Experimental Design and Basic Statistics

Dr. Arief Gusnanto

Department of Statistics
University of Leeds

IMforFUTURE Network meeting

Zagreb 25-28 March 2018

Outline

1. Population vs Sample
2. Types of data (exploration)
3. Normal distribution
4. One-sample and two-sample tests
5. Introduction to linear regression
6. Experimental design

Why statistics?

- Research on evaluating existing hypotheses and exploring new ones often requires appropriate statistical analysis of numerical information
- Numerical information often consists of a sample from the population
- Inferences made from the sample are based on uncertainty
- Statistics cannot **prove** – it can only give guidance on the most plausible conclusion

The role of statistics

- Summary of numerical information
- Model situations involving uncertainty
  - Allow decisions to be made on the basis of incomplete information
  - Model relationships between variables
  - Model variability in measurements
- Making decisions based on this information
- A support for many types of enquiry
  - Scientific, sociological, criminal etc.

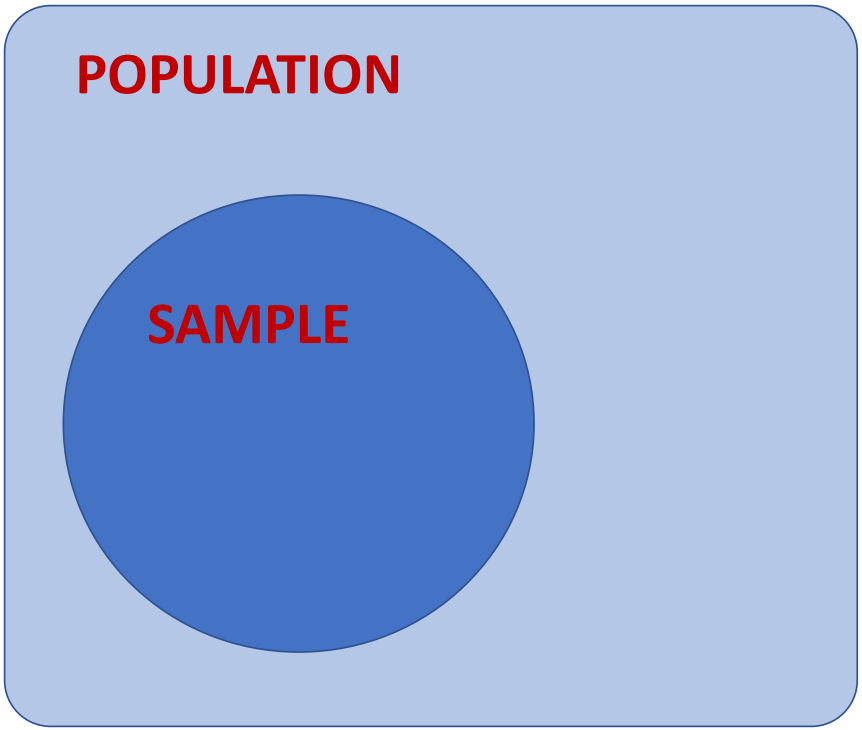
What Can Statistics Prove?

- Since data tends to be only a representation of the whole, *actual proof* is not possible
- The data can be used to make an assessment of the relative plausibility of competing hypotheses
- There will always be uncertainty in the decision we have made
- The cause of the trouble is *variability*

Components of statistical analysis

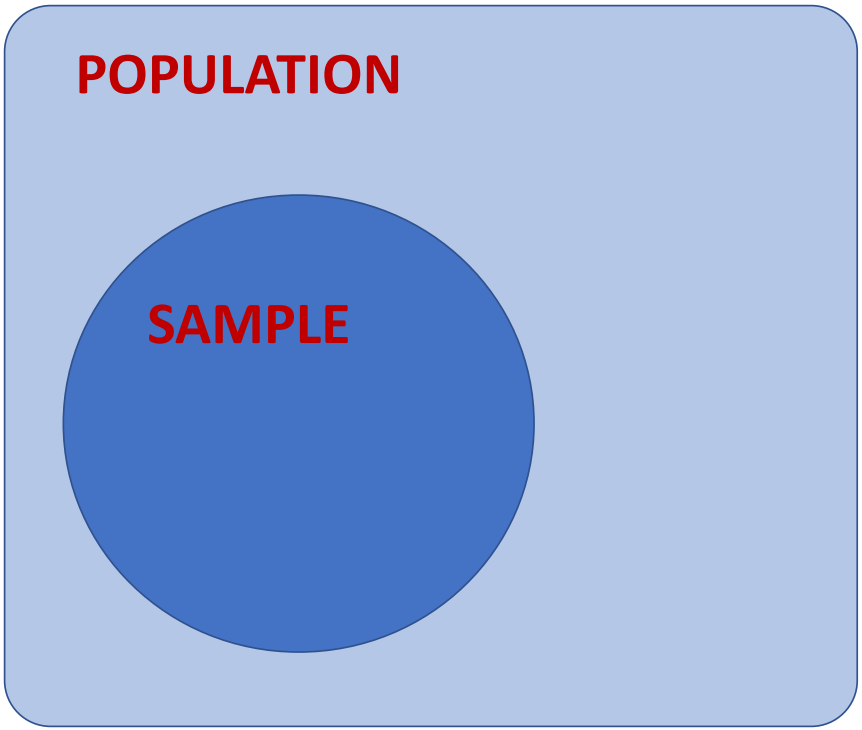
- Formulation of Problem
- Identifying Appropriate Technique and Model
- Deciding How, What and Quantity of Data to Gather
- Analysing and Interpreting Data
- Making Decisions
- Validation of Model Chosen
- Presentation of Information and Conclusions

Population vs Sample



- **Population** (in statistical sense) is a set of items or entities that is the focus of your study
- It defines the scope or limits of your study (depending on the objective of the experiment)
- You define/set it
- Any conclusion that you draw applies to the population
- You will not be able to get data from population

Population vs Sample



- **Sample** is a subset of the population
- It should represent the population
- You obtain/observe them through experiment or sampling
- Any conclusion that you draw based on the sample applies to the population, and not the sample

Population vs Sample: Examples

- Suppose, you are interested to identify the average height of Leeds University students
- Define the population (Ask yourselves these questions:)
  - Leeds University students or Leeds students?
  - Current students only or current and former students?
- Suppose, you take a random sample of 10 students from the Faculty of Engineering
  - Is this a representative sample?

Data exploration

- MANDATORY!
- Why?
  - Identify shape and structure
  - Identify possible problems (e.g. missing values, outliers)
- How?
  - Numerical summaries (means, medians, standard deviations etc.)
  - Visual summaries (histograms, boxplots, scatterplots, etc.)

Types of data

- Suppose some measurements are observed or collected from your sample
- The type of data dictates what analysis is possible (bearing in mind the objective of the experiment)
- Most decision making strategies are based on assumptions about, and properties of, data
- In general, we have two main types of data: qualitative and quantitative

Quantitative data

1. Discrete
   - Integer (e.g. heart rate, count data)
   - Categorical (e.g. pain scores, likert scale)
2. Continuous
   - Height, weight, age, rainfall, temperature, ozone measurement, sulphur dioxide concentrations, etc.

Qualitative data

- Ordinal (order does matter)
- Nominal
  - Order does not matter
  - Blood group, gender,

Visual Summary

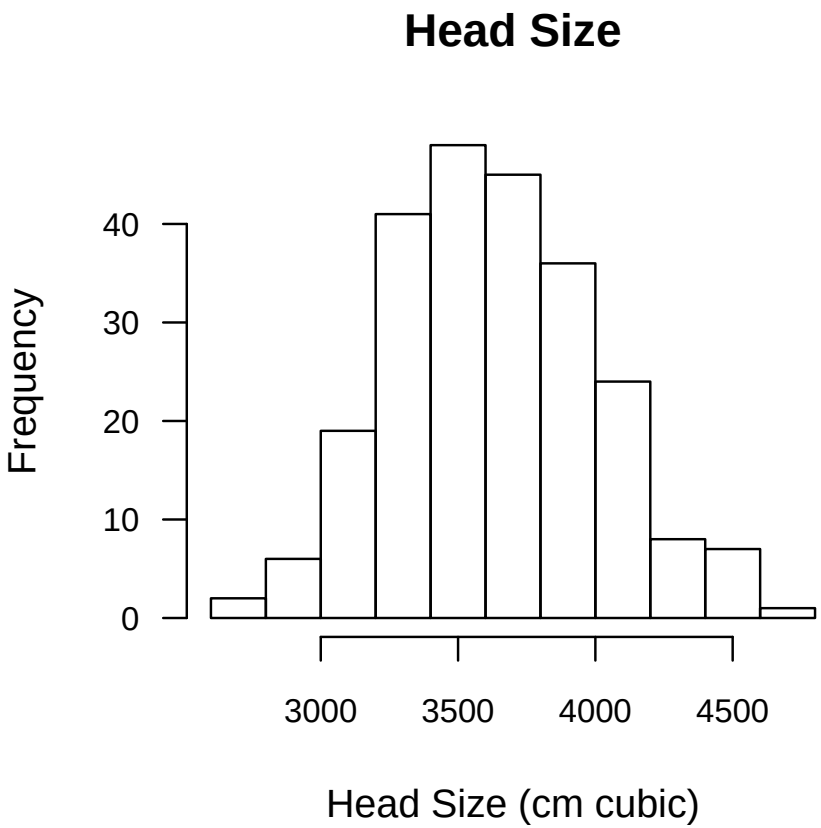
- A graph often makes **the most powerful impact** when trying to summarise data
- Summarising position, variability and shape
- Making visual comparisons between samples (E.g. Histograms and Box Plots)
- Visual exploration of relationships between two related sets of measurements (E.g. Scatter Plots)
- Identifying a plausible model (E.g. Probability Plots)

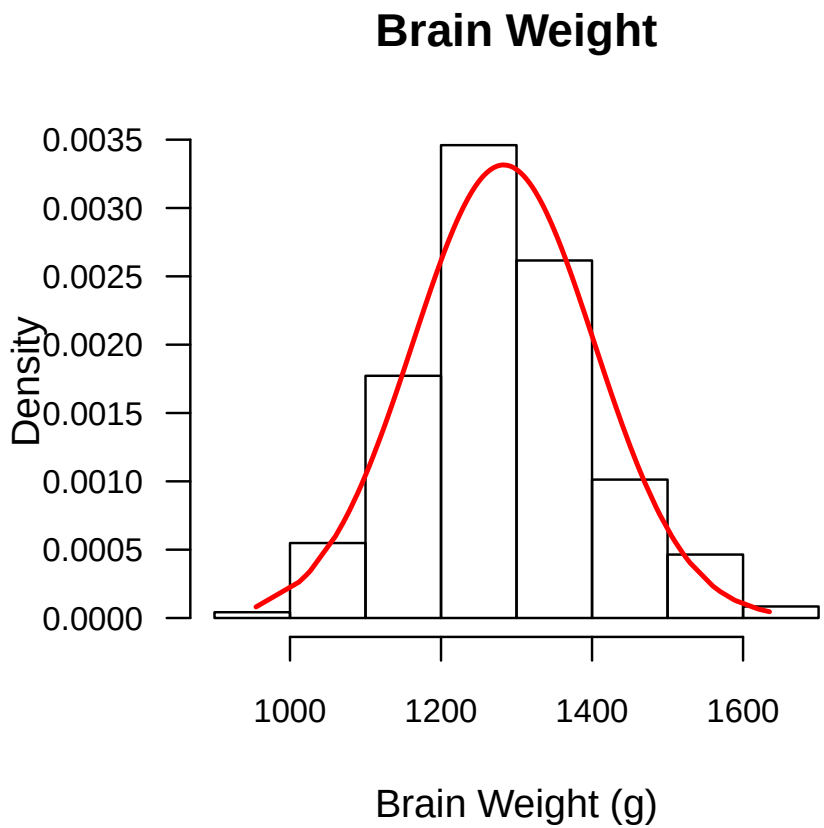
Example

- Data from R.J. Gladstone (1905). "A Study of the Relations of the Brain to to the Size of the Head", Biometrika, Vol. 4, pp105-123
- Brain weight (grams) and head size (cubic cm) for 237 adults classified by gender and age group.

Histograms

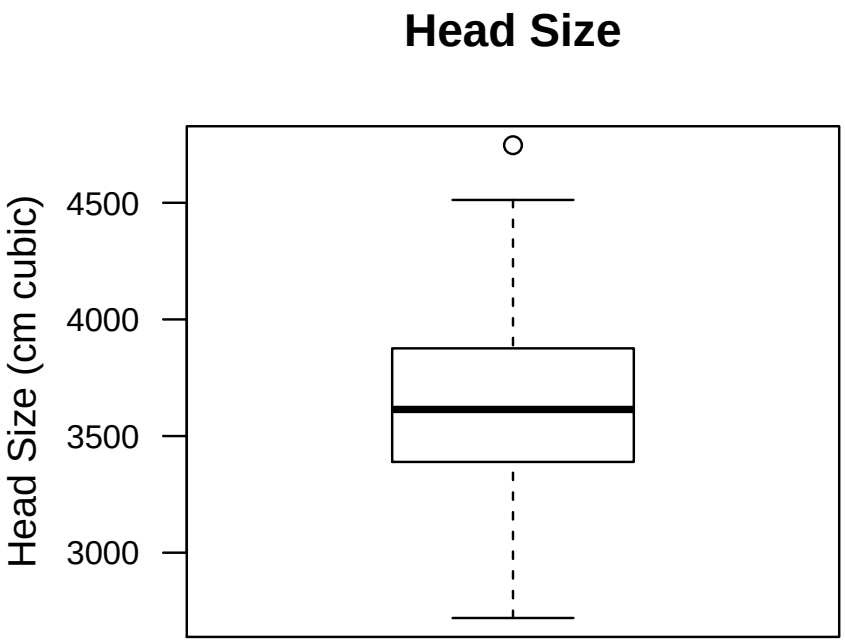
- Useful for continuous data and large samples





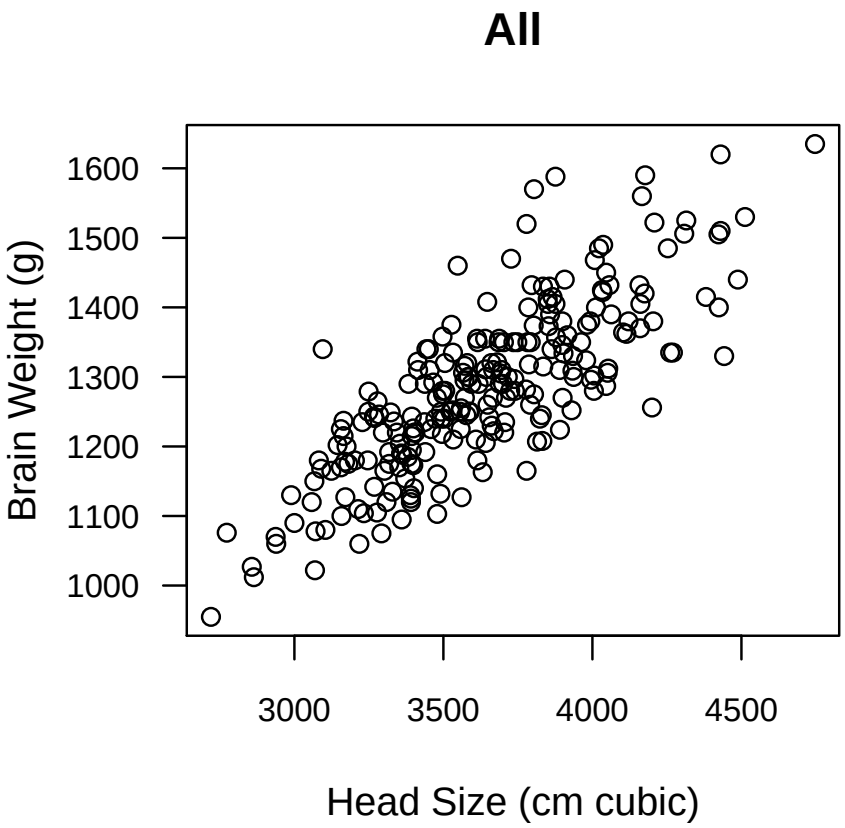
Box Plots

- Useful for small samples, and for comparing samples



Scatter Plots

- Used to identify or illustrate possible relationships between measurements made on the same individual



Numerical summary

- Frequency Distributions
- Stem and Leaf Plots
- Measures of Centre
  - Mean and Median
- Measures of Variability
  - Variance, Standard Deviation, Quartiles
- Measures of Skewness and Shape

Example: frequency distribution

- For categorical data (e.g. Gender)

```
Female   Male
   103    134
```

- For continuous data (e.g. Head Size)

```
  Head Size Freq
1 2500-2999    7
2 3000-3499   85
3 3500-3999  105
4 4000-4499   38
5 4500-4999    2
```

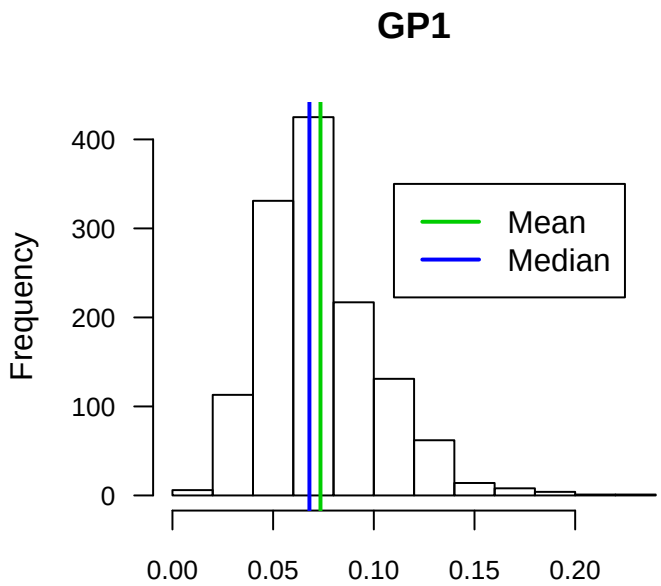
Stem and Leaf Plots

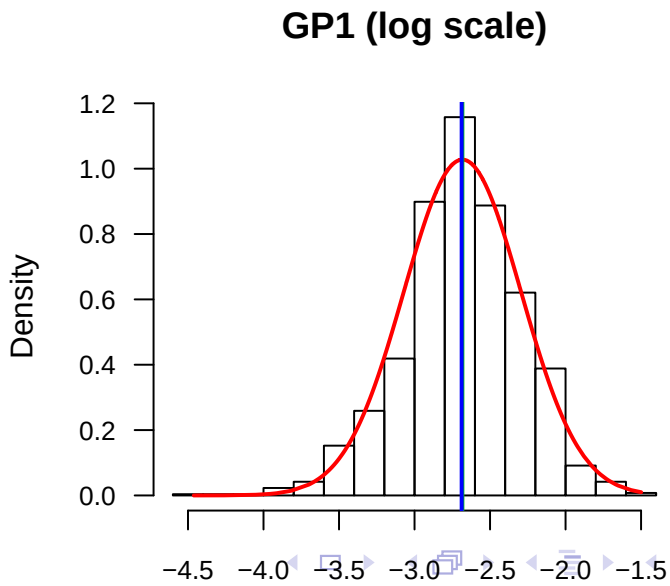
- Representation of data in a frequency diagram, which retains some information about original values
- Useful for small samples (for large samples, use histogram)

```
> stem(dat[,1])# male, 20-49 yo only

  The decimal point is 2 digit(s) to the right of the |

  30 | 07
  32 | 3
  34 | 4567901369
  36 | 14441458999
  38 | 0001368899689
  40 | 12345606688
  42 | 16712
  44 | 221
  46 | 5
```

Mean and Median

- *Mean* is basically average ('centre of gravity')
- *Median* is the middle observation
- Mean is better than median if distribution of data is close to symmetric
- Mean is distorted by outliers and hard to interpret if data are skew
- Example: Glycan released from human plasma protein separated using hydrophilic interaction high-performance liquid chromatography ($n = 1,313$)





Variance and standard deviation

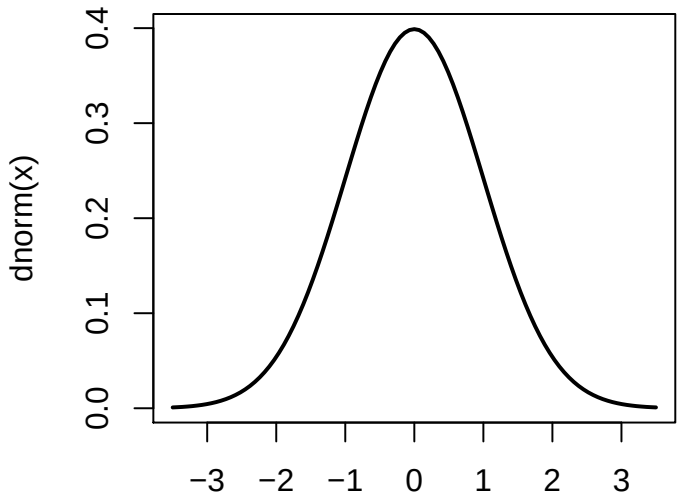
- The variance is the average of the squared deviations from the mean
- The standard deviation is the square root of the variance
- If a distribution is symmetric, approximately 95% of the data lie within two standard deviations of the mean
- In the head size example, we have a mean of 3,634, variance of 133,415.9, and standard deviation of 365.26.
- Approximately 95% of observations are expected between 2,903.5 and 4,364.5 ($3634 \pm 2 \times 365.26$)
- There are 224 observations within that interval (94.5% out of 237)

Modelling data

- Objectives:
  1. To produce a plausible way of summarising the pattern in the distribution based on some simple assumption
  2. Enable calculations to be performed on the potential behaviour of data if the model is correct

Normal distribution

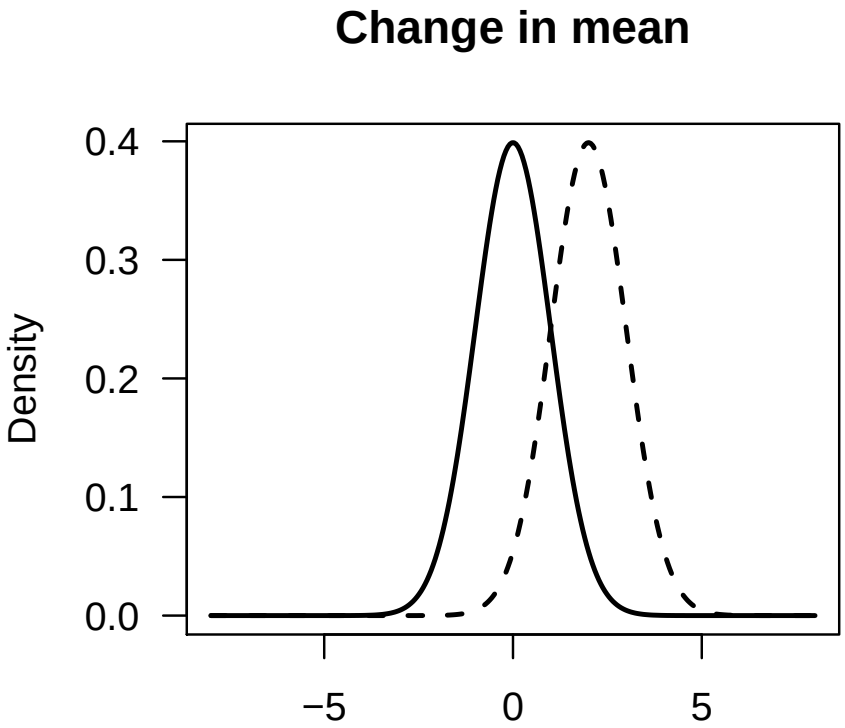
- Most common distribution used in statistics
- Properties: Symmetric and Bell-shaped

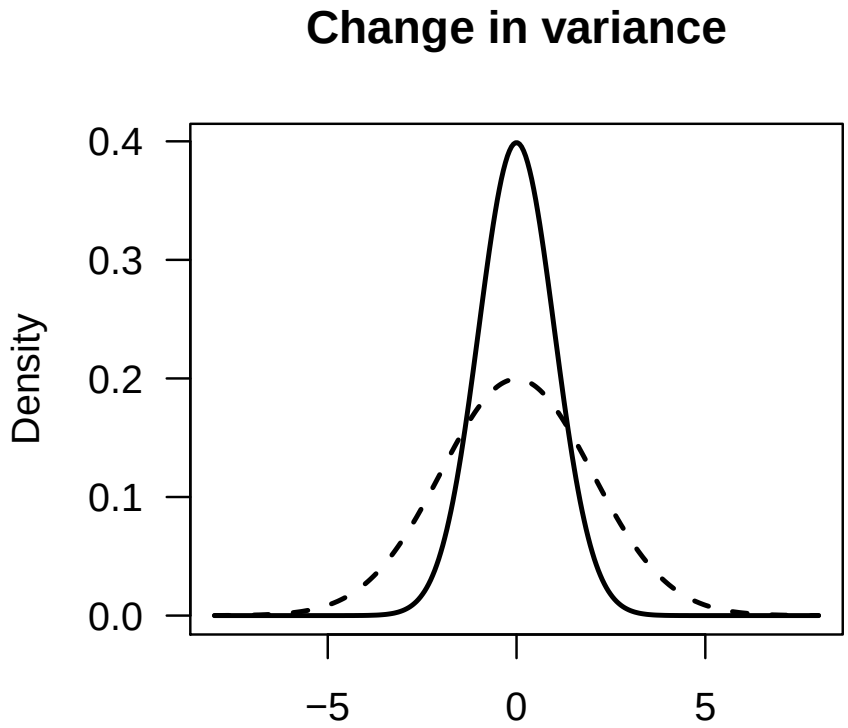


- Characterises shape of many samples gathered at random from populations
- Ideal distribution of errors

Normal distribution

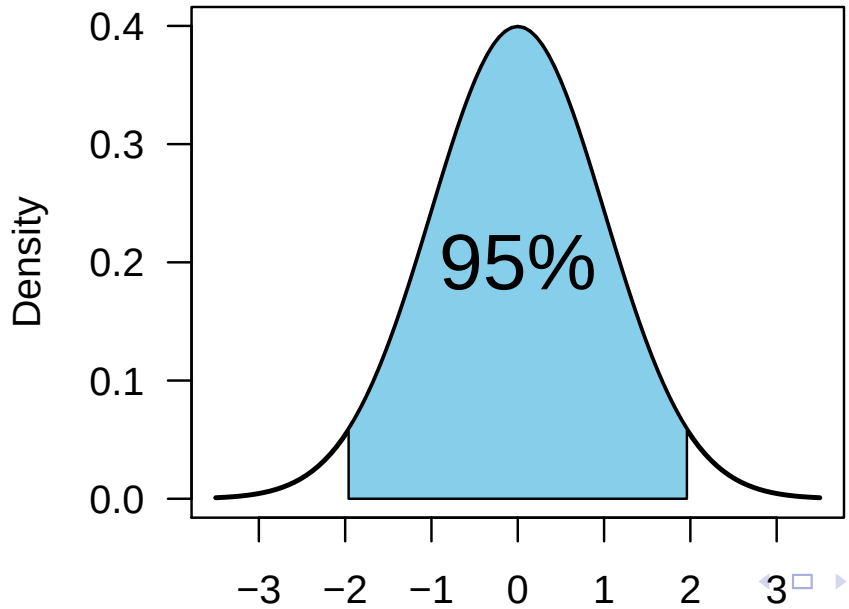
- Position and spread are characterised by two parameters: mean (μ) and variance (σ^2)
- Standardised form is Z=(observation - mean of distribution)/standard deviation to have mean zero and standard deviation one





Probabilities associated with normal distribution

- Probabilities are represented by the area under the bell-shaped curve
- For standard normal distribution, area under the curve between -1.96 and 1.96 is 95%
- In general, area under the normal curve between $\mu - 1.96\sigma$ and $\mu + 1.96\sigma$ is 95%
- The complementary areas account for 5%, 2.5% on each 'tail'



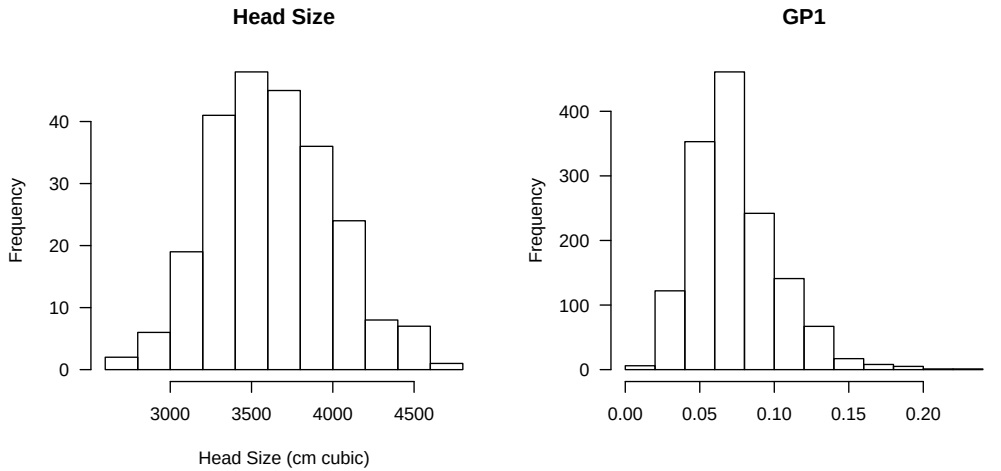
How to check that the data follow a normal dist.?

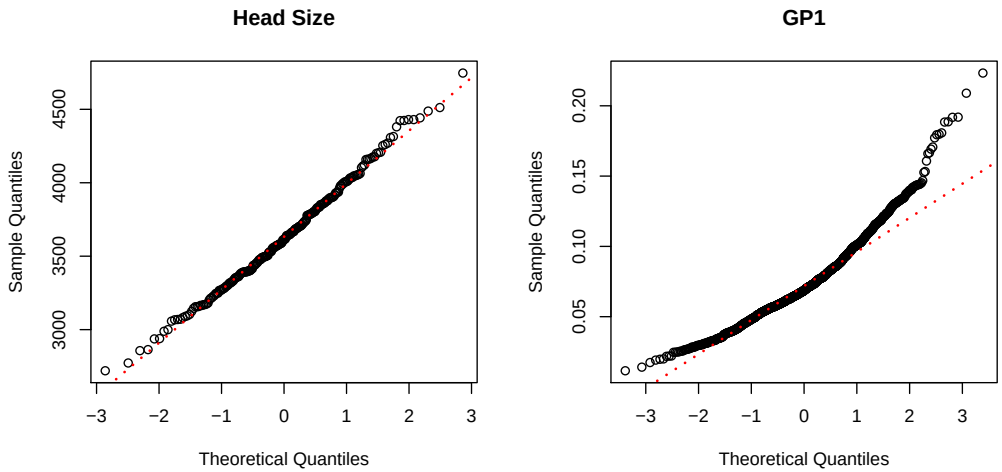
- Visual check: quantile-quantile plots (next slide)
- Formal tests: Kolmogorov-Smirnov test (sensitive), and Shapiro-Wilk test (preferred)

```
> shapiro.test(a[,3]) # Head Size

Shapiro-Wilk normality test

data:  a[, 3]
W = 0.99435, p-value = 0.5203

> ks.test(a[,3], pnorm)

One-sample Kolmogorov-Smirnov test

data:  a[, 3]
D = 1, p-value < 2.2e-16
alternative hypothesis: two-sided

Warning message:
In ks.test(a[, 3], pnorm) :
  ties should not be present for the Kolmogorov-Smirnov test
```

Normal quantile-quantile plot

- If the data are approximately normal, then we expect a straight line pattern



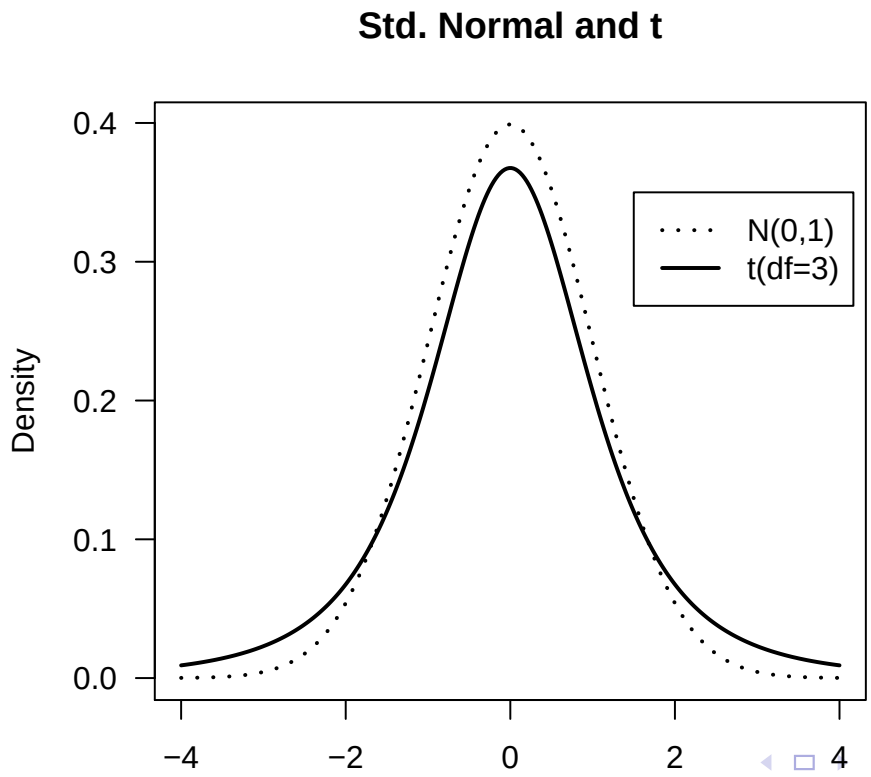


Sampling distribution and standard error

- From the sample, we calculate *estimates* of (unknown) population parameters
- For example, the mean head size is an estimate to the (unknown) mean of head size of the population
- As sample is a representation of the population, estimates differ between samples
  - There is uncertainty attached to the estimates
- The **standard error** of the mean is a measure of the variability in the mean
- It equals the sample *standard deviation* (of data) divided by the square root of the number of observations
- Note: Some papers report standard deviation and others the standard error of the mean

Confidence interval

- Confidence intervals for the sample mean assume either
  - population has a Normal distribution if sample is small, or
  - sample is sufficiently large for us to assume that distribution of sample mean is normal
- Interval is constructed using the t-distribution, with $n - 1$ *degrees of freedom*



Confidence interval

- 95% CI for the population mean is given by

$$\text{Sample mean} \pm t_{n-1}(2.5\%) \times \text{standard error}$$

  where $t_{n-1}(2.5\%)$ is the 2.5 percentile of t distribution with $n-1$ degrees of freedom
- From the head size data among the male 20-46 years old only, we have a mean of 3864.8, standard deviation 336.3, $n = 57$, and $t_{56}(2.5\%) = 2.0$
- The 95% confidence interval for the mean is given by

$$3864.8 \pm 2.0(336.3/\sqrt{57})$$

  or (3,375, 3,954)
- Interpretation: We are 95% sure that the true mean of head size in the 20-46 year old male is between 3,375 and 3,954 cm^3

Hypothesis testing

- Hypotheses are statements we want to be able to test
  - For example: The effectiveness of the new drug is better than the current one
- Formulate two hypotheses
- (1) Null Hypothesis (H_0): One on which we are able to perform statistical calculations
  - The mean response to treatment is the same for both drugs
- (2) Alternative Hypothesis (H_1): A second hypothesis which we accept if the null hypothesis is rejected
  - The mean response for the new drug is greater than that for the old drug

Making Decisions Between Hypotheses

- Assume null hypothesis is true
- Calculate an estimate of parameter we are testing
- Evaluate how far it is from value we have hypothesised (standardising the distance to allow for variability)
- If it is sufficiently far from this value, reject null hypothesis and accept the alternative

Is it significant?

- *p*-value is defined as the probability of observing an estimate further from the value under the null hypothesis than the one we have observed in our sample
- If this probability is less than 0.05 (or 5%) say, then we say that our observed estimate is sufficiently far from the value under the null hypothesis for us to reject that hypothesis
- The 5% value used here is called the *significance level*, and is the normal acceptable size of the *Type-I Error*
- Type-I error is *rejecting* the null hypothesis (H_0) when in fact it is true
- The 5% level is general convention – in omics research, there is a multiplicity problem

One-sample test

- Assumptions
  - If sample is small, assume that random sample is from a population with a Normal distribution
  - If sample is large enough ($n > 50$, say), assume random sample is from population; mean will have approximate normal distribution
- Test the *null hypothesis* H_0: Population Mean = Specified Value
- Against the *alternative hypothesis* H_1: Population Mean $\neq$ the Specified Value

- The test statistic is given by

$$\frac{\text{Sample mean} - \text{Specified value under } H_0}{\text{Standard error of mean}}$$

- This value is called the Student's t statistic, and has a t-distribution with $n - 1$ degrees of freedom
- Example: Consider the head size data, only for 20-49 year old males
- Consider the null hypothesis H_0 : (Pop.) mean $= 4,000$ cm^3 against H_1 : (Pop.) mean $\neq 4,000$ cm^3

One-sample test: Example

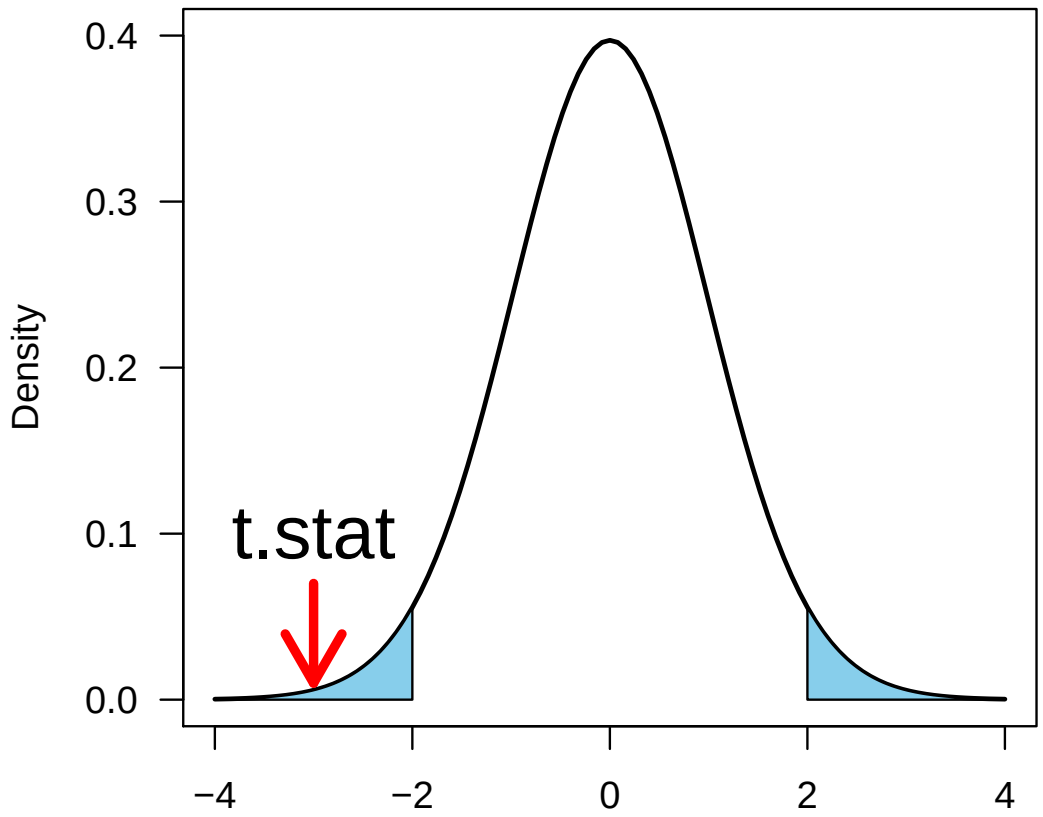
- From our previous calculation we have the mean of 3864.8, standard deviation 336.3, $n = 57$, and $t_{56}(2.5\%) = 2.0$
- The test statistic is given by

$$\frac{3864.8 - 4000}{336.3/\sqrt{57}} = -3.04$$

- In absolute value, -3.04 is greater than the critical value $t_{56}(2.5\%) = 2.0$, so we reject the null hypothesis
- Equivalently, we obtain the p-value of 0.0037 (< 0.05, the significance level) using statistical software to reject the null hypothesis
- We conclude that the true mean of head size for 20-46 year old males is not 4,000 cm^3 (at 5% significance level)

One-sample test: Example

- It is usually useful to visualise the position of test statistic



Two-sample test

There are two cases we need to consider

1. The samples are **independent**
   - Observations taken on two unrelated groups of individuals
   - For example, the effect of drinking alcohol on heart disease
   - Two groups: those who drink and those who don't
2. The samples are **related**
   - Observations taken on the same individual
   - For example, *before* and *after* an operation

Two **independent** samples

- Hypotheses involved:
  1. H_0: Mean of Pop. 1 = Mean of Pop. 2
  2. H_1: Mean of Pop. 1 $\neq$ Mean of Pop. 2
- Test statistic is given by

$$\frac{\text{Mean of Sample 1} - \text{Mean of Sample 2}}{\text{Estimated standard error of the difference}}$$

- There are several different versions of calculating the standard error of the difference and the degrees of freedom

- In `R`, the function `t.test` by default calculate Welch t statistic, where the standard error of the difference is calculated as

$$\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}$$

  where s_1^2, s_2^2, n_1, and n_2 are variance and number of observations in Sample 1 and Sample 2, respectively

- The degrees of freedom (ν) is calculated as (may not be integer!)

$$\nu \approx \frac{\left(\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}\right)^2}{\frac{s_1^4}{n_1^2(n_1-1)} + \frac{s_2^4}{n_2^2(n_2-1)}}$$

Example

- Consider two samples from the head size data: 20-46 y.o. and 46+ y.o. males

Our interest is to know whether the head size between the two groups are different

- H_0 : Mean of head size for 20-46 y.o. is equal to that of 46+ y.o. males
- H_1 : Mean of head size for 20-46 y.o. is not equal to that of 46+ y.o. males

Example

- R output is given below

```
Welch Two Sample t-test

t = 2.0267, df = 116.11, p-value = 0.04498
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
   2.637369 229.098790
sample estimates:
mean of x mean of y
 3864.842  3748.974
```

- As p-value is less than 0.05, we reject the null hypothesis
- Conclusion: The mean head size of 20-46 y.o. males is different from that of 46+ y.o. males at the 5% significance level

Two *related* samples

- When the observations in the two samples are related, (eg taken on the same individual), the test should be constructed on the observed differences within an individual
- The sample of observations consists of n differences
- Thus, our problem is reduced to a one-sample problem

Hypotheses and test statistic

- H_0 : Mean of difference is zero, against
  H_1 : Mean of difference is not zero
- Test statistic (reduces to one sample test on the *differences*)

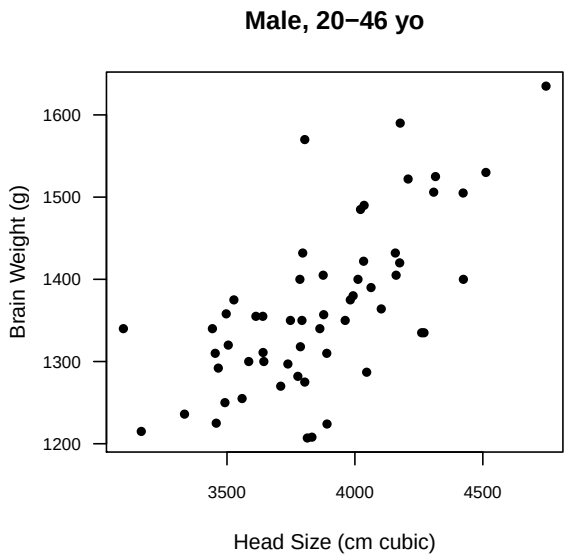
$$\frac{\text{Mean of differences } - 0}{\text{Std. Error of the mean of difference}}$$

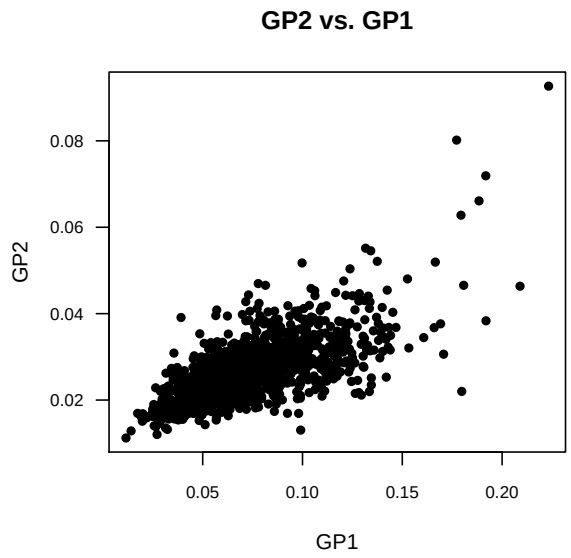
- The standard error is calculated as the (observed) standard deviation of the differences divided by square root of n
- Under the null hypothesis, the test statistic is compared to the quantile of t_{n-1} distribution

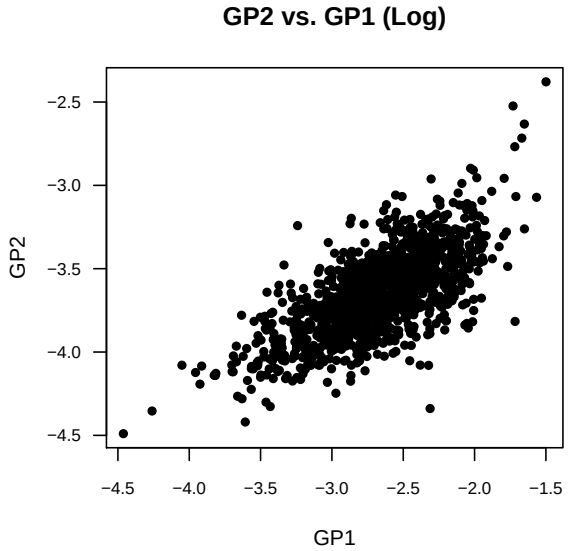
Introduction to linear regression

- This is only to introduce briefly linear regression
- The topic will be covered separately in a different course
- We are interested in the *linear* relationship between two (a pair) *continuous* variables
- Purpose: explain the dependencies or prediction
- First steps: explore the data, scatter plots (!)

Introduction to linear regression







- Suppose we have the i-th pair of data (y_i, x_i), $i = 1, 2, \ldots, n$
- Consider a *response variable* y_i
- We wish to model y_i as a function of *predictor* x_i

$$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$$

where

  - β_0 is an intercept or 'constant'
  - β_1 is a 'slope' – it determines how 'fast' any changes in x would affect the mean of y
  - ε_i is the error term

- The error term is assumed to follow a normal distribution with mean zero and a variance

Introduction to linear regression

- In the head size data, 20-46 y.o. males, suppose we model

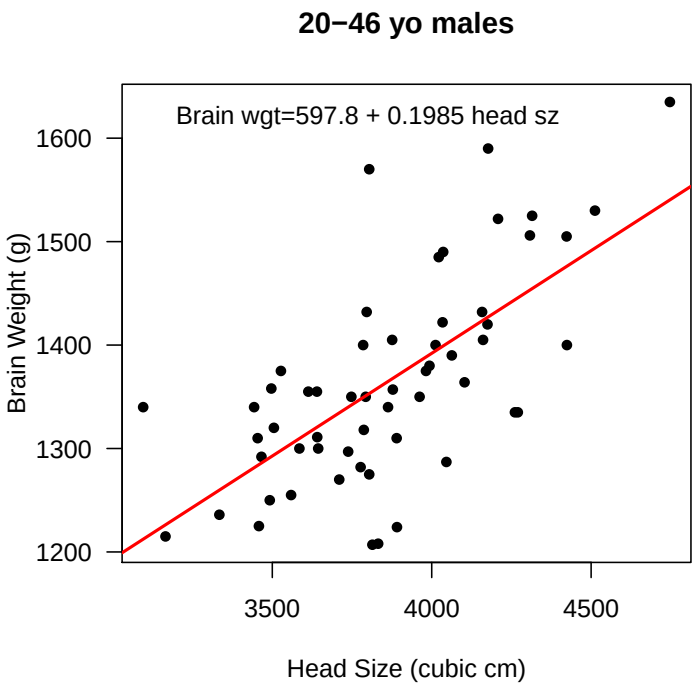
$$i\text{-th brain weight} = \beta_0 + \beta_1 \times i\text{-th head size} + i\text{-th } \varepsilon$$

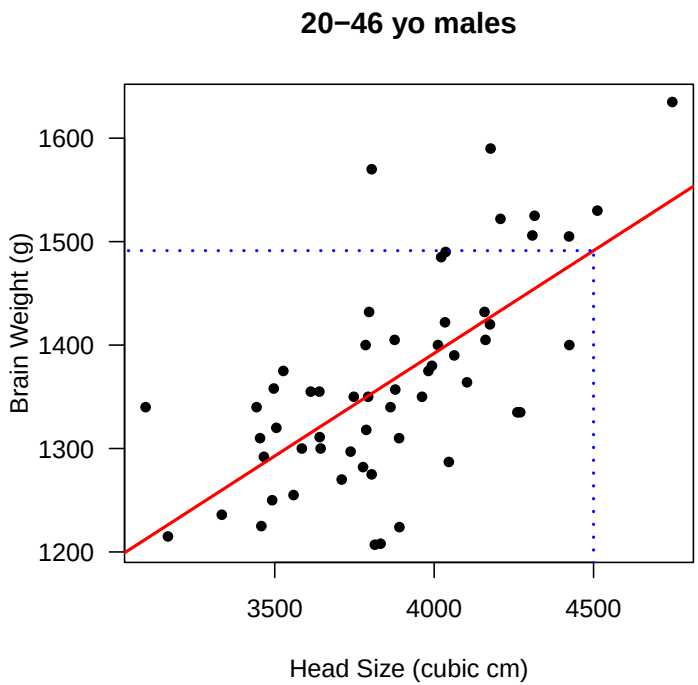
- It can be shown that the (resulting) fitted regression line is given by

$$\text{brain weight} = 597.8 + 0.1985 \times \text{head size}$$

- We can predict the mean brain weight at head size 4,500cm^3 (say) as

$$597.8 + 0.1985 \times 4500$$





Basics of experimental design

- **Objective** is critical – it determines all actions that you (should) do
  - The main objective may be expressed in several intermediate objectives
- Each step in experiment will add to variability (**accumulate**!)
  - Control is necessary
  - Otherwise the error variability will cloud any meaningful (biological) variability of interest
- Under or over design is wasteful
  - Designed at the appropriate level

Randomisation

- A process of assigning treatments to experimental units
- To protect us against the effects of factors that cannot be controlled perfectly
- Consider a research whose aim is to investigate the effect of a certain food supplement to cognition
- Suppose, 50 healthy males and 50 healthy females volunteered in the study
- You assigned the treatment (food supplement) to all of the females and placebo to all of the males
- (What is the problem?)

- You don't know whether any difference between treatment and placebo that you observed is due to the food supplement or the (participants') gender

Neuroscience and Biobehavioral Reviews 35 (2011) 565–572

Contents lists available at ScienceDirect

Neuroscience and Biobehavioral Reviews

journal homepage: www.elsevier.com/locate/neubiorev

Review

Sex bias in neuroscience and biomedical research

Annaliese K. Beery [a], Irving Zucker [b,c,*]

[a] Robert Wood Johnson Health & Society Scholar at University of California, San Francisco and University of California, Berkeley, CA, USA
[b] Department of Psychology, and Helen Wills Neuroscience Institute, University of California, 3210 Tolman Hall, 1650 Berkeley, 94720 CA, USA
[c] Department of Integrative Biology, University of California, Berkeley, 94720 CA, USA

ARTICLE INFO

Article history:
Received 4 May 2010
Received in revised form 29 May 2010
Accepted 1 July 2010

Keywords:

ABSTRACT

Female mammals have long been neglected in biomedical research. The NIH mandated enrollment of women in human clinical trials in 1993, but no similar initiatives exist to foster research on female animals. We reviewed sex bias in research on mammals in 10 biological fields for 2009 and their historical precedents. Male bias was evident in 8 disciplines and most prominent in neuroscience, with single-sex studies of male animals outnumbering those of females 5.5 to 1. In the past half-century, male bias in non-

Control

- Aim: Minimise experimental error
- Consistent handling of sample
  - Lab technician
  - Protocols
  - "Supplies"
  - Environments/ambience

Replicates

- The number of experimental units that receive each treatment
- A repetition of the basic experiment
- To have confidence that the effect we see is not due to chance
- Provide experimental error variability
- Repeated measurements on the *same* experimental unit may or may not constitute true replications
- Treating dependent observations as if they were independent is one of the most common statistical errors

END

Any Questions?