The University of Manchester

Statistical Inference

PCHN62121
Image Analysis

Dr Martyn McFarquhar

The University of Manchester

Introduction

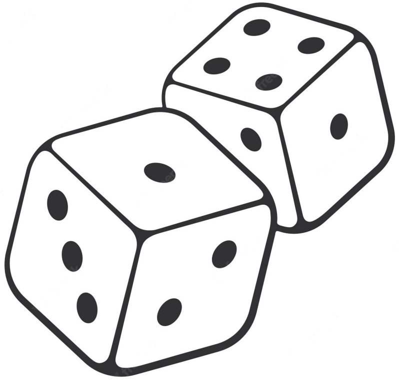
- Statistics lie **at the heart** of everything we will be doing during this module
- Our ability to reach conclusions about our fMRI and M/EEG data depends entirely upon **statistical modelling** and **statistical inference**
- Poldrack, Mumford & Nichols (2011) name *Probability and Statistics* as their **number 1** prerequisite for fMRI data analysis

> 1. *Probability and statistics.* There is probably no more important foundation for fMRI analysis than a solid background in basic probability and statistics. Without this, nearly all of the concepts that are central to fMRI analysis will be foreign.

- In this session we will review the **fundamentals** of statistical inference to prepare you for the content that is to come on this module

Probability

- Probability is the foundation of everything in statistics.
- Statistics is the **science of uncertainty** - of reaching conclusions based on **noisy** or **incomplete** information
- Probability is the **language of uncertainty**
- Statistics uses probability to **describe the nature of data** and how we can reach **general conclusions** about a phenomena by examining **a small part of it**



- Probability provides a mechanism for **inductive reasoning** - going from the **specific** to the **general**
  - **Induction** is a big philosophical problem that is not fully resolved - hence why we cannot *prove* anything in science

Probability

Kolmogorov axioms

- Mathematically, for a number to be called a probability it must adhere to some **rules** – known as the Kolmogorov axioms
- Imagine rolling a six-sided die:
    - There are 6 **mutually exclusive** events – the numbers 1 to 6
    - Each event needs to be assigned a number that is ≥ 0 (Axiom 1)
    - They cannot all be 0 (Axiom 2)
    - The sum of all the probabilities must be equal to 1 (Axiom 3)
- Therefore, if we believe all the outcomes to be equally probable we can define

$$P(E) = \frac{\text{number of favourable events}}{\text{total number of events}}$$

Kolmogorov axioms

- The probability of rolling a 5 would be

$$P(5) = \frac{1}{6}$$

- The probability of rolling an even number would be

$$P(2 \cup 4 \cup 6) = \frac{3}{6} = \frac{1}{2}$$

- All these examples satisfy the Kolmogorov axioms and thus can be called probabilities
- Notice that any notion of **what probability means** is completely absent – Kolmogorov tells us how to calculate the numbers, but he does not tell us **what they mean**

Interpreting Probability

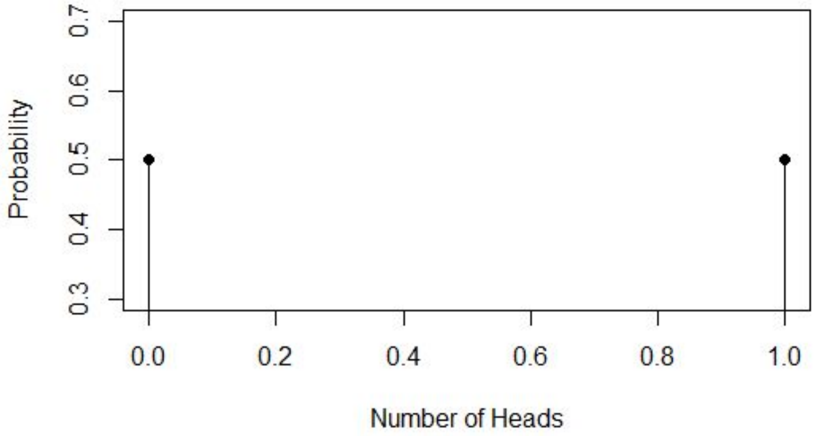
- One of the greatest divides in modern science between the **Frequentist** and **Bayesian** approaches to statistics
- For the Frequentist, probabilities represent **physical phenomena** that can be **counted**
  - A probability is the **long-run frequency** of an event
- For the Bayesian, probabilities represent **degrees of belief**
  - A probability indicates, based on the available evidence, how likely an event is to occur
- A Bayesian can apply probability to events that cannot be counted (e.g. the probability of rain tomorrow)
- The Bayesian view leads to a **much more flexible analysis framework** – the notion of **degree of belief** has been criticised as **too subjective**
- The development during the 20th century of inferential statistics by **Ronald Fisher** was motivated by his **deep disdain** for the Bayesian perspective on probability

MANCHESTER 1824
The University of Manchester

Random variables

- Irrespective of your philosophical views on interpretation, one of the most important concepts from probability for statistics is the **random variable**
- A random variable is
  - A variable whose value is dependent upon the outcome of some random processes
  - A variable where we will measure a different value every time we observe it
  - A variable where each possible values can be associated with a probability
- A basic example would be the outcome of flipping a coin

Outcome	Probability
H	1/2
T	1/2



Random variables

- Another example would be counting the number of **heads** after 3 flips of a coin

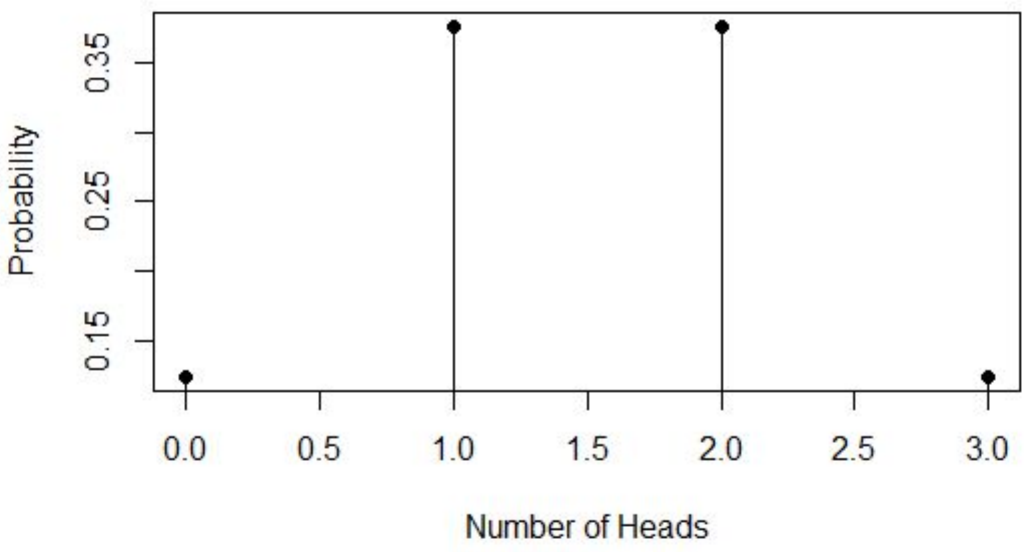
Outcome	**Number of Heads**
HHH	3
HHT	2
HTH	2
THH	2
HTT	1
THT	1
TTH	1
TTT	0

Number of heads	**Probability**
3	1/8
2	3/8
1	3/8
0	1/8

Random variables

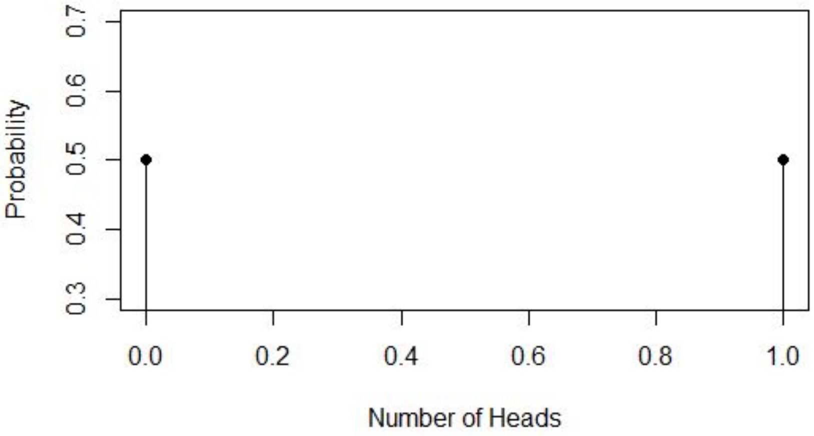
- Another example would be counting the number of **heads** after 3 flips of a coin

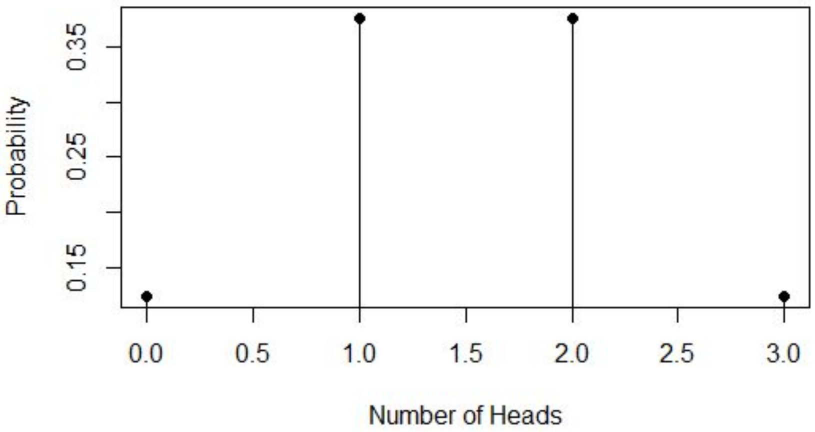
Number of heads	Probability
3	1/8
2	3/8
1	3/8
0	1/8



Probability distributions

- These shapes are known as **probability distributions** - they tell us the probability of **all possible values** of the random variable



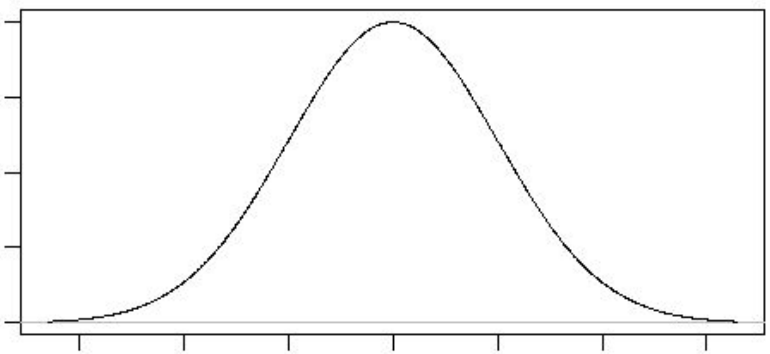


- These shapes are examples of the *binomial distribution* $y \sim \text{Binomial}(n, p)$
- Each probability distribution is controlled by **parameters** that describe the **shape**
  - n = the number of trials, p = the probability of success on a single trial

Probability distributions

The Normal Distribution

- The binomial distribution is an example of a **discrete** probability distribution because the measurements are **whole numbers**
- In the real world we often deal with random variables that can take on an **infinite** number of possible values
    - Time, height, weight, reaction time, BOLD signal etc.
- In these cases we have to use a **continuous** probability distribution
- Although there are many continuous distributions available, the most commonly used is the **normal distribution**
- Also known as the **Gaussian** distribution



Probability distributions

The Normal Distribution

- The normal distribution is fully described by the formula

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

- The important point is that this is parameterised by **two** values:
  - The **mean** (μ) - the centre of the distribution
  - The **standard deviation** (σ) - the width of the distribution

$$y \sim \mathcal{N}\left(\mu, \sigma\right)$$

- If we assume our random variable of interest comes from a normal distribution, our aim is to estimate the **mean** and **standard deviation** and how these **change** under **different experimental conditions**

The University of Manchester

Random samples

- Imagine we have an interest in the **weight** of **males** who are suffering from **major depressive disorder** in the **UK**
- **Weight** is a **continuous random variable** with some distribution **-** if we assume this is a **normal distribution** then

$$\text{weight} \sim \mathcal{N}\left(\mu, \sigma\right)$$

- This distribution represents the **entire population** under study
- We want to know the **parameter values** of this distribution – we would need to weight **every male who has major depression in the UK**
- Instead we take a **sample** and use this to **infer** something about the population
- Using a **sample** to say something about the **population distribution** lies at the heart of **parametric** statistical methods

Random samples

- A **random sample** of size *n* from a **population** can be conceptualised as a sequence of *n* independent random variables (y_1, y_2, y_3, ..., y_n), where each random variable is drawn from the **same distribution** (i = 1,...,n)

$$y_i \sim \mathcal{N}(\mu, \sigma)$$

- These are known as **independent and identically distributed** (i.i.d.) random variables
- This random sampling model describes an experimental situation where **repeated observations** are made on the **same variable** *y*
- For our current example, each observation represents the **weight** of a **different subject** in our experiment

Estimating population parameters

- Our aim is now to use our **random sample** to **estimate** values for the **population mean** and the **population standard deviation**
- It would seem that what we want to calculate is

$$P(\mu, \sigma | y)$$

- This cannot be calculated without using Bayesian methods (which Fisher hated) so classical statistical methods instead use something called the **likelihood**

$$\mathcal{L}(\mu, \sigma | y) = P(y | \mu, \sigma)$$

- An **optimisation algorithm** is used to search through different values of the parameters to find those that **maximise the likelihood**
- In some cases, optimisation is not needed because there are **closed form** solutions to finding the estimates that maximise the likelihood

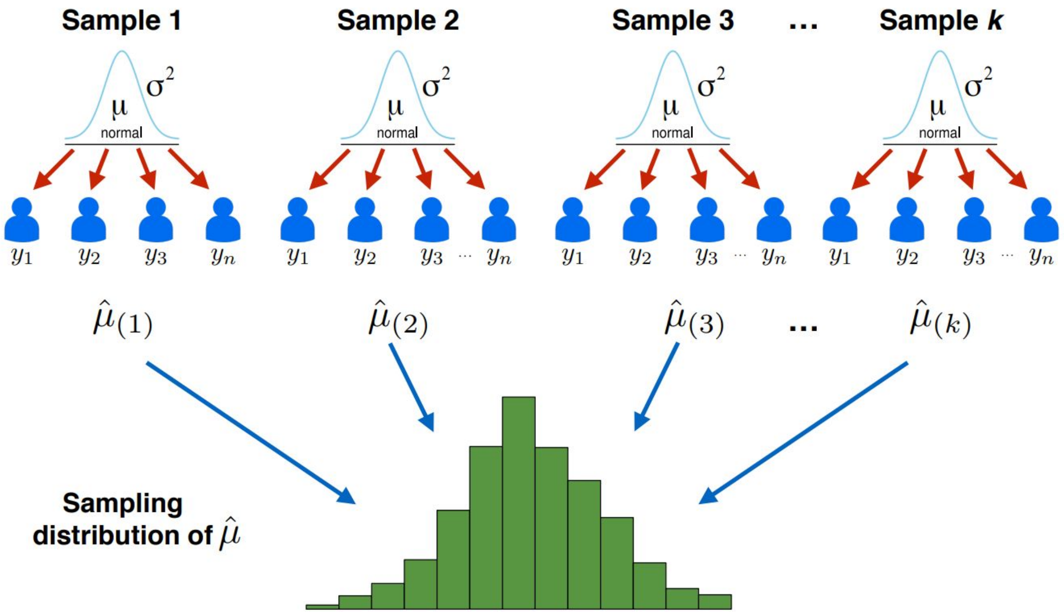
Uncertainty in the parameter estimates

- So let us take a step back:
  - We have a phenomena of interest characterised as a **random variable** from a **normal population distribution**
  - We want to know the **population parameter values**, but **cannot measure** the **whole population**
  - We **take a sample** and use the method of **maximum likelihood** to **estimate** the population parameters
  - For a normal distribution, this involves calculating the **sample mean** and **sample standard deviation**

- There is a problem with doing this:
  - What happens if we take a **different sample**? Will we get the same estimates?
  - No! Because a **different sample** will contain **different data -** so which estimates do we use?
  - We need some way of characterising the **uncertainty** in our estimates.

Uncertainty in the parameter estimates

- The key insight is to recognise that with **each new sample** we will get **different parameter estimates -** our estimates are **also random variables**
- This means they have an associated **probability distribution** – the **sampling distribution**



MANCHESTER 1824
The University of Manchester

Uncertainty in the parameter estimates

- The key insight is to recognise that with **each new sample** we will get **different parameter estimates -** our estimates are **also random variables**
- This means they have an associated **probability distribution** – the **sampling distribution**
- For a **normal population distribution** the sampling distribution of the mean is **also normal**

$$\hat{\mu} \sim \mathcal{N}\left(\mu, \frac{\sigma}{\sqrt{n}}\right)$$

- The **mean of this distribution** is the **true population mean** – on average, we should estimate this correctly across samples
- The **standard deviation** of this distribution depends upon the **sample size -** the more data the more accurate we will be – this is known as the **standard error**

The University of Manchester

Hypothesis testing

- By this point we have successfully managed to:
  - Characterise our phenomena of interest as a **random variable** with a **distribution**
  - Use formulas derived from **maximum likelihood** to estimate the **parameters** of this distribution based on a single sample
  - Calculate the **standard error** of these estimates as a means of characterising their **uncertainty**
- So we now have **estimates** and **standard errors** – how do we use these to reach conclusions about the population under study?
- This is where the process of **null hypothesis significance testing** comes in

Hypothesis testing

Test Statistics

- Trying to draw conclusions based on the **parameter estimates** has two issues:
    - The estimates are on the **same scale** as the data (e.g. weight) so depend upon our **domain knowledge** to interpret
    - The estimates alone do **not** take the **uncertainty** into account

- Both of these issues can be solved by **dividing** the estimate by the standard error

$$t = \frac{\text{estimate}}{\text{standard error}} = \frac{\hat{\mu}}{\sigma\left\{\hat{\mu}\right\}}$$

- The quantity *t* is now a **standardised** variable – same units irrespective of the data
- The quantity *t* contains both the **estimate** and **uncertainty** – the value will increase as the uncertainty decreases

Hypothesis testing

Test Statistics

- Using *t* for **hypothesis testing** involves comparing our estimate with some **hypothesised value** for the population parameter

$$t = \frac{\hat{\mu} - \mu^{H_1}}{\sigma\left\{\hat{\mu}\right\}}$$

- The **larger** the value of *t*, the greater the **discrepancy** between our estimate and our hypothesised value
- So **big values** of *t* suggest that our **hypothesised population** value is **incorrect**
- In this example, the hypothesised value of the mean would depend upon domain knowledge (e.g. average weight of males in the UK)

Hypothesis testing

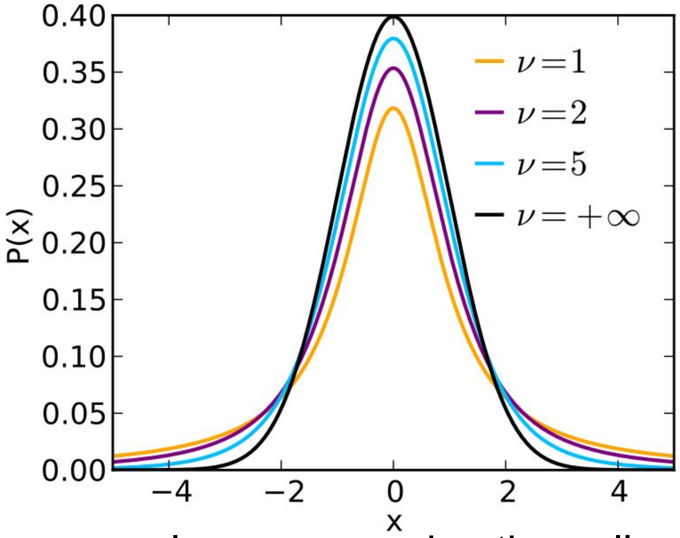
Null Hypothesis Significance Testing

- The insight that Ronald Fisher provided was that our test should form a **null hypothesis**
- In this instance, it would be there the **difference** between the **true mean** and the **hypothesised mean** is **0** in the population
- To see why this is useful, consider that *t* is **also a random variable** because it is calculated from two other random variables
- This means that *t* has a **distribution** that can be derived from knowing the sampling distribution of the estimates
- If we assume that the null hypothesis is **true** then the *t*-distribution will be centered on **0** with a width that depends upon the sample size

Hypothesis testing

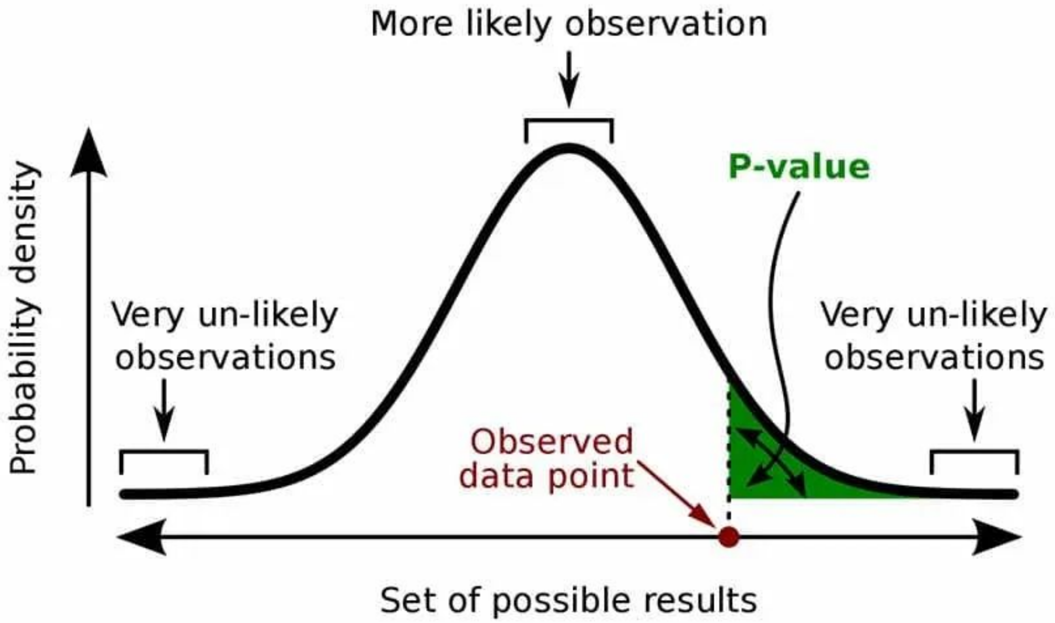
Null Hypothesis Significance Testing

- This distribution tells us the various values of *t* we would expect to calculate **if the null hypothesis were true**
- So what we can do is use this distribution to calculate the **probability** of obtaining our particular value of *t*
- This gives the *p*-value
  - The probability of obtaining a test statistic as larger, or larger, assuming the null hypothesis is true
- A **small** *p*-value suggests that our calculated test statistic is **unlikely**, if the null were true – either we observed a **rare event** or the **null hypothesis is not accurate**



Hypothesis testing

P-values



A **p-value** (shaded green area) is the probability of an observed (or more extreme) result assuming that the null hypothesis is true.

Hypothesis testing

P-values

- How do we use this information?
  - Fisher's recommendation was to count any $p < 0.05$ as **evidence against the null**
  - In our example, the null was that the population mean is the same as the hypothesised mean (their difference was **0**)
  - If $p < 0.05$
    - We would call this a **significant** result and **reject** the null hypothesis – it is unlikely that the population mean is the same as the hypothesised mean
  - If $p > 0.05$
    - We would call this a **non-significant** result and **fail to reject** the null hypothesis – it is possible that the population mean is the same as the hypothesis mean
- The *p*-value is a way of reaching binary conclusions from our results

Application to multiple groups

Two sample tests

- To see how this method applies to more complex experiments, consider comparing the weights of depressed individuals taking **two different drugs**

$$y_{ij} \sim \mathcal{N}(\mu_j, \sigma_j)$$

- We now have **two** population distributions

$$y_i^{(\text{Drug A})} \sim \mathcal{N}\left(\mu^{(\text{Drug A})}, \sigma^{(\text{Drug A})}\right)$$
$$y_i^{(\text{Drug B})} \sim \mathcal{N}\left(\mu^{(\text{Drug B})}, \sigma^{(\text{Drug B})}\right)$$

- Our aim is still to **estimate** the **parameters** of these distributions – we want to **compare** the means to see whether average weight changes due to the drug

MANCHESTER 1824
The University of Manchester

Application to multiple groups

Two sample tests

- We use the same procedure as before to estimate the **means** and **standard deviations** of these populations, as well as the **standard errors** of the estimates
- The *t*-statistic then involves comparing the **mean difference** to a **hypothesised mean difference** – typically taken to be **0**

$$t = \frac{(\hat{\mu}_1 - \hat{\mu}_2) - D^{H_1}}{\sqrt{\sigma\{\hat{\mu_1}\} + \sigma\{\hat{\mu}_2\}}} = \frac{\hat{\mu}_1 - \hat{\mu}_2}{\sqrt{\sigma\{\hat{\mu}_1\} + \sigma\{\hat{\mu}_2\}}}$$

- We can then use the **same** null t-distribution to calculate a *p*-value to provide evidence for or against the null hypothesis of the population distributions having the **same mean**
- The process is **the same** – assume a population distribution, estimate the parameters from a sample, form a hypothesis test about the parameters, calculate a *p*-value

The University of Manchester

Application to continuous variables

Regression models

- We can also use the same framework to reach conclusions about the relationship between our random variable of interest and other continuous measures

- Imagine that we are interested in how the **weight** of our **depressed males** relates to the **severity of their symptoms**

- In this situation, we might start with the normal distribution model

$$\text{weight} \sim \mathcal{N}(\mu, \sigma)$$

- But then specify a more complex form for the mean

$$\mu = \beta_0 + \beta_1 \text{severity}$$

- So the value of the mean depends upon the severity of symptoms

Application to continuous variables

Regression models

- Assuming a **mean function** of

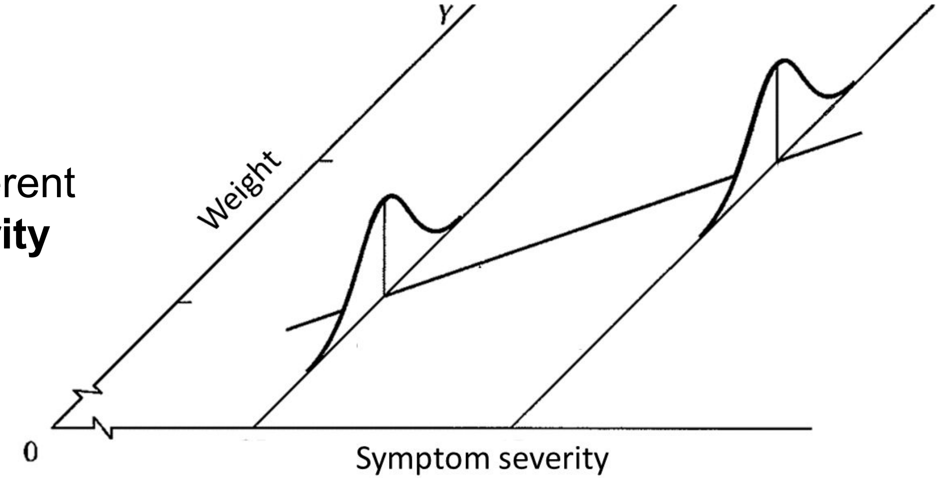
$$\mu = \beta_0 + \beta_1 \, \text{severity}$$

  is an example of a **linear regression model**

- This assumes that the relationship between **weight** and **symptom severity** is a **straight-line**
    - β_0 is the **intercept**
    - β_1 is the **slope**

- The probability model is that there is a different normal distribution for **each value** of **severity**

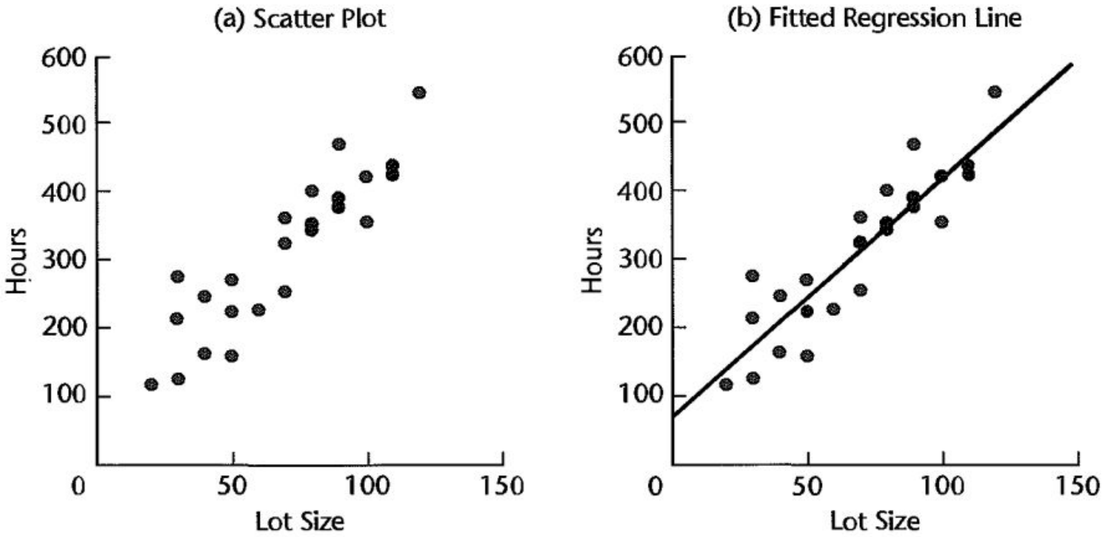
- The standard deviations are the **same** and the **means** sit along a **straight line** defined by the two parameters



Application to continuous variables

Regression models

- In order to estimate the **mean** of our **population distribution** we need to estimate the values of the **intercept** and the **slope** – in this example the mean depends upon two further parameters
- **Maximum likelihood** can do this for us



Application to continuous variables

Test Statistics

- We can again calculate a *t*-statistic, but this time on the *intercept* and the *slope*

$$t = \frac{\hat{\beta}_1 - \beta_1^{H_0}}{\sigma\{\hat{\beta}_1\}}$$

- The hypothesised value for the slope is usually taken as **0 – no relationship** between **weight** and **severity**

$$t = \frac{\hat{\beta}_1 - 0}{\sigma\{\hat{\beta}_1\}} = \frac{\hat{\beta}_1}{\sigma\{\hat{\beta}_1\}}$$

- So this is the same approach as **before** – the only difference is that the **mean function** is more complex – this is the difference between different **statistical models**

The University of Manchester

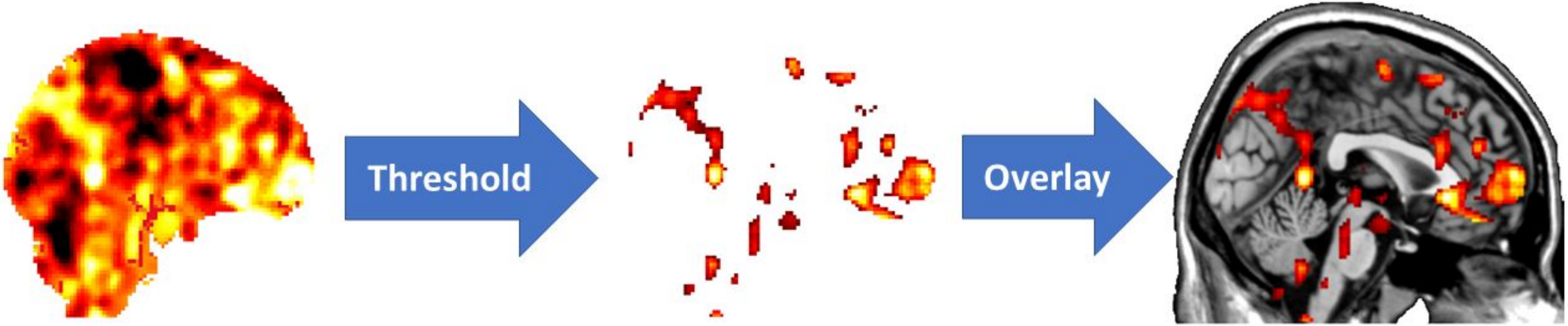
Summary

- We have now seen the process of **statistical inference**, from **first principles** about probability, all the way up to ***p*-values** and **hypothesis testing**

- This is a somewhat complex process:
  - Our data are conceptualised as random variables drawn from a distribution
  - This distribution has parameters that characterise the whole population
  - We want to know these parameters but cannot use the whole population
  - Instead, we take a sample and estimate the population parameters
  - These estimates are random variables with an associated sampling distribution
  - The standard deviation of the sampling distribution is known as the standard error
  - Dividing the estimates by the standard error produces a test statistic
  - This test statistic is also a random variable with a distribution
  - We can calculate the shape of this distribution under the null hypothesis of no effect
  - We can then calculate a *p*-value to tell us how likely it would have been to obtain our test statistic if the null hypothesis were true
  - $p < 0.05$ is evidence against the null

Summary

- This may take some time to sink in if it is new to you – it is the fundamental process used to reach conclusions about fMRI and M/EEG data
- We will see all of this in action as we learn how statistical modelling and inference works inside of SPM



SPM{*t*}

t-values with $p < 0.05$

Final results