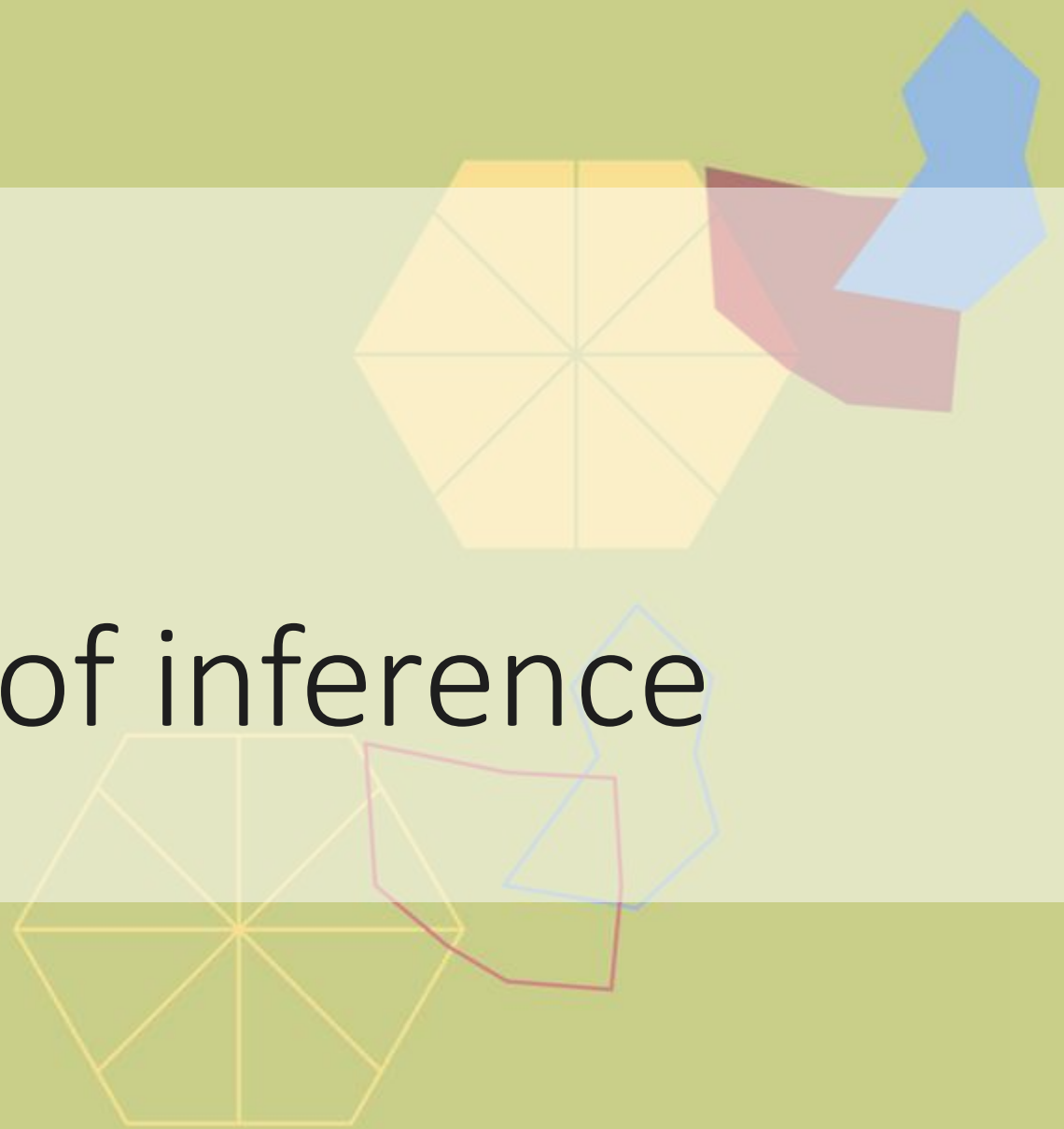
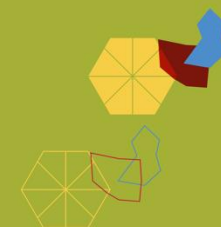


Foundations of inference

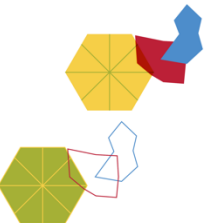


Probability



Why do we need probability?

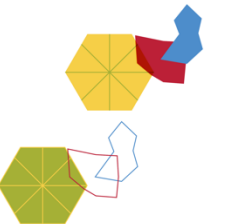
- Probability theory offers a systematic and organised way to describe and quantify uncertainty
- This helps us understand how likely events are to occur
- It can help with decision making
- In science, importantly, it helps us weigh evidence in order to understand the systems we study
- Human intuition is surprisingly terrible when it comes to probabilities...



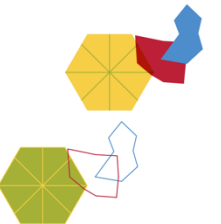
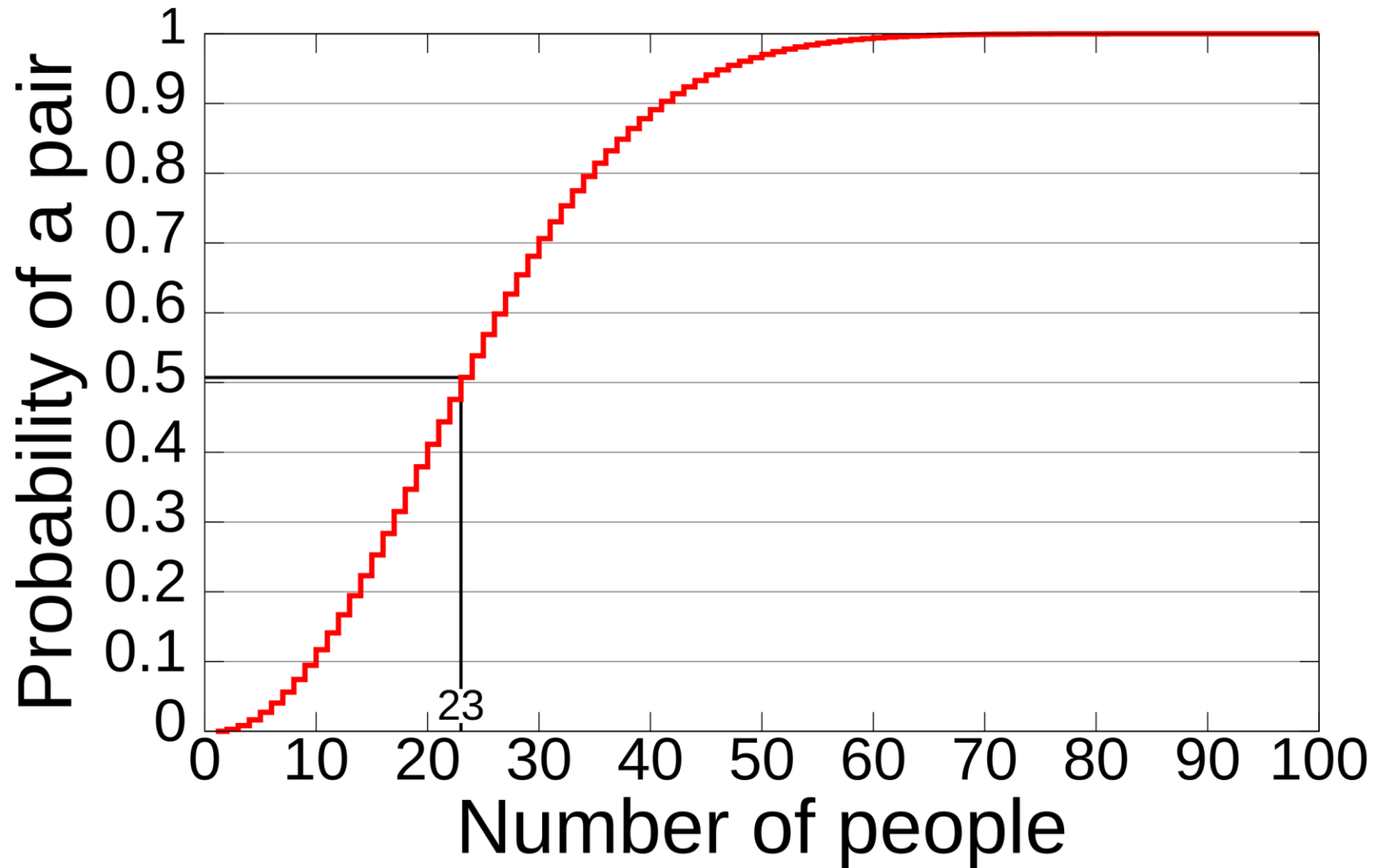
The birthday problem

How many people do we need to have in a room to have a 50% probability that two of them share a birthday?

- A) 5 – 10
- B) 10 – 30
- C) 30 – 100
- D) 100 – 200



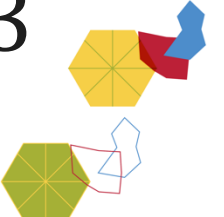
The birthday problem



The birthday problem

- The reason why this is counterintuitive is because people don't realise just how many different pairs there are for 23 people

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

$$C_2^{23} = \binom{23}{2} = \frac{23!}{2! (23 - 2)!} = \frac{23 \times 22 \times 21!}{2! \times 21!} = 253$$
A collection of small, colorful geometric shapes including a yellow hexagon, a red pentagon, a blue hexagon, a green hexagon, and a white hexagon with a red outline, arranged in a cluster at the bottom right of the slide.

The birthday problem

- With 253 possible pairs, and 365 (or 366) days in a year, getting through the 50% threshold is more obvious, even if non intuitive
- And if you think people are not interested in those probabilities you'd be mistaken!

Message

Hi there,

This is a very unimportant, jovial request.

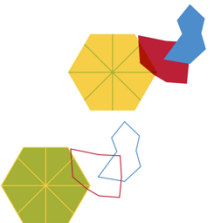
Our research group (of approximately 20 members) has some interesting birthday patterns. Three people share the same birthday, and then there are three separate pairs of people who also share birthdays - each pair on a different day.

We have tried to calculate the probability of this but our answers seem too small (too improbable).

Would you help put our minds to rest?

Many thanks,

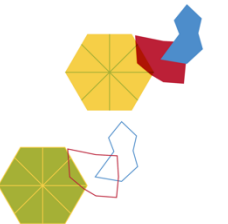
Probability of 0.00003 if anyone is wondering!



The three axioms of probability

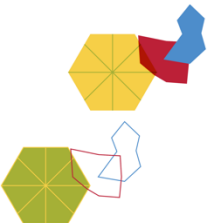
1. The probability of an event A is a value $P(A)$ between 0 and 1
2. $P(\Omega)=1$
3. If A_1, A_2, \dots are mutually exclusive events, then

$$P(A_1 \cup A_2 \cup \dots) = P(A_1) + P(A_2) + \dots$$



Random variables

- Variables whose possible values are numerical outcomes of a random phenomenon
- Subject to variation due to chance
- A probability is associated to each of the possible values the variable can take
- There are two types of random variables:
 - Discrete
 - Continuous



Discrete random variables

- Can take only a number of distinct values
- Have a probability mass function
 - Assigns probabilities to the possible values of the random variable $P(X=x)$, where X is the random variable and x the possible value

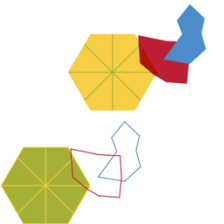
❗ The probabilities must follow some requirements:

$$p_i \geq 0 \forall i$$

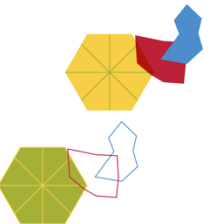
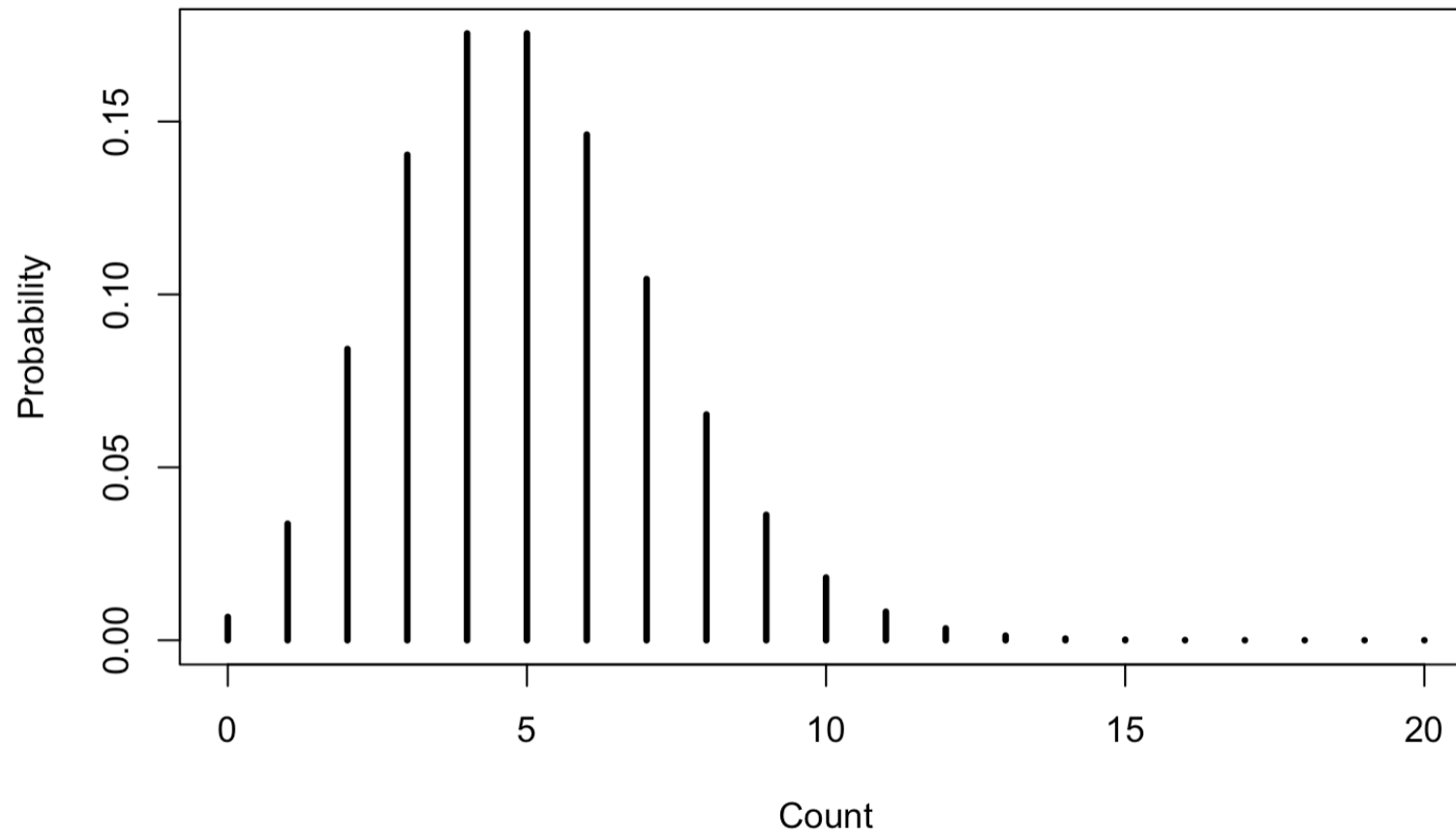
Probabilities must be positive numbers

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

The sum of the probabilities must be 1



Example: Poisson distribution



Example: Poisson distribution

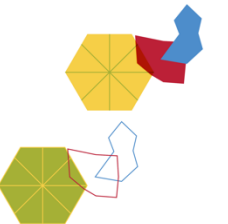
The probability mass function for the Poisson distribution is:

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

where:

x is the number of events

λ is the rate (expected number of events in an interval)

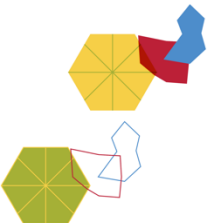


Example: Poisson distribution

- Say we have a production line, which makes 600 units an hour and on average 5 of them are faulty
- If we want to know the probability of finding 0 faulty parts in 15min, we can use the Poisson distribution
- In this case, our rate λ which is originally 5 parts an hour needs to be scaled to a quarter of an hour:

$$\lambda = 5 \times \frac{15}{60} = 1.25$$

$$P(X = 0) = \frac{1.25^0 e^{-1.25}}{0!} = 0.287$$



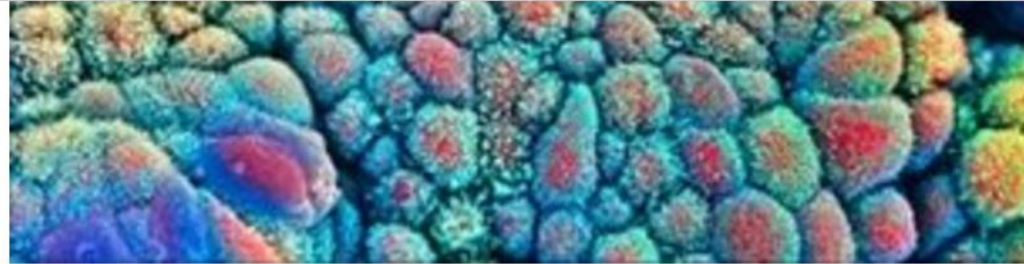
NEWS

[Home](#) | [Israel-Gaza war](#) | [Cost of Living](#) | [War in Ukraine](#) | [Climate](#) | [UK](#) | [World](#) | [Business](#) | [Politics](#) | [Culture](#)

[Health](#)

'Three-fold variation' in UK bowel cancer death rates

© 12 September 2011



Bowel cancer is the second biggest cause of cancer deaths in the UK after lung cancer

By Dominic Hughes

Health correspondent, BBC News

There is a big variation across the UK in the number of people who die from bowel cancer, figures show.

The death rate is lowest in the town of Rossendale, Lancashire, at nine in 100,000 people, while the highest is found in Glasgow, at 31 in 100,000.

Beating Bowel Cancer researchers say taking part in screening, awareness of symptoms and unhealthy diets probably all play a role in the variation.

The disease is the UK's second most common cause of cancer death.

The charity Beating Bowel Cancer said that its research took into account the number of elderly people living in a particular area as the risk of bowel cancer increases with age.

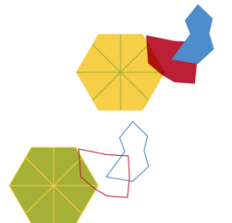
The average death rate from bowel cancer across the UK is 17.6 per 100,000.

Mark Flannagan, chief executive of Beating Bowel Cancer, said that no matter where people lived, too many people were dying from bowel cancer.

"Deaths from bowel cancer could, and should, be much less common," he said.

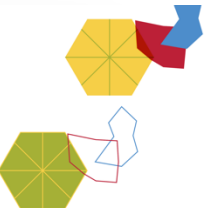
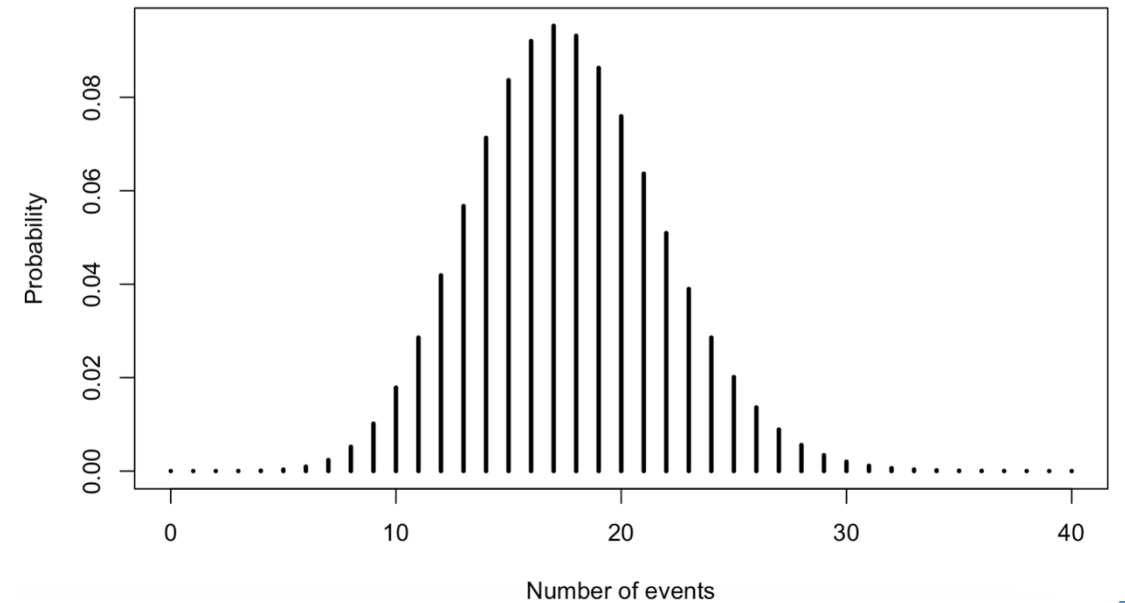
"Early diagnosis is key. People can give themselves a life-saving chance by being aware of bowel cancer symptoms and taking part in bowel cancer screening when it is offered to them.

"The figures are intriguing. It will be extremely important for local NHS organisations to examine information for their own areas and use it to inform potential changes in delivery of services."



Bowel cancer statistics

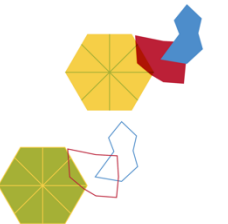
- Can we quantify how unlikely the values reported in this article are?
- What we know:
 - The average death rate per 100,000 is 17.6
 - Rossendale has a rate of 9 per 100,000
 - Glasgow has a rate of 31 per 100,000



Poisson distribution to the rescue

- First of all, let us acknowledge that some of the assumptions are unrealistic
 - For example, Rossendale and Glasgow are not comparable
- Let us work out the probability for Rossendale:

$$P(X = k) = \frac{\lambda^k e^{-\lambda}}{k!} = \frac{17.6^9 e^{-17.6}}{9!} = 0.01015$$

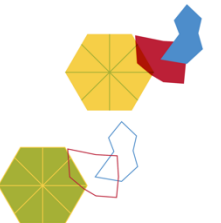


Poisson distribution to the rescue

- Now, let us work out the probability for Glasgow:

$$P(X = k) = \frac{\lambda^k e^{-\lambda}}{k!} = \frac{17.6^{31} e^{-17.6}}{31!} = 0.00113$$

- So it seems the Glasgow observation ($p=0.00113$) is less likely than the Rossendale ($p=0.01015$)
- What could be happening there?



Glasgow effect

🌐 9 languages ▾

Article [Talk](#)

Read [Edit](#) [View history](#) [Tools](#) ▾

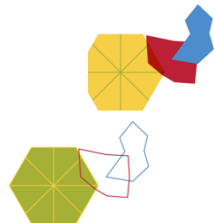
From Wikipedia, the free encyclopedia

The **Glasgow effect** refers to the lower life expectancy of residents of [Glasgow](#) compared to the rest of the United Kingdom and Europe.^{[1][2]} The phenomenon is defined as an "[e]xcess mortality in the West of Scotland (Glasgow) after controlling for deprivation."^[3] Although lower income levels are generally associated with poor health and a shorter lifespan, [epidemiologists](#) have argued that poverty alone does not appear to account for the disparity found in Glasgow.^{[2][4][5][6][7][8]} Equally deprived areas of the UK such as [Liverpool](#) and [Manchester](#) have higher life expectancies, and the wealthiest ten percent of the Glasgow population have a lower life expectancy than the same group in other cities.^[9] One in four men in Glasgow will die before his sixty-fifth birthday.^[10]

Several hypotheses have been proposed to account for the ill health, including the practice in the 1960s and 1970s of offering young, skilled workers in Glasgow social housing in [new towns](#), leaving behind a demographically "unbalanced population".^{[11][12]} Other suggested factors have included a high prevalence of premature and low birthweight births, land contaminated by toxins, a high level of derelict land, more deindustrialisation than in comparable cities, poor social housing, religious [sectarianism](#), lack of [social mobility](#),^[13] [vitamin D deficiency](#), cold winters, higher levels of poverty than the figures suggest, [adverse childhood experiences](#) and childhood stress, high levels of stress in general, and social alienation.^[14]



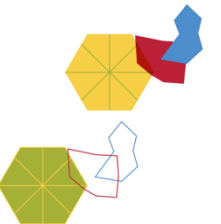
[Buchanan Street](#), one of the main shopping areas in [Glasgow city centre](#)



Negative binomial distribution

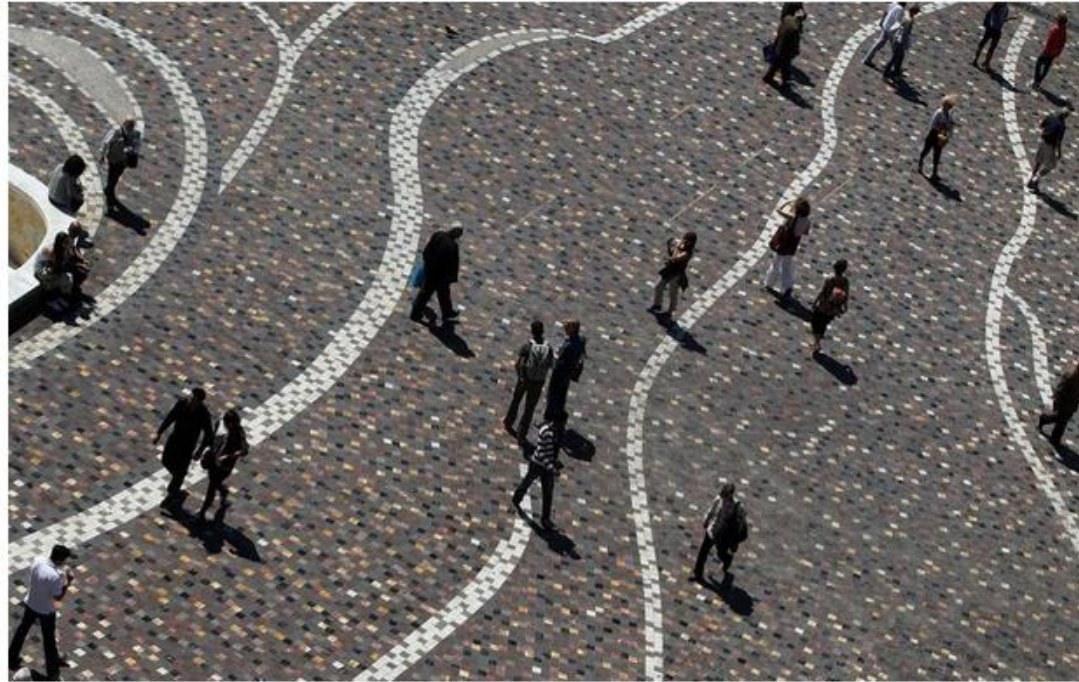
- The negative binomial is a slightly different setup to the Poisson
- In this case we are thinking of successful encounters versus unsuccessful ones
- Consider k as the successes with probability $1-p$, r as the failures with probability p , then the probability of exactly k successes is:

$$f(k; r, p) \equiv \Pr(X = k) = \binom{k + r - 1}{r - 1} (1 - p)^k p^r$$



NEWS

Giorgos and Maria most popular names in Greece

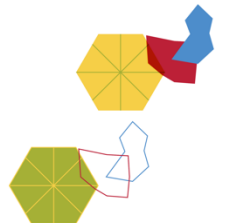


■ Newsroom

21.05.2019 • 14:52



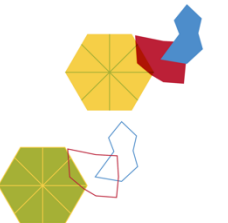
Giorgos and Maria, after the dragon-slaying saint and the mother of Christ, are the most popular names in Greece, representing 8.8 and 8.3 percent, respectively, of the country's male and female populations, data released on Tuesday by ELSTAT showed.



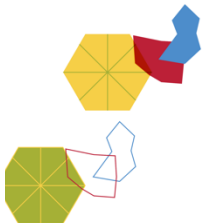
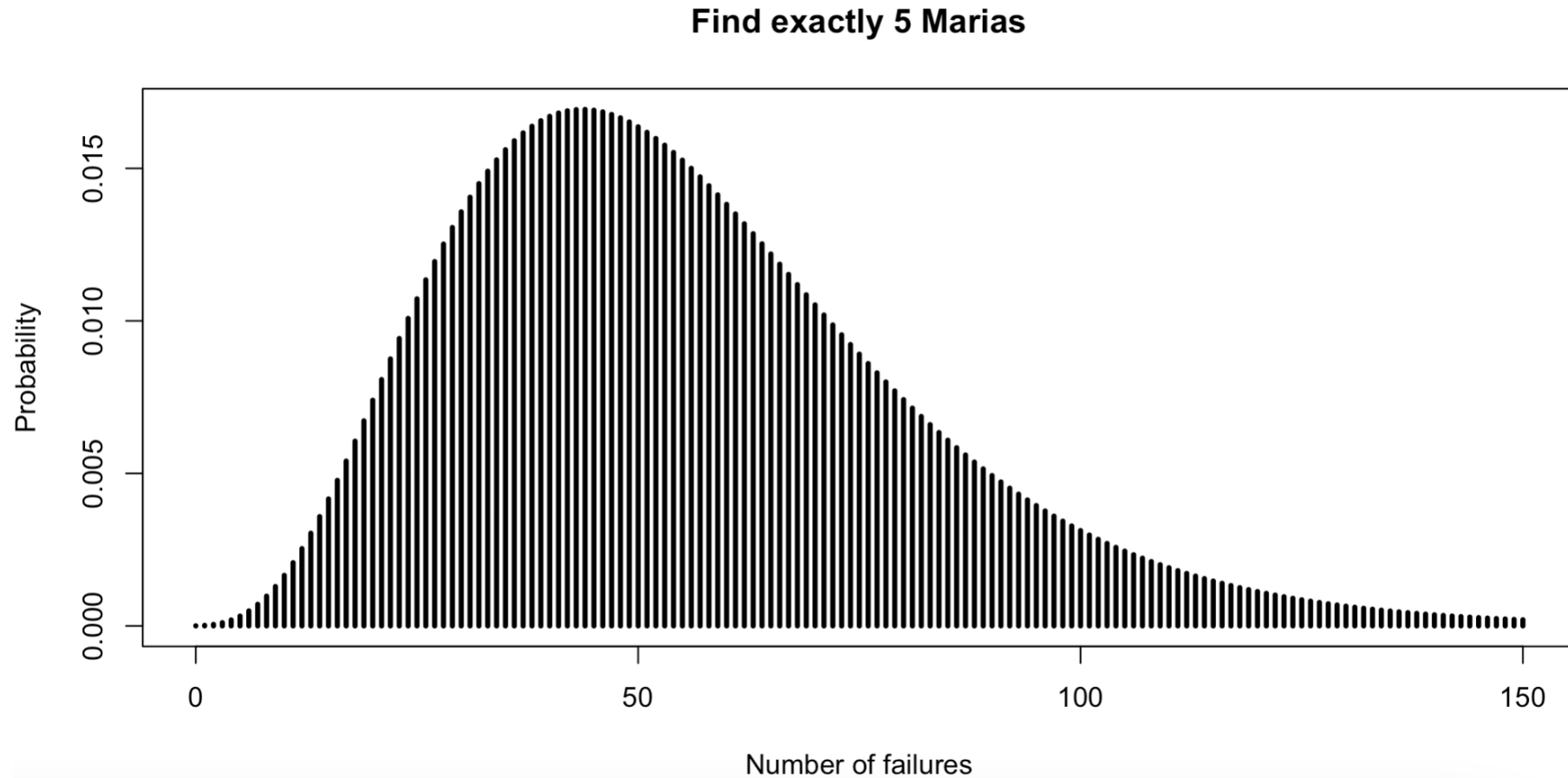
How do you solve a problem like Maria?

- Let us say we're standing in Syntagma square and stopping people and asking them if their name is Maria
- If we stop 20 people, what is the probability of encountering exactly 5 Marias

$$f(5; 20, 0.083) = P(X = 5) = \binom{5 + 20 - 1}{20 - 1} (1 - 0.083)^5 0.083^{20} \\ = 0.074$$



Example: Negative binomial distribution



Continuous random variables

- Can take an infinite number of possible values
- Have a probability density function
 - specifies the probability that the value of the random variable falls within a specific range
 - it is represented by the area under the density function (integral)

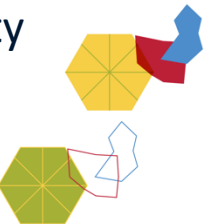
! The PDF needs to satisfy the following requirements:

Probabilities must be positive numbers

$$f(x) \geq 0 \quad \forall x$$

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

The area under the entire density curve must be 1



Continuous random variables

❗ The PDF needs to satisfy the following requirements:

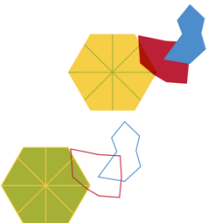
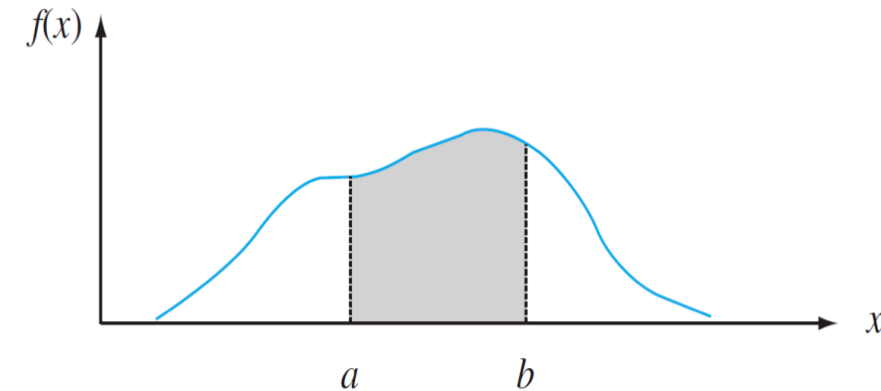
Probabilities must be positive numbers

$$f(x) \geq 0 \quad \forall x$$

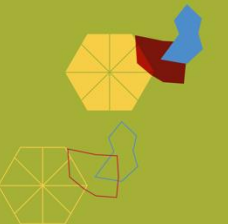
$$\int_{-\infty}^{+\infty} f(x) dx = 1$$

The area under the entire density curve must be 1

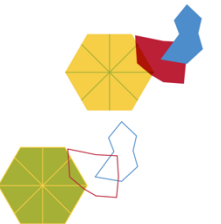
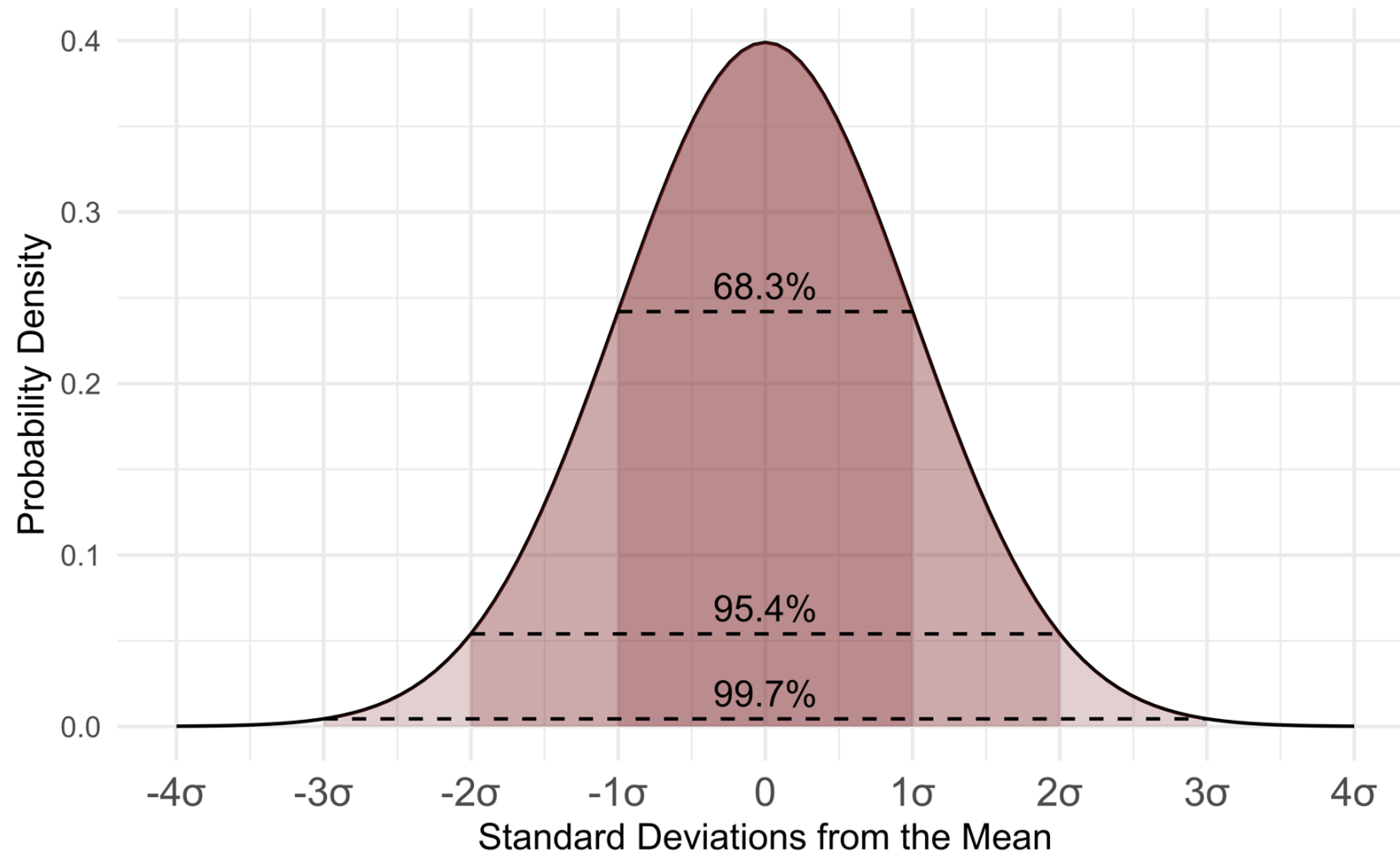
- For example, considering a random variable X measuring the depth of a lake in various spots
- The probability that X takes on a value in the interval $P(a \leq X \leq b)$ is the area under the density function curve
- The probability that X takes an exact value of x is 0



Don't be normal, it's a trap...



The normal distribution



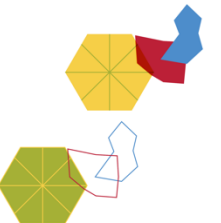
What is the normal distribution?

- A continuous probability distribution that is symmetric around the mean
- You may recognise it as the bell curve
- It needs two parameters to be pinpointed:
 - the mean (μ) and
 - the standard deviation (σ).
- It has some useful properties:
 - Symmetry about the mean
 - Mean = Median = Mode
- And because it's a probability distribution: **Area under the curve = 1**



Why is it important in biology?

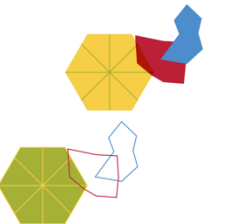
- The common reason you will hear about this is: “Many biological variables (e.g., height, weight, blood pressure, enzyme activity) follow a normal distribution or approximately follow it”
- That’s *sort of/kind of/-ish/if you squint really hard* true but actually real data don’t quite do this
- So why do we keep bothering with it?...



Why do we focus on it so much?

Many reasons, some better than others:

- It's intuitive (*okay reason*)
- It has a useful rule of thumb (also known as **the empirical rule**) that states that:
 - 68%, 95%, and 99.7% of the data are within 1, 2, and 3 standard deviations from the mean (*good reason if you know what you're doing*)
- People are lazy (*baaaaaaad reason*)
- The central limit theorem (**YES!**)



You're on a boat

- Let us say you're on a boat and you want to measure the depth of a body of water
- You use an echo sounder to send a sound pulse down from the boat's hull and you measure the time it takes for the sound to bounce back
- If you do that for a bit, it can give you a collection of measurements
- However...a lot of things can go wrong in the process
 - Inaccuracy because the boat keeps moving
 - Reflections
 - Calibration
 - etc...

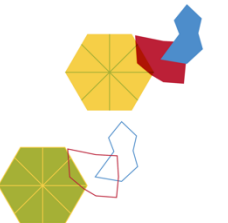


You're still on a boat...

- Errors are unavoidable, but importantly:

ERRORS ARE USUALLY THE SUM OF SEVERAL FACTORS

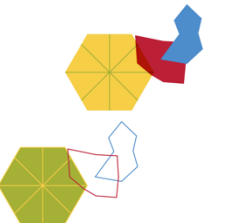
- Helpfully, some of the behaviour of variables which are the sum of several other variables is described by the Central Limit Theorem



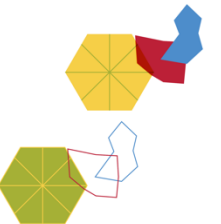
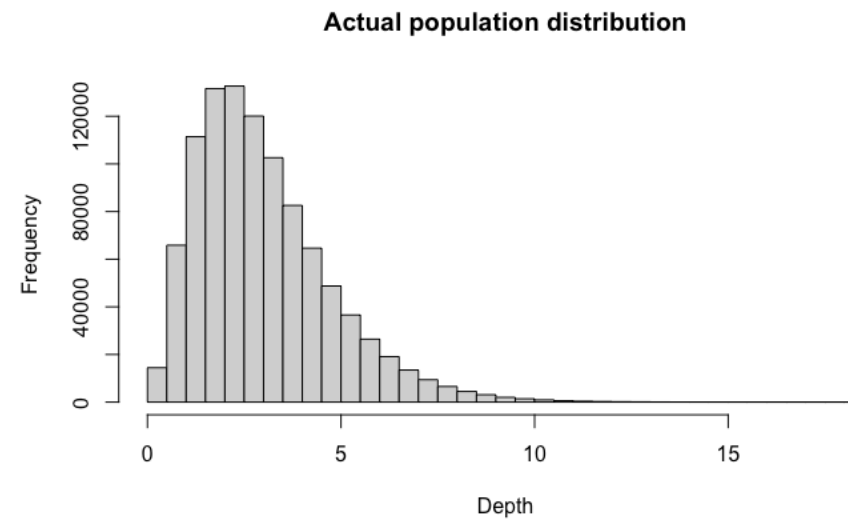
The central limit theorem

Central Limit Theorem:

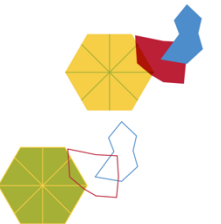
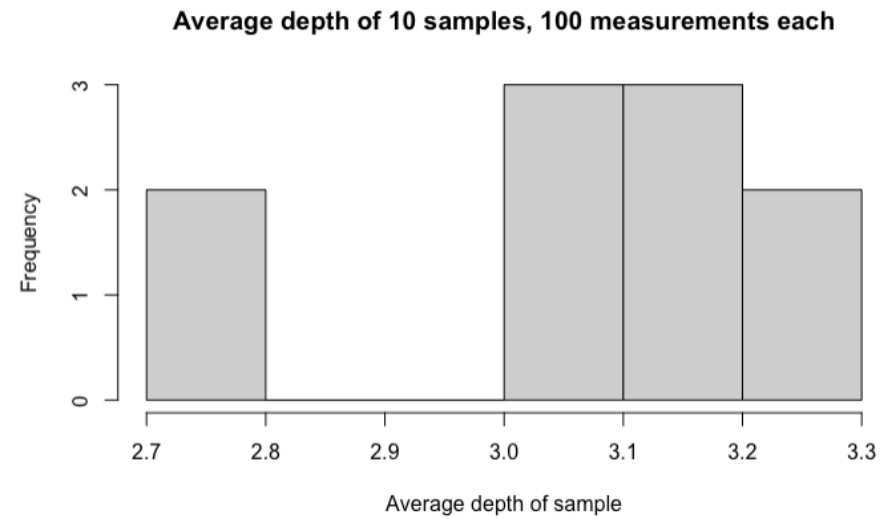
For a large enough sample size, the distribution of the **sample means** (or averages) of independent, identically distributed (i.i.d.) random variables will approximate a normal distribution, regardless of the shape of the original population distribution



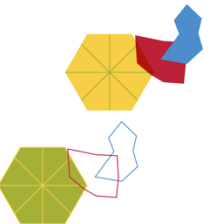
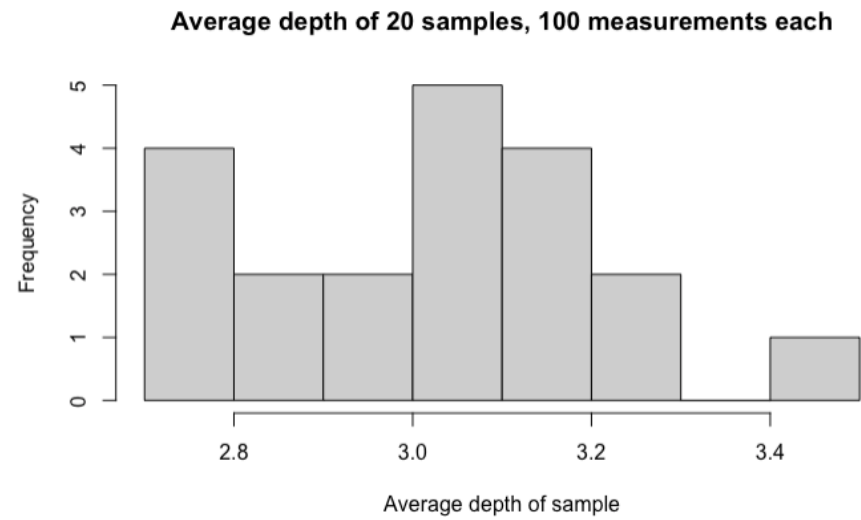
Back on the boat...



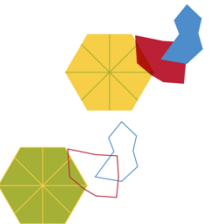
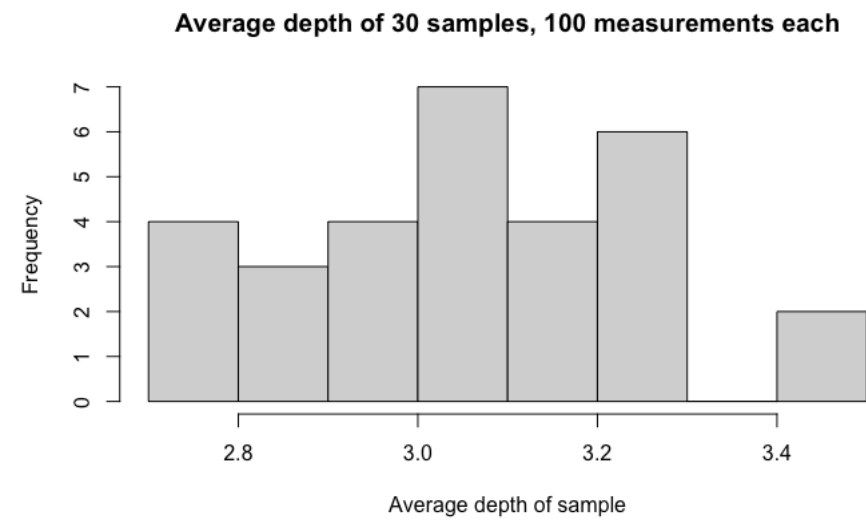
Back on the boat...



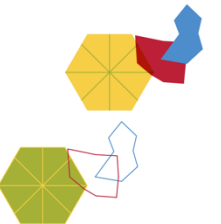
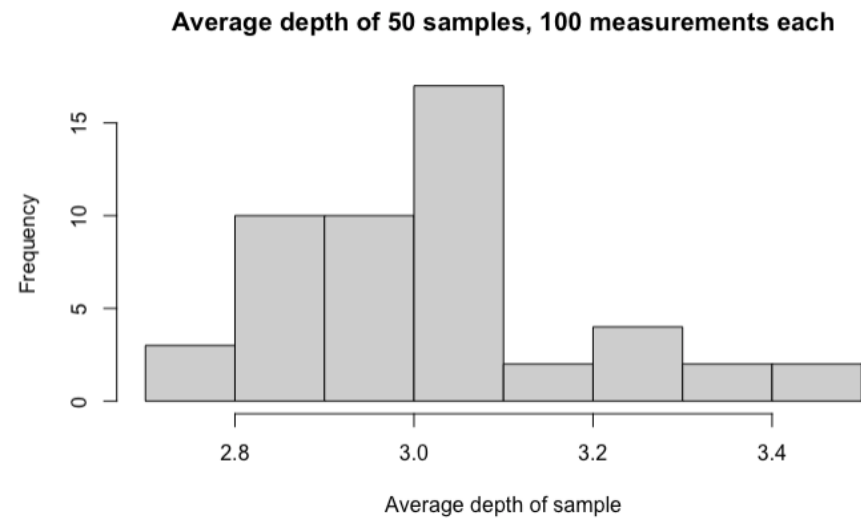
Back on the boat...



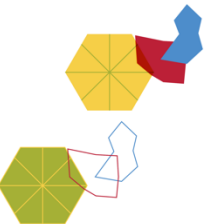
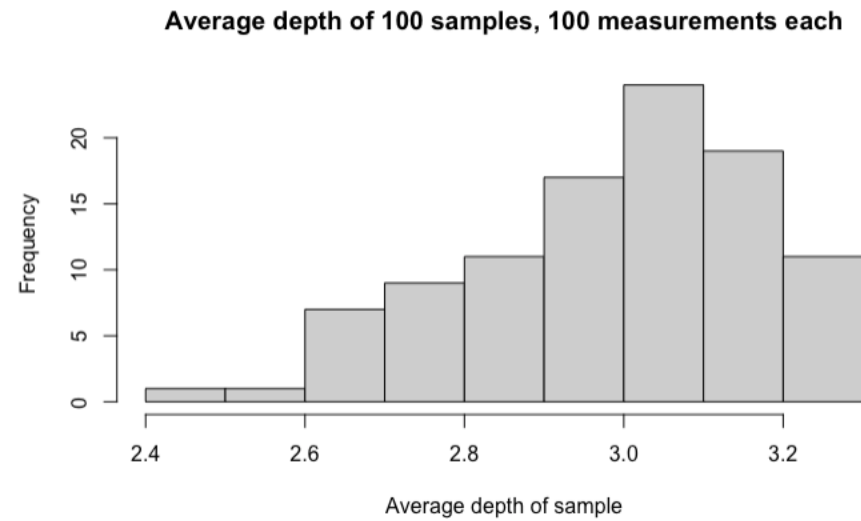
Back on the boat...



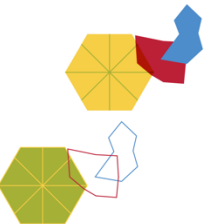
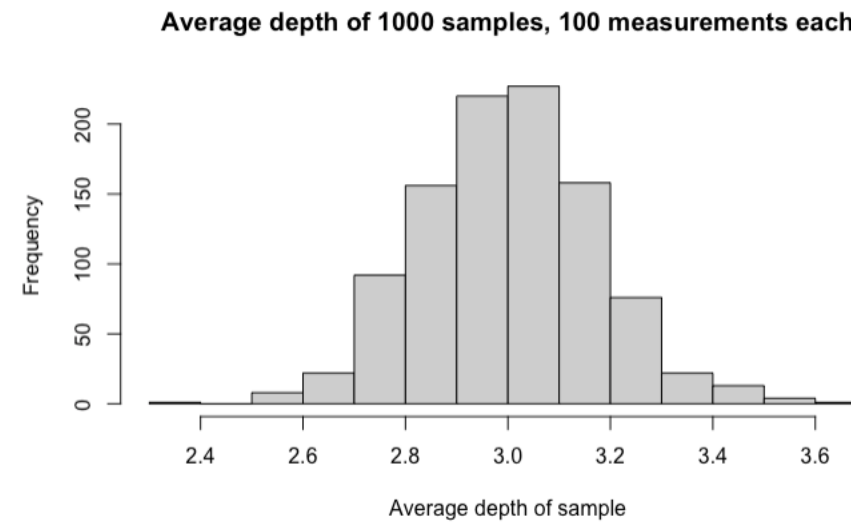
Back on the boat...



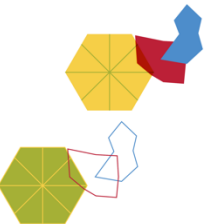
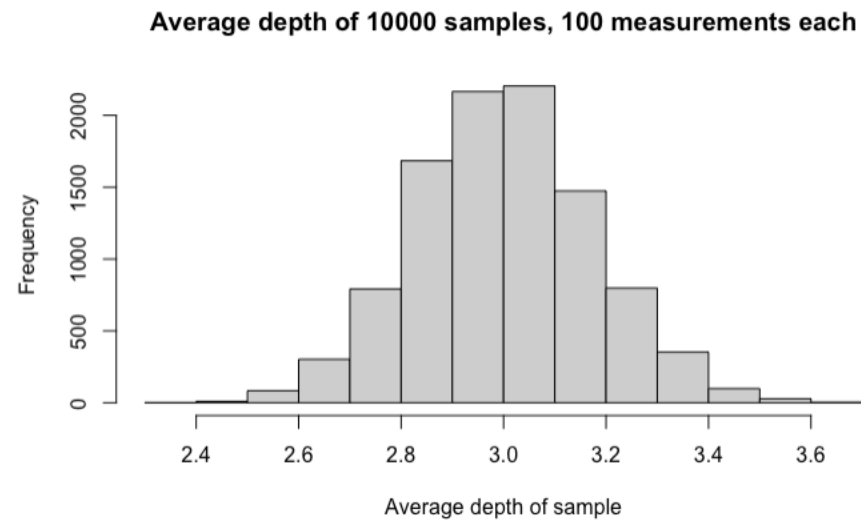
Back on the boat...



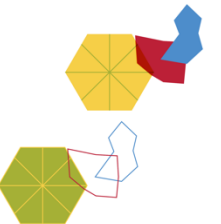
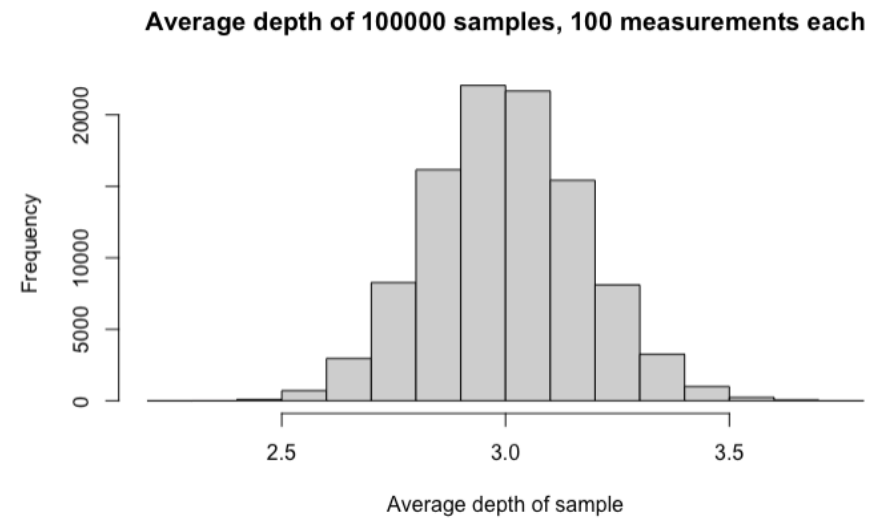
Back on the boat...



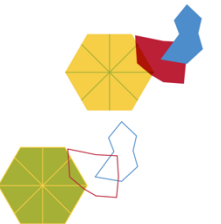
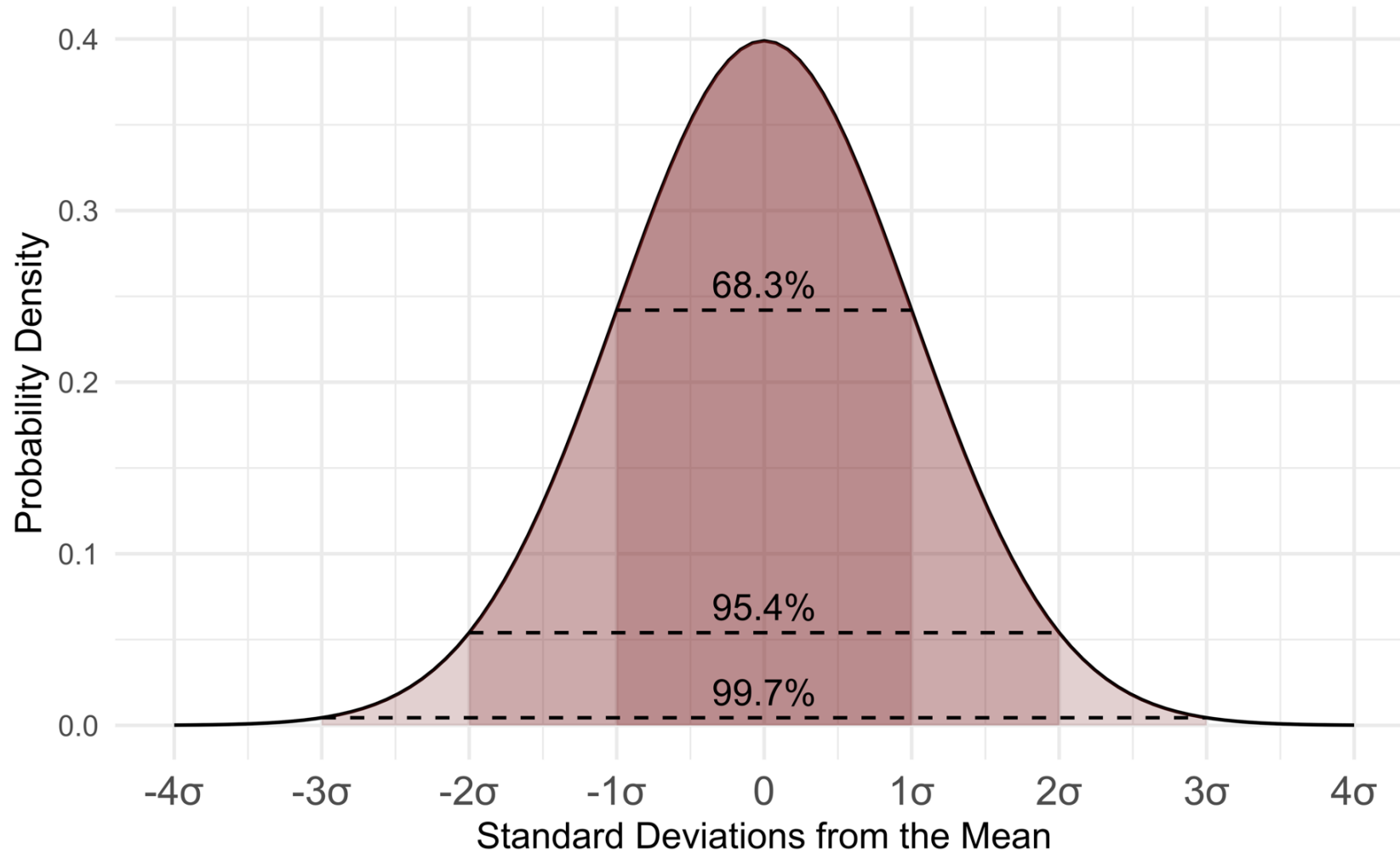
Back on the boat...



Back on the boat...

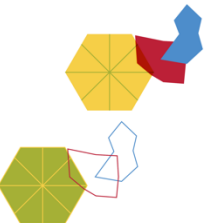


A special kind of normal distribution



The standard normal distribution

- A special case of the normal distribution, with mean 0 and standard deviation 1
- Useful when comparing values from different samples
- This is done by subtracting the mean from each sample and dividing it by the standard deviation creating a score for each sample
- These are known as standard scores or z-scores



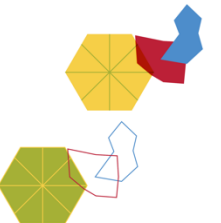
Applications of the normal distribution

Hypothesis Testing:

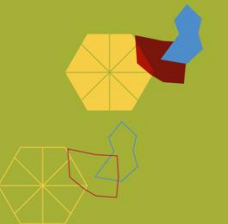
- Many statistical tests (e.g., t-tests, ANOVAs) assume data are normally distributed

Confidence Intervals:

- The normal distribution is central to constructing confidence intervals around population estimates (mean, proportion, etc.)



Summary statistics



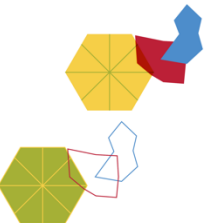
Summary statistics – Describing univariate distributions

Usually we have three types of descriptors

Measures of central
tendency

Measures of variability

Shape of the distribution



Summary statistics – Measures of central tendency

- They give us an indication of the typical score in our sample
- An effective estimate of the middle point in our data distribution
- The most common central tendency measures are:

Mean

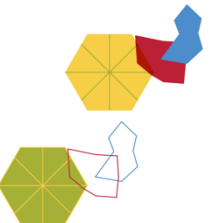
Median

Mode



In a normal distribution: Mean = Median = Mode

In a standard normal distribution: Mean = Median = Mode = 0

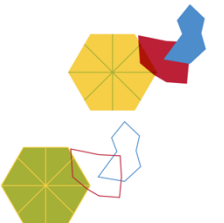


Summary statistics – Central tendency

- Mean: The sum of all the scores in a sample divided by the number of scores in that sample

 The value of the mean is highly influenced by extreme values

- Median: The value which lies in the middle of the sample – it has the same number of scores below and above it
- Mode: The most frequently occurring score in a sample



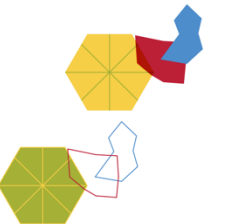
Summary statistics – Central tendency

- An example – Commute distance of 15 students:

ID	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
Dist. (km)	2.5	2.0	1.0	2.5	5.0	0.5	0.5	4.0	5.0	1.0	0.5	4.0	3.0	0.5	4.0

- To calculate the mean, we need to add all the entries together and then divide this sum by 15 (the number of students):

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

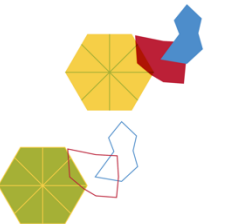


Summary statistics – Central tendency

ID	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
Dist. (km)	2.5	2.0	1.0	2.5	5.0	0.5	0.5	4.0	5.0	1.0	0.5	4.0	3.0	0.5	4.0

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

$$\bar{x} = \frac{1}{15} \times (2.5 + 2.0 + \dots + 4.0) = 2.4$$

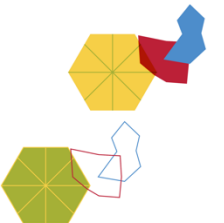


Summary statistics – Central tendency

ID	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
Dist. (km)	2.5	2.0	1.0	2.5	5.0	0.5	0.5	4.0	5.0	1.0	0.5	4.0	3.0	0.5	4.0

- To calculate the median, we need to order all entries from smallest to largest and find the middle value:

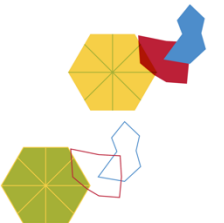
0.5, 0.5, 0.5, 0.5, 1.0, 1.0, 2.0, **2.5**,
2.5, 3.0, 4.0, 4.0, 4.0, 5.0, 5.0



Summary statistics – Central tendency

ID	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
Dist. (km)	2.5	2.0	1.0	2.5	5.0	0.5	0.5	4.0	5.0	1.0	0.5	4.0	3.0	0.5	4.0

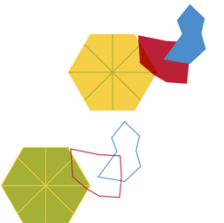
- To calculate the median, we need to order all entries from smallest to largest and find the middle value
- If the number of entries is even, we pick the midpoint between the two middle values



Summary statistics – Central tendency

ID	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
Dist. (km)	2.5	2.0	1.0	2.5	5.0	0.5	0.5	4.0	5.0	1.0	0.5	4.0	3.0	0.5	4.0

- To calculate the mode, we need to find the value that most commonly shows up in our dataset
- **0.5, 0.5, 0.5, 0.5**, 1.0, 1.0, 2.0, 2.5, 2.5, 3.0, 4.0, 4.0, 4.0, 5.0, 5.0
- Mode is 0.5



Summary statistics – Variability

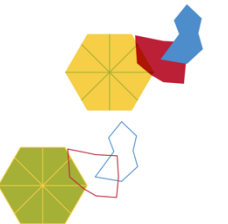
Most common measures of variability are:

Range

Variance

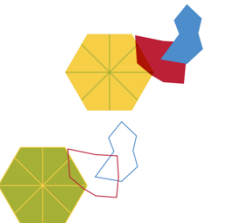
Standard Deviation

Interquartile Range



Summary statistics – Variability

- Range: The lowest and highest values in the distribution
 - ⚠ Just focuses on extreme values
- Variance: The average squared difference of the values from the mean
 - ⚠ Expressed in square units of measurement
- Standard deviation: The degree to which the scores deviate from the mean – the square root of variance
- Interquartile range: The range of the middle 50% of the values in the distribution

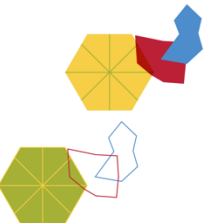


Summary statistics – Variability

ID	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
Dist. (km)	2.5	2.0	1.0	2.5	5.0	0.5	0.5	4.0	5.0	1.0	0.5	4.0	3.0	0.5	4.0

- For the range we need to find the smallest and largest value:

0.5, 0.5, 0.5, 0.5, 1.0, 1.0, 2.0, 2.5, 2.5, 3.0, 4.0, 4.0, 4.0, 5.0, 5.0

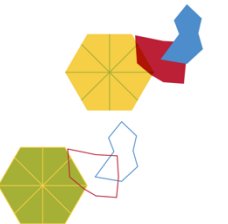


Summary statistics – Variability

ID	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
Dist. (km)	2.5	2.0	1.0	2.5	5.0	0.5	0.5	4.0	5.0	1.0	0.5	4.0	3.0	0.5	4.0

- To calculate the variance, we need to use the following function (remember from earlier: $\bar{x}=2.4$)

$$var = \frac{1}{n-1} \sum_{i=1}^n (x_i - \bar{x})^2 = 2.86$$

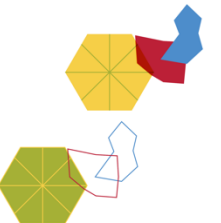


Summary statistics – Variability

ID	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
Dist. (km)	2.5	2.0	1.0	2.5	5.0	0.5	0.5	4.0	5.0	1.0	0.5	4.0	3.0	0.5	4.0

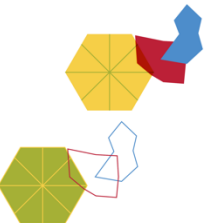
- To get the standard deviation, we need to calculate the square root of the variance:

$$sd = \sqrt{var} = \sqrt{2.86} = 1.69$$



Summary statistics – Variability

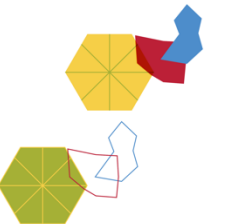
- The interquartile range is slightly more involved
- Start by ordering your values
- Find the median (overall median)
- Find the median of all the values smaller than the overall median (lower median) excluding the minimum
- Find the median of all the values greater than the overall median (upper median) excluding the maximum
- Calculate the difference between upper and lower median



Summary statistics – Variability

- Start by ordering your values

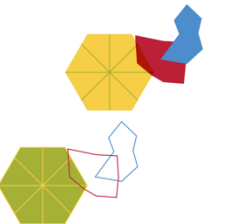
0.5, 0.5, 0.5, 0.5, 1.0, 1.0, 2.0, 2.5, 2.5, 3.0, 4.0, 4.0, 4.0, 5.0, 5.0



Summary statistics – Variability

- Find the median (overall median)

0.5, 0.5, 0.5, 0.5, 1.0, 1.0, 2.0, **2.5**, 2.5, 3.0, 4.0, 4.0, 4.0, 5.0, 5.0

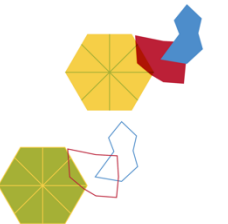


Summary statistics – Variability

- Find the median of all the values smaller than the overall median (lower median) excluding the minimum

0.5, 0.5, 0.5, **0.5**, **1.0**, 1.0, 2.0, **2.5**, 2.5, 3.0, 4.0, 4.0, 4.0, 5.0, 5.0

$$(0.5+1.0)/2=\mathbf{0.75}$$

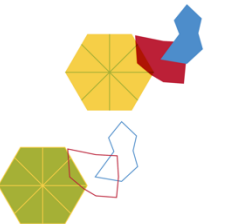


Summary statistics – Variability

- Find the median of all the values greater than the overall median (upper median) excluding the maximum

0.5, 0.5, 0.5, **0.5**, **1.0**, 1.0, 2.0, **2.5**, 2.5, 3.0, **4.0**, **4.0**, 4.0, 5.0, 5.0

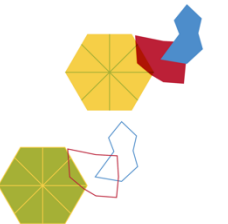
$$(4.0+4.0)/2=\mathbf{4.0}$$



Summary statistics – Variability

- Calculate the difference between upper and lower median

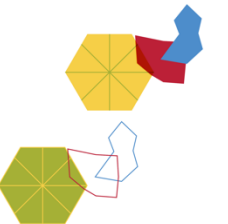
$$4.00 - 0.75 = 3.25$$



Summary statistics – Shape of distribution

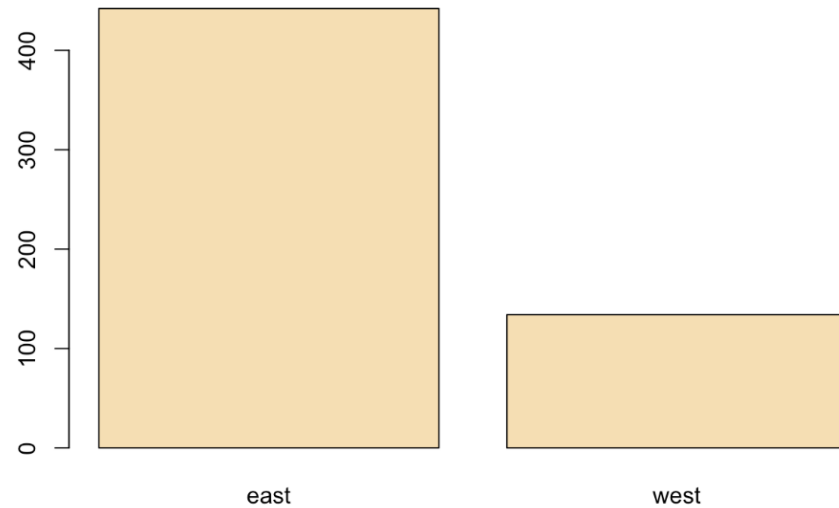
When describing the shape of a distribution we normally discuss:

- Symmetry
- Skew
- Central tendency
- Variability

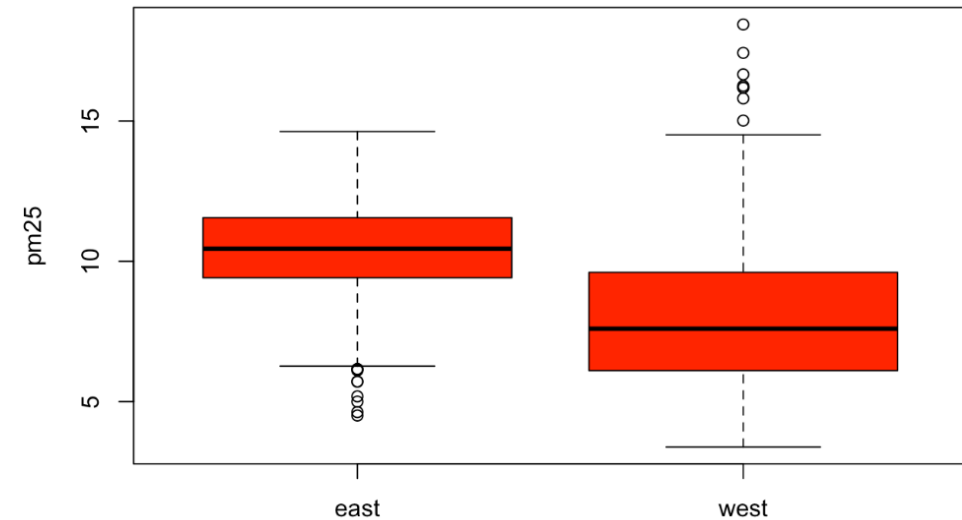


Summary statistics – Shape of distribution

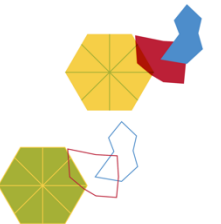
- Categorical/Discrete variables



Distribution of the # of counties per U.S. region



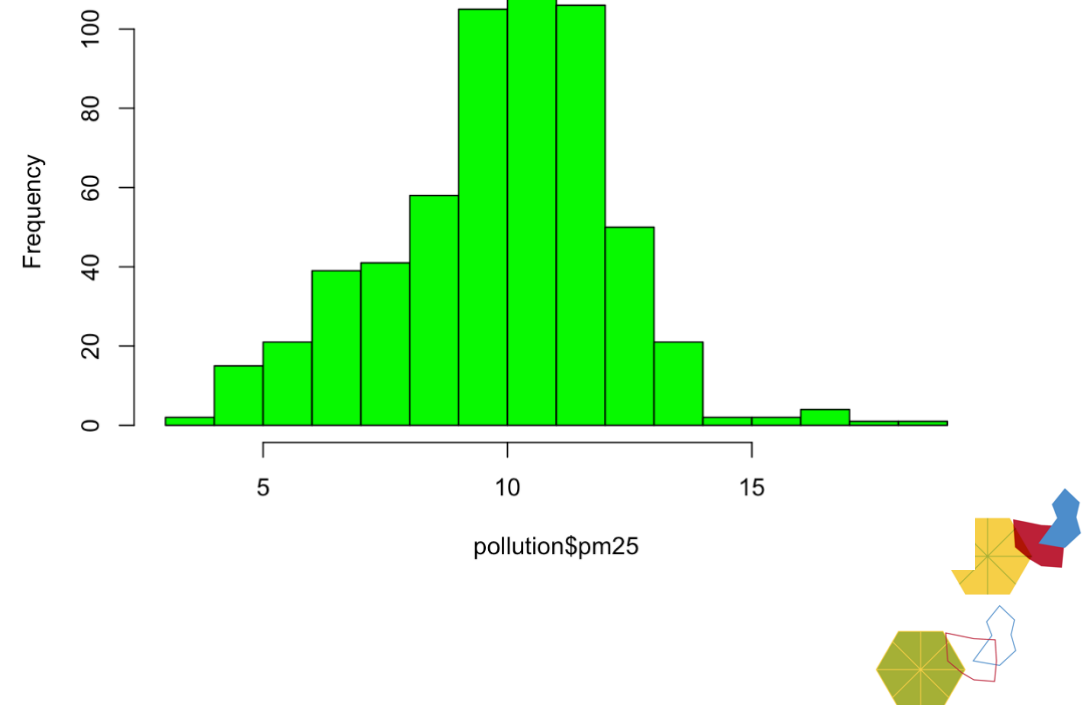
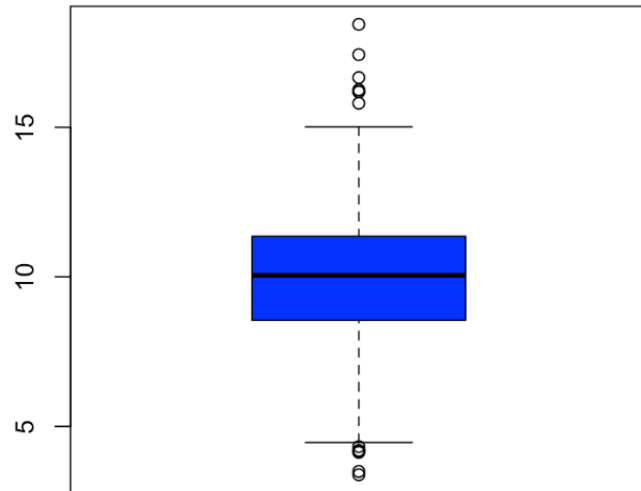
Distribution of PM2.5 per U.S. region



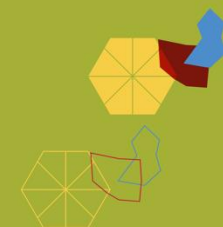
Summary statistics – Shape of distribution

- Continuous variables

Distribution of PM2.5 in U.S. counties

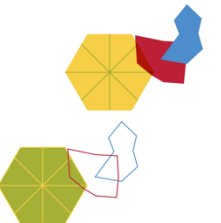


Estimators and bias



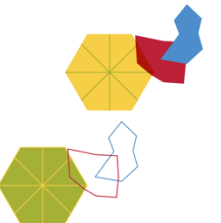
How we go about modelling patterns

- Quite often the purpose of our data collection and modelling steps is to describe the patterns in the population
- In simple terms, the entire point of analysis is to be able to make a general statement about an underlying population using a small snapshot
- The connection between data and population is done through a process of estimation
- How good our estimation is depends on a lot of things
- But first things first...



What is an estimator?

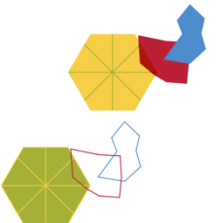
- An ***estimator*** is a rule for calculating the value of interest
 - For example, the process of calculating a sample average
- An ***estimand*** is the true value of interest
 - The true value for the actual unknown population we want to understand
- An ***estimate*** is the outcome of the calculation
 - For example, the result of the sample average calculation



What is an estimator?

We are interested in the human height of a population

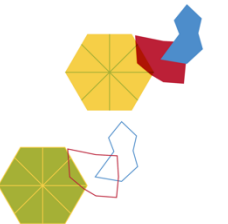
- We take a sample of 20 heights
- We calculate the average of the sample
- The calculation is the ***estimator***
- The result, \bar{x} , is the ***estimate***
- The true average height of the population, μ , is the ***estimand***



Estimand, estimator, estimate




Estimand



Estimand, estimator, estimate




Estimand





Hedgehog cake

★★★★☆ By Valerie Barrett

A celebration cake with a touch of woodland style, this chocolate cake is decorated with edible spikes, buttercream and chocolate details

 **Prep:** 1 hrs 30 mins
Cook: 1 hrs






 Serves 16

 Easy

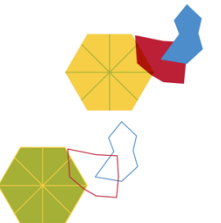
Nutrition per serving

kcal	fat	saturates	carbs	sugars	fibre	protein	salt
478	27.7g	16.9g	53.5g	45.1g	1.6g	4.5g	0.3g

[Add to favourite +](#) [Print](#)

Estimator



Estimand, estimator, estimate



Estimate...!

Marie Barrett

Such of woodland style, this chocolate cake is decorated with edible spikes, buttercream and

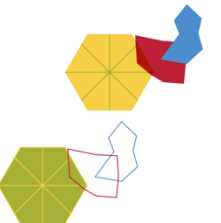
Serves 16

Easy

saturated	carbs	sugars	fibre	protein	salt
16.9g	53.5g	45.1g	1.6g	4.5g	0.3g

+

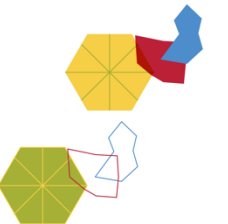
Estimator



Estimate

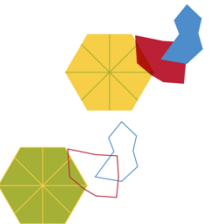
So what can cause terrible estimates?

- A lot of things can impact the process from sample to estimate
 - Bad luck
 - Too small a sample size
 - Outliers
 - Confounders
 - Clustering
 - Model misspecification
 - Bias
- That last one is particularly painful...



What is bias in statistics?

- Bias is anything that leads to a systematic difference between the true population and the samples used for inference
- It can arise at any stage of your scientific work
- There are too many different types of bias...



Spectrum bias

Selection bias

Sampling bias

Instrument bias

Observer bias

Lead time bias

Cultural bias

Estimator bias

Omitted variable bias

Performance bias

Volunteer bias

Survivorship bias

Reporting bias

Confirmation bias

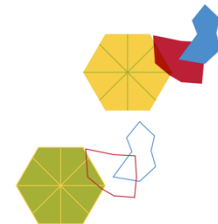
Social desirability bias

Publication bias

Attrition bias

Time lag bias

Recall bias



Spectrum bias

Selection bias

Estimator bias

Omitted variable bias

Social desirability bias

Publication bias

Attrition bias

Time lag bias

Recall bias

Confirmation bias

Survivorship bias

Cultural bias

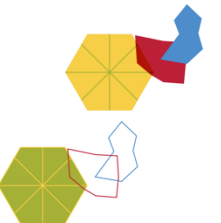
Lead time bias

Observer bias



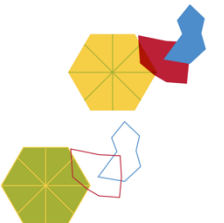
Some common types of bias

- Attrition:
 - Selective dropout of participants from a study
- Confirmation:
 - Favouring information that confirms existing beliefs and ignoring the rest
- Cultural:
 - Assessment of phenomena based solely on one's own cultural views
- Estimator:
 - The systematic difference between the estimator's expected value and the parameter being estimated



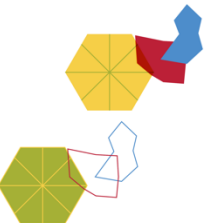
Some common types of bias

- Instrument:
 - When the instrument we are using to collect data is inappropriately calibrated
- Lead time:
 - When survival looks longer because someone received an earlier diagnosis
- Observer:
 - When the person collecting the data is influencing the collection
 - Bonus bias: Hawthorne effect – When the presence of an observer alters the behaviour of those observed
- Omitted variable:
 - When an important explanatory variable has been left out of a model



Some common types of bias

- Performance:
 - When knowledge of the treatment allocation leads to differences in the care provided
- Recall:
 - When there is a difference in accuracy of information relating to past events due to memory
- Reporting:
 - When only part of the findings from a study are reported



Catalogue of Bias



Admission rate bias

Arises when the variables under study are affected by the selection of hospitalized subjects leading to a bias between the exposure and the disease under study.

[Read More](#)

All's well literature bias

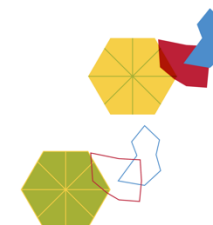
Occurs when publications omit or play down controversies or disparate results.

[Read More](#)

Allocation bias

Systematic difference in how participants are assigned to comparison groups in a clinical trial.

[Read More](#)



Catalogue of Bias



Differential Reference bias

When not all participants receive the same reference test in a diagnostic accuracy study.

[Read More](#)

Hawthorne effect

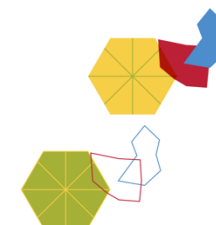
When individuals modify an aspect of their behaviour in response to their awareness of being observed.

[Read More](#)

Hot stuff bias

When a topic is fashionable ('hot') investigators may be less critical in their approach to their research, and investigators and editors may not be able to resist the temptation to publish the results.

[Read More](#)



Catalogue of Bias



Non-response bias

A bias that occurs due to systematic differences between responders and non-responders

[Read More](#)

Novelty bias

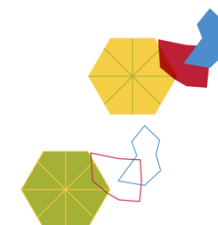
The tendency for an intervention to appear better when it is new.

[Read More](#)

Observer bias

The process of observing and recording information which includes systematic discrepancies from the truth.

[Read More](#)



Catalogue of Bias



Publication bias

When the likelihood of a study being published is affected by the findings of the study.

[Read More](#)

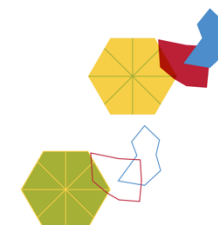
Racial bias

A distortion arising from systemic, institutional, interpersonal or individual forms of explicit (conscious) or implicit (unconscious) prejudice against individuals or groups based on social constructs of race or ethnicity that influences the planning, methods, results, interpretation, dissemination and application of health research.

[Read More](#)

Recall bias

Systematic error due to differences in accuracy or completeness of recall to memory of past events or experiences.

[Read More](#)

What can you do?

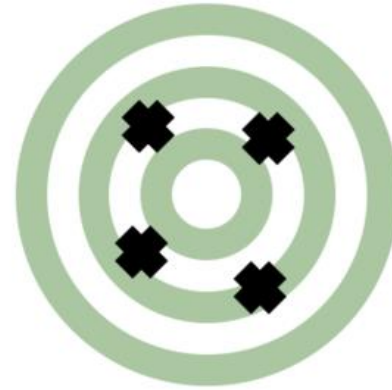
- Design
- Record
- Examine
- Correct
- Acknowledge
- Report



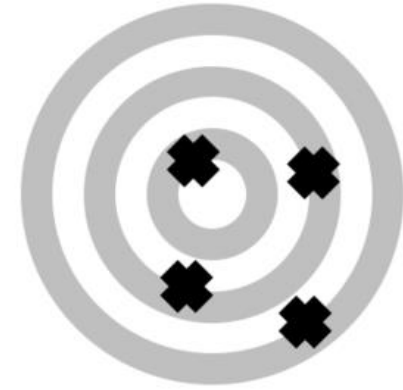
a) Accurate



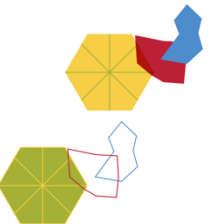
b) Biased



c) Noisy



d) Biased & noisy



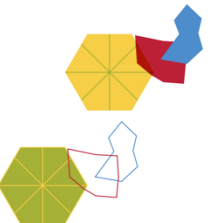
A biased estimator...

ID	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
Dist. (km)	2.5	2.0	1.0	2.5	5.0	0.5	0.5	4.0	5.0	1.0	0.5	4.0	3.0	0.5	4.0

- To calculate the variance, we need to use the following function (hint: $\bar{x}=2.4$)

$$var = \frac{1}{n-1} \sum_{k=1}^n (x_i - \bar{x})^2 = 2.86$$

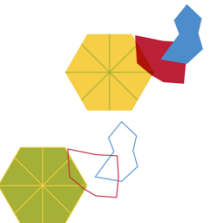
What's with the n-1??



Let's think about what sample variance does

- Sample variance is trying to work out the population variance
- The only knowledge we have about the population is derived from a sample
- We don't actually know the population mean μ , we only know the sample mean \bar{x}

$$var = \frac{1}{n-1} \sum_{k=1}^n (x_i - \bar{x})^2$$

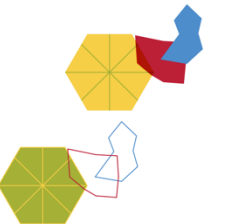


Bias...bias everywhere...

- That sample mean was calculated from our limited sample
- Any bias present in that sample will carry over to the sample variance
- How do we correct for it?
- This idea connects with degrees of freedom

$$var = \frac{1}{n-1} \sum_{k=1}^n (x_i - \bar{x})^2$$

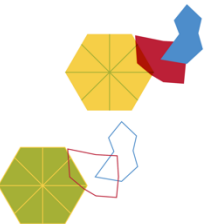
Bessel's correction



Let's return to an earlier point: Too few samples

Why would this be a problem?

- In simple terms: if you have too few samples then you have no way of knowing whether you have observed complete patterns
- Your measurements will be subject to the quirks of very few components



Let's return to an earlier point: Too few samples

Why would this be a problem?

- Think of the average of 3 numbers:

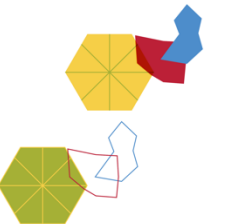
2,4, and 126 (avg=44)

- Now let's add just one more number:

2,3,4, and 126 (avg=33.75)

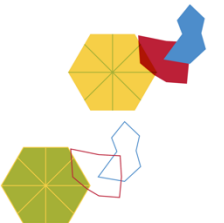
- And now 16 numbers:

2,3,4,2,3,4,2,3,4,2,3,4,2,3,4 and 126 (avg=10.69)



Impact of sample sizes

- In most cases, the more samples you have, the more certain you can be of your estimates
- Sample size plays a crucial part in hypothesis testing, but before we get to this, we need to consider simpler ways in which the sample size enters our process (from data to estimate)
- The first one is the **standard error**
- The second one is the idea of **degrees of freedom**

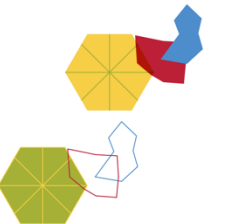


Standard error of the mean

- The standard error of the mean (SEM) measures the precision of the sample mean as an estimate of the population mean

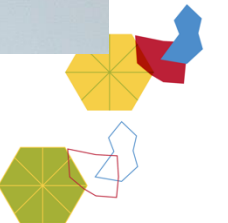
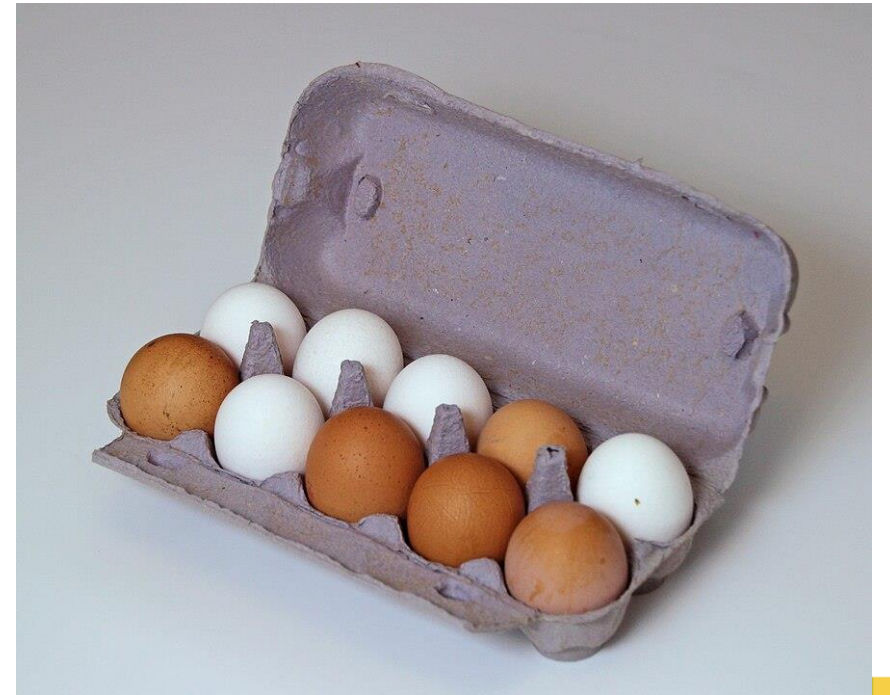
$$SEM = \frac{s}{\sqrt{n}}$$

- A smaller SEM means the sample mean is likely to be a more accurate estimate of the population mean



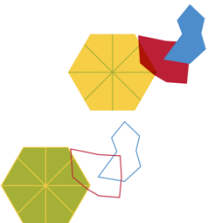
What are degrees of freedom?

- Degrees of freedom are the number of independent units in our sample that are free to vary without constraints
 - E.g., you have a carton of 10 eggs
 - You know that the average egg weight for the carton is 50g
 - How many individual eggs do you need to weigh before you know the exact weight of every egg?

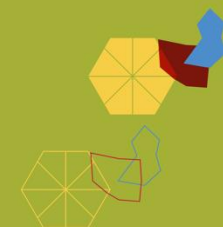


What are degrees of freedom?

- In the egg example, knowledge of the average weight for the carton placed a constraint on our sample
- Every egg was free to vary except for the last egg
- Typically, the degrees of freedom equal our sample size minus the number of parameters we need to calculate during an analysis

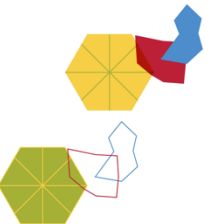
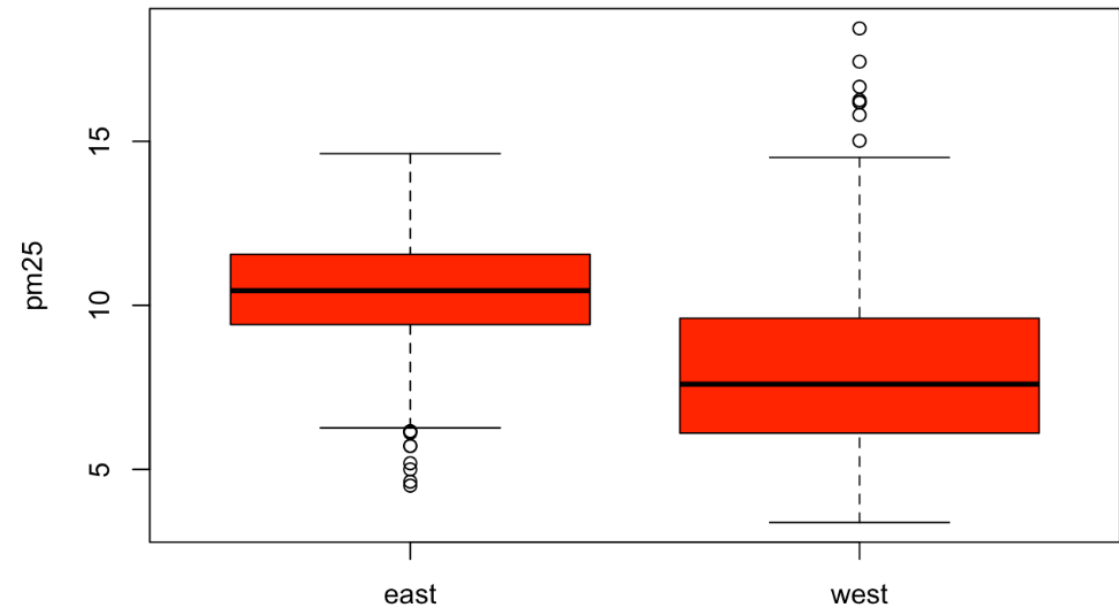


So, why do we need statistics?



Why should we care about statistics?

- Separation of signal from noise
- Measurement of uncertainty
- Evaluation of evidence
- Quantification of effect
- Modelling of patterns

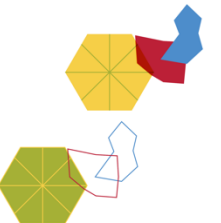


What is inference

Statistical inference is the process of forming judgements about the population based on a sample drawn from the population itself

(Verzani, 2005)

In general, we don't have access to the entire population – we need therefore to extract a sample and conduct our analyses on it



Hypothesis testing

- A hypothesis starts with a connection between a scientific question and what can be measured
- Experimental design aims to control the effect of various independent variables
- Statistics help specify the possible connection between dependent and independent variables



Hypothesis testing – What makes a good hypothesis?

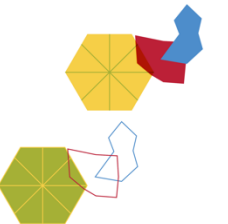
If Hypothesis A results in Outcome B



then observing that Outcome B is unlikely

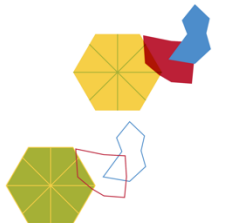
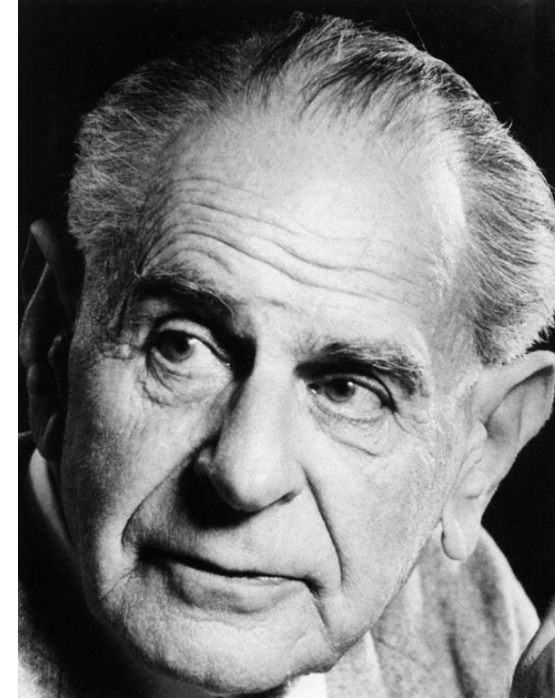


tells us that Hypothesis A is improbable



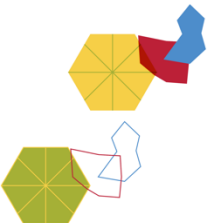
Hypothesis testing – What makes a good hypothesis?

- A good hypothesis is a testable hypothesis
- Usually this means it needs to be a ***falsifiable hypothesis***
- This idea forms the basis of “null hypothesis”
- A null hypothesis states that nothing interesting happens
 - No difference in performance between two groups of students
 - No difference in plant development between two treatments
- Absence of evidence is not evidence of absence



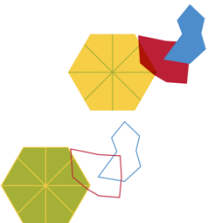
Hypothesis testing – Null vs Research hypotheses

- **Null hypothesis:** usually states that there is no effect in the underlying population
- **Research or alternate hypothesis:** it is our prediction of the effect in the population
- How can we choose one hypothesis over the other?
- We compute a conditional probability of observing a certain outcome by chance given that there is no effect in the population
- This conditional probability value is the **p-value**



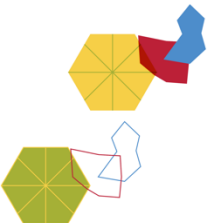
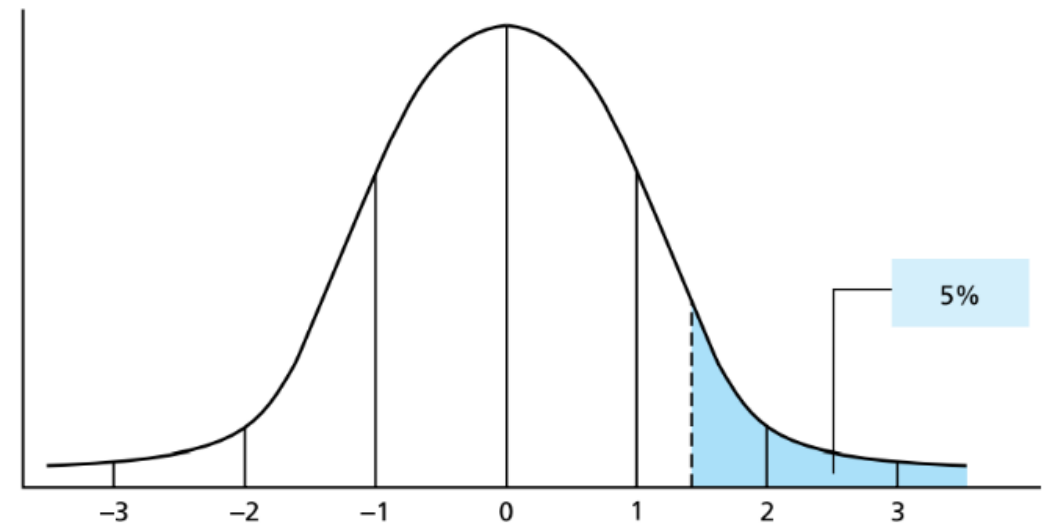
Hypothesis testing – p-value and significance level (α)

- We normally define a cut-off for p-values
- In life sciences this is 5% and it is called significance level (α)
- The significance level is the probability of erroneously rejecting the null hypothesis
- If the p-value is smaller than 0.05 we normally reject the null hypothesis
- If it is greater than 0.05 we fail to reject the null hypothesis



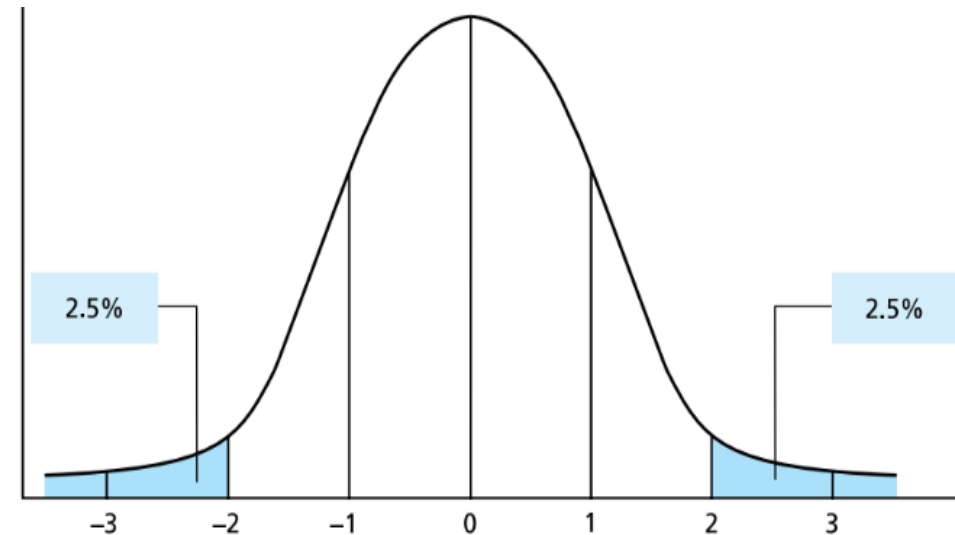
Hypothesis testing – One-tailed or two-tailed hypotheses

- One-tailed or directional hypothesis : when we specify the direction of the effect.
- E.g. as hours of study increase, so would exam grades
- The direction of the effect determines the tail of the distribution in which the resulting score will be located

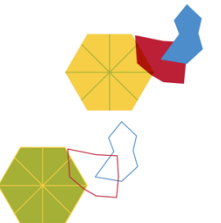


Hypothesis testing – One-tailed or two-tailed hypotheses

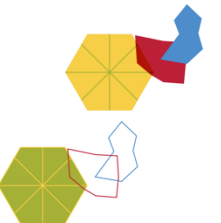
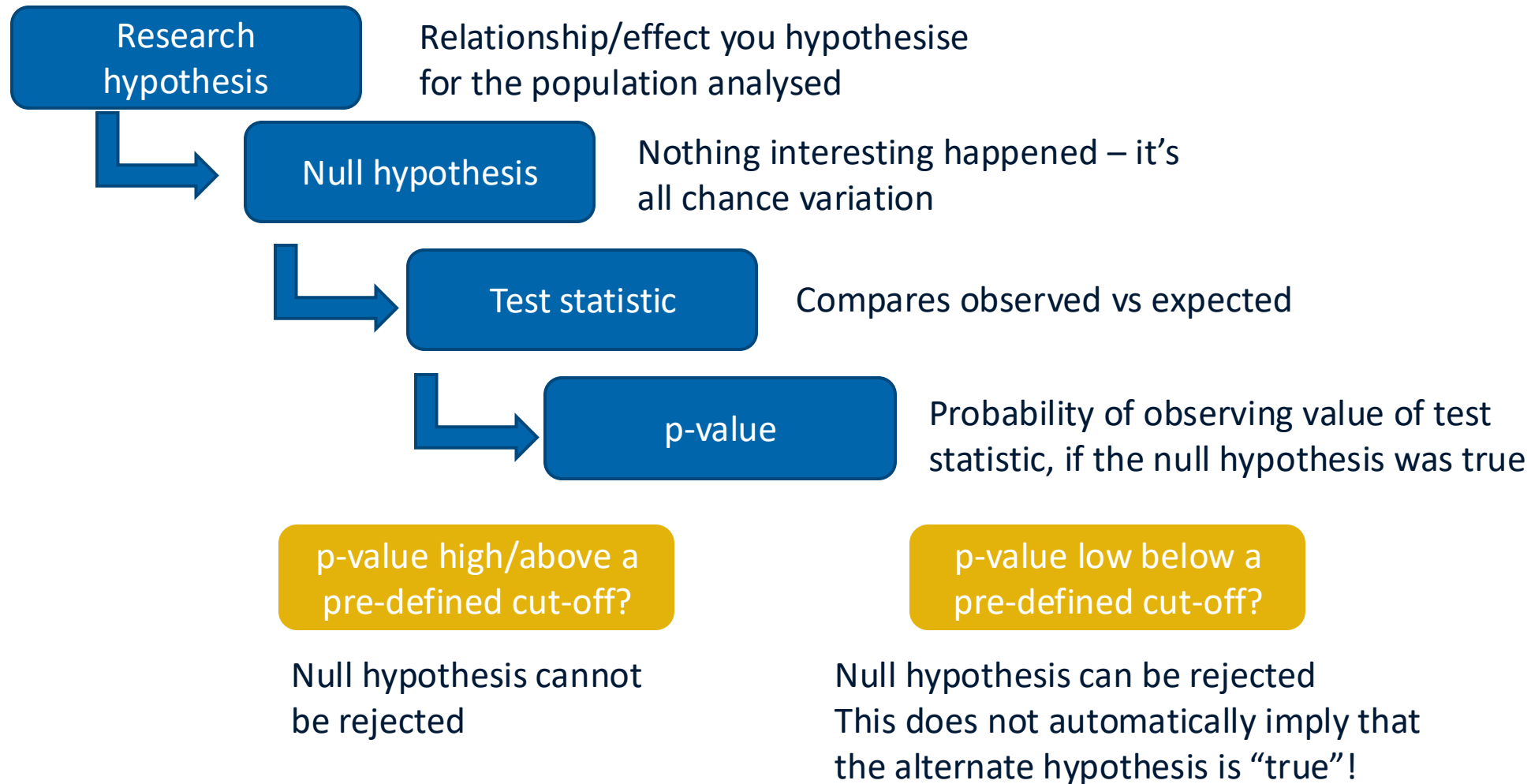
- Two-tailed or bi-directional hypothesis: there is an effect, but we don't know the direction
- E.g. relation between statistics anxiety and procrastination
- To achieve an overall $\alpha = 5\%$, we reject the null if we are either in the lowest or highest 2.5%



Two-tailed tests are more commonly used, unless there is already substantial evidence regarding the direction of the effect



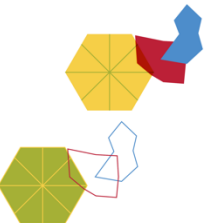
Hypothesis testing – Summary







Hypothesis testing – Error types

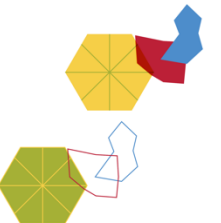
There are two main types of errors in hypothesis testing

- **Type I error** (False positive): we erroneously reject the null hypothesis
 - It is the probability of obtaining an effect as result of sampling error alone (α)
- **Type II error** (False negative): we erroneously fail to reject the null hypothesis (β)
 - We conclude there is no effect in the population, but in reality there is



Hypothesis testing – Error types

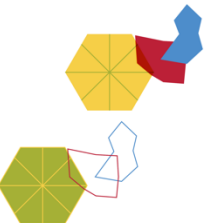
HYPOTHESIS TESTING OUTCOMES		Reality	
Research		The Null Hypothesis Is True	The Alternative Hypothesis is True
	The Null Hypothesis Is True	Accurate $1 - \alpha$ 	Type II Error β 
	The Alternative Hypothesis is True	Type I Error α 	Accurate $1 - \beta$ 



Hypothesis testing – Error types





Let's think of an example:

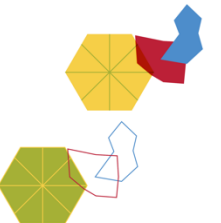
- Null hypothesis: All fruit belongs in a fruit salad
- Test: Is it a botanical fruit?
- Examples:
 - Banana – botanical fruit, belongs in a fruit salad
 - Potato – tuber (not a botanical fruit), does not belong in a fruit salad
 - Tomato - botanical fruit, does not belong in a fruit salad (**False positive**)
 - Rhubarb - leaf stalk (not a botanical fruit), belongs in a fruit salad if you're into it (**False negative**)



What is the power of a test?

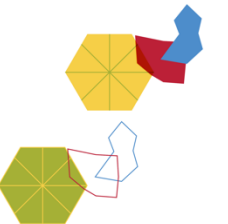
- Power is the probability that the test correctly rejects a false the null hypothesis
- It is essentially $1 - \beta$
- The higher the power, the lower the Type II error
- This only is relevant when the null hypothesis is false

HYPOTHESIS TESTING OUTCOMES		Reality	
		The Null Hypothesis Is True	The Alternative Hypothesis is True
R e s e a r c h	The Null Hypothesis Is True	Accurate $1 - \alpha$ 	Type II Error β 
	The Alternative Hypothesis is True	Type I Error α 	Accurate $1 - \beta$ 



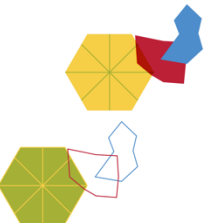
What is the power of a test?

- Power is a probability
 - Therefore it goes from 0 to 1
- If power is 1 then we always reject a false null hypothesis
- Commonly β is set at 0.2
 - Therefore power is often at 0.8
- This is translated as a 0.8 probability of correctly rejecting a false null hypothesis







What affects power? Significance level

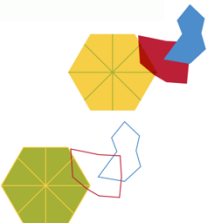
- Significance level is the probability of erroneously rejecting the null hypothesis
- If we want to increase our certainty of not rejecting a null hypothesis that should not be rejected we can reduce α
- Lower significance level means lower overall chances of rejecting a null hypothesis
- This means lower overall chances of rejecting a “**false**” null hypothesis
- Therefore lowering significance level reduces power



Return of Hypothesis testing – Why is $p < 0.05$?

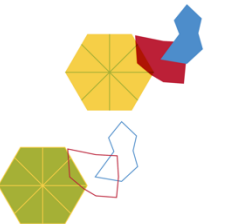
- This is a fairly arbitrary yardstick, but it aims to balance between different types of error
- To illustrate this let us consider $\alpha = 0.2$
 - Type I error tolerance 1/5
 - Type II error would decrease
- Now let us consider $\alpha = 0.001$
 - Type I error 1/1000
 - Type II would increase

HYPOTHESIS TESTING OUTCOMES		Reality	
		The Null Hypothesis Is True	The Alternative Hypothesis is True
R e s e a r c h	The Null Hypothesis Is True	Accurate $1 - \alpha$ 	Type II Error β 
	The Alternative Hypothesis is True	Type I Error α 	Accurate $1 - \beta$ 



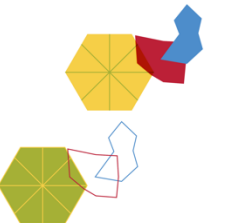
What affects power? Sample size

- If everything else is kept equal, the more samples you have the higher the chances of consistently detecting subtle differences in populations
- Essentially, larger samples result in narrower sampling distributions
- Narrower sampling distributions increase the chance of correctly rejecting a false null hypothesis
- More samples increase power



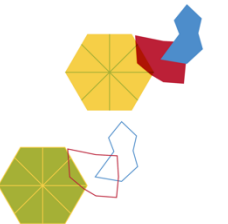
What affects power? Effect size

- Effect size captures the magnitude of difference between groups or the magnitude of the interaction between variables
- You can have situations where the difference between groups is consistent and highly significant but the effect size is very small
- Compare a situation where difference between two groups is large against one where one it is more subtle
 - The distributions in the first case will separate without needing many samples
 - Higher chance of rejecting a false null hypothesis

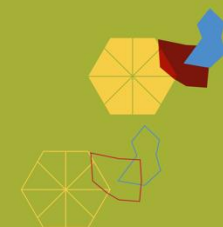


What affects power?

<https://rpsychologist.com/d3/nhst/>

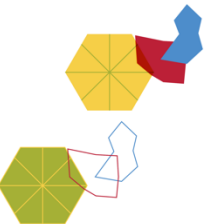


Confidence intervals



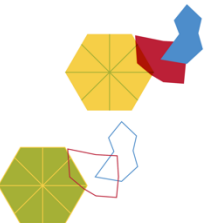
Why do we need confidence intervals?

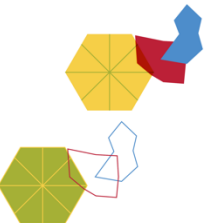
- Going from the sample to the population is a big step
- Estimates will inadvertently depend on our sampling
- We obviously should be careful when we sample, but we cannot control for everything
- It is more useful to think of regions that have a high degree of probability they contain the real population parameter instead of focusing on our estimate



Confidence intervals

- A confidence interval is actually a very tricky concept
- You will hear a lot of “almost right but not quite” definitions for it
- Confidence intervals are actually probability statements on long term frequencies of repeated samples
- They are **NOT** a statement about whether the true population value is in your calculated confidence interval with a certain probability (e.g., 95%)



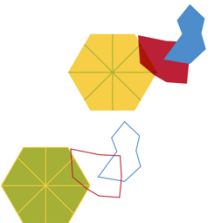


Confidence intervals

Confidence intervals are actually probability statements on long term frequencies of repeated samples

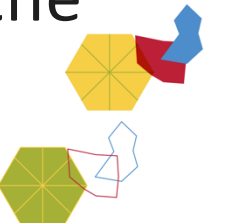
That...sounds like gibberish, but here is what it means:

- If you sample a population, and calculate the confidence interval, and then you sample again, and calculate another one, and you do that a lot, you will have a lot of confidence intervals
- Out of all those intervals, X% of them (commonly that's 95%) will contain the true population parameter



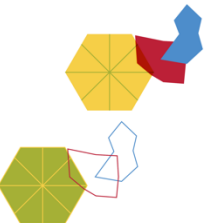
Confidence intervals

- I know, it sucks, and feels like a useless construction but there are good reasons for this awkward definition AND why we bother using them
- Once you calculate a confidence interval, the idea of a probability of it containing the true value is pointless!
- The interval either contains it, or it doesn't
- Our interval is a fixed result, and just because we don't know the whether the true population value is included that does not mean the answer to this is not a binary



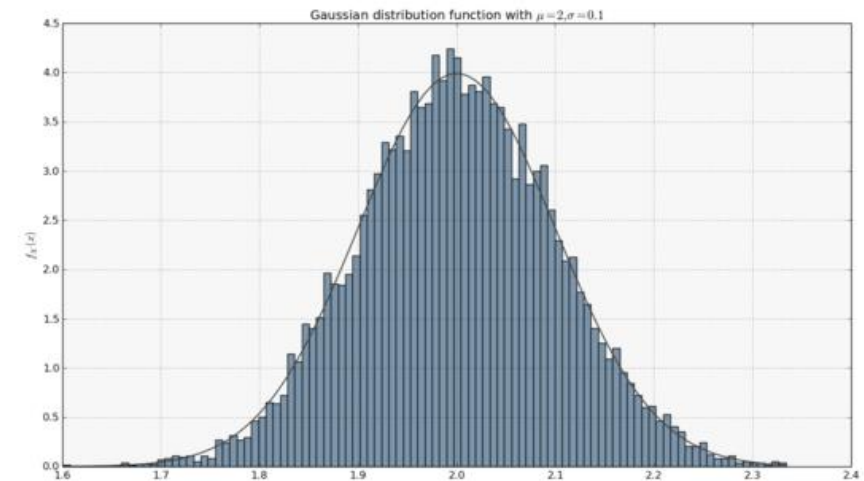
Confidence intervals

- Even with this tricky definition, confidence intervals help us quantify the uncertainty surrounding our estimates
- If we believe that the process we used to generate this confidence interval, when repeated a lot, would include in its vast majority the population parameter, then gain some confidence on the value of our point estimate
- A narrow interval is connected to more certainty than a wider confidence interval
- So how do we calculate them?



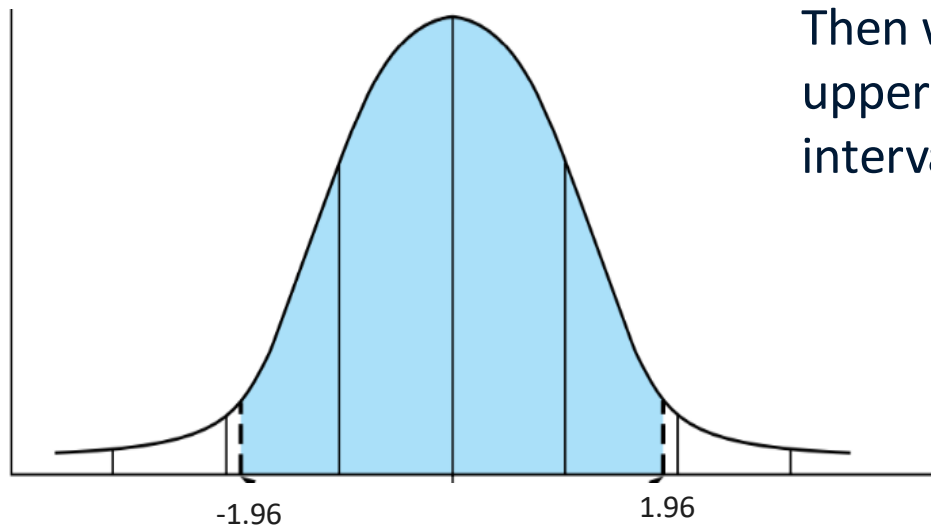
Confidence intervals

- We start with the sampling distribution of the estimator
- For example, a normal distribution
- A normal distribution is described by its mean and standard deviation
- With those, we can sketch the curve and use it to extract the confidence intervals
- Why can we do this?



Confidence intervals

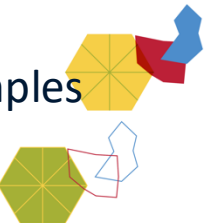
- The 68-95-99.7 rule tells us that 68%, 95%, and 99.7% of the data are within 1, 2, and 3 standard deviations from the mean
- Technically, 95% is 1.96 standard deviations away from the mean



Then we can compute lower and upper bounds of the confidence interval as:

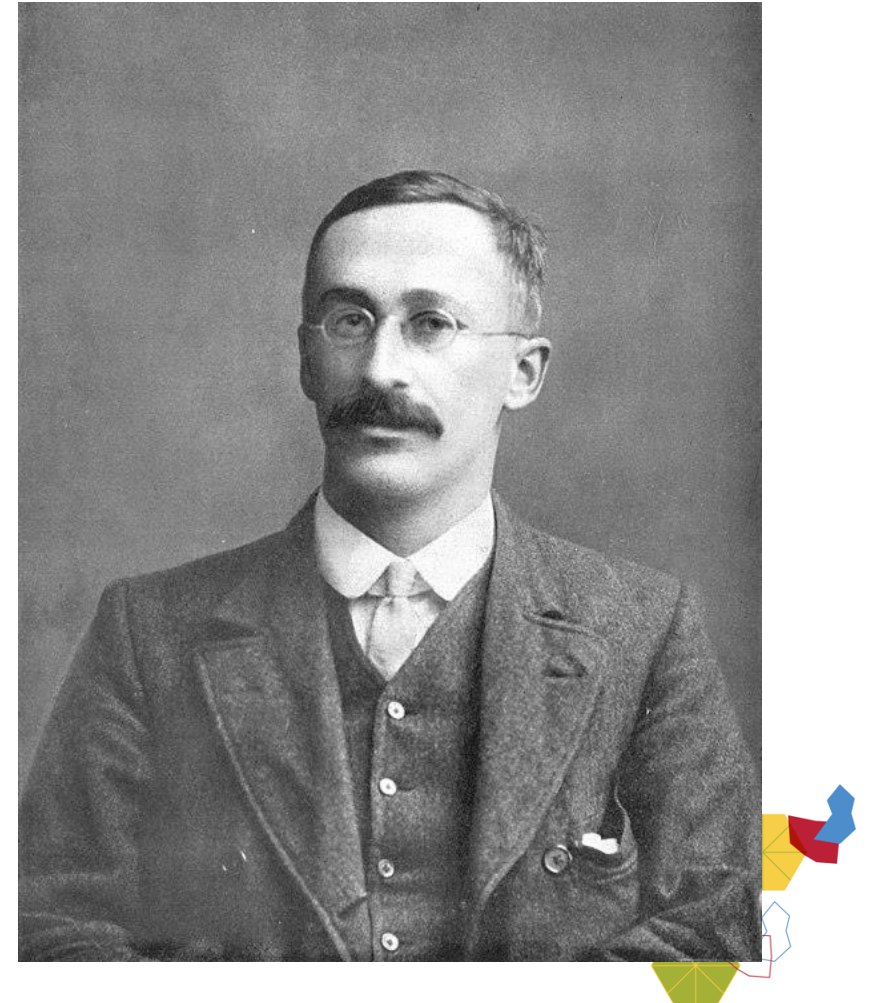
$$CI = \bar{x} \pm z \frac{\sigma}{\sqrt{n}}$$

- \bar{x} is the sample mean
- z is the critical value (z-score)
- σ is the population standard deviation
- n are the number of samples



Hello bias my old friend...

- That σ on the previous slide is quite troublesome...
- In most cases we don't actually know the population standard deviation
- Replacing with the sample standard deviation is questionable, especially for small sample sizes
- This also troubled William Gosset, at the Guinness factory



A Student's perspective

- Gosset developed what is known as the t-distribution
- The t-distribution is very similar in shape to the normal distribution, with symmetry around zero
- The primary parameter is the degrees of freedom, when this becomes very large then the t-distribution becomes a normal distribution
- Amongst other things, the t-distribution gives us more accurate estimates of the confidence interval



THE PROBABLE ERROR OF A MEAN

BY STUDENT

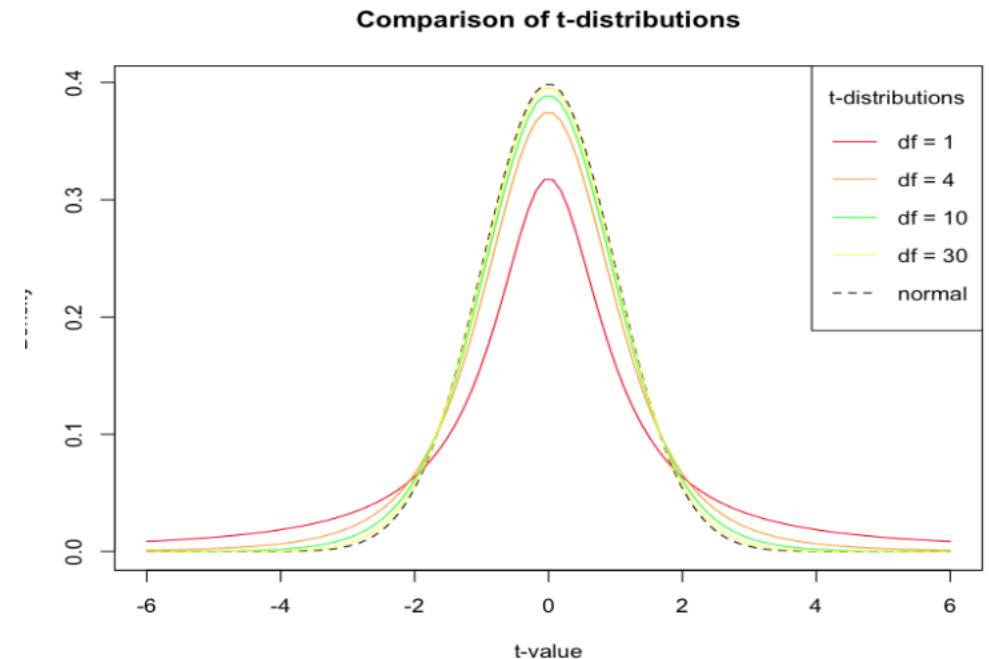
Let's try this again...

- When both the population mean and the population standard deviation is unknown, and the sample size is not extremely large, we're better off using the t-distribution for our CI

Then we can compute lower and upper bounds of the confidence interval as:

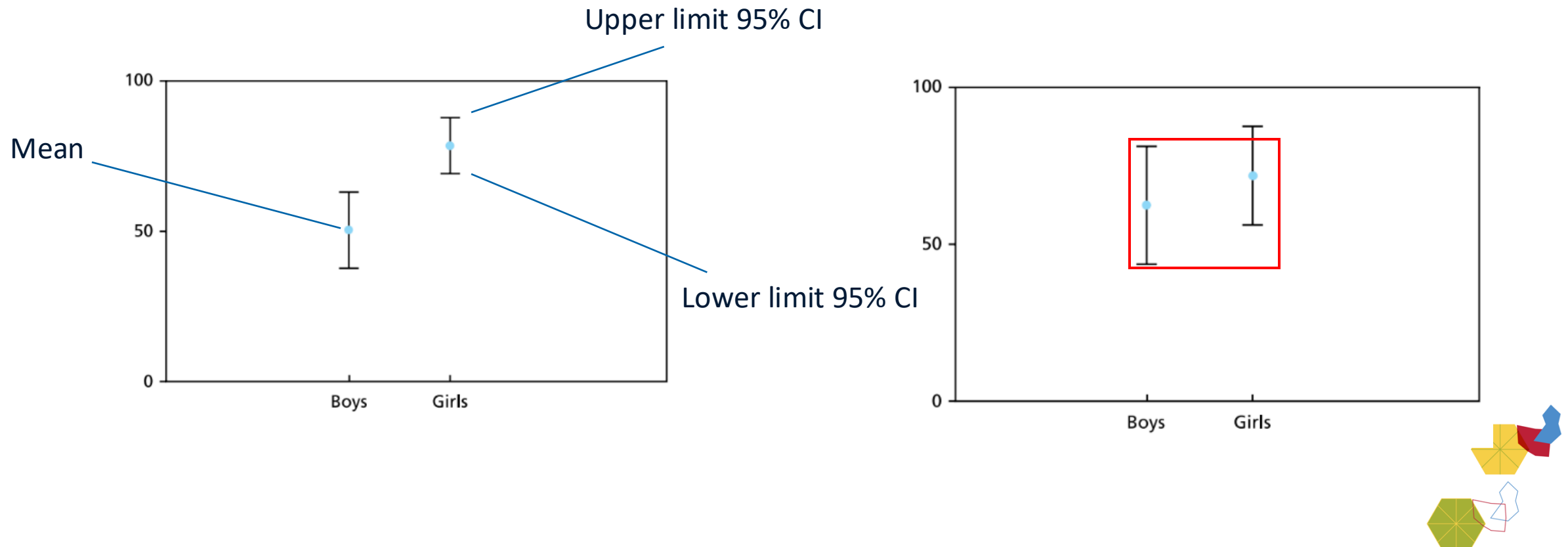
$$CI = \bar{x} \pm t \frac{s}{\sqrt{n}}$$

- \bar{x} is the sample mean
- t is the critical value (t-score)
- s is the sample standard deviation
- n are the number of samples



Confidence intervals

- Confidence intervals are particularly helpful in graphical displays



If the CLT holds, why don't we always use normal distribution...?

- Because what the CLT does well is take a continuous population distribution, and help us model the sampling distribution for the average
- Modelling the average is not sensible when we are interested in
 - binary outcomes,
 - counts,
 - bimodal distributions
 - survival times
 - SO MUCH more

