A soft introduction to statistical inference

30 November 2015

Ole Christian Lingjærde, Biomedical Informatics, UiO

The purpose of statistics

Statistics can help us with many tasks involving data:

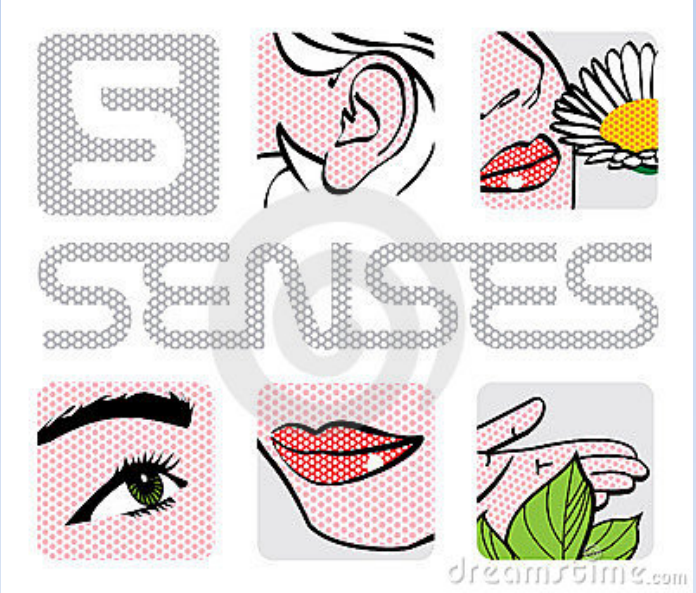
- How to design an experiment to maximize the information output
- How to extract information from observations
- How to guard against drawing false conclusions

How to make sense of chaos

Most of what we observe through our senses is irrelevant to the tasks we seek to perform.

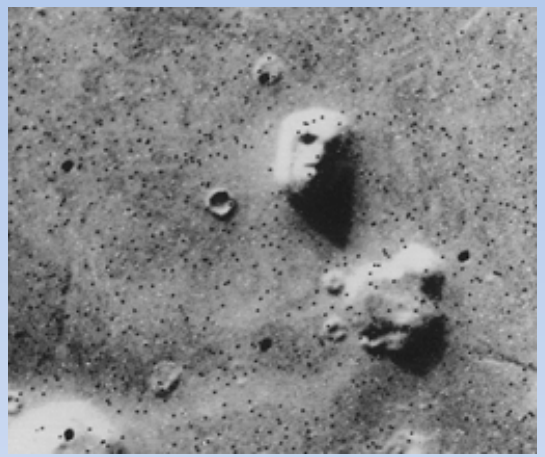
We are (mostly) incredibly good at extracting the relevant part of the data we receive and discarding the irrelevant parts.

Example: keeping track of a single conversation in a room filled with chattering people.

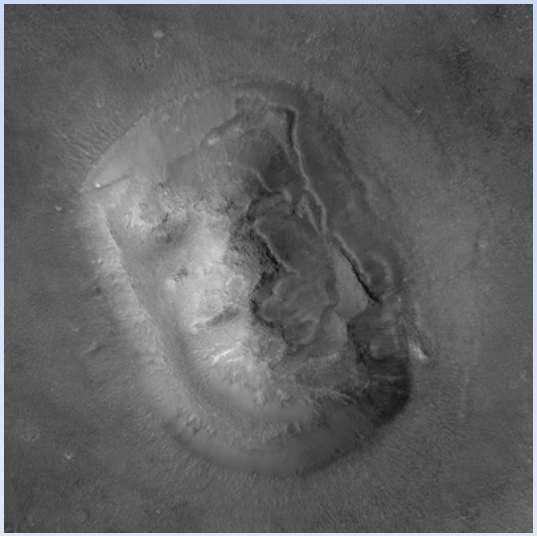


We fill in the gaps when data are missing

Sometimes our brain is too good at this!



1976



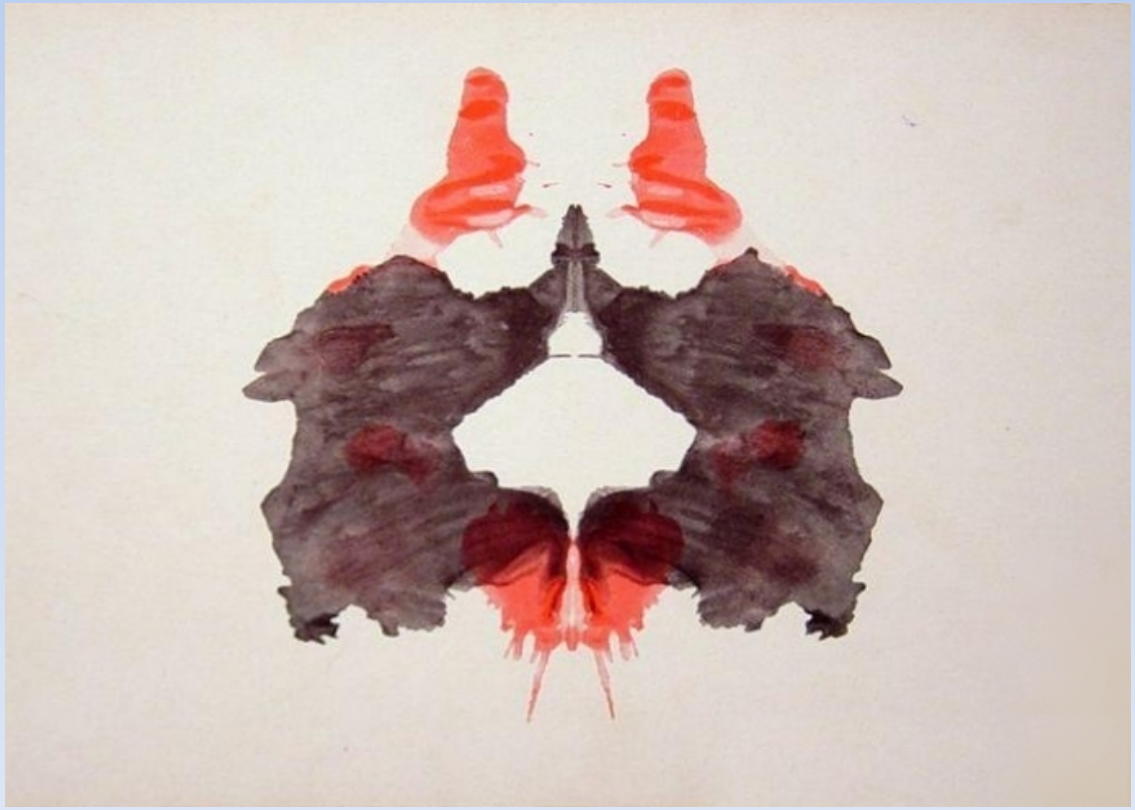
2001

NASA's web site:

NASA's Viking 1 spacecraft was circling the planet, snapping photos of possible landing sites for its sister ship Viking 2, when it spotted the shadowy likeness of a human face. An enormous head nearly two miles from end to end seemed to be staring back at the cameras from a region of the Red Planet called Cydonia.

http://science1.nasa.gov/science-news/science-at-nasa/2001/ast24may_1/

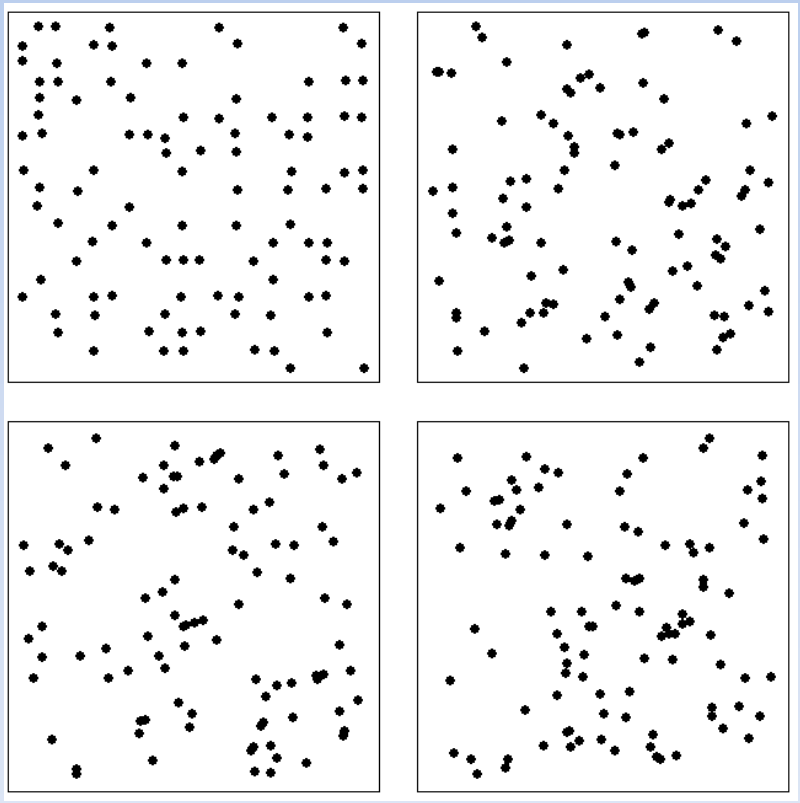
How we interpret things is partly a result of our own personality: subjectiveness.



Card no. II in the Rorschach test

http://www.dagbladet.no/2009/08/20/magasinet/psykologi/helse/7707225/

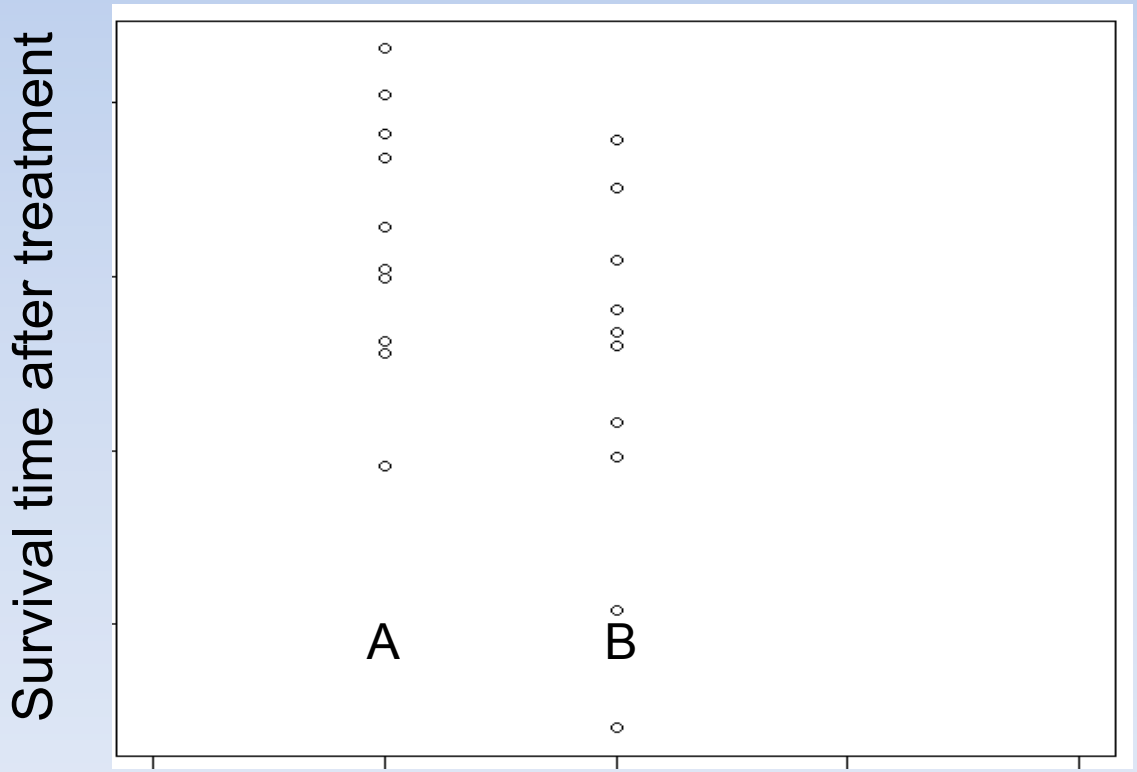
Subjectiveness can play a major role in interpretation of abstract patterns



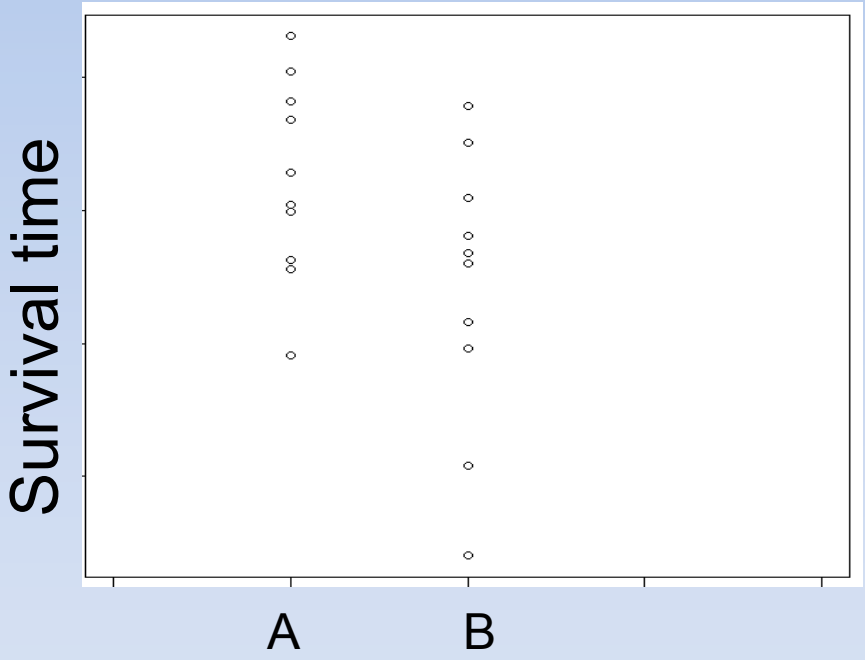
Which of these four points clouds are drawn at random with uniform distribution within the square?

Correct answer: all except the one in the upper left corner.

Should you expect to live longer after treatment A than after treatment B?



Important note



It is easy enough to conclude that the average survival time in group A exceeds that of group B. This is a summary of the data.

What statistics may help us with is to conclude whether this results holds in general (for new cases).

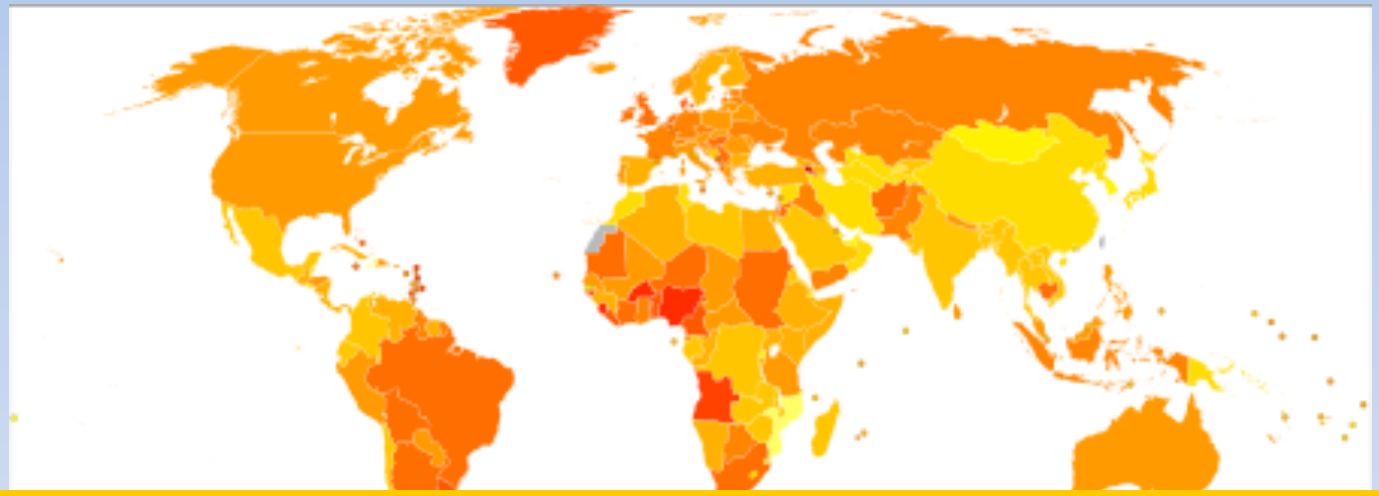
Thus we want to generalize from a (small) sample of observations to the whole population (most of whom are not observed).

Morale:

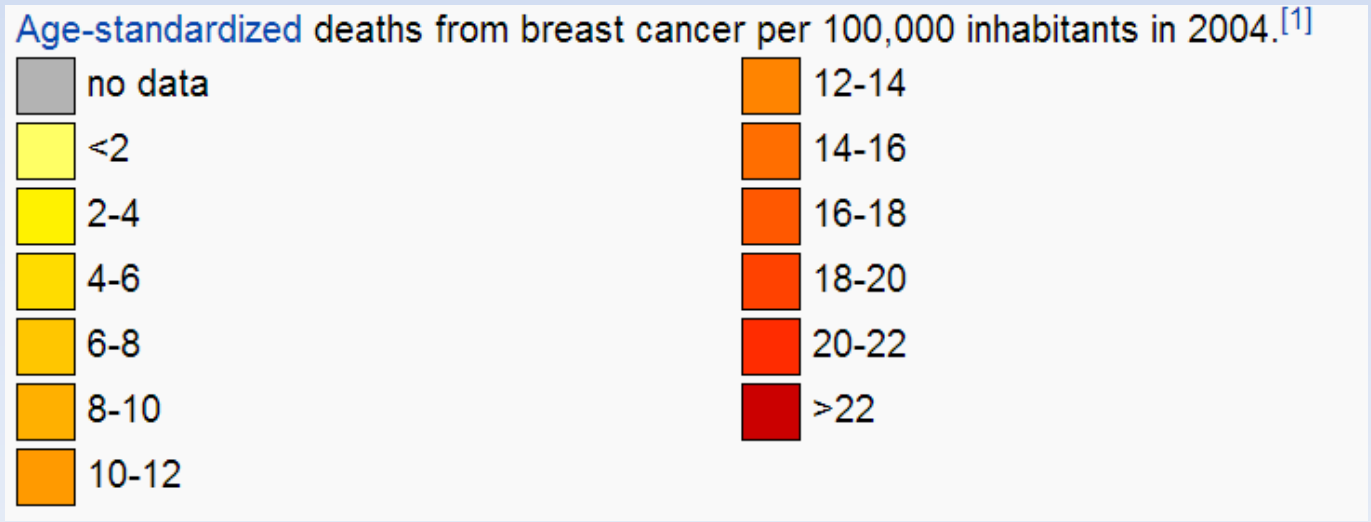
Our brain is typically very efficient at finding patterns in data

Sometimes these patterns are not real – i.e. they are merely the result of the observations not being a totally representative subset of the population.

So what patterns should we trust?



Which of the differences represent more than simply random variation in the observations?



Example: power lines

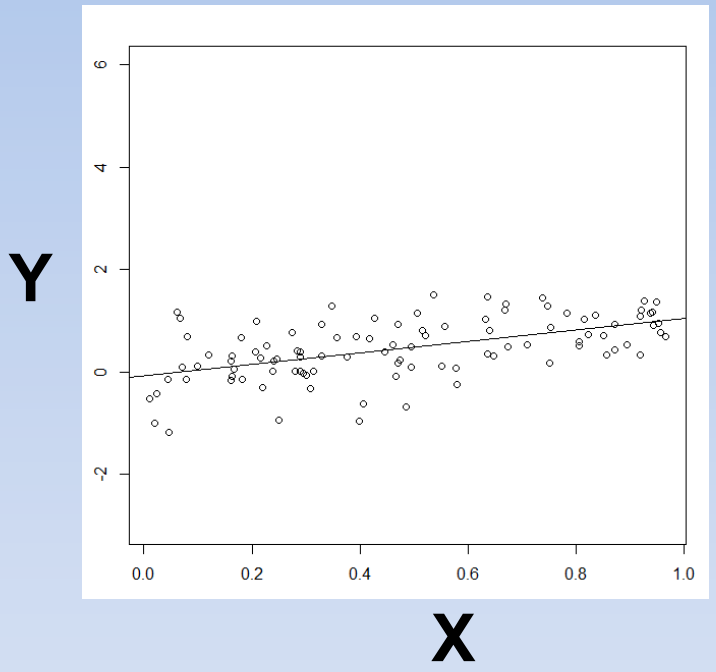
Some studies have found a weak link between living near electric power lines and risk of blood cancer (leukemia).

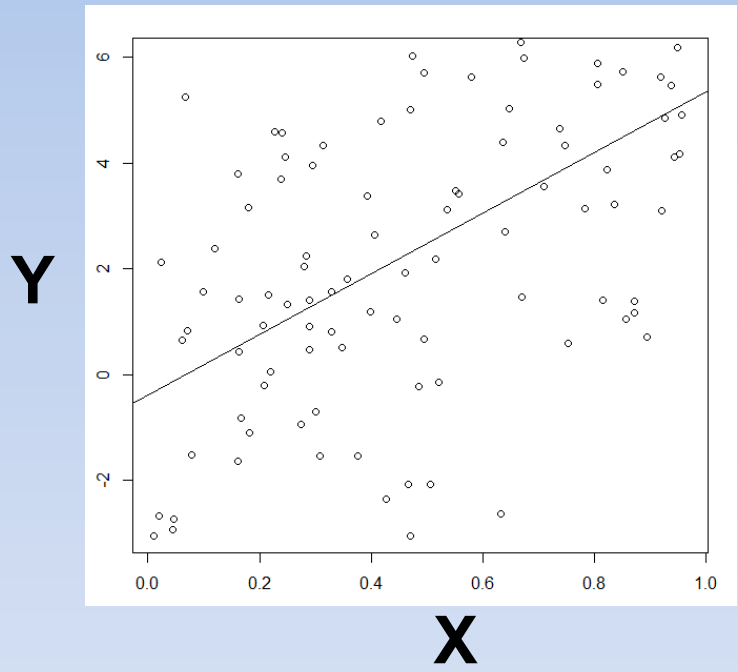
Other studies have not been able to establish such a link.

A study in 1979 pointed to a possible association between living near electric power lines and childhood leukemia (3). Among more recent studies, findings have been mixed. Some have found an association; others have not. These studies are discussed in the following paragraphs.

> How would you interpret the phrase ”weak link”?
> How come some studies find a link and others not? Any ideas?

What is a strong link (strong association)?





The correlation between X and Y is the same in both plots (0.55). The scale on the X and Y axes are identical in the two figures.

If X = the amount of tea a person drinks per day
and Y = risk of kidney cancer
then which of the two scenarios should scare tea drinkers the most?

Seemingly unlikely incidents may be random

How likely is it that two pupils in a class of 30 have birthday on the same day?

In fact the probability > 70%!

Analogy: you have a bag with 365 coins. You draw a coin at a time from the bag thirty times, every time putting the coin back in the bag before the next draw. What is the probability that the same coin is drawn at least twice?

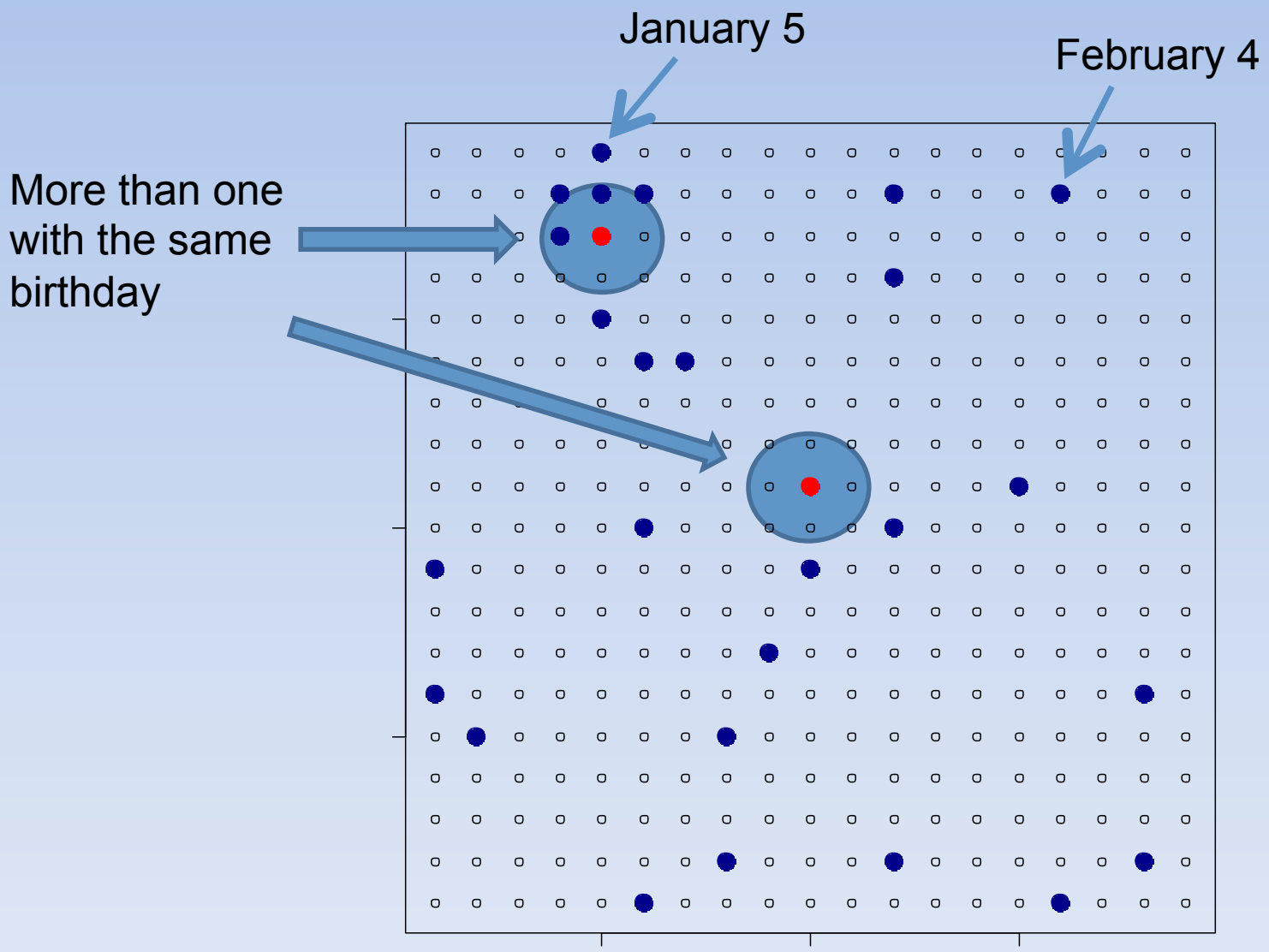
Let’s do an experiment

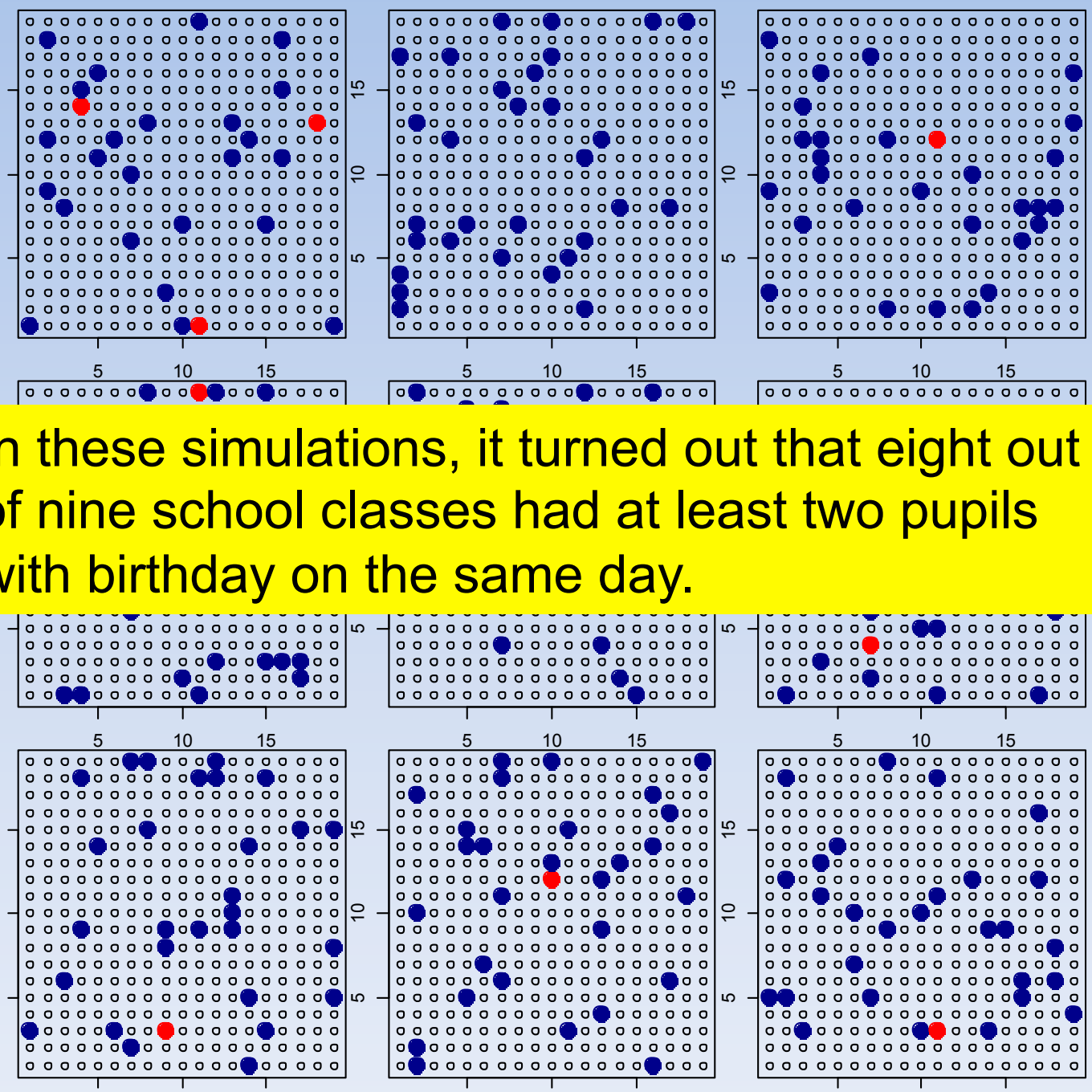
1) Draw a checkerboad with 365 squares, each representing a day of the year.

2) Draw a random number between 1...365 and color the corresponding square blue.

3) If the square is already blue, color the square red.

4) Repeat points 2) and 3) thirty times.


January 5
February 4
More than one with the same birthday



In these simulations, it turned out that eight out of nine school classes had at least two pupils with birthday on the same day.

Morale: out intuition cannot be trusted with regard to which events are likely to be caused by random variation and which are not likely to be caused by such variation.

Rare diseases are common!

Scleroderma is a rare human disease characterized by over-production of collagen and hardening of the skin. It can also affect internal organs.

Suppose 1 out of 5000 get this disease. That sounds pretty rare, and you would think that any person getting this disease must have very bad luck.

Rare diseases are common!

However, there are more than 7000 so-called rare diseases (affects fewer than 2.5 out of 5000 in the general population).

In Europe alone, there are more than 30 million people with a rare disease. On a global scale, roughly 5-10% of all humans have a rare disease.

Thus, it is in fact quite common to have a rare disease! It is just not common to any specific rare disease.

Example

Peter goes to his doctor every year to get a check of his health.

This year, his physician takes five blood tests. All five tests are negative (meaning they are within the 95% normal population reference level).

Peter believes his doctor is not taking his health sufficiently serious, and decides to go to another one to take more tests.

The new doctor takes 100 blood tests, and somewhat to the surprise of both him and his doctor, one test is positive even though Peter shows no signs of disease.

> How would you interpret this?

The number of positive findings increases with the number of tests performed

It is certainly possible that the doctor found a condition in Peter that has to be treated.

However, the chance of a false positive increases with the number of tests performed.

- Suppose Peter is healthy
- According to specification, every test has a probability of 95% of being negative for a normal person.
- The probability of all 100 tests being negative for a normal person is therefore (assuming independent tests):

$$\underbrace{0.95 * 0.95 * \ldots * 0.95}_{\text{100 ftimes}} < 1\%$$

- Thus it is > 99% probability of a positive test result!

What statistics can do for us

Provide compact descriptions of large or huge amounts of data

Help us decide whether what we see is a likely result of randomness, or whether it cannot be explained by this.

Central concepts

Observations:	the measurements that we do and that we want to analyse statistically
Population:	The group of individuals that we want to draw inferences about
Sample:	the individuals that we actually observe
Random sample:	A sample that is representative of the whole population (in the sense that every member of the population had the same probability of being chosen).

Example

Population: Norwegian voting population	**Observation**: Answer from N people when asked what they will vote
Sample: Thousand persons in the voting population	**Random sample:** Thousand randomly selected persons in pop.

> Suggest how one could achieve a sample that was (a) not random; (b) random.

Representative sample?

Suppose we want to consider coffee intake.

Population: all persons between 18 and 23 year

Observation: no of cups of coffee per day

Sample: the first 100 persons in line at the coffee bar at the university library

> Is this a representative sample (why/why not)?

Example

A sample is only good as far as it is representative for the population as a whole.

Sounds banal – but a major reason for the ”bad reputation” of statistics is unrepresentative sampling!

Example: Vulnerability analysis of planes returning from bombing missions during World War II.

Data types

Ratio	Numerical values, equal intervals represent equal distances, may talk about small/large
Interval	Numerical values, equal intervals represent equal distances
Ordinal	categories with a natural sequence (e.g. small, medium, large)
Nominal	categories (e.g. yes/no)

Examples

Nominal County (Asker, Oppland)	**Ordinal** Seriousness (small, medium, high)
Interval Temperatures	**Ratio** Height above sea level

Categorical data

Numerical data

What data types do we have?

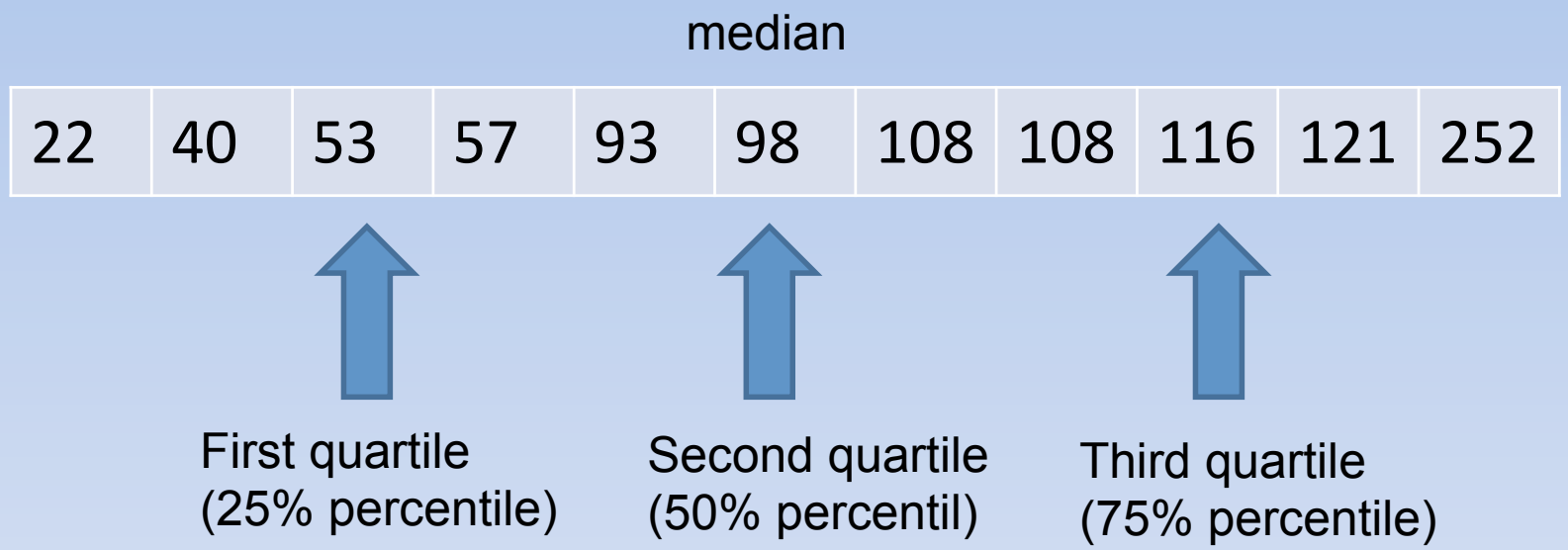
- Body weight
- Calory consumption (Joule)
- Gene expression (log-scale)
- Genotype (AA, AB, BB)
- Mutation (A>C, A>G, A>T, C>A, ...)
- Cluster assignment (cluster1, ..., clusterN)

Center measures

- **Mean**
  Sum of all N measurements divided by N

- **Median**
  Sort all N observations from the smallest to the largest and take the middle one if N is odd and the average of the two middle ones if N is even.

- **Mode**
  Take the most frequent observation



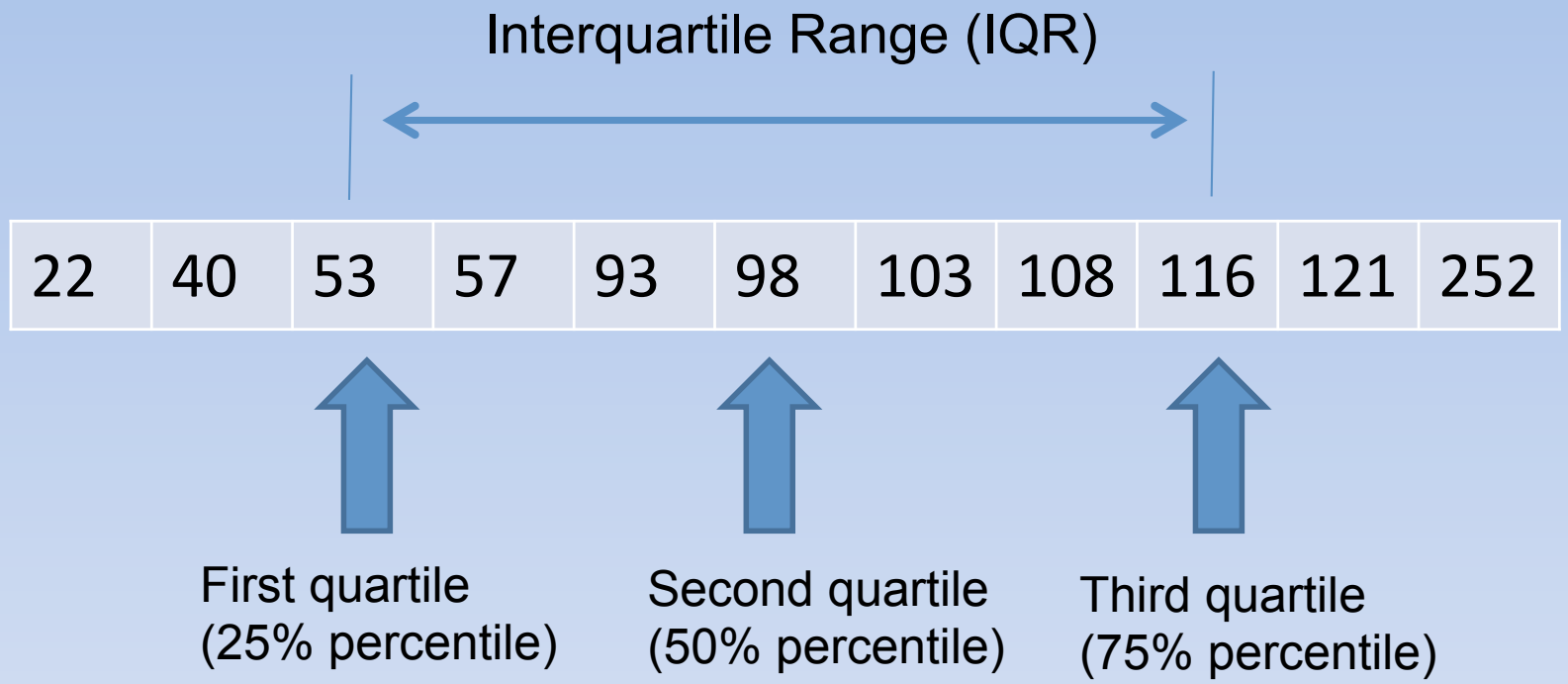
Mean = (22 + 40 + ... + 252) / 11

Median = 98

Mode = 108

Spread measures

- **InterQuartile Range (IQR)**
  The difference between the third quartile and the first quartile (116-53 = 63 on the previous slide)
- **Median Absolute Deviation (MAD)**
  First find the median. Then find the absolute difference between each observation and the median. Then take the median of all the resulting numbers.
- **Standard Deviation (SD)**
  First find the mean. Then calculate the squared difference between each observation and the mean. Then take the mean of all the resulting numbers (dividing by N-1 rather than by N).

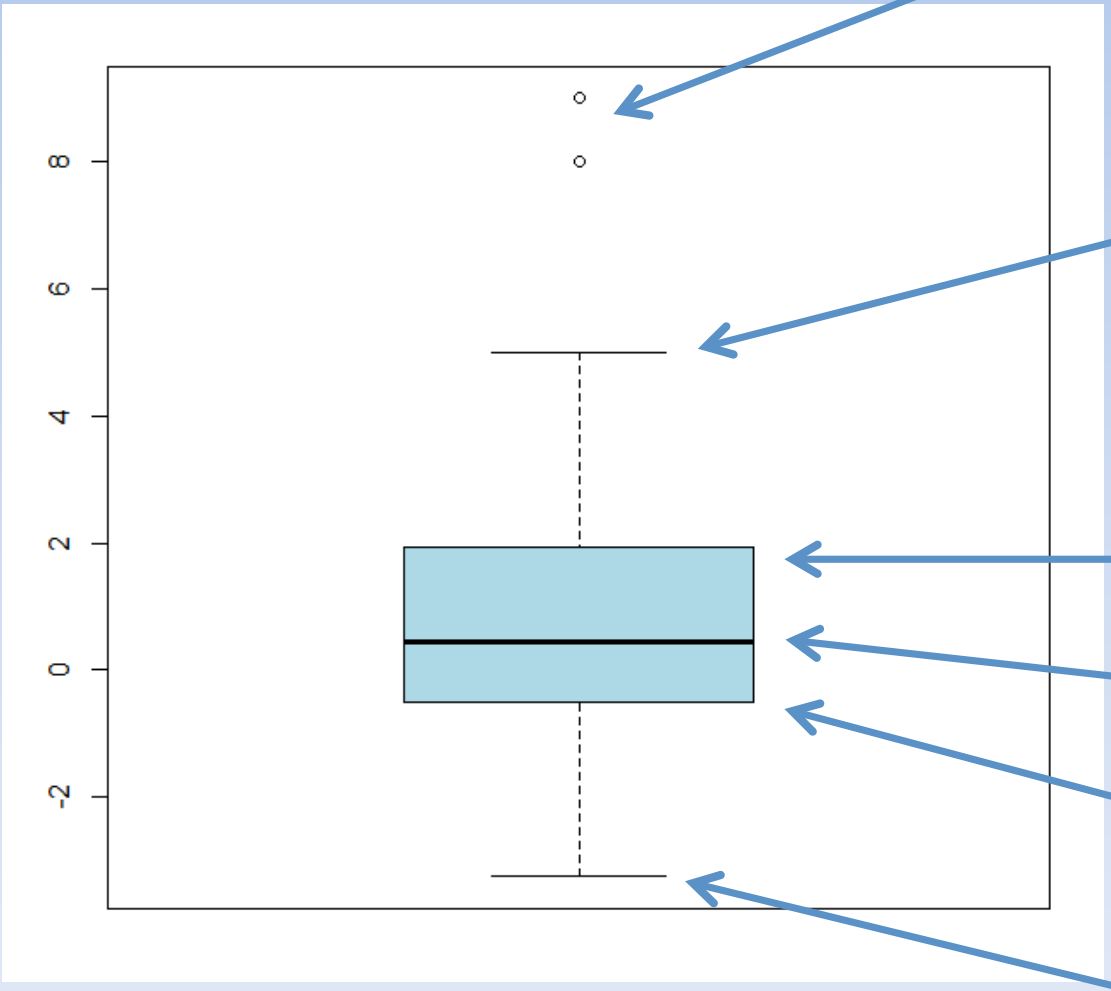


$$IQR = 116\text{-}53 = 63$$

$$MAD = median(|22 - 98|, |40 - 98|, |53 - 98|, \ldots, |252 - 98|) = 23$$

$$SD = \frac{1}{10}\sqrt{(22 - 96.63)^2 + (40 - 96.63)^2 + \cdots + (252 - 96.63)^2} = 61.27$$

Box plot



'outliers'

largest observed value not more than 1.5 IQR from the box

third quartile

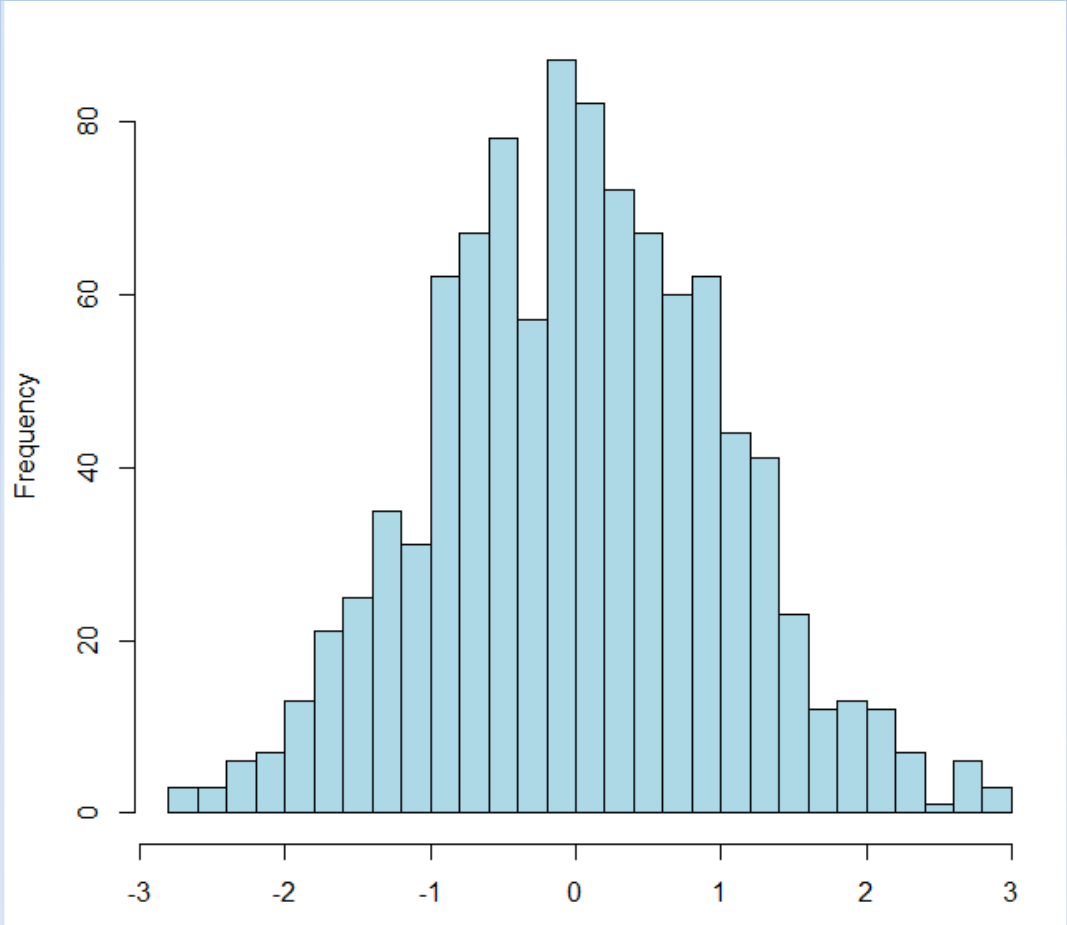
second quartile (= median)

first quartile

least observed value not more than 1.5 IQR from the box

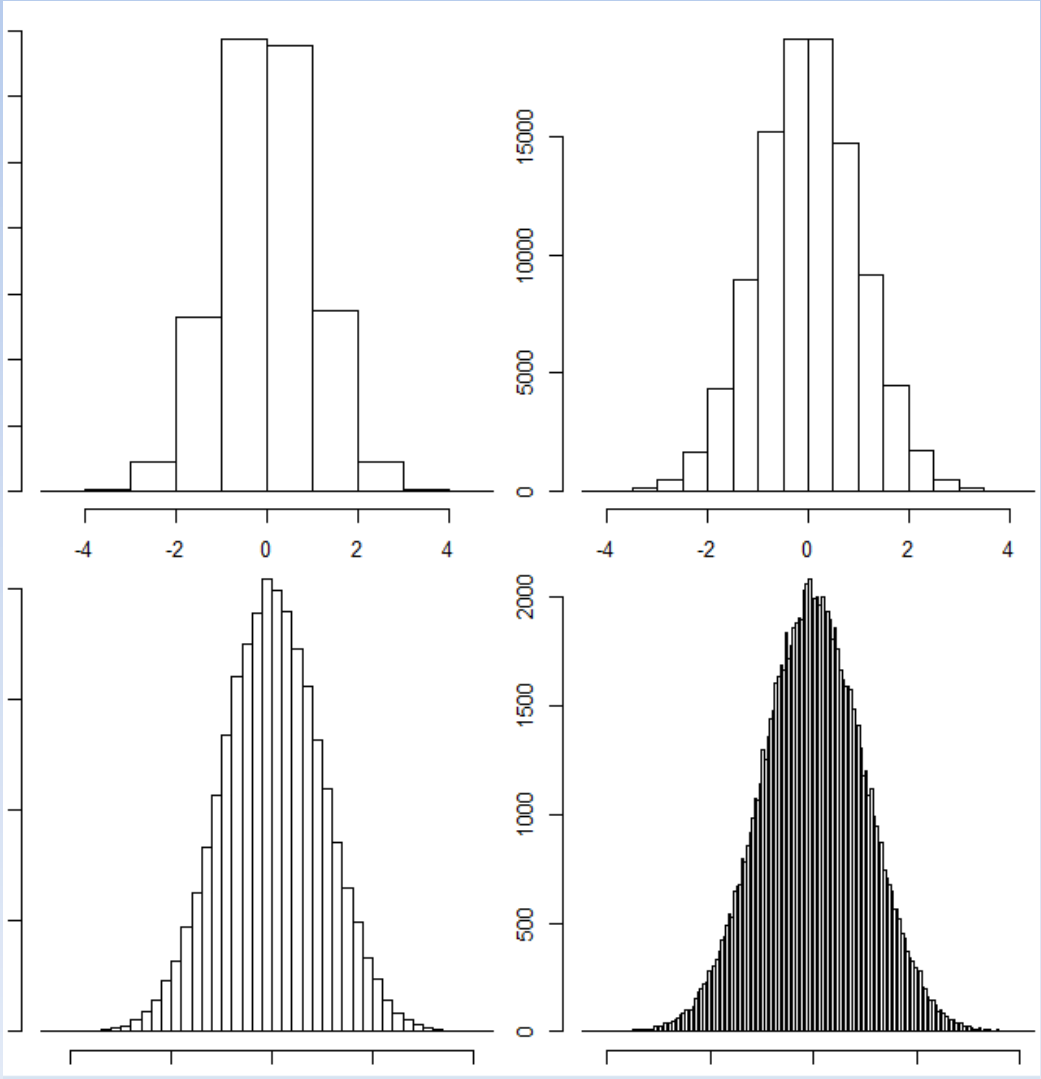
Very useful visualization of the empirical distribution of a numerical sample, when the data are reasonably well spread out. > When would you not use it?

Histogram



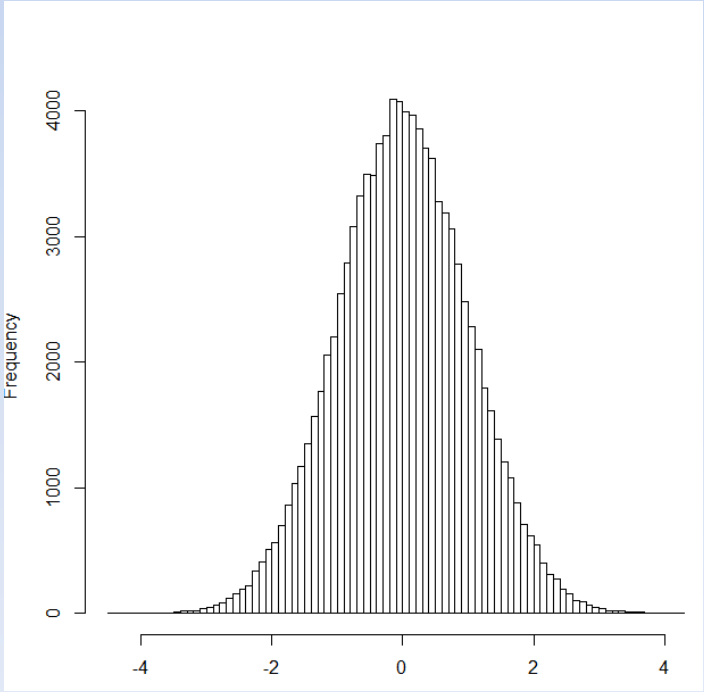
A more detailed picture of the empirical distribution of a numerical sample. Number of bins should not be too large (why?). The default in R tends to produce quite few bins and you may want to overrule this.

The same sample in four ways



Bell shaped histograms

For many types of data, the resulting histogram will have an approximate bell shape (possibly after a transformation), suggesting some similarity to a normal (Gaussian) distribution.



Why bell shape?

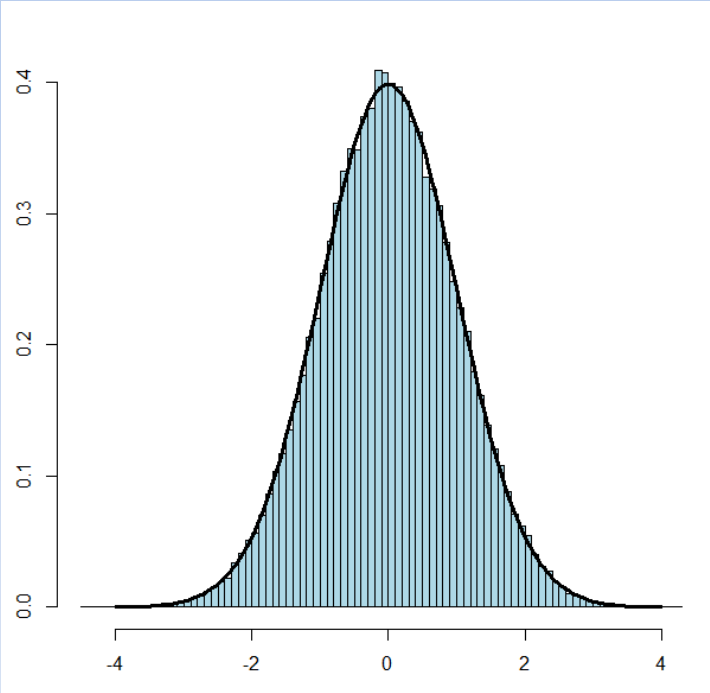
You can show mathematically that when you average over many quantities, each of which follows the same distribution, then the resulting distribution will be approximately bell shaped.

- Some values occur more often (close to mean)
- Values occur with equal frequency at each side of the mean – the distribution is symmetric
- Values very far from the mean do not occur at all in practice (although theoretically they may)

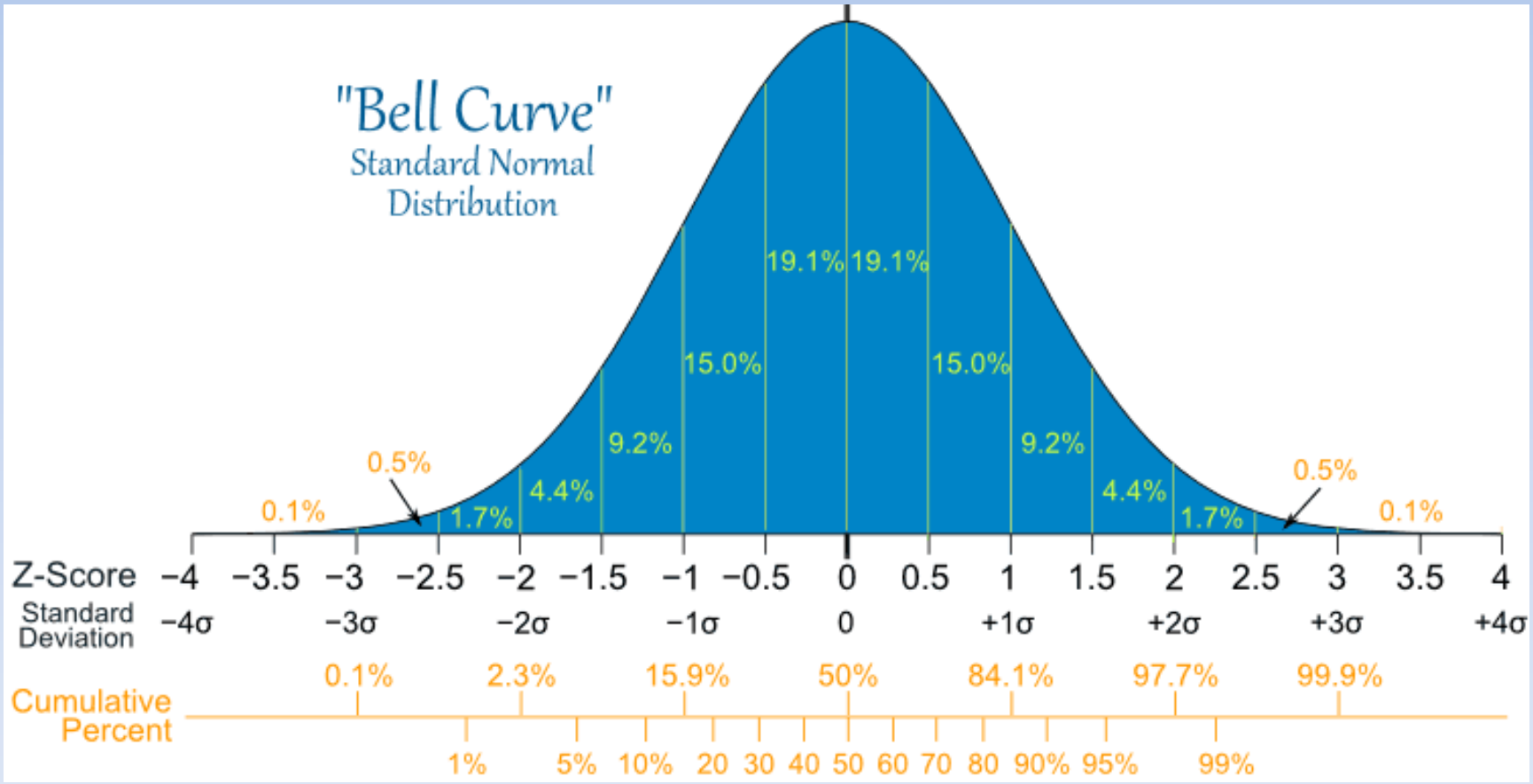
The normal distribution

The distribution most commonly associated with the bell shape is of course the normal (or Gaussian) distribution.

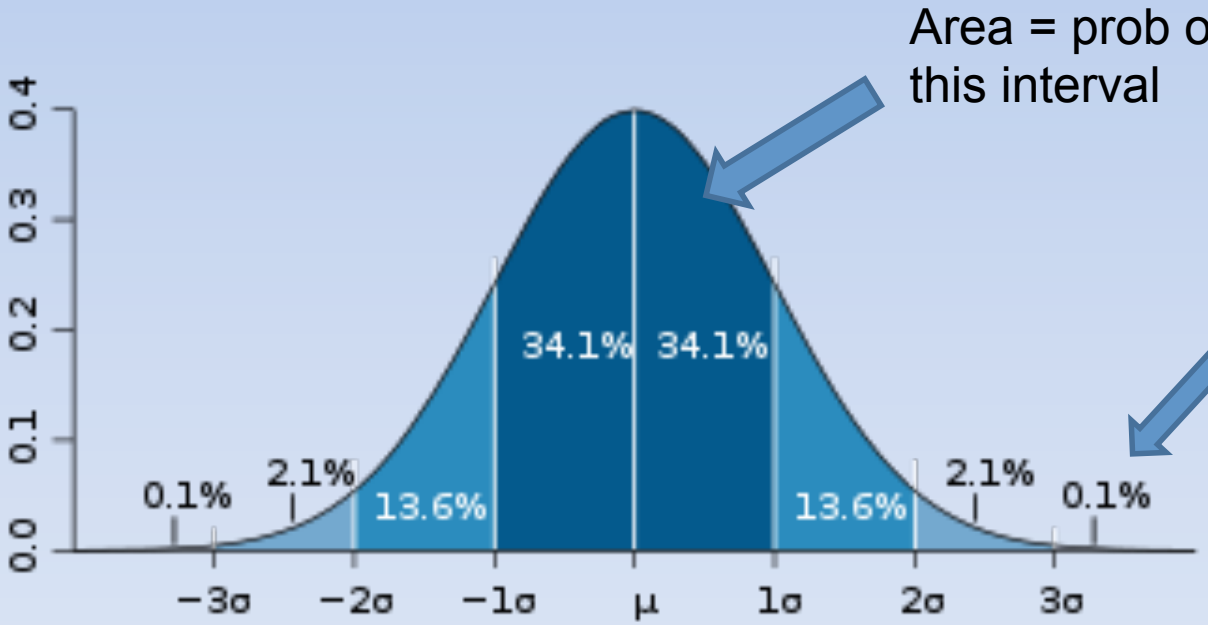
The normal distribution is very commonly encountered in statistics – for the reasons just given AND because it is mathematically tractable in many theoretical calculations.



$$f(x) = C \cdot e^{-x^2/2}$$


"Bell Curve"
Standard Normal Distribution
19.1%
19.1%
15.0%
15.0%
9.2%
9.2%
4.4%
4.4%
1.7%
1.7%
0.5%
0.5%
0.1%
0.1%
Z-Score
−4 −3.5 −3 −2.5 −2 −1.5 −1 −0.5 0 0.5 1 1.5 2 2.5 3 3.5 4
Standard Deviation
−4σ −3σ −2σ −1σ 0 +1σ +2σ +3σ +4σ
Cumulative Percent
0.1% 2.3% 15.9% 50% 84.1% 97.7% 99.9%
1% 5% 10% 20 30 40 50 60 70 80 90% 95% 99%

The general shape of the normal



Area = prob of falling within this interval

The prop of observations larger than this point

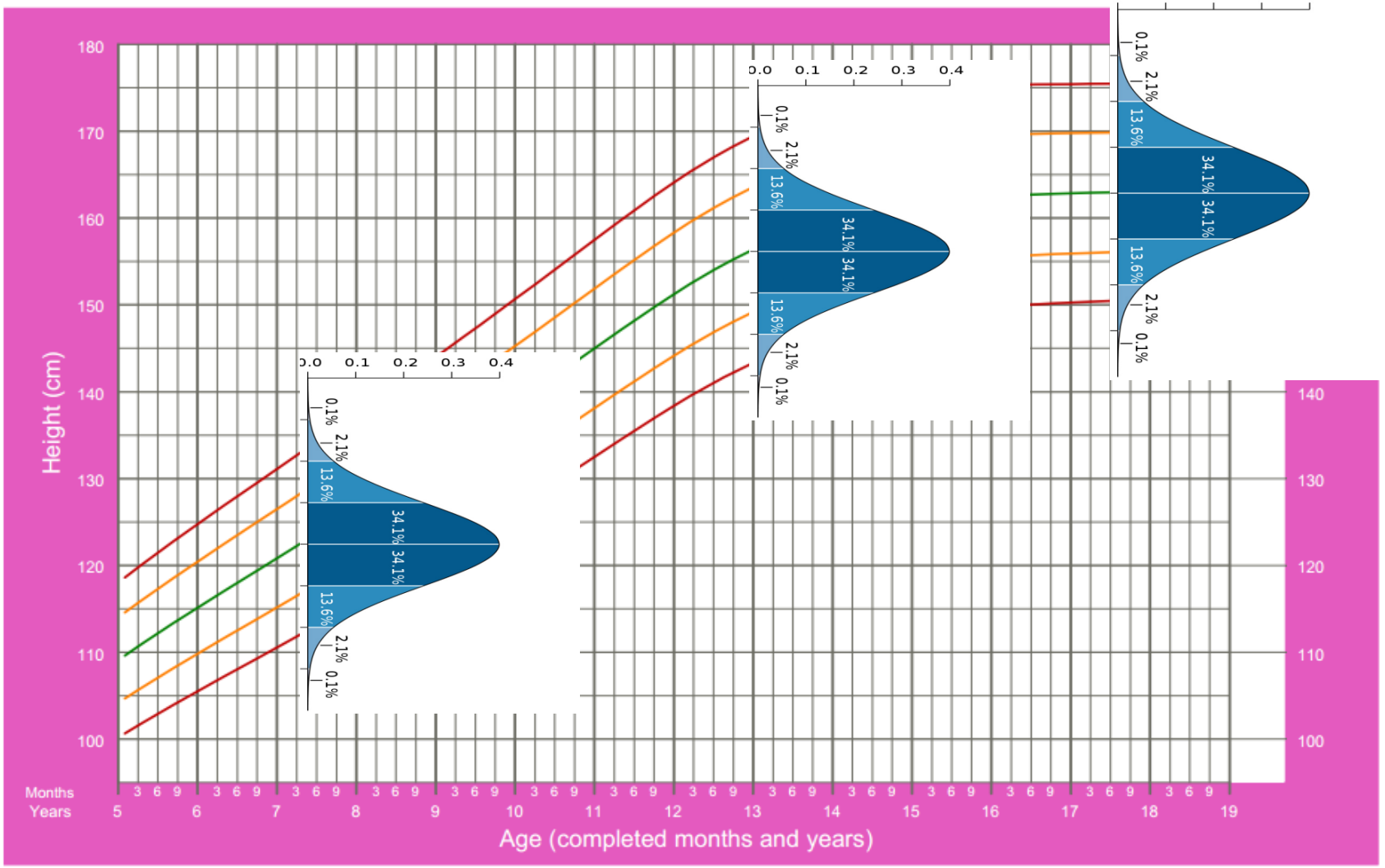
Standard deviation (SD, sd)

Mean = median = mode

Height-for-age GIRLS

5 to 19 years (percentiles)

World Health Organization



2007 WHO Reference

Inference

It has been stated before and it should be stated again that the purpose of statistics (and in particular statistical inference) is to find properties of the underlying population and not the observations at hand.

Of course, in most everyday situations (and some not so everyday) we are mostly concerned with and must respond to the observations at hand!

Inference

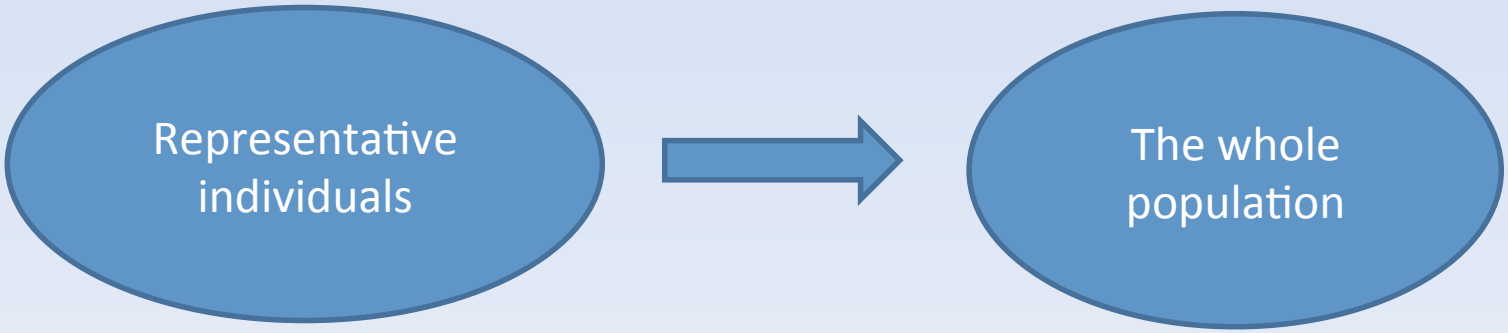
In the empirical sciences including biology the sample at hand is only a tool to learn properties of the underlying population.

In medical trials, the patients that are included in the study are not only irrelevant for us, they are anonymous. Any unexpected findings (rare fatal disease etc) cannot even be communicated to the patient.

Inference

Inference means concluding something for the population as a whole, based on a collection of samples from the population.

Inference is always a stretch – from known individuals to unknown individuals. It will always be based on some assumptions (that may be wrong).

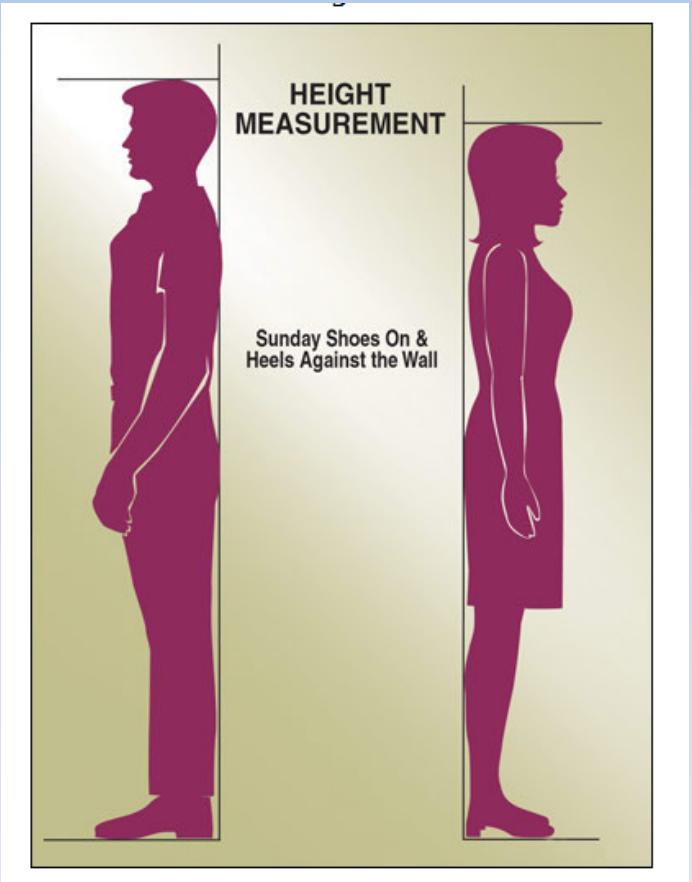


Example

Suppose the height of 16 year old boys with Diabetes I follows a normal distribution with unknown mean and known standard deviation.

A normal distribution seems quite reasonable in this case.

We measure the heights of 20 boys in the population of 16 year olds with Diabetes I.



Example cont.

It seems reasonable to calculate average height.

In fact it can be shown that

average height $\longrightarrow$ true mean height in pop

as the number of observations increases.

The average is the maximum likelihood estimator (MLE) for the mean of a normal distribution.

Maximum likelihood

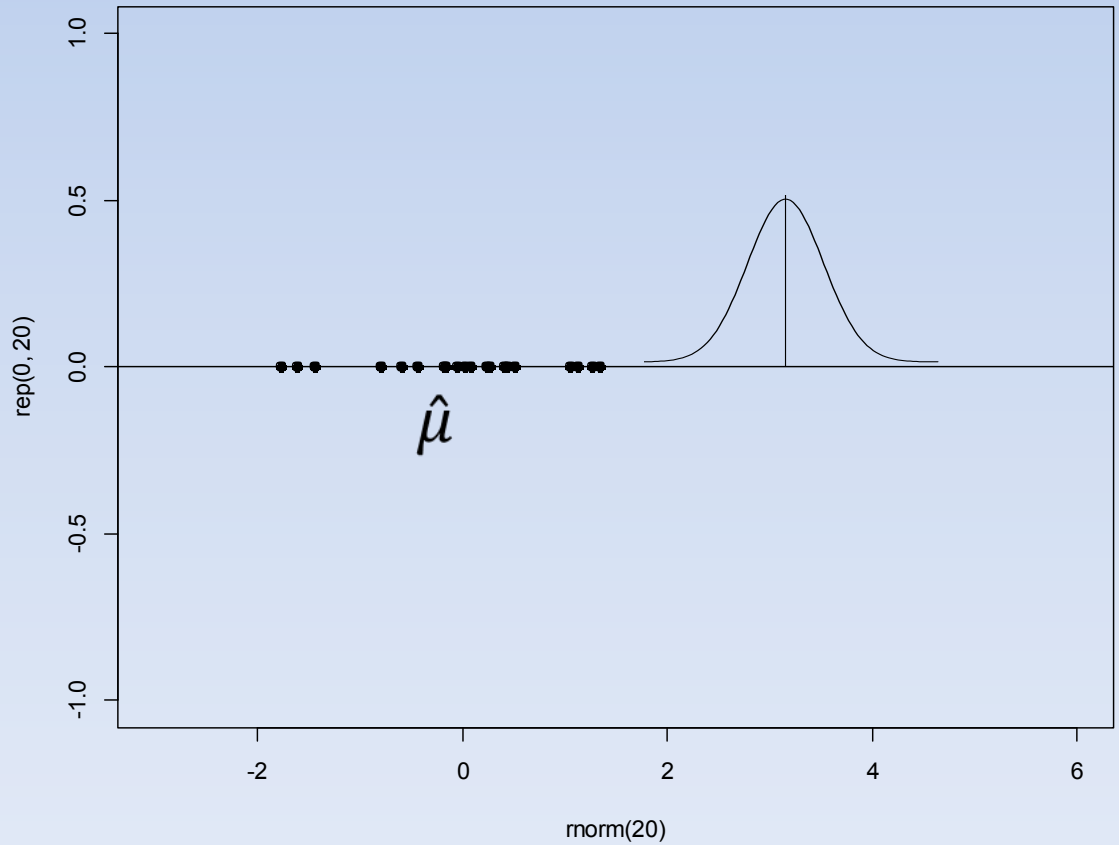
The maximum likelihood method is based on this principle:

We assume that the heights of individuals in the population follow the standard normal distribution N(mu,sigma^2).

We do not know mu and sigma, but we can try different choices and thus create different “scenarios”.

The observations we do are more likely under some scenarios than under others. We select the most likely scenario = the one that leads to the maximal probability of observing what we actually have observed.

Maksimum likelihood "by eye"



Hypothesis testing

Hypothesis testing means selecting between two stated hypotheses:

H0 (the null hyp): first scenario

H1 (the alternative hyp): second scenario

Which of the two are most compatible with what we have actually observed?

Hypothesis testing

Hypothesis testing involves a compromise:

• To avoid concluding H1 by mistake, we should have solid evidence that H0 is false

• But if we require too strong evidence to leave the null hypothesis H0, we will never be able to discover that H1 is true.

Type I error means concluding H1 by mistake.
Type II error means concluding H0 by mistake.

Example

H0: The new drug Fluxomab has the same effect as the old drug already on the market

H1: The new drug has a better effect that the old

Type I error: We conclude that Fluxomab is better when in fact the two drugs are equally effective

Type II error: We conclude that the two drugs are equally effective when in fact the new one is better

Kompromisset

Prøver vi å redusere sjansen for type I-feil, øker vi samtidig sjansen for type II-feil.

Prøver vi å redusere sjansen for type II-feil, øker vi samtidig sjansen for type I-feil.

Vi må gjøre et kompromiss. I hypotese-testing gjør vi det ved å legge et absolutt krav på en av de to typene feil (type I).

Testprosedyre

Vi forlater (eller forkaster) H0 dersom observasjonene er lite forenlige med H0.

I praksis: definere en testobservator T basert på observasjonene, som tar lave verdier når H0 er sann. Vi forkaster H0 dersom $T > a$.

Vi velger terskelen a slik at sannsynligheten for type I-feil er (f.eks.) 5% eller 1%.

Hypotesetesting (oppsummering)

- Definer en observator som har mer ekstreme verdier under H1 enn under H0.

- Beregn verdien til observatoren, og forkast H0 dersom verdien er ekstrem.

Vi kan beregne hvor ekstrem observatorens verdi er ved å finne sannsynligheten for å observere en så ekstrem (eller mer ekstrem) verdi når H0 er sann. Det gir **p-verdien** til testen.