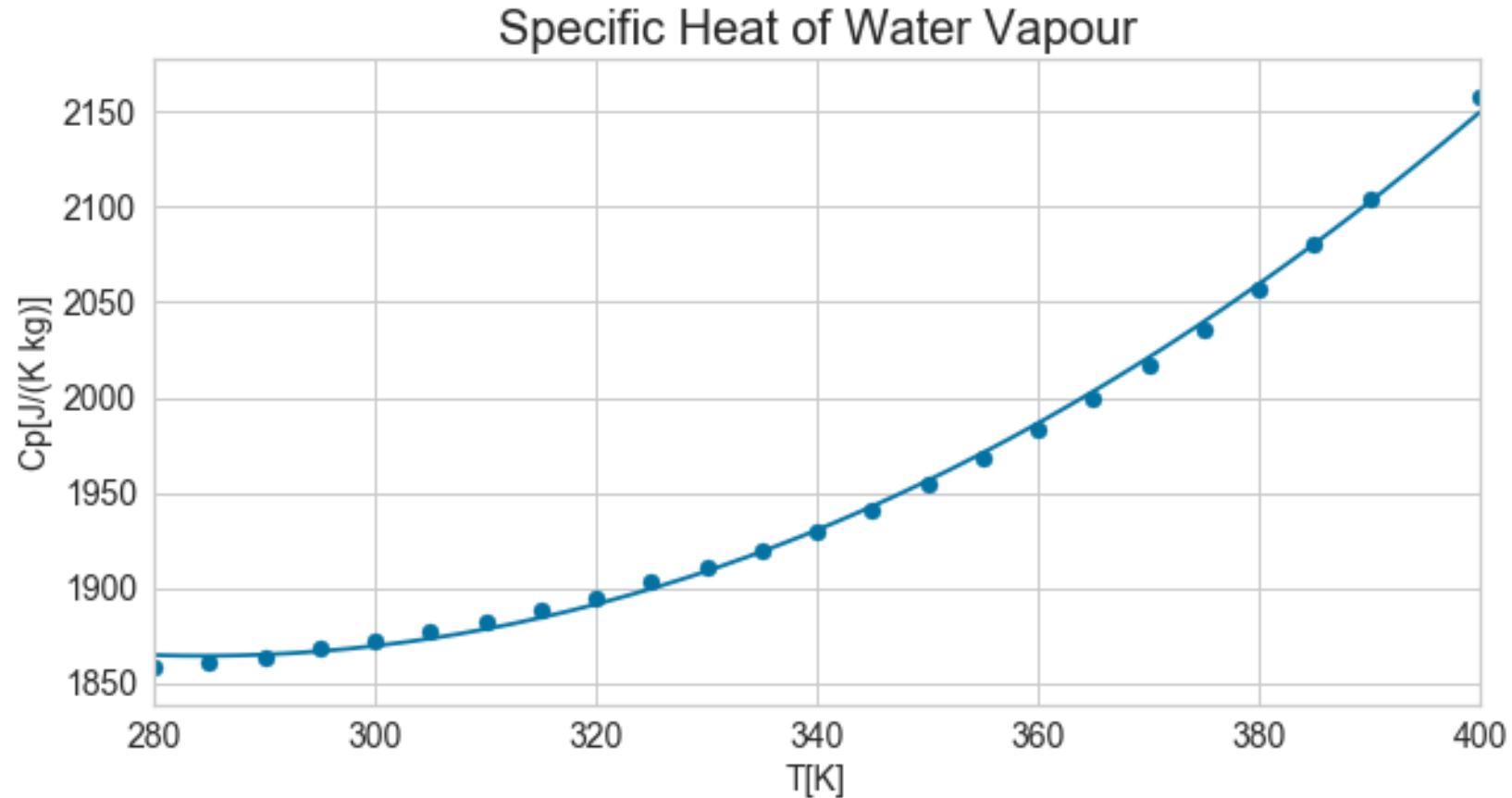
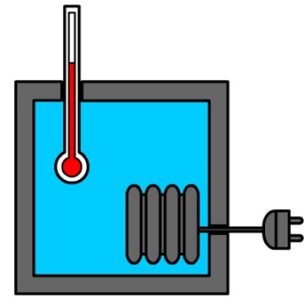


Testing assumptions is even more important when performing regression analysis because there are many more assumptions.

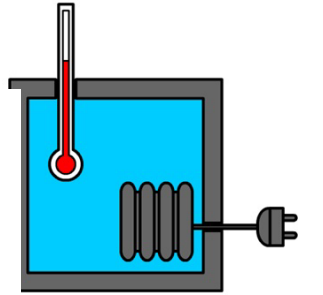
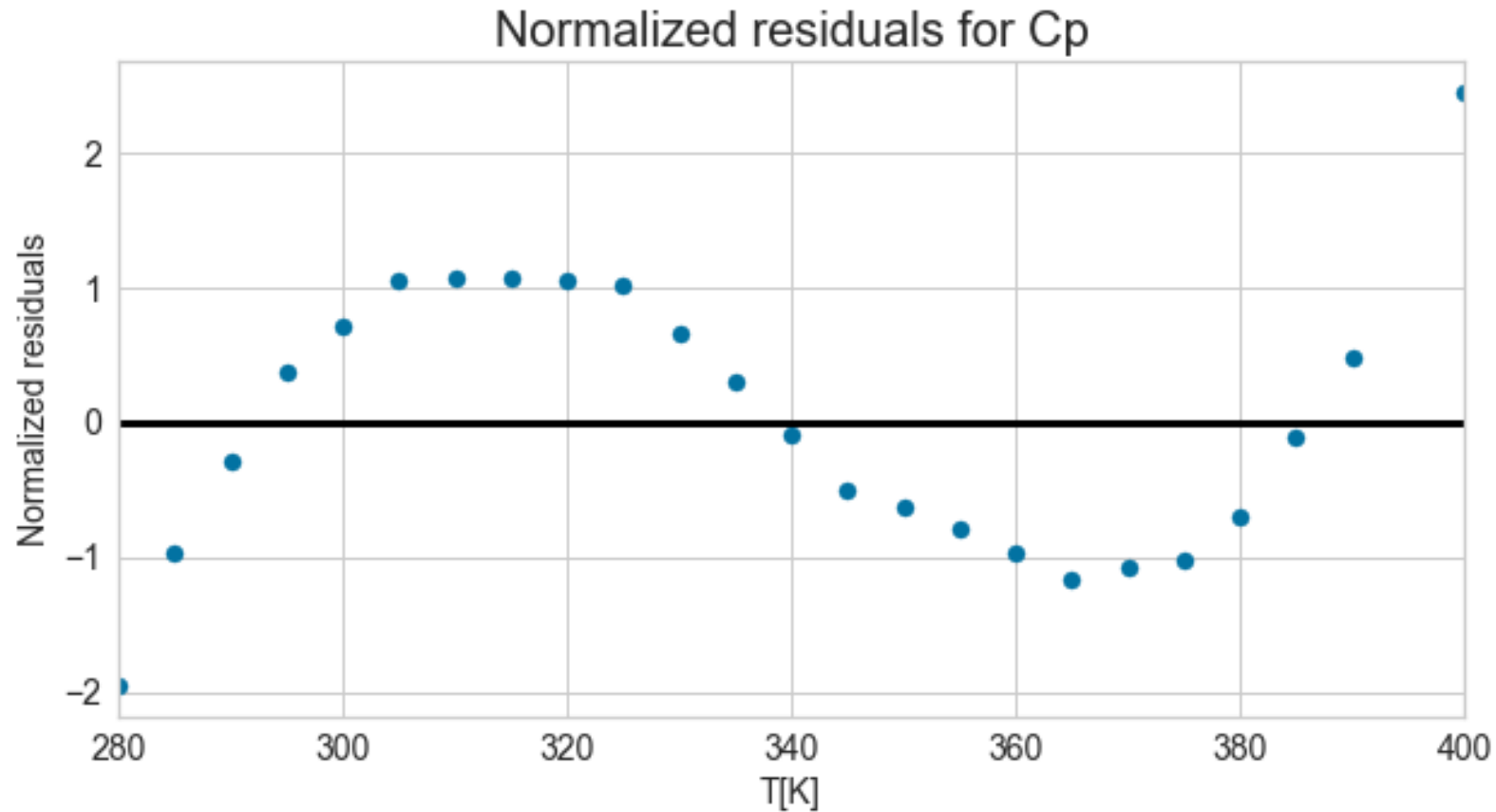


Textbook theory suggest quadratic model  $C_p = \beta_0 + \beta_1 T + \beta_2 T^2$

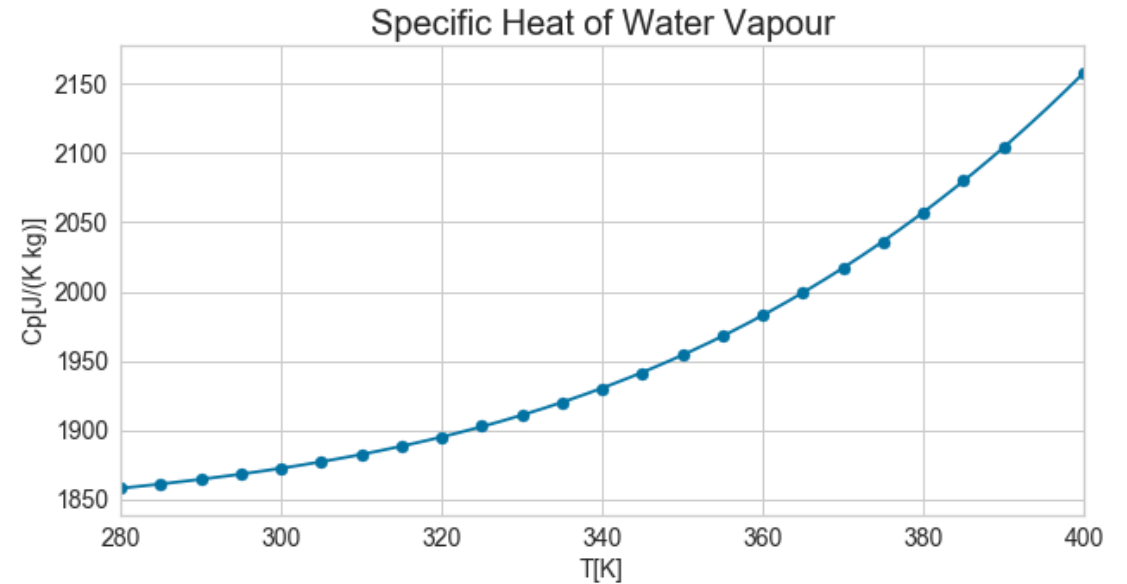
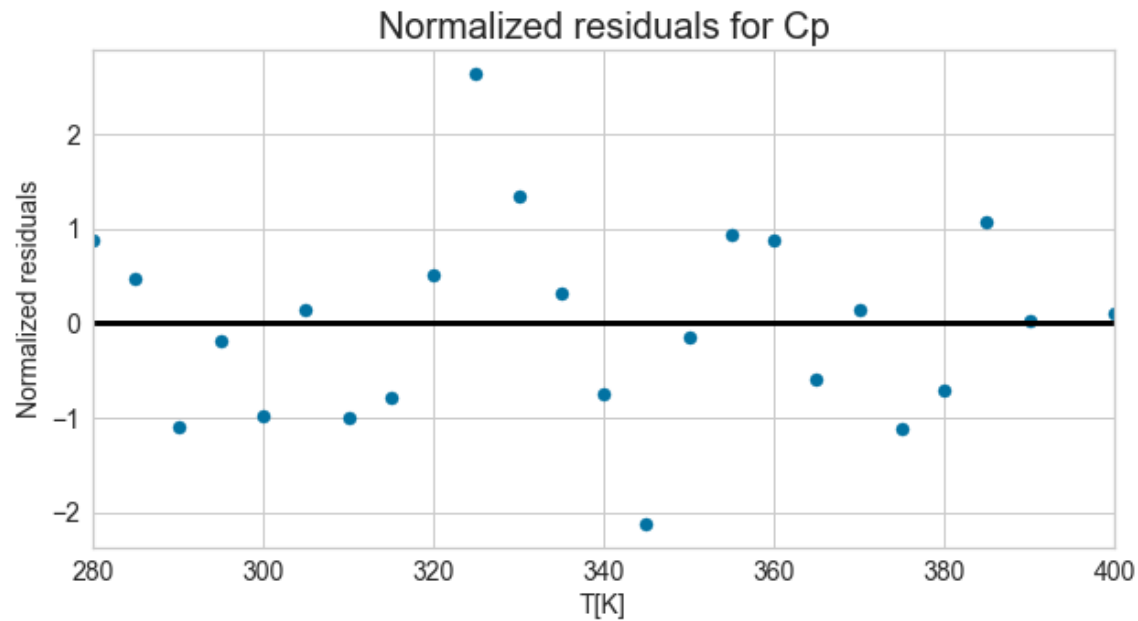
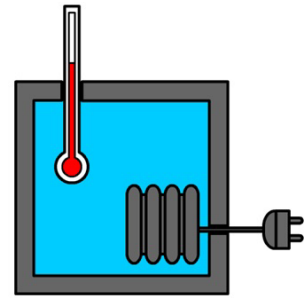
# Second-order polynomial model



# Residuals second-order polynomial model



# Third-order polynomial order



# Modelling assumptions linear regression

Model:  $Y = \beta_0 + \beta_1 x_1 + \cdots + \beta_p x_p + \varepsilon, \varepsilon \sim N(0, \sigma^2)$

Assumptions:

1. expectation is a linear function of parameters:  
$$E(Y) = \beta_0 + \beta_1 x_1 + \cdots + \beta_p x_p$$
2. additive error term  $\varepsilon$  (since error term is written as  $+\varepsilon$ )
3. normality of the error term
4. independence of observations
5. equal variance  $\sigma^2$  of all observations



# Diagnostics

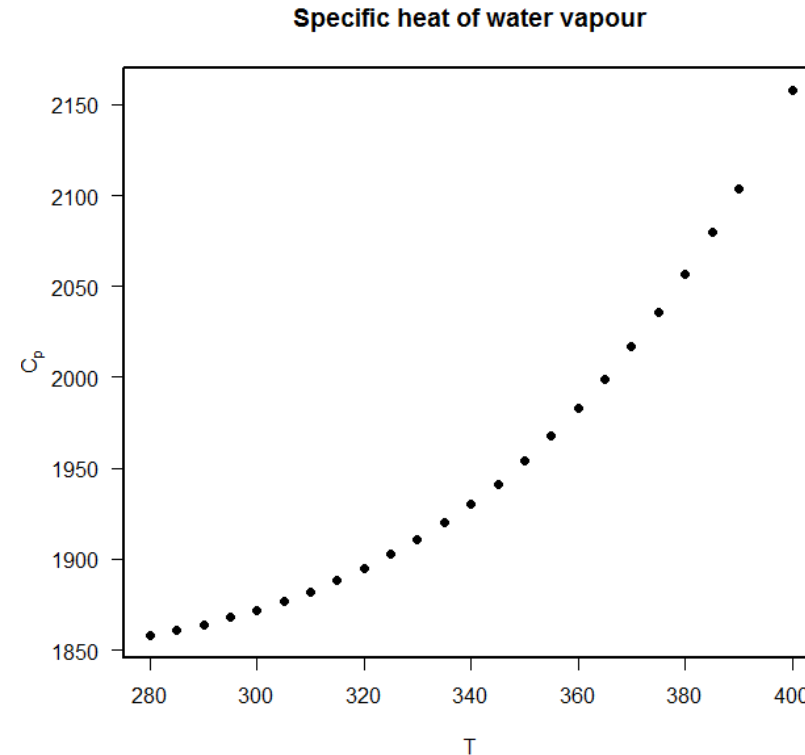
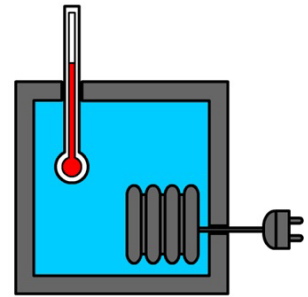
Each of these model assumptions for linear regression may not be met.

They are often hard to check upfront, so one checks them **after** a model has been fitted (“model diagnostics”).

Tools for diagnostics:

1. scatter plots (for assumption 1.)
2. plots of residuals (  $e_i = y_i - \hat{y}_i$  ) versus variables and predicted values (for assumptions 1. and 2.)
3. normal probability plots (assumption 3., based on residuals)

# Linear regression – scatter plot



The scatter plot show that a straight line does not seem to be appropriate.

# Linear regression – residual plots

Raw residuals are the differences  $y_i - \hat{y}_i$  of the observations and the fitted values corresponding to them (the regression model).

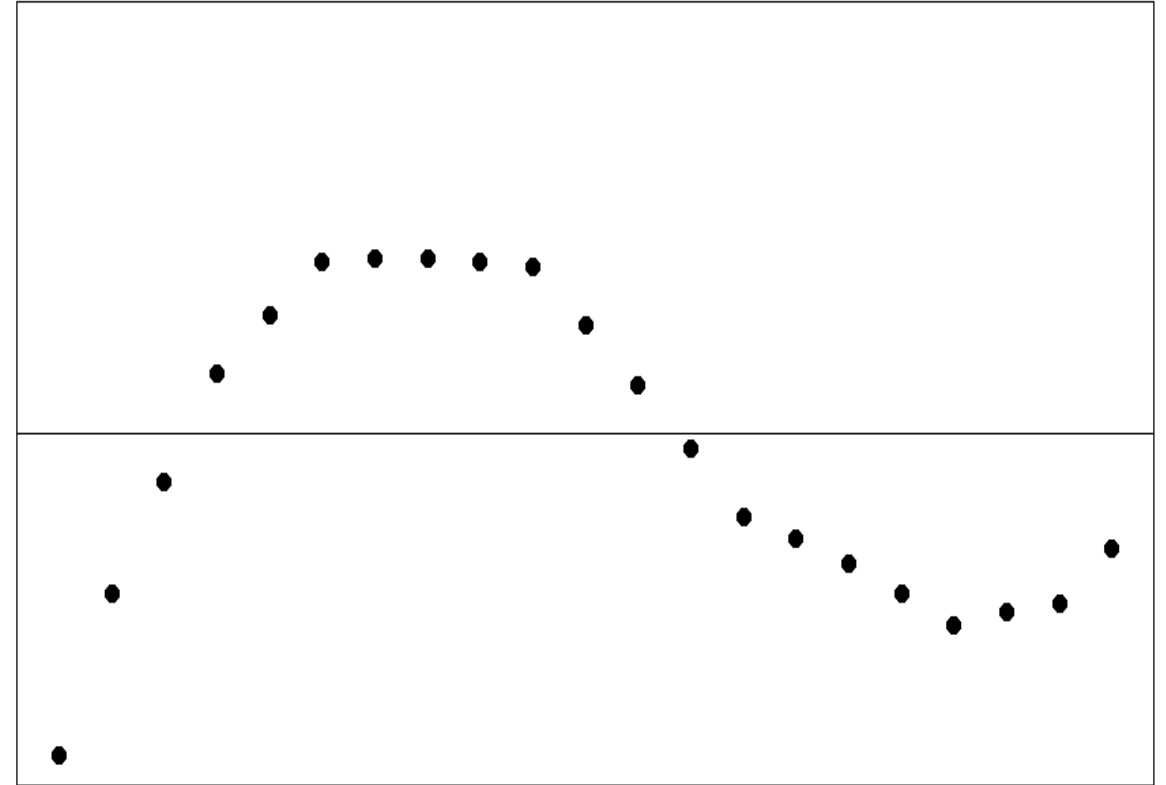
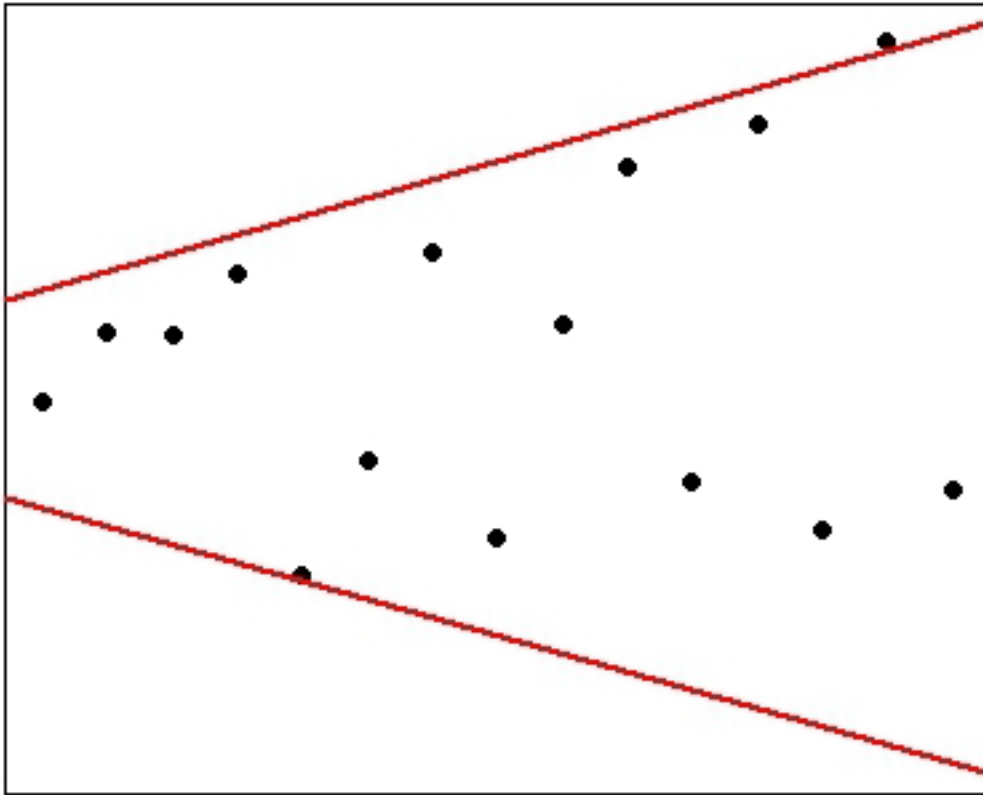
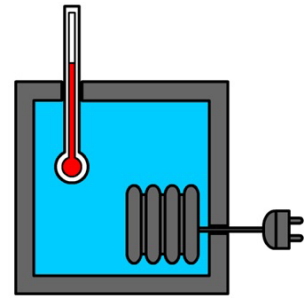
For diagnostics it is better to scale them (studentized residuals) in order to get a universal scale

Residual plots have on the x-axis fitted values or independent variables.

Interpretation rules:

1. no clear pattern (see next slide)
2. all (studentized) residuals within range  $(-2.5, 2.5)$

# Examples of patterns in residual plots



# Linear regression – normality

## Tools

1. normal probability plot of standardized residuals (optimal for detecting deviations)
2. kernel density plot of standardized residuals (better in detecting where deviations occur)
3. Anderson-Darling test applied to standardized residuals

## Remedies

1. remove suspect observations (after checking the context)
2. apply transformation of the data (e.g. log-transform)

# Summary Assumptions Linear Regression

---

- Linear regression depends on several assumptions
- Residuals are an important tool to check assumptions:
  - normality through the tools mentioned for hypothesis tests
  - patterns in residual plots to check for constant variance and model deviations

# Summary of Lecture

## Statistical Formulation of Hypothesis Testing

# Important concepts that you should know

---

- probability distribution
- empirical cumulative distribution function
- cumulative distribution function
- binomial distribution
- density
- expectation
- variance
- normal distribution



# Important concepts that you should know (2)

---

- null hypothesis and alternative hypothesis
- $t$ -test
- confidence interval
- $p$ -value
- type I,II error / significance
- goodness-of-fit test
- residual plot
- outlier

# What should you be able to do?

---

- formulate statistical hypotheses from a simple practical situation
- choose the correct confidence interval (means/proportions) and compute it using software
- choose the correct hypothesis test (means/proportions) and perform it using software
- interpret the  $p$ -value of a hypothesis test
- perform a hypothesis test when you are given a confidence interval
- perform normality check on data

# Key insights

---

- Probability distributions are a mathematical way to model uncertainty
- Hypothesis testing is a way to obtain scientific answers to questions using data
- Confidence intervals show how certain you are and may also be used to perform hypothesis tests
- Statistical analyses come with assumptions about the data. It is the responsibility of the user to check these assumptions in order to obtain conclusion validity
- Checking assumptions of statistical analyses can often only be done after analysing the data. There is a wide range of tools for this.

# Improbably Happy New Year!

