Introduction to Statistics

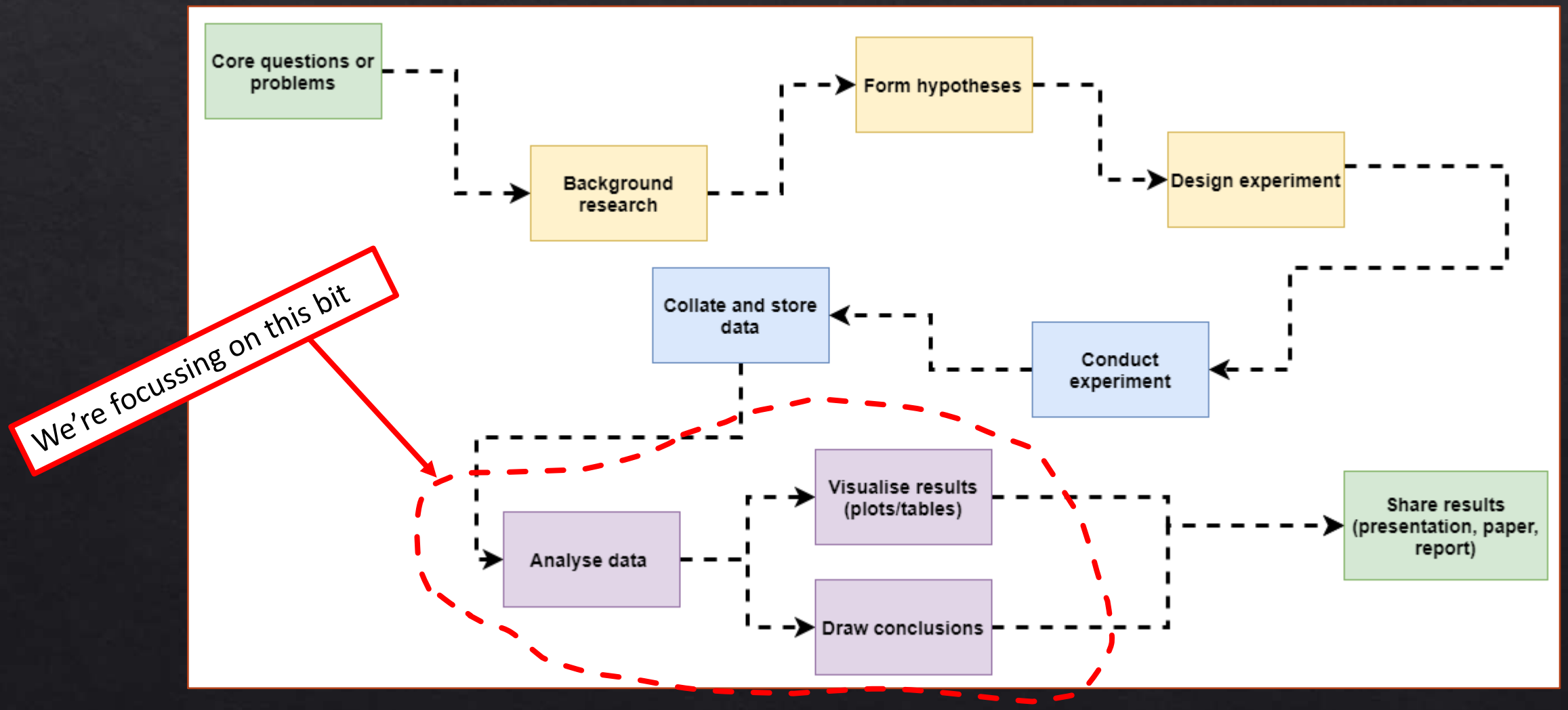
Session 1: The Hard Part is Getting Started

Institute of Zoology

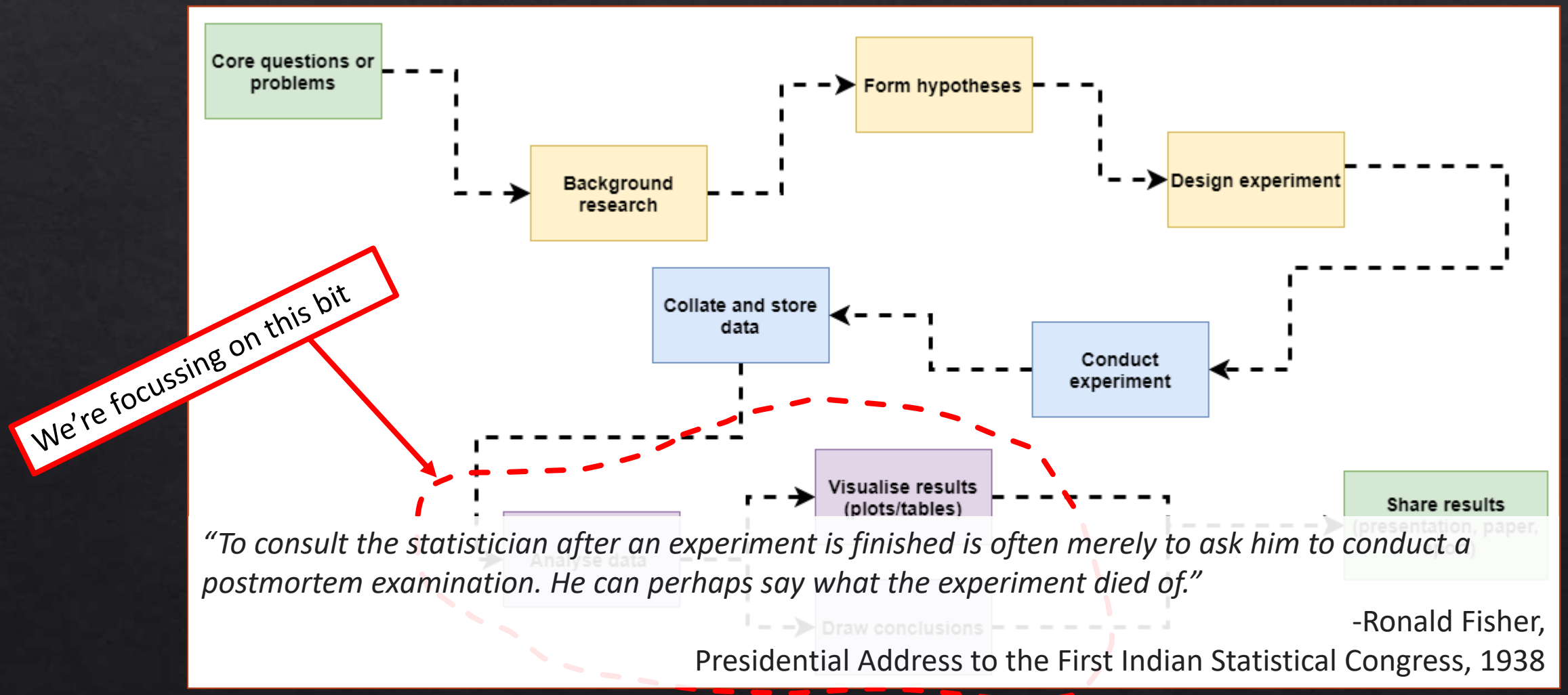
Intro

- **Henry Häkkinen**
  - Post-Doc studying the effects of climate change on seabirds and how to plan effective conservation. Latterly branched out into impacts of renewable energy on biodiversity.

What are statistics?



What are statistics?



"To consult the statistician after an experiment is finished is often merely to ask him to conduct a postmortem examination. He can perhaps say what the experiment died of."

-Ronald Fisher,
Presidential Address to the First Indian Statistical Congress, 1938

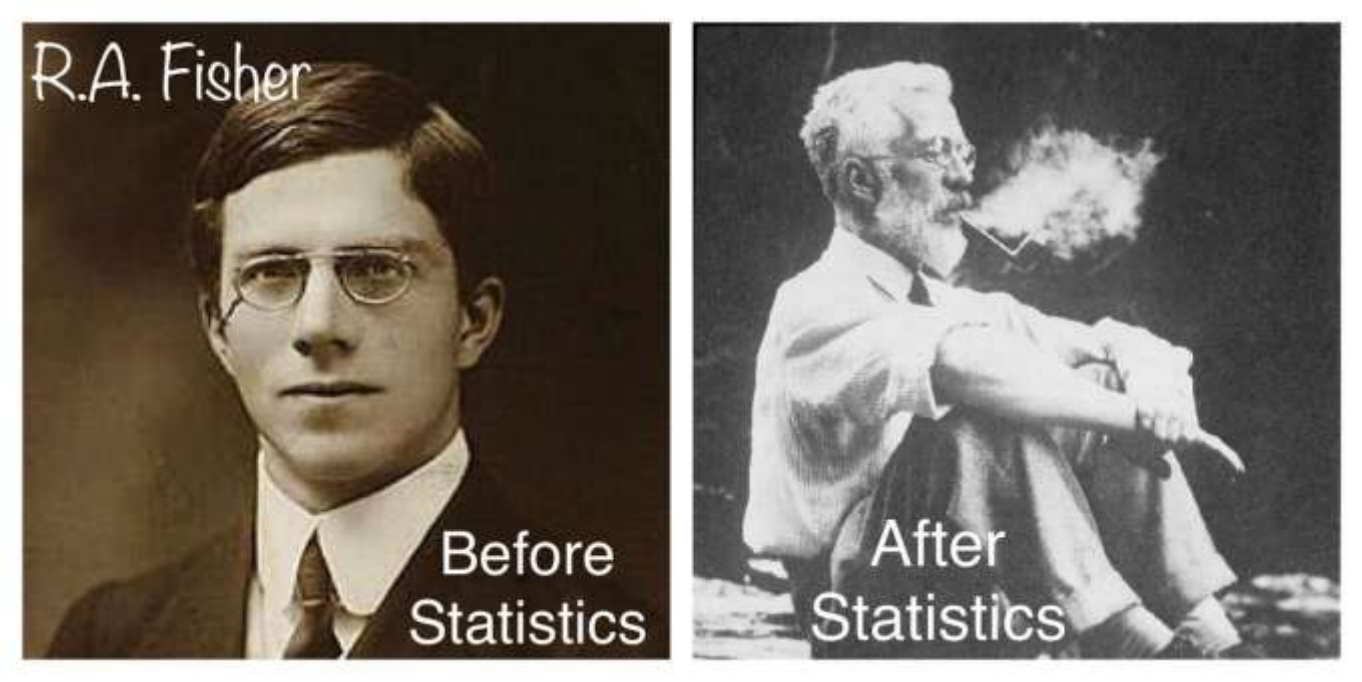
In this session…

- Introduction to statistical theory
- Creating a statistical toolkit
  - Understanding basic tests
  - Making a basic statistical workflow
  - Interpreting output
- Building statistical literacy and next steps
- Work through examples in R.

What is not in this session…

- Lots of mathematics
- Detail on lots of individual tests
- Comprehensive coverage of all statistical concepts.

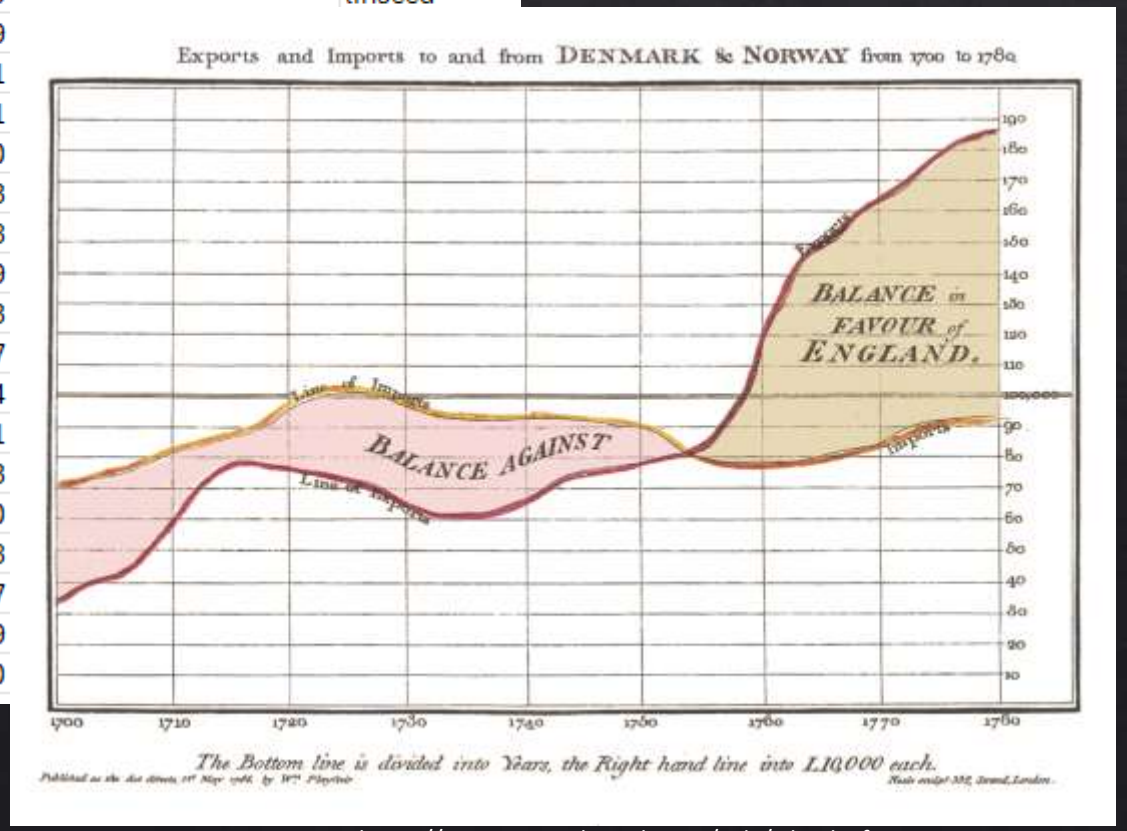
PART 1: Why do we learn statistics?



What are statistics?

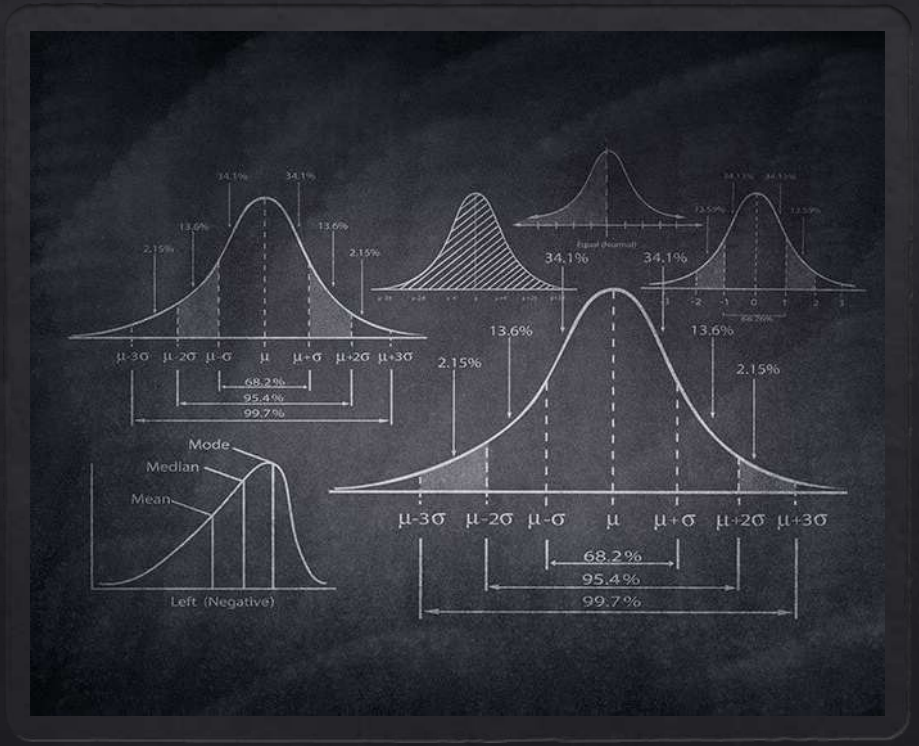
- The collection, organization, analysis, interpretation, and presentation of data
- It is not easy to make conclusions from raw data alone, even with plots
- We *could* use common sense and nothing else to make conclusions.
- BUT human judgement is prone to error and biases. We use statistics as a tool to limit the impact of our own biases.

	A	
1	weight (g)	feed
2	179	horsebean
3	160	horsebean
4	136	horsebean
5	227	horsebean
6	217	horsebean
7	168	horsebean
8	108	horsebean
9	124	horsebean
10	143	horsebean
11	140	horsebean
12	309	linseed
13	229	
14	181	
15	141	
16	260	
17	203	
18	148	
19	169	
20	213	
21	257	
22	244	
23	271	
24	243	
25	230	
26	248	
27	327	
28	329	
29	250	



https://commons.wikimedia.org/wiki/File:Playfair_TimeSeries-2.png

What are statistics?



- In science, we often use “statistics” as short-hand for “inferential statistics” which examines properties of data in relation to underlying probability distributions
  - This includes hypothesis testing! And most forms of classical models!
- “Descriptive statistics” summarise data in a compact, easily understood fashion
  - Include things like mean, median, mode
  - They do not tell us anything about probability or confidence

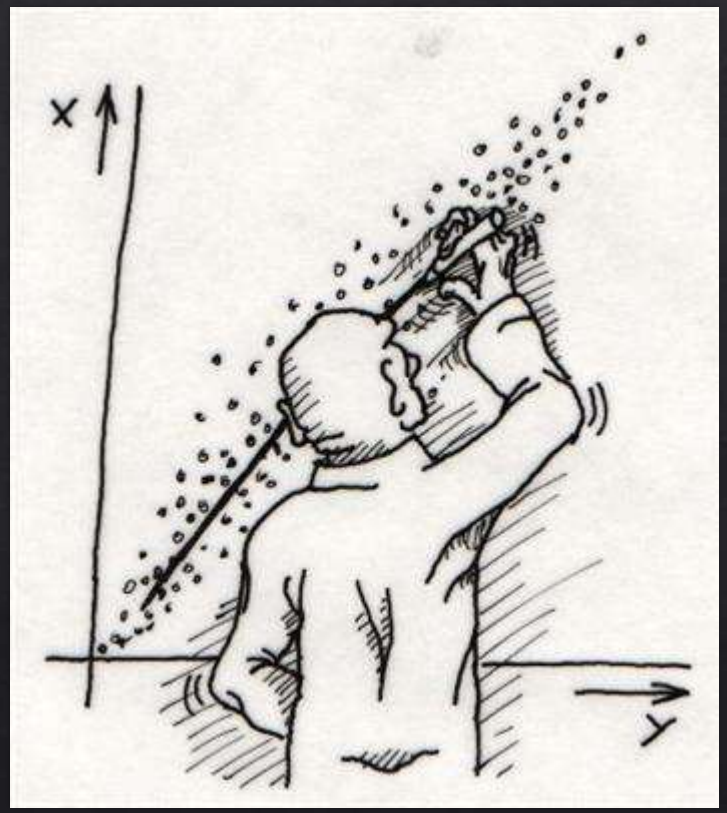
Why do we learn statistics?

- Literacy is needed to understand existing literature
- Bad practice is common; you need to be able to spot it!
- Helps you design experiments and data collection.
- Ecological data tends to be messy, and having a lot of available “tools” makes analysis a lot easier

The Good News

- Nearly all statistical tests follow shared fundamental principles
- Learning your first few tests can be difficult, but over time this gets easier and easier.
- Building a solid theoretical base will make your life a LOT easier later on.

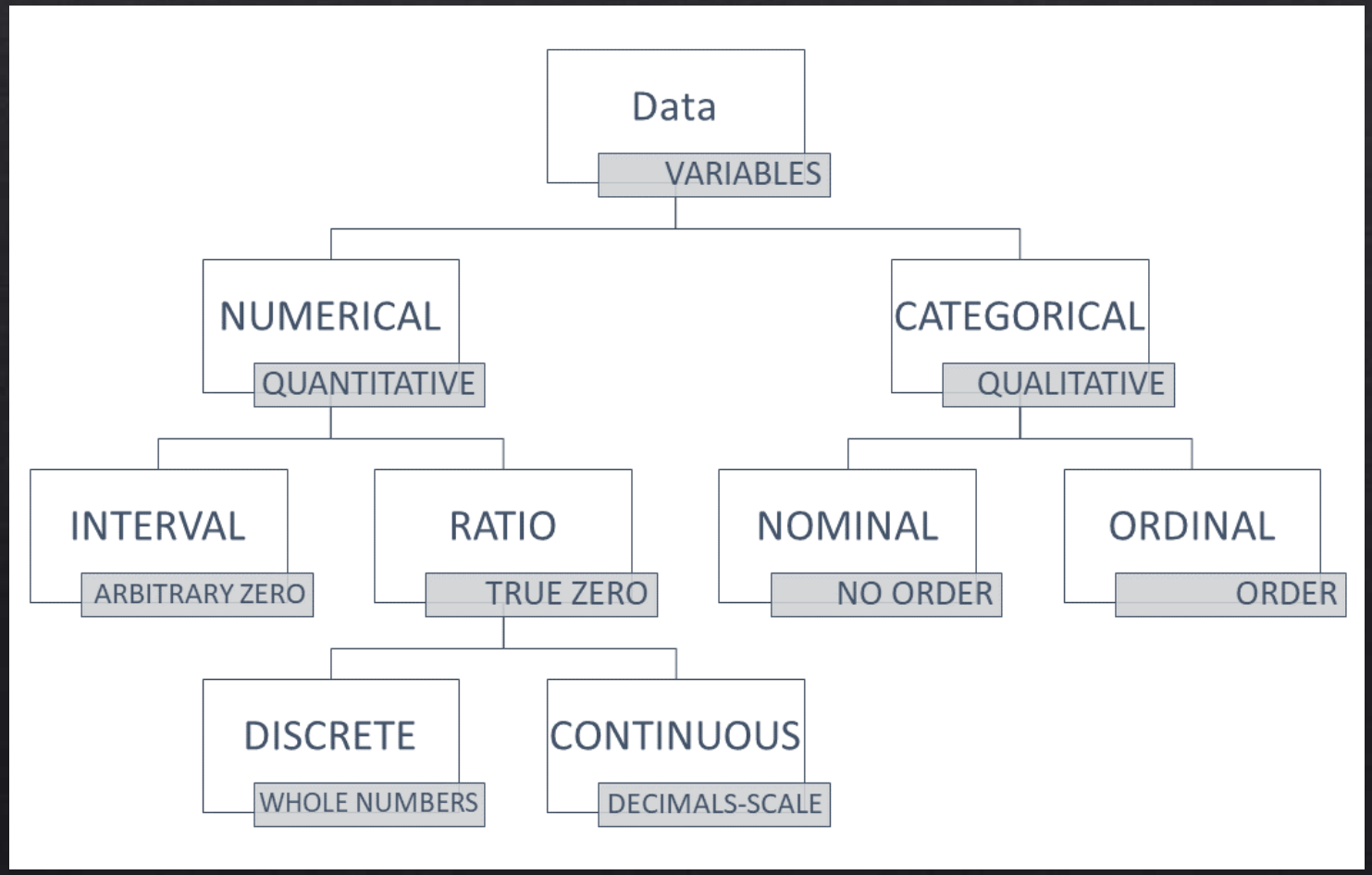
PART 2: Basic Theory



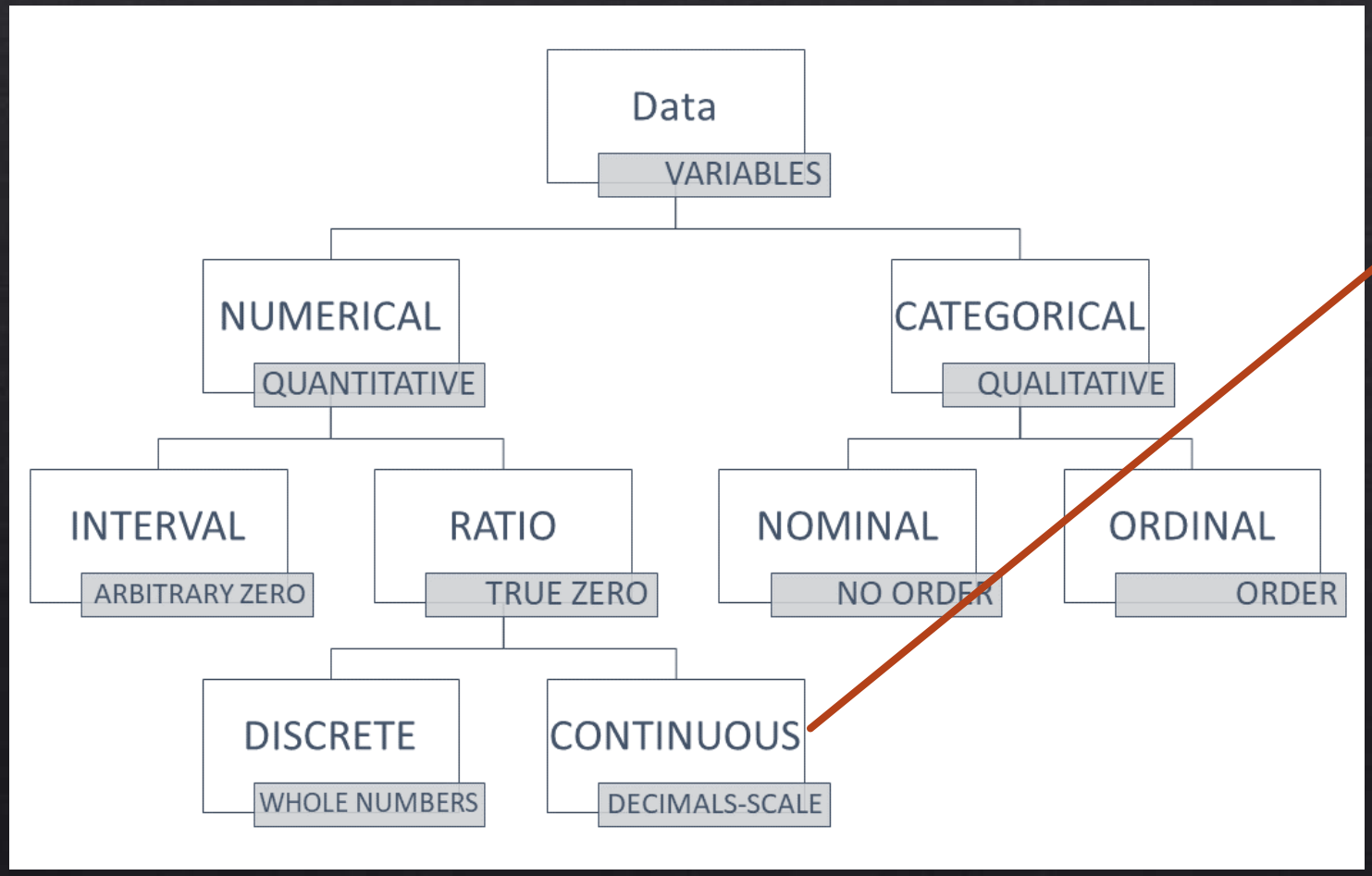
Before you run your test

- Understand your data
- Understand your hypothesis
- Select a sensible statistical framework

Types of Data



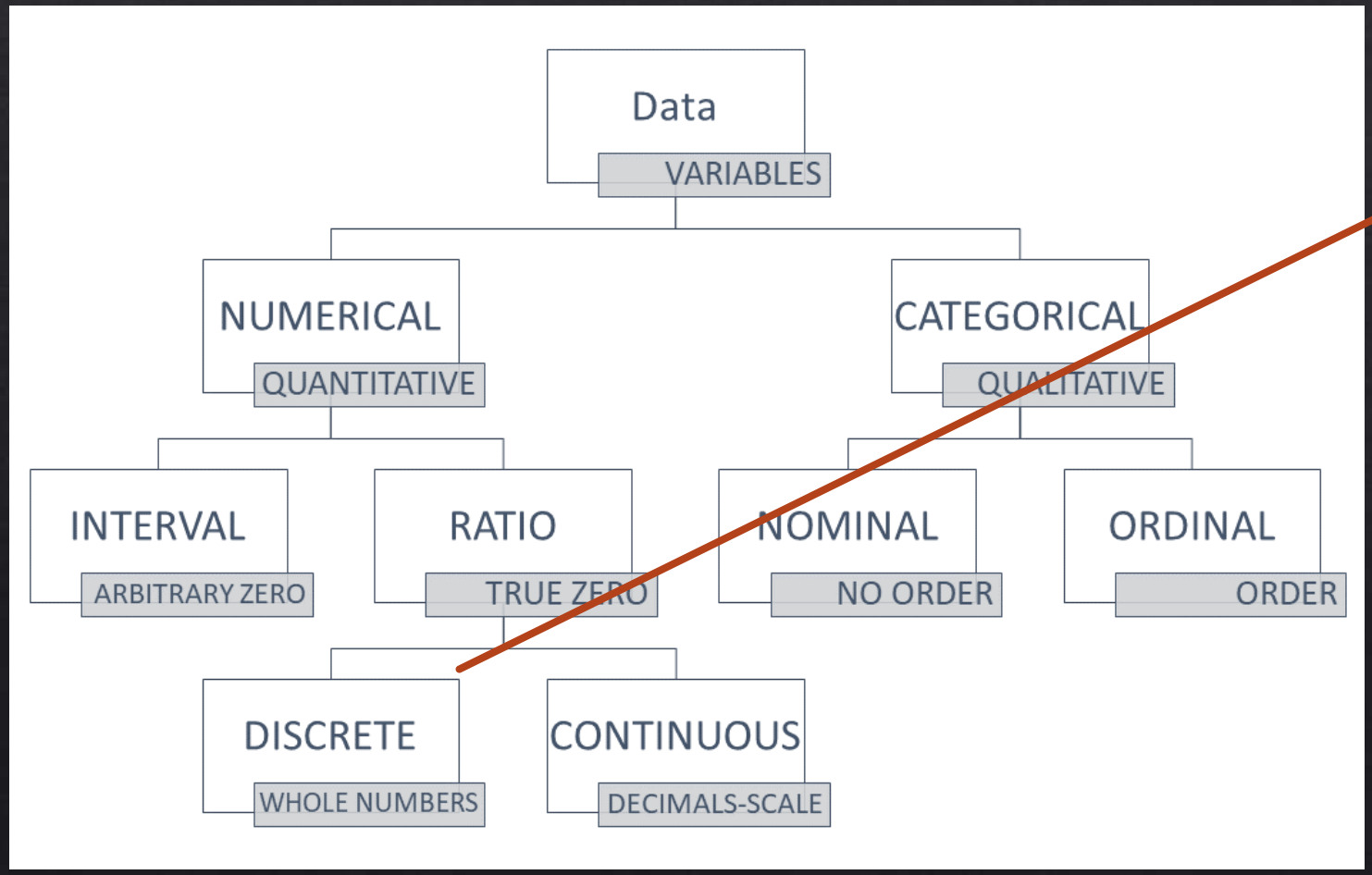
Types of Data



Can take any numerical value, including decimals. May be bounded at 0.

- Length
- Height
- Distance
- Speed

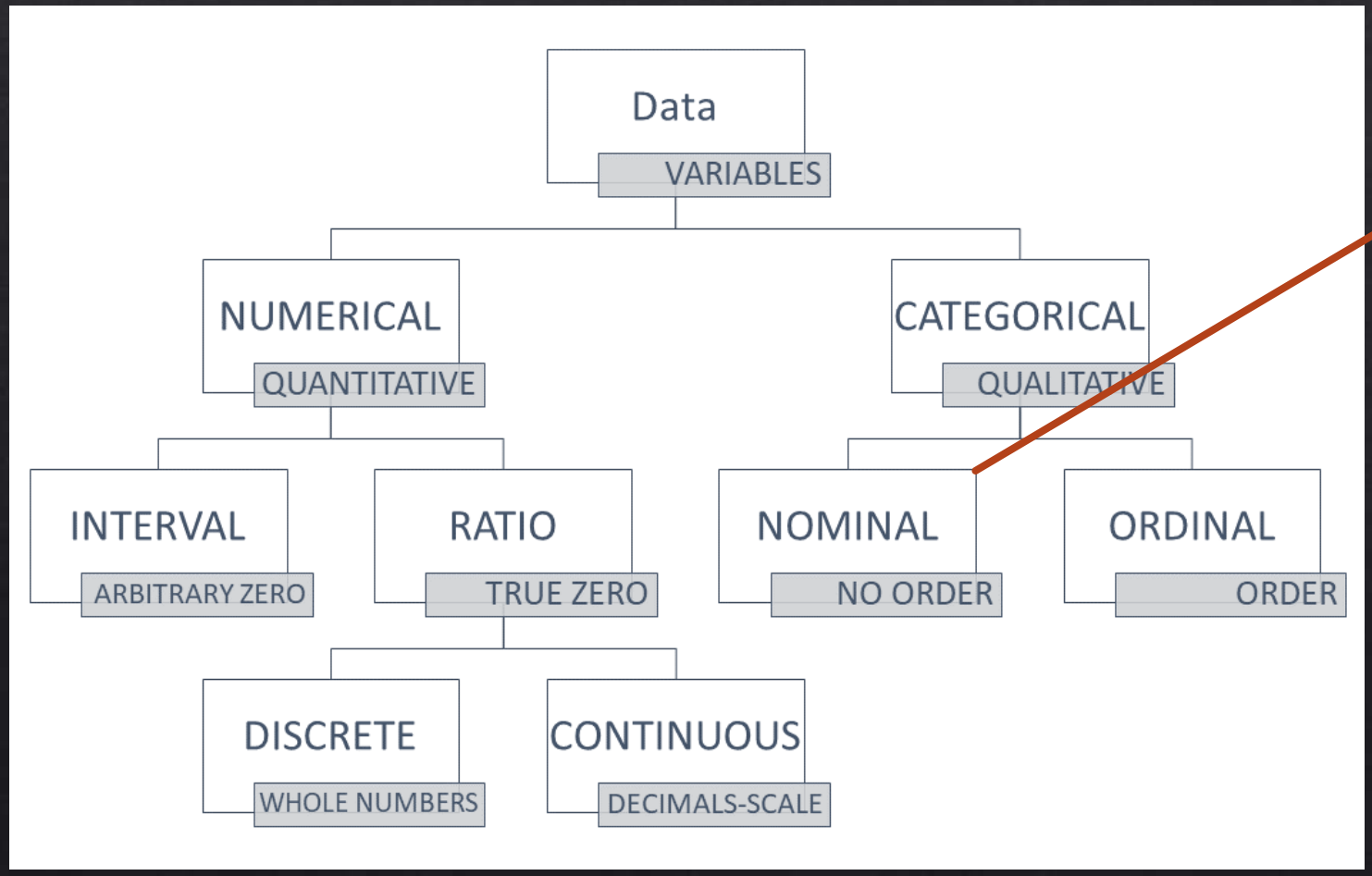
Types of Data



Can take any numerical value, but must be whole number. “Integer” in coding terminology

- Count data
- Number of days/years
- Shoe size

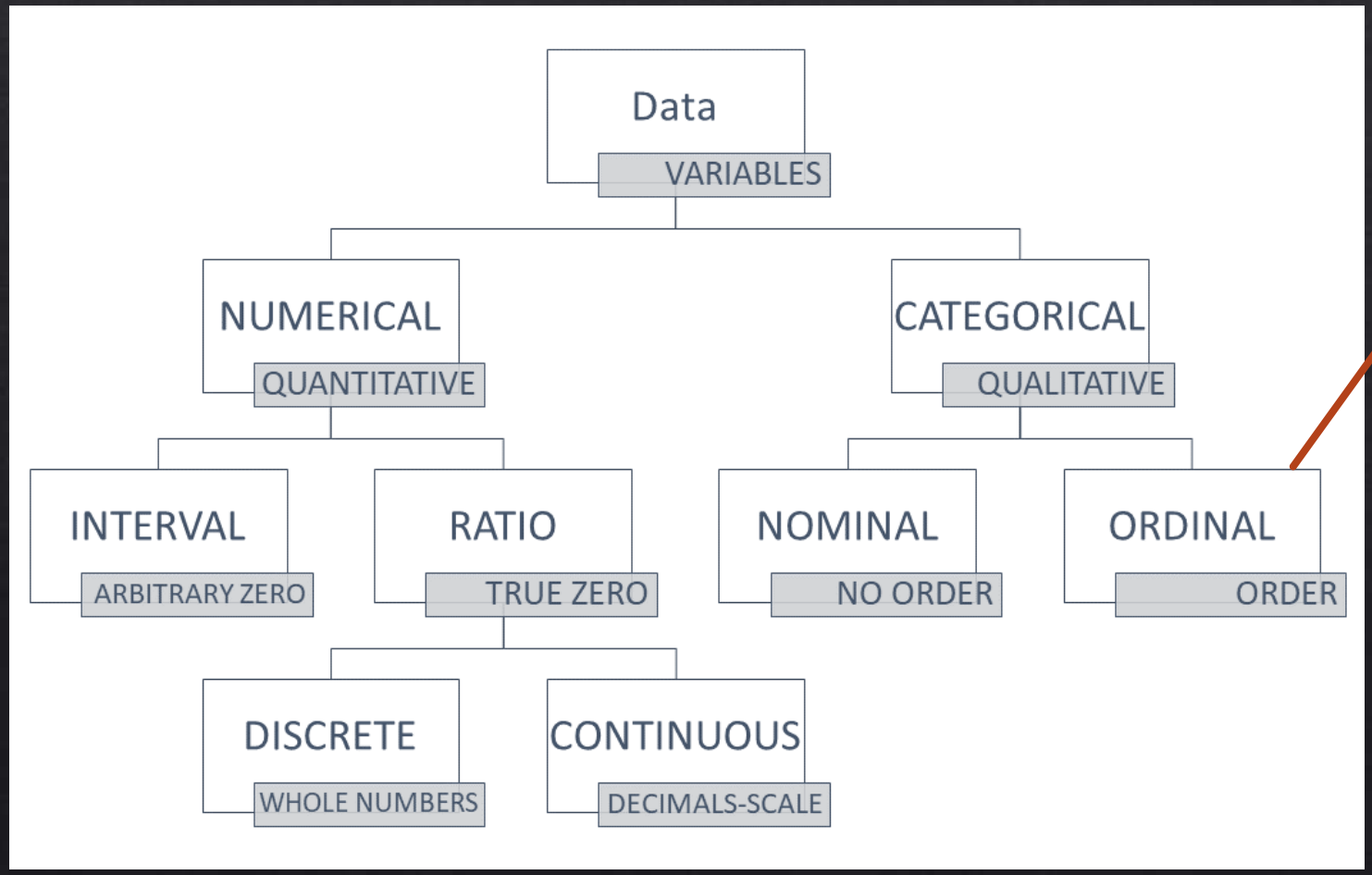
Types of Data



Categories that are functionally independent and mutually exclusive

- Sites
- Eye colour
- Favourite food
- Blood type
- Success/Failure (1/0) binary is a special case!

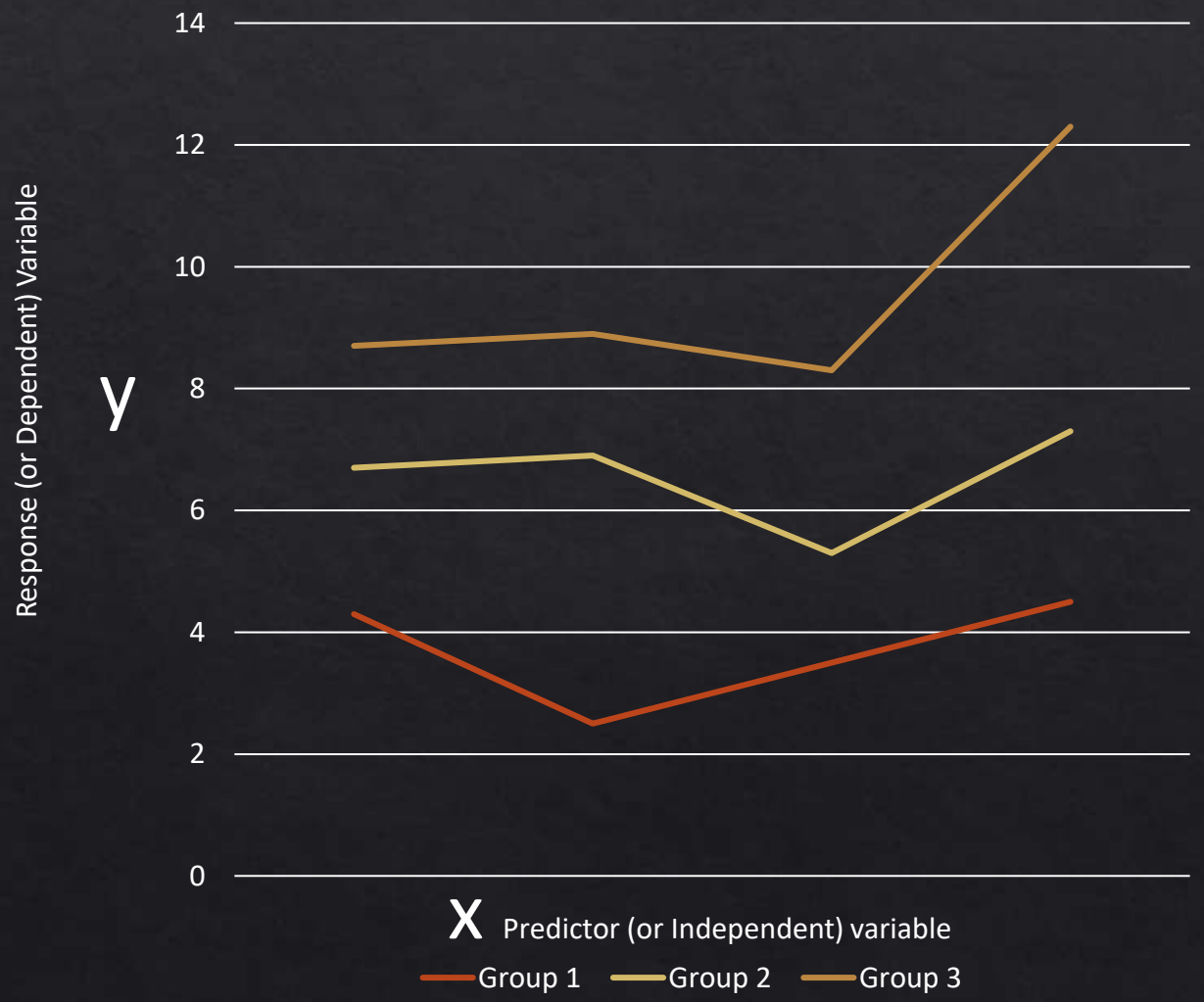
Types of Data



Categories that have a clear order

- Highest qualification (school to PhD)
- Likert scale (Strongly Disagree to Agree)
- Rank

Basic Notation



- Standard notation is to have our response variable (what we are measuring) on the y axis
- Our predictor (what we are varying/controlling) is shown on the x axis
- This plot also has a second, categorical predictor named “group”
- Standard notation for this statistical model would be:
  - y ~ x + group

ACTIVITY

- We gather a number of radioactive spiders of different species, each of which vary in size, radioactivity, age and leg number. Each volunteer is bitten by a radioactive spider, we monitor them, and we observe any superpowers they develop
- QUESTIONS:
  - what is our response variable? What are some reasonable predictors?
  - What type of data is our response variable?

Descriptive Statistics

- Raw data is often difficult to interpret or explain easily.
- Descriptive statistics provide easy to understand summaries and overviews

	A	
1	weight (g)	feed
2	179	horsebean
3	160	horsebean
4	136	horsebean
5	227	horsebean
6	217	horsebean
7	168	horsebean
8	108	horsebean
9	124	horsebean
10	143	horsebean
11	140	horsebean
12	309	linseed
13	229	linseed
14	181	linseed
15	141	linseed
16	260	linseed
17	203	linseed
18	148	linseed
19	169	linseed
20	213	linseed
21	257	linseed
22	244	linseed
23	271	linseed
24	243	soybean
25	230	soybean
26	248	soybean
27	327	soybean
28	329	soybean
29	250	soybean

Descriptive Statistics

- There are many ways to summarise data, but two very common methods are to study averages and variance
- Averages (central tendencies) include mean, median, and mode
- Variance describes the spread of data around the average (typically the mean).
- Other key descriptions include range, interquartile range and various other quartiles.

	A	
1	weight (g)	feed
2	179	horsebean
3	160	horsebean
4	136	horsebean
5	227	horsebean
6	217	horsebean
7	168	horsebean
8	108	horsebean
9	124	horsebean
10	143	horsebean
11	140	horsebean
12	309	linseed
13	229	linseed
14	181	linseed
15	141	linseed
16	260	linseed
17	203	linseed
18	148	linseed
19	169	linseed
20	213	linseed
21	257	linseed
22	244	linseed
23	271	linseed
24	243	soybean
25	230	soybean
26	248	soybean
27	327	soybean
28	329	soybean
29	250	soybean

Descriptive Statistics

- *Variance* of the data is simply how far (on average) is each observation from the mean (squared).
  - 6095.503
- Very mathematically useful, but completely impossible to interpret.
- So we use the *Standard Deviation,* which is the square root of the Variance
  - 78.07

$$\frac{1}{N-1}\sum_{i=1}^{N}\left(X_i-\bar{X}\right)^2$$

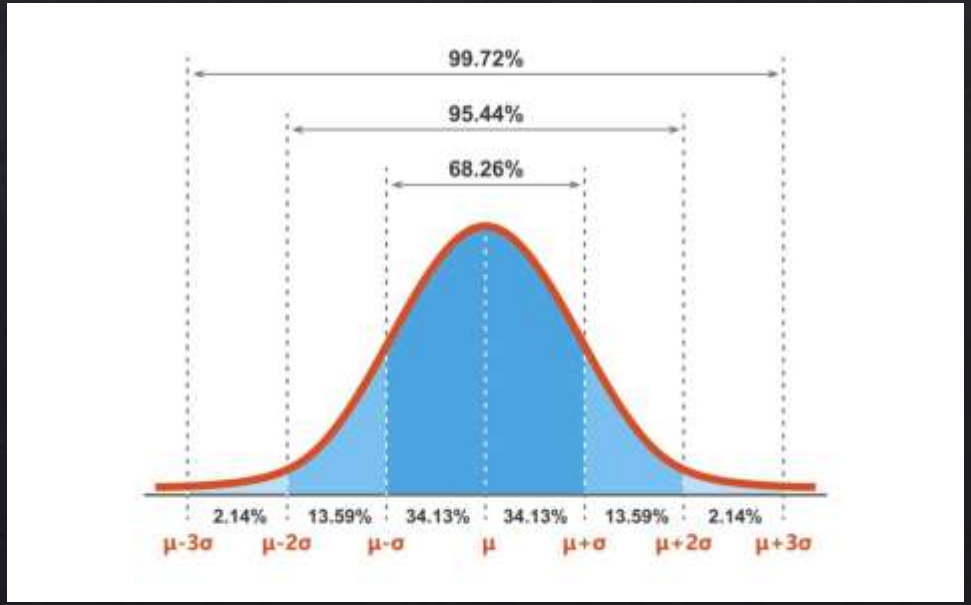
$$\hat{\sigma}=\sqrt{\frac{1}{N-1}\sum_{i=1}^{N}\left(X_i-\bar{X}\right)^2}$$

weight (g)	Mean	Deviance
179	282.6	-103.6
160	282.6	-122.6
136	282.6	-146.6
227	282.6	-55.6
217	282.6	-65.6
168	282.6	-114.6
108	282.6	-174.6
124	282.6	-158.6
143	282.6	-139.6
140	282.6	-142.6
309	282.6	26.4
229	282.6	-53.6
181	282.6	-101.6
141	282.6	-141.6
260	282.6	-22.6
203	282.6	-79.6
148	282.6	-134.6
169	282.6	-113.6
213	282.6	-69.6
257	282.6	-25.6
244	282.6	-38.6
271	282.6	-11.6
243	282.6	-39.6
230	282.6	-52.6
248	282.6	-34.6
327	282.6	44.4
329	282.6	46.4

Descriptive Statistics

- The Standard Deviation has some very useful properties, especially if your data resembles a bell curve
  - 68% of data will be within 1 SD of mean
  - 95% of data will be within 2 SD

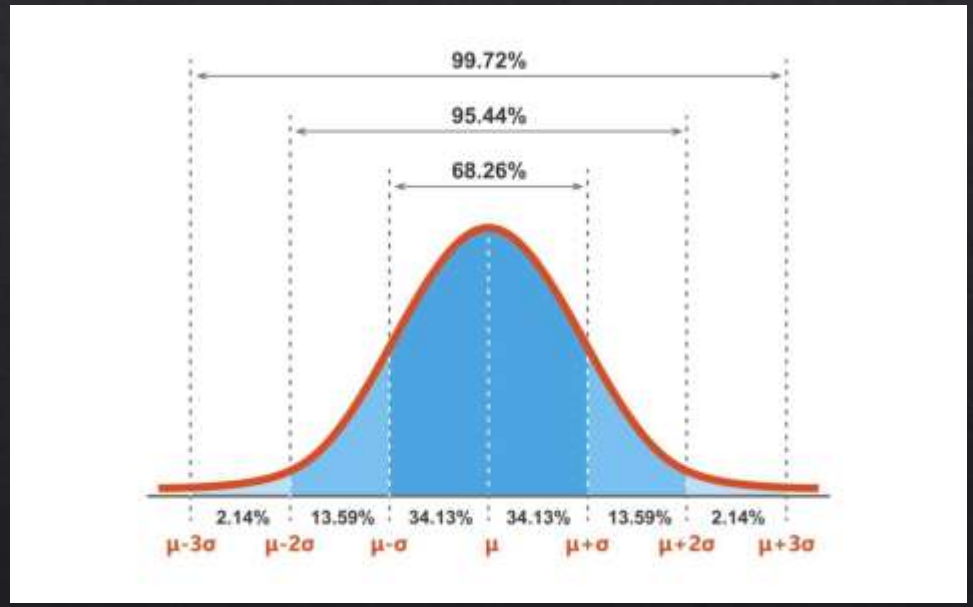
$$\hat{\sigma} = \sqrt{\frac{1}{N-1}\sum_{i=1}^{N}(X_i - \bar{X})^2}$$



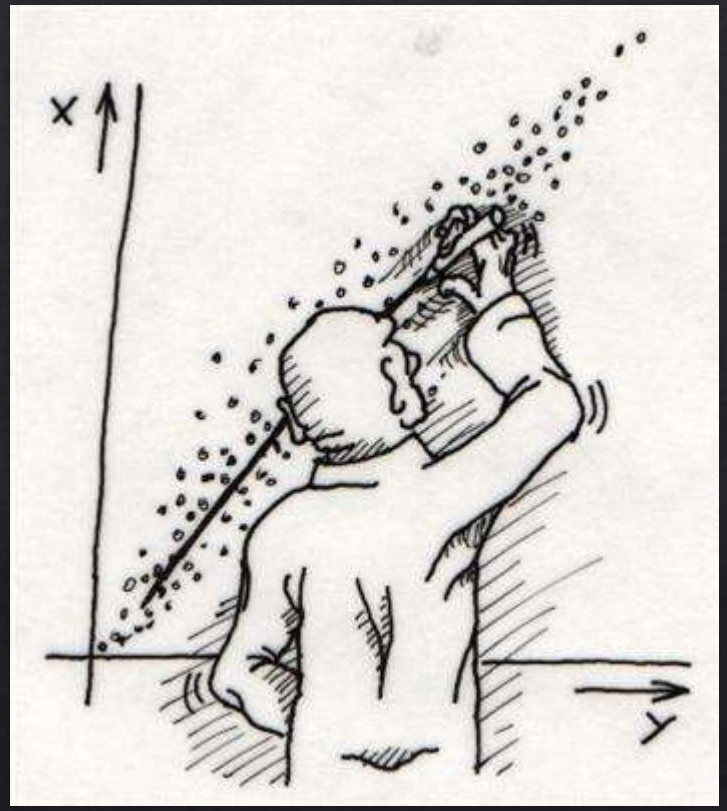
weight (g)	Mean	Deviance
179	282.6	-103.6
160	282.6	-122.6
136	282.6	-146.6
227	282.6	-55.6
	282.6	-65.6
	282.6	-114.6
	282.6	-174.6
	282.6	-158.6
	282.6	-139.6
140	282.6	-142.6
309	282.6	26.4
229	282.6	-53.6
181	282.6	-101.6
141	282.6	-141.6
260	282.6	-22.6
203	282.6	-79.6
148	282.6	-134.6
169	282.6	-113.6
213	282.6	-69.6
257	282.6	-25.6
244	282.6	-38.6
271	282.6	-11.6
243	282.6	-39.6
230	282.6	-52.6
248	282.6	-34.6
327	282.6	44.4
329	282.6	46.4

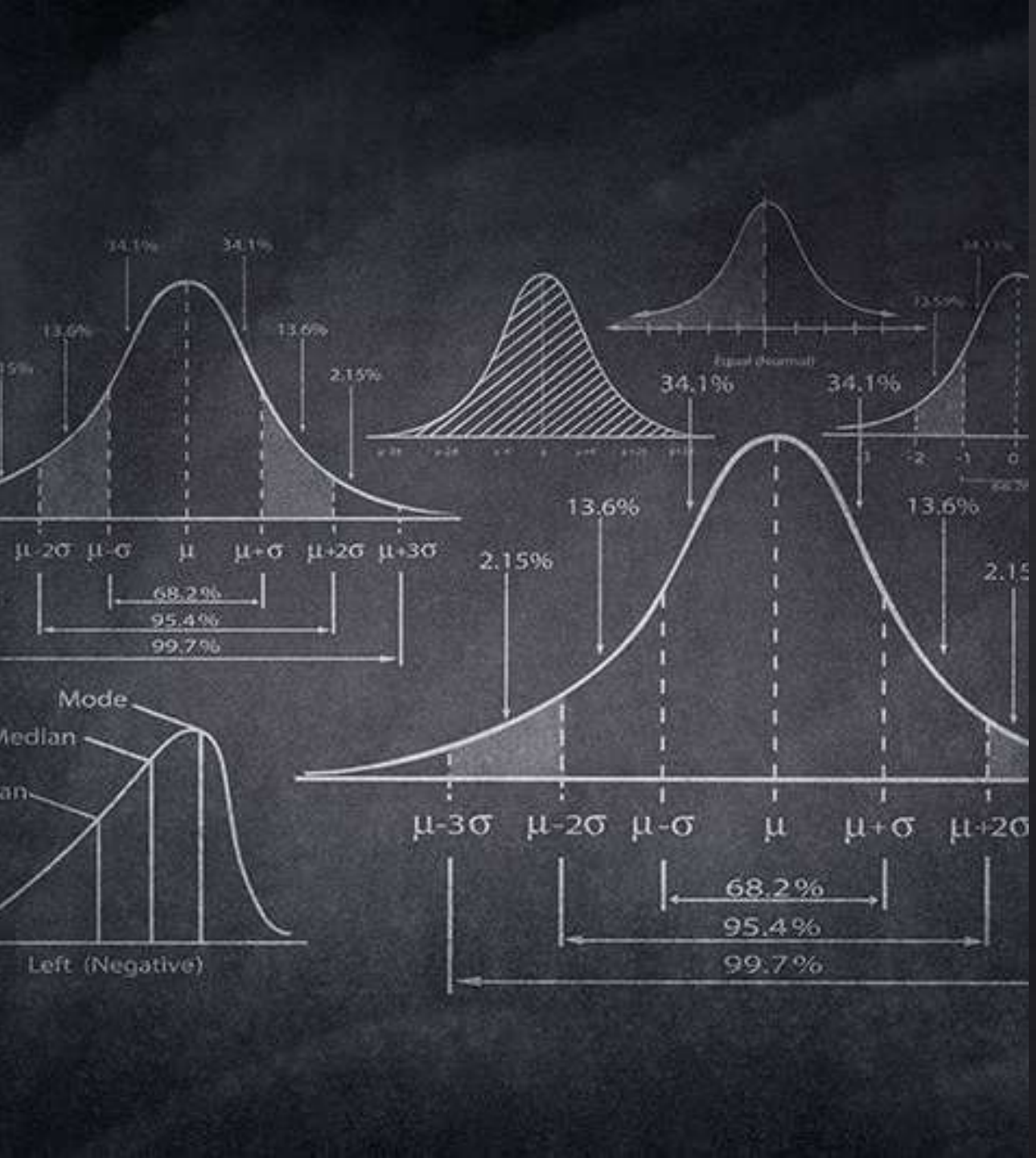
Descriptive Statistics

- Intuitive description:
  - A low SD means most individuals are close to the mean
  - A high SD means most individuals are very different to the mean



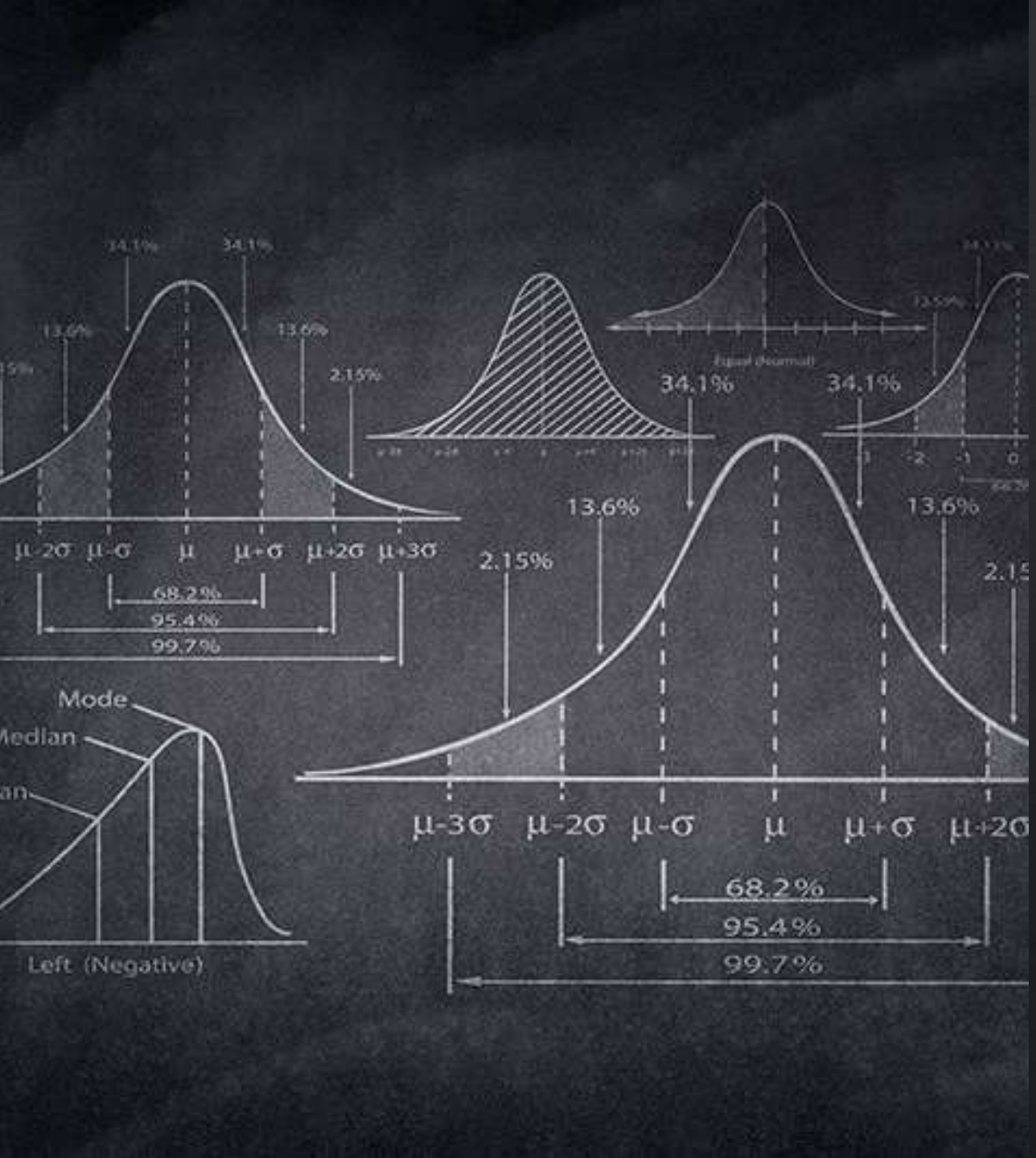
PART 3: Probability and Distributions





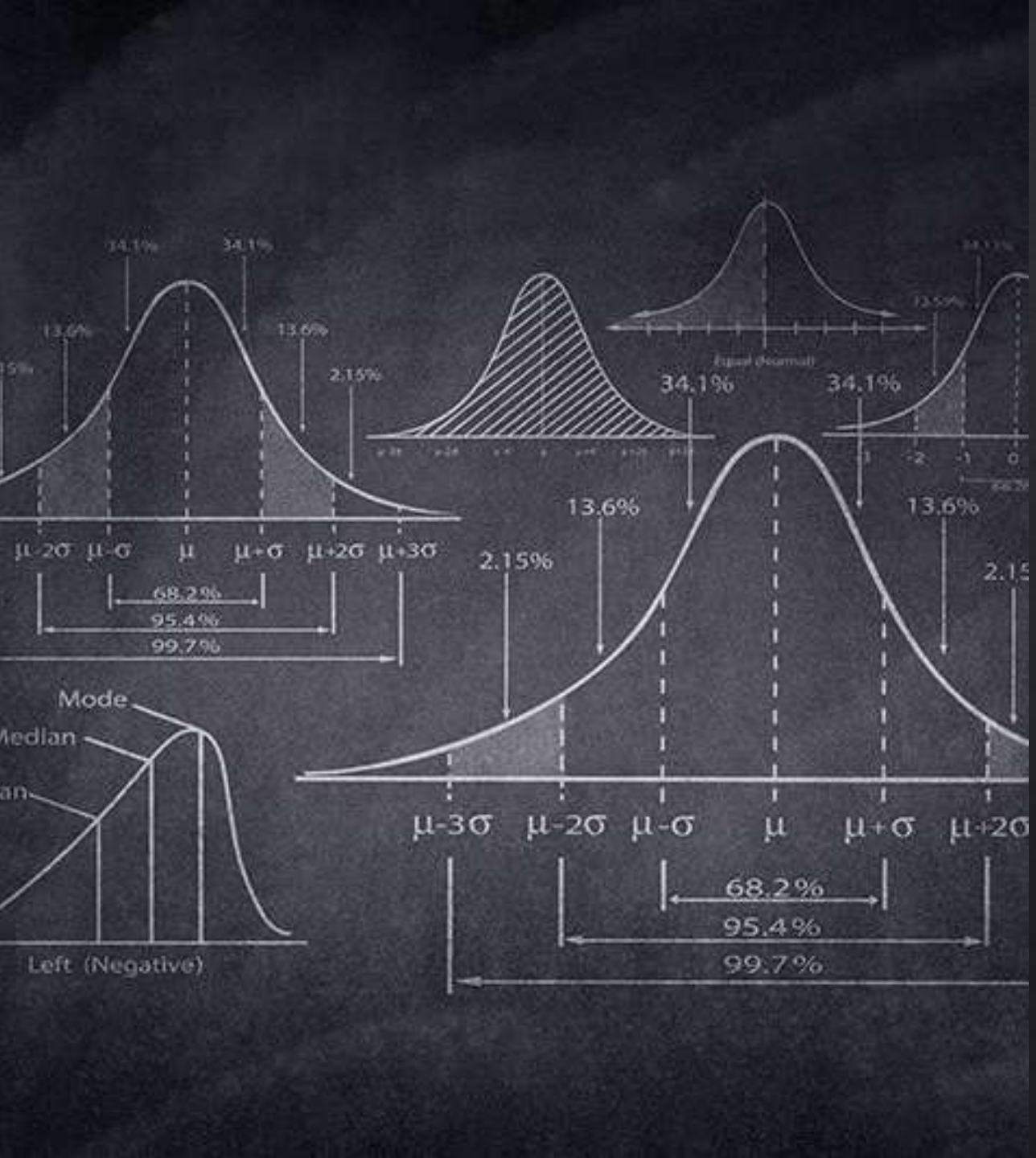
Inferential Statistics

- *Probability* describe the chance of an event, or set of events, occurring
  - It assumes we know all the relevant underlying information about the event
  - We predict the data from what we know
- *Inference* estimates the probability of observed data
  - We have the data, but we don't know the relevant underlying information about what made it
  - We *infer* information about the events



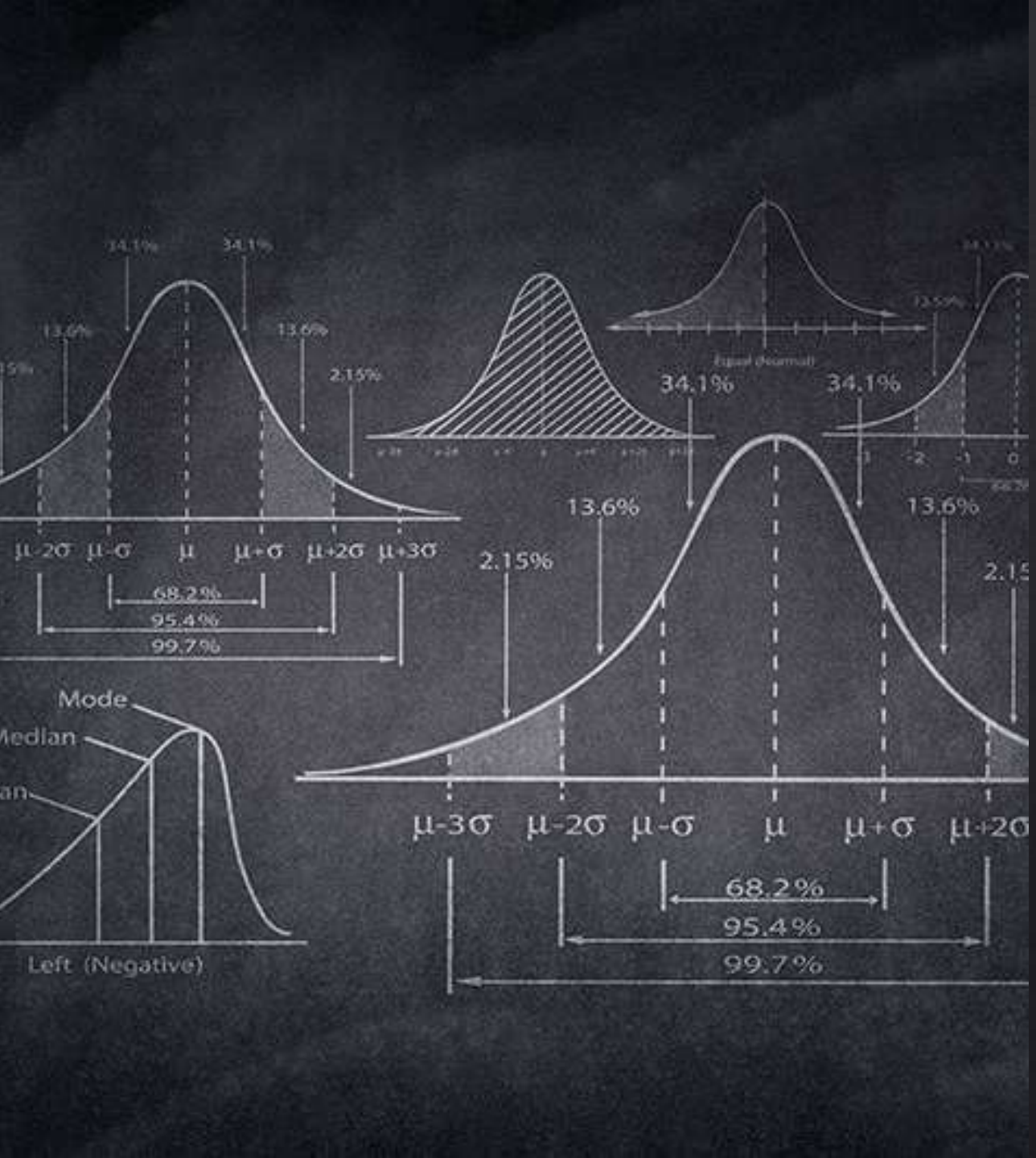
Probability

- *For example:*
- If we have a standard coin, what is the probability that we flip 10 heads in a row?
- We know all the relevant information about the system, i.e. the probability of flipping heads once is 0.5 (50%)
- From there we can calculate:
  - 0.5^10 = 0.009765625 -> ~1%



Inferential Statistics

- A very simple example of inference:
- A coin is flipped 5 times. We get heads every time.
- There are two possible reasons:
  - The coin is weighted
  - We were very lucky
- We have the observation data, but do not know about the underlying probability
- We have to *infer* information about this system.



Inferential Statistics

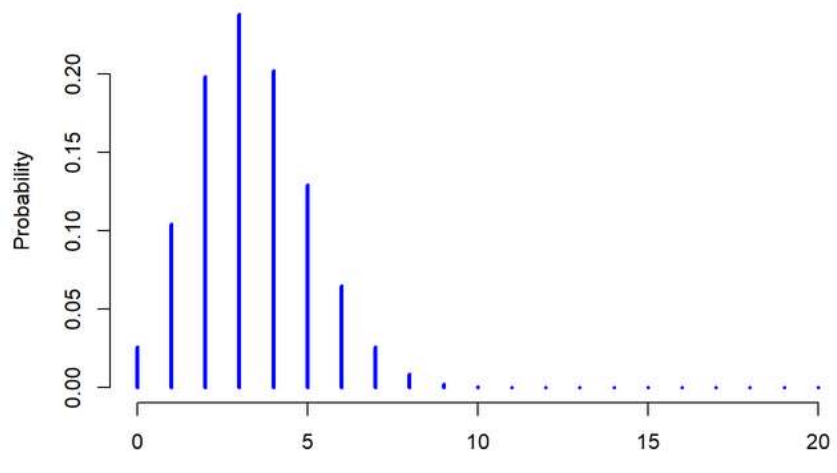
- We can work out the probability we were very lucky:
  - p = 0.5^5 = 0.03125
- There is a ~3% chance we would see 5 heads in a row on a fair coin.
- Is this enough to make a conclusion?
- In statistics we often use a threshold for significance, known as a p-crit.
  - This is often 0.05 (5%) but this is arbitrary!
- Since 0.03 is less than our p-crit (0.05), we conclude there is strong evidence the coin is weighted
- BUT we cannot be certain either way.

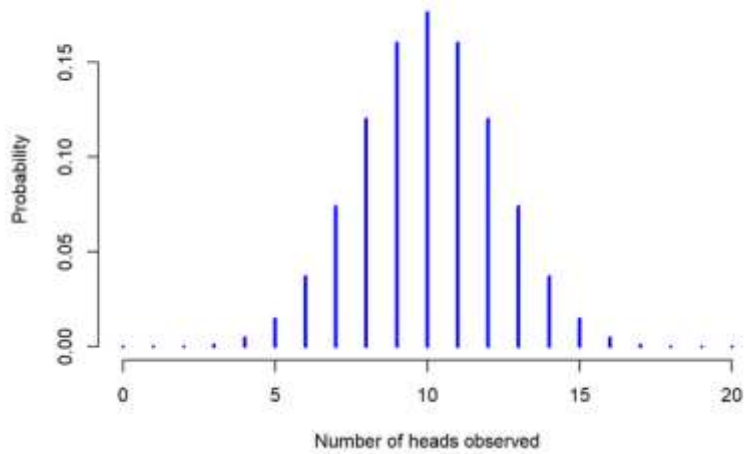
The Binomial Distribution

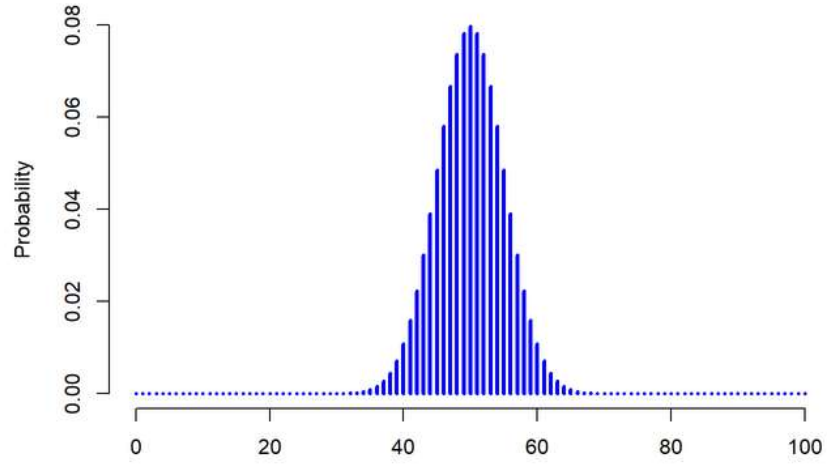
- What you just saw was a very simple example of binomial distribution. With it we can calculate the probability of an event (0/1 outcomes) based on our data.

The Binomial Distribution

- Every distribution can be formally described, and has some key parameters and assumptions.
- The general form of the binomial distribution is $X \sim Binomial(\vartheta, N)$
  - X is the observed results
  - N is the number of trials (e.g. number of time the coin was flipped)
  - Θ is the probability of success (e.g. probability of flipping heads)
- We can plot the outcomes of our trials to see what a binomial distribution looks like:







The Binomial Distribution

- We can go further!
- We can work out the probability of seeing our data, for a given value of θ.
- Often, we want to estimate the probability our null hypothesis is true. To do this, we estimate *p,* which is the probability that the data we observe is completely random and there are no underlying “weighted” parameters.

Binomial

$$P(X|\theta, N) = \frac{N!}{X!(N-X)!}\theta^X(1-\theta)^{N-X}$$

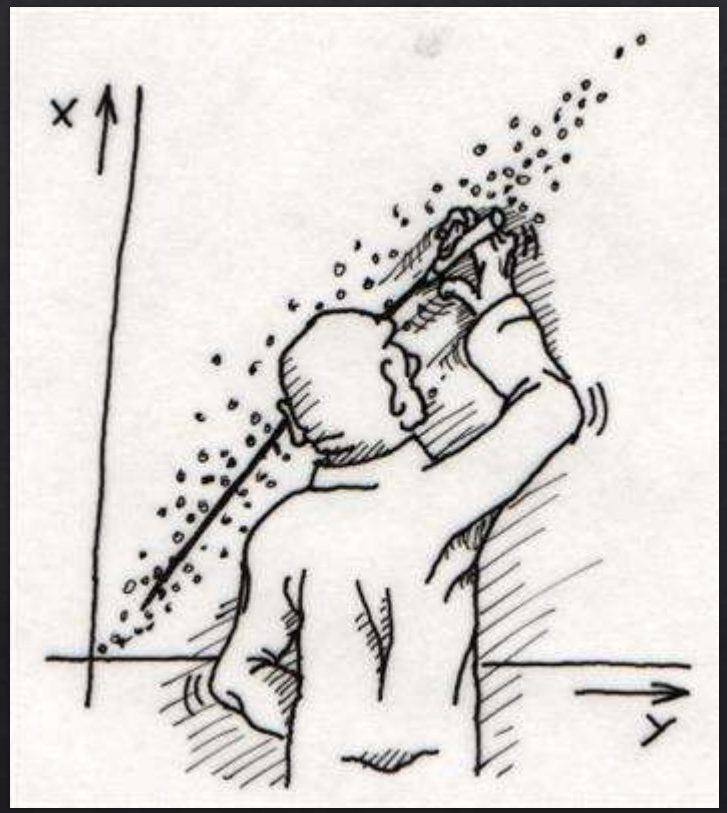
Distributions

- For a binomial distribution, converting from data to probability and back again is relatively easy
- But this is a lot more complicated with most other types of data. “Distributions” are designed to allow to convert between properties of our data and probability.
- BUT this means we have to make sure our data fits our selected distribution. If it doesn’t, it screws up all our calculations of probability.

Quick Summary

- We have covered a number of core concepts:
  - Types of data
  - Thinking about standard notation, null and alternative hypotheses
  - Working out our first p-value
  - And looked at our first distributions.
- Let’s look at another example, closer to actual ecological data!

PART 4: Running a Model



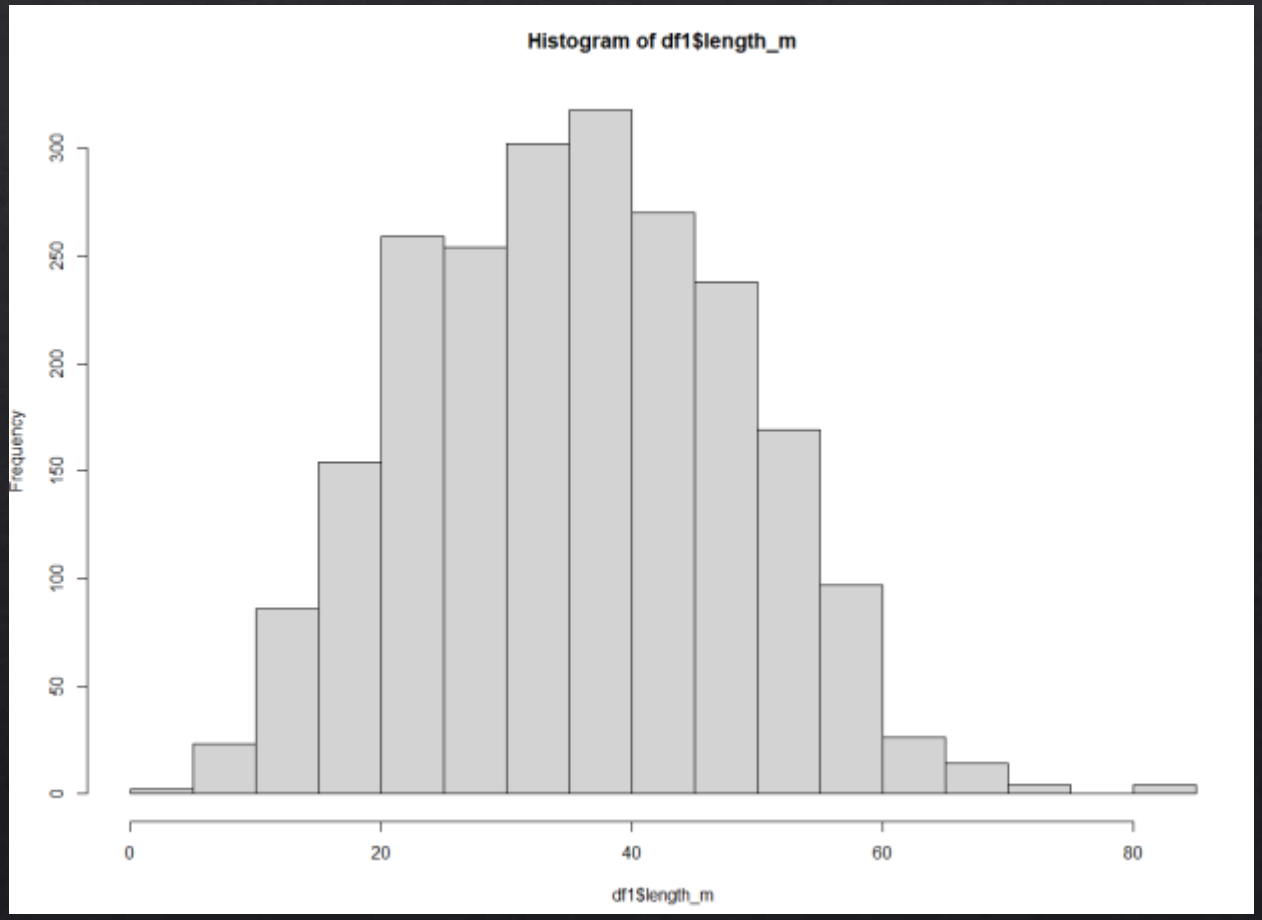
Data example!

- A research team has found two previously undiscovered populations of flying whales in Antarctica.
- The team have gathered data on their length, age, location, population and local intensity of background magic (i.e. an environmental variable).
- They wish to investigate whether the two populations may be different species. As a simple start they want to know if the average lengths of the two populations differ

Source: Sampo Jumisko

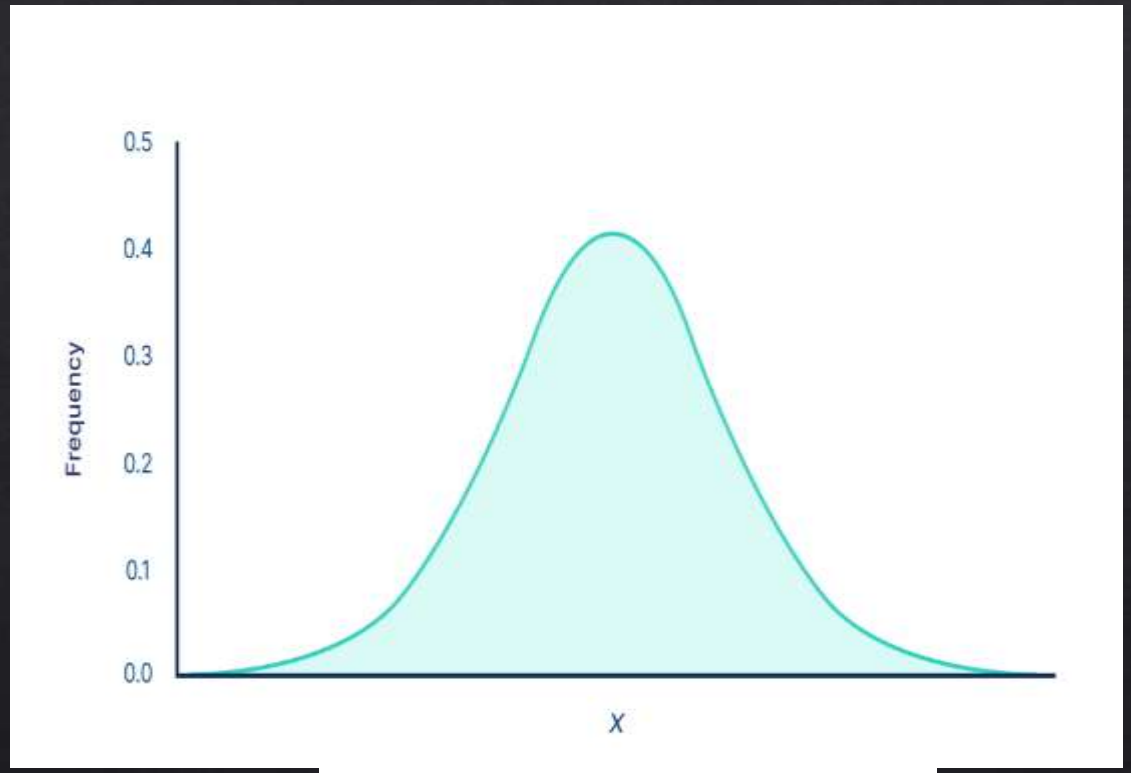
Data example!

- Our model equation:
  - Length ~ population
  - Note: there are two populations
- Histogram of response variable:



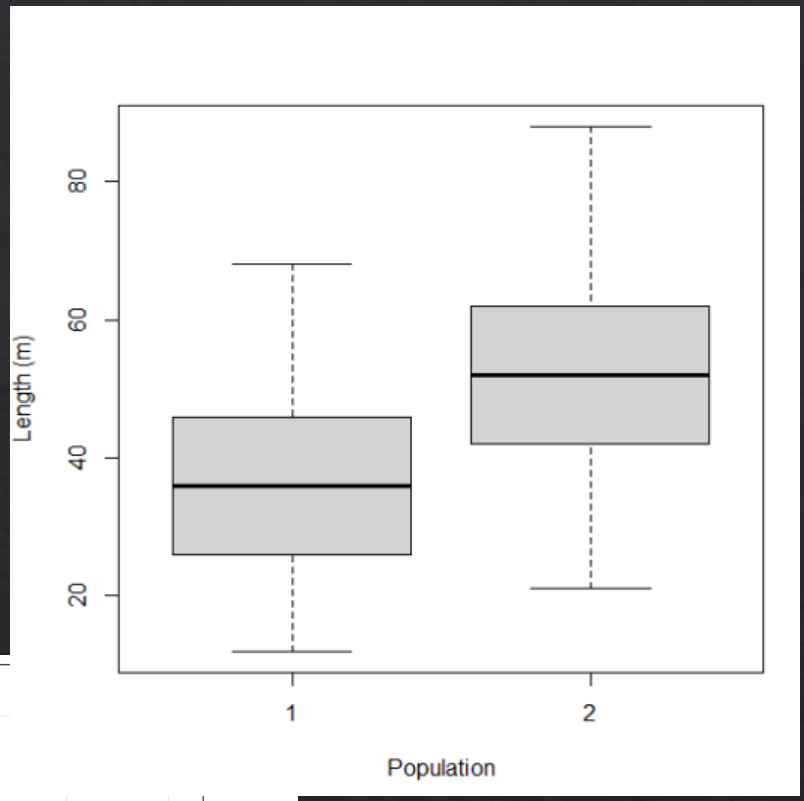
Data distributions

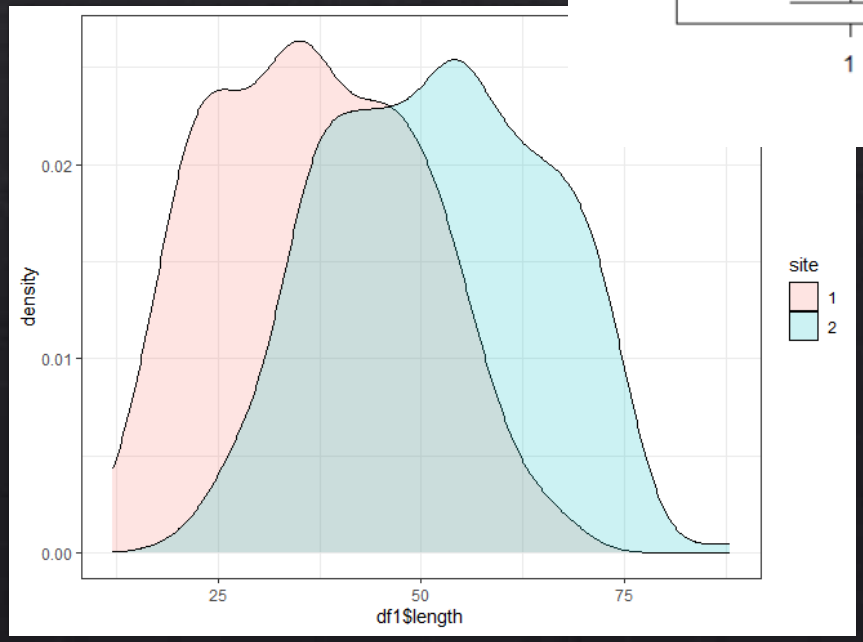
- **Normal (aka Gaussian) distribution**
- Continuous response variable:
  - Length of population
  - Time of migration arrival
  - Intelligence
- Two parameters: mean (μ) and standard deviation (σ)
- General form:
  - X∼Normal(μ,σ)



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

- We want to compare the length of flying space whales in population 1 and population 2.
- Our null hypothesis is: the average length of a whale in population 1 is the same as in population 2. i.e. there is no difference (on average) between populations
- We assume that our data approximately fits a normal distribution






```
lm.pop <- lm(length_m ~ site, data = df1)
```

Am I using the right distribution?

- Make sure you know the type of data you are working with (continuous, count etc.)
- Plot the data. Plot the data. PLOT THE DATA!
- Use experience/text books/search engines to find possible distribution in R
- How closely your data matches a distribution matters less if your dataset is large
- If in doubt, run the test, you can check afterwards whether the model “fits” your data.
  - We will come back to this later….

Running a test!

- We assume the response variable (length) is approximately a normal distribution
- Our predictor is population, which is categorical
  - There are 2 categories (Site 1 and Site 2)

- So let's feed our statistical formula into R:
  - Length ~ Population

```
lm.pop <- lm(length_m ~ site, data = df1)
```

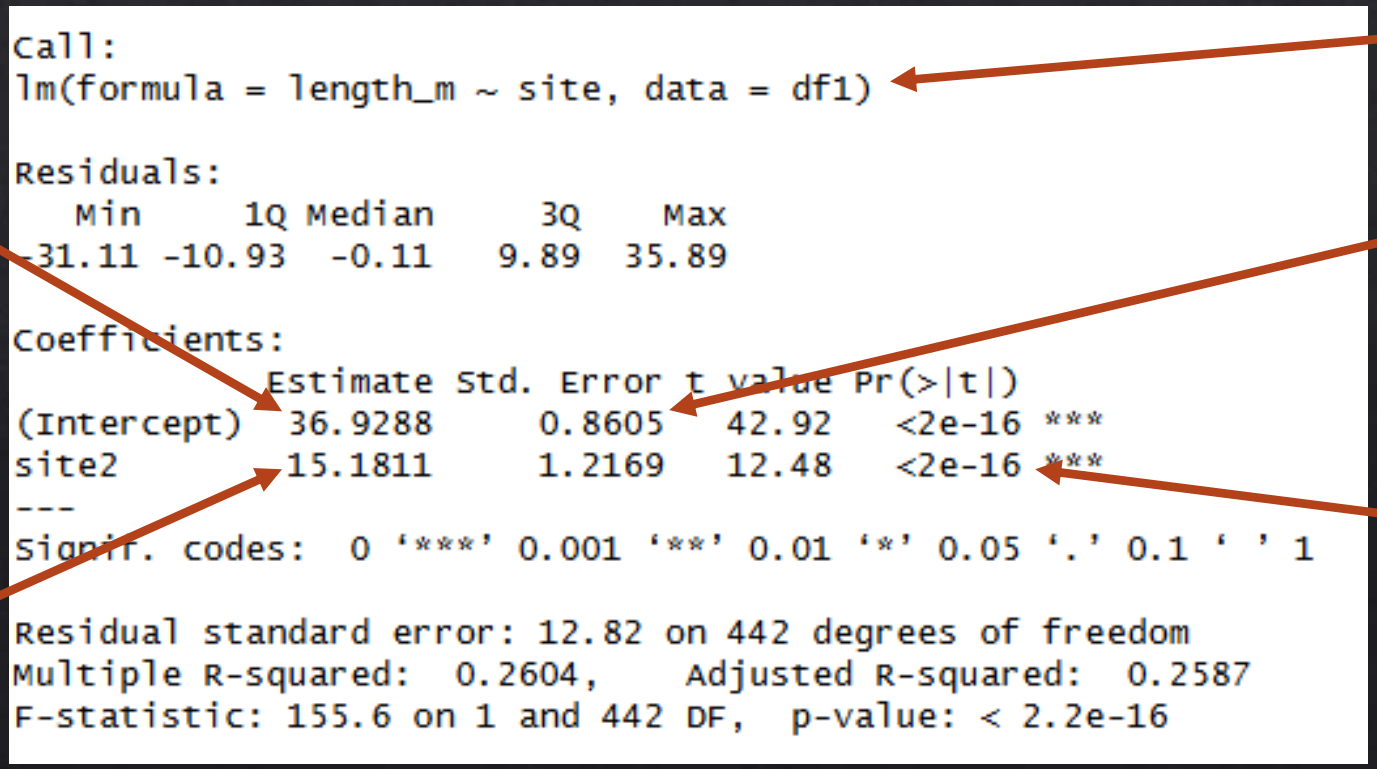
Running a test!

- The output:

```
Call:
lm(formula = length_m ~ site, data = df1)

Residuals:
   Min     1Q Median     3Q    Max
-31.11 -10.93  -0.11   9.89  35.89

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  36.9288     0.8605   42.92   <2e-16 ***
site2        15.1811     1.2169   12.48   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 12.82 on 442 degrees of freedom
Multiple R-squared:  0.2604,    Adjusted R-squared:  0.2587
F-statistic: 155.6 on 1 and 442 DF,  p-value: < 2.2e-16
```

Running a test!

- The output:



Model formula

Mean length of site 1

Standard error

p-value

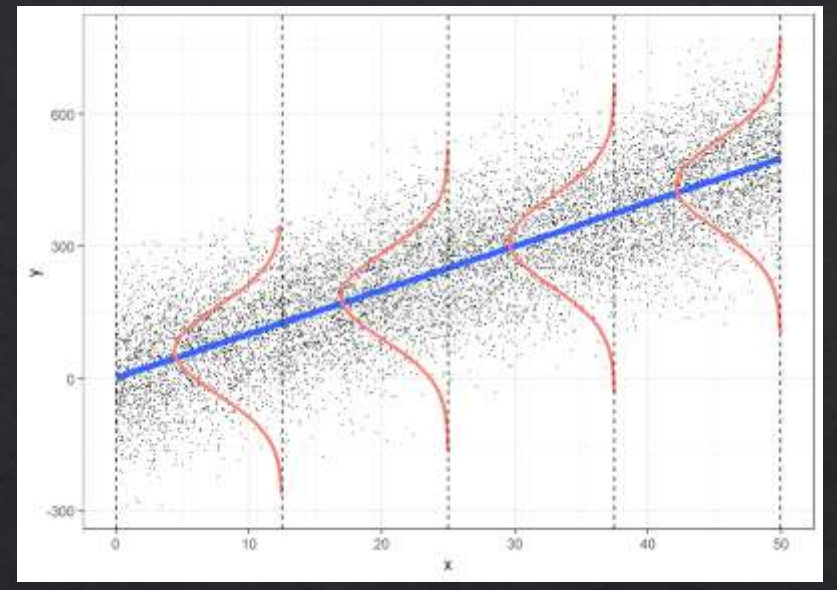
Difference between mean of site 1 and site2

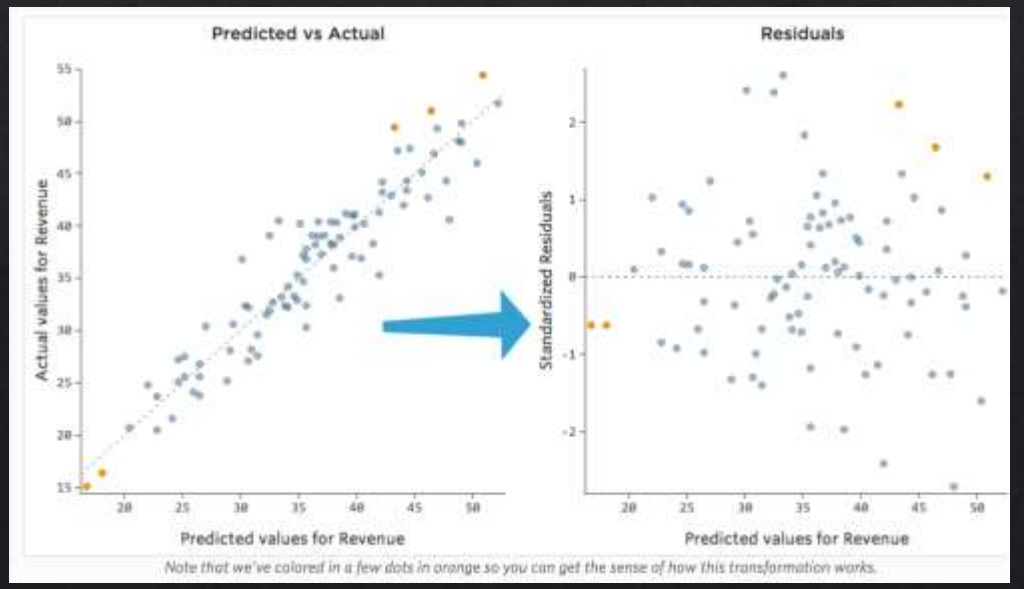
Diagnosing a test

- Is our job done? No!
- We have to check the model is actually working properly, and our distribution is the correct choice
- To do this we can run diagnostics after performing a statistical test
- But before that we have to explain “residuals”:

Diagnosing a test

- Residuals measure the difference between “what the model predicts” and “what does the data says”
- No model is perfect so you will always have some residuals
- Residuals should be *random* and (typically) are *normally distributed.*
- Any systemic patterns in your residuals indicate there is a mismatch between your model design and your data

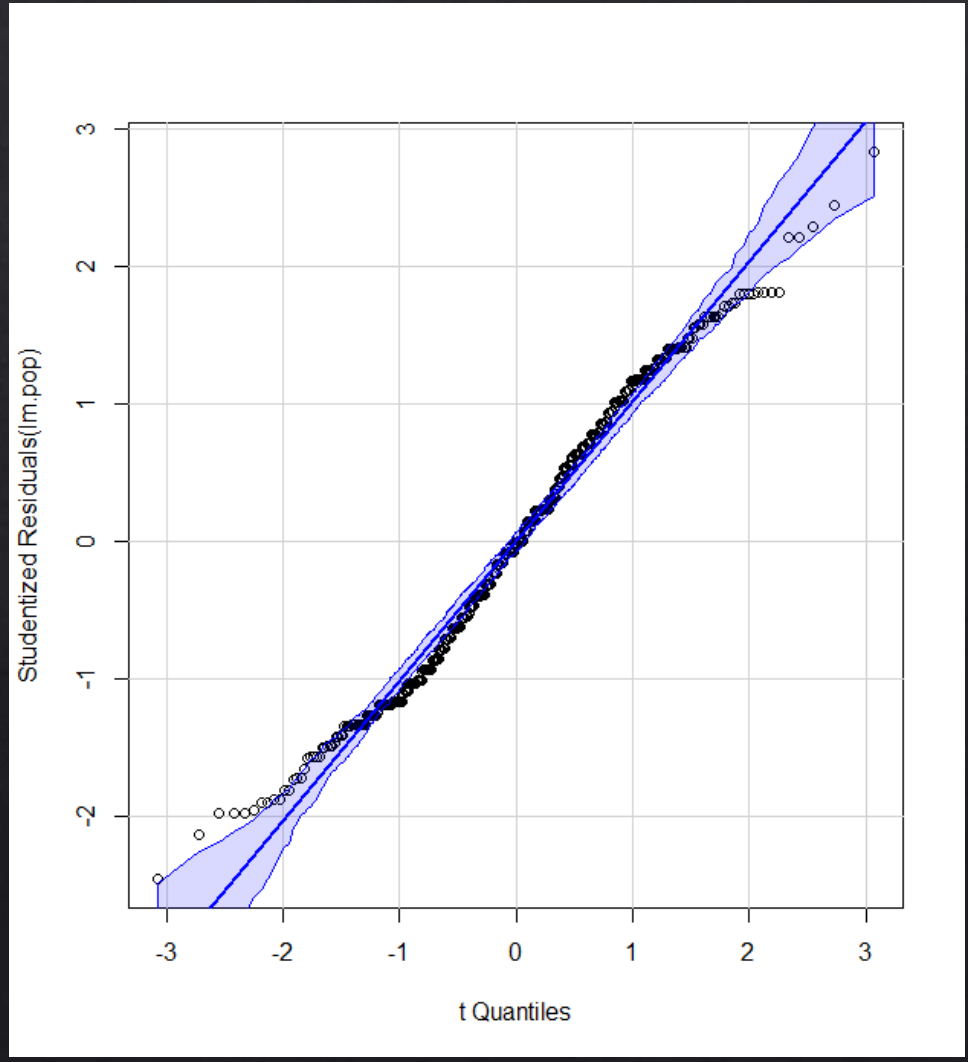




https://statistics4ecologists-v2a.netlify.app/linreg#lmassumptions

Diagnosing a test

- R has a lot of tools to visualise residuals, here is one of them for our simple test
- In this very simple model, the only source of error is the choice of distribution.
- A perfect normal distribution would fit on the dotted line
- Deviations away from this line indicate this is not a normal distribution
- This is not a great plot, but good enough I would consider it acceptable!

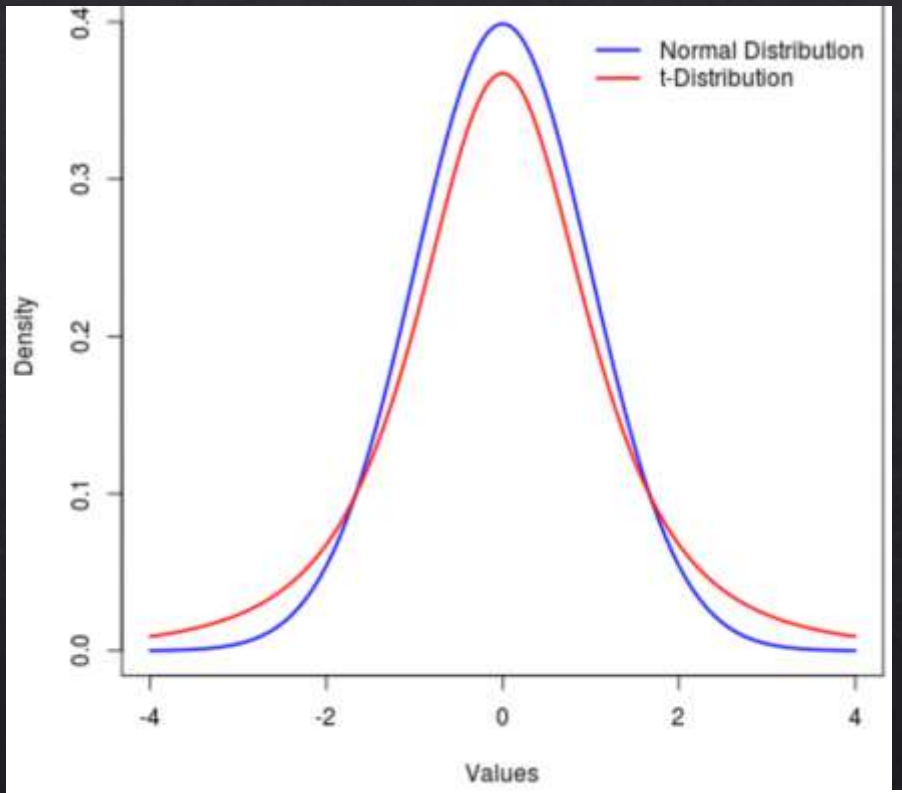


When modelling goes wrong

- Remember a key assumption is our data fits our chosen distribution. If it does not, then the test is flawed
- We check that our model fits our data well, diagnostic plots and inspecting residuals are key to this. But what do we do when we find our model design is bad?
- One option: pick another distribution
- How do we know which to use?
  - Experience
  - Searching online or in R documentation

Data distributions

- **Student's t-distribution**
- Extremely similar to normal distribution
- BUT better at handling wider variance and small samples
- Continuous response variable:
  - Length of population
  - Time of migration arrival
  - Intelligence
- Two parameters: degrees of freedom (v) and gamma function (Γ)



$$f(t) = \frac{\Gamma\left(\frac{\nu+1}{2}\right)}{\sqrt{\pi\nu}\,\Gamma\left(\frac{\nu}{2}\right)}\left(1+\frac{t^2}{\nu}\right)^{-(\nu+1)/2},$$

Data distributions

- **Student's t-distribution**
- In this case gives same answer
- Makes sense as t-distribution is very similar to a normal distribution
- Surprise: we just did a t-test!

```
Call:
lm(formula = length_m ~ site, data = df1)

Residuals:
   Min     1Q Median     3Q    Max
-31.11 -10.93  -0.11   9.89  35.89

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  36.9288     0.8605   42.92   <2e-16 ***
site2        15.1811     1.2169   12.48   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 12.82 on 442 degrees of freedom
Multiple R-squared:  0.2604,    Adjusted R-squared:  0.2587
F-statistic: 155.6 on 1 and 442 DF,  p-value: < 2.2e-16
```

```
        Welch Two Sample t-test

data:  length_m by site
t = -12.476, df = 441.29, p-value < 2.2e-16
alternative hypothesis: true difference in means between group 1 and group 2 is not
equal to 0
95 percent confidence interval:
 -17.57266 -12.78950
sample estimates:
mean in group 1 mean in group 2
       36.92883        52.10991
```

Reporting statistical tests

- Describe the result, state the hypothesis result, and state the details of the test.

- *"On average, whales from population 1 were 36.9m in length (n=222; SD=12.6), while those from population two were 52.1m (n=222; SD=13.1). A Student's t-test found a significant difference in length between the two populations (t=-12.5, df=441.39, p<0.001)."*

Recap

- Steps to making a statistical test:
  1. Think about our study system, identify our response and our predictors and make a formula
  2. Make plots of our data (especially our response variable), and select a distribution.
     i. Check our data fulfils the assumptions of the distribution
  3. Run chosen statistical test
  4. Check the diagnostics and residuals
     i. Consider alternative distributions if necessary
  5. Interpret our output

Side note: sub-tests

- One of the reasons learning statistics can be confusing is because we have to sift through decades of work and a lot of confusing terminology.
- I generally advise people to ignore names, and focus on
  - the type of data you have
  - the distribution of the response variable

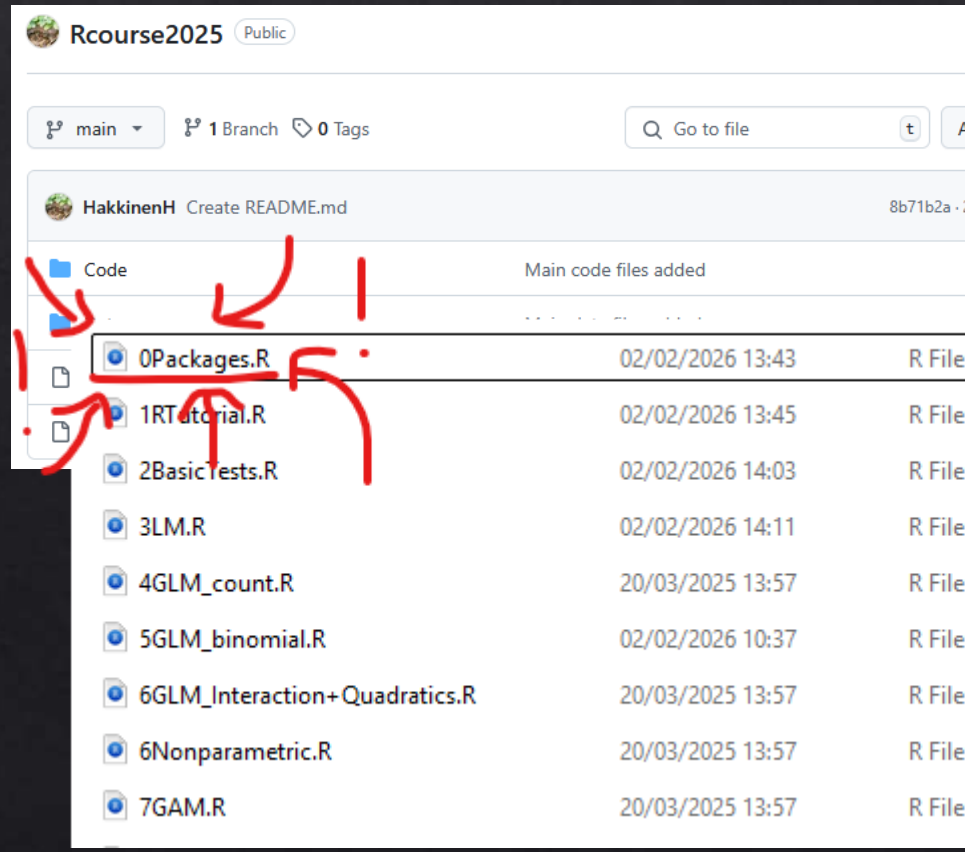
Side note: sub-tests

- Many tests have variants that account for different scenarios, unfortunately they often have different names and (I think) needlessly confusing
- For example, there are many sub-types of t-test:
  - One-sample t-test
  - Student's (Independent Samples) t-test
  - Paired Sample t-test
  - Welch's t-test
  - One/Two tailed t-test
- If you come across a test with many sub-types, learn the basic form first. Then investigate if a sub-type is specifically needed for your data/hypothesis.

Response variables is…	Predictor	Response variable distribution	Possible test
Binary (0/1)	1 or more categories	binomial	Binomial test
Continuous	2 categories	Normal OR t-distribution	Linear model with 2 categories OR t-test
Continuous	>2 categories	Normal	Linear model with >2 categories OR ANOVA
Count	>=2 categories	Chi-squared	Chi-squared test
Continuous	*Continuous*	*Normal*	*Linear Regression*

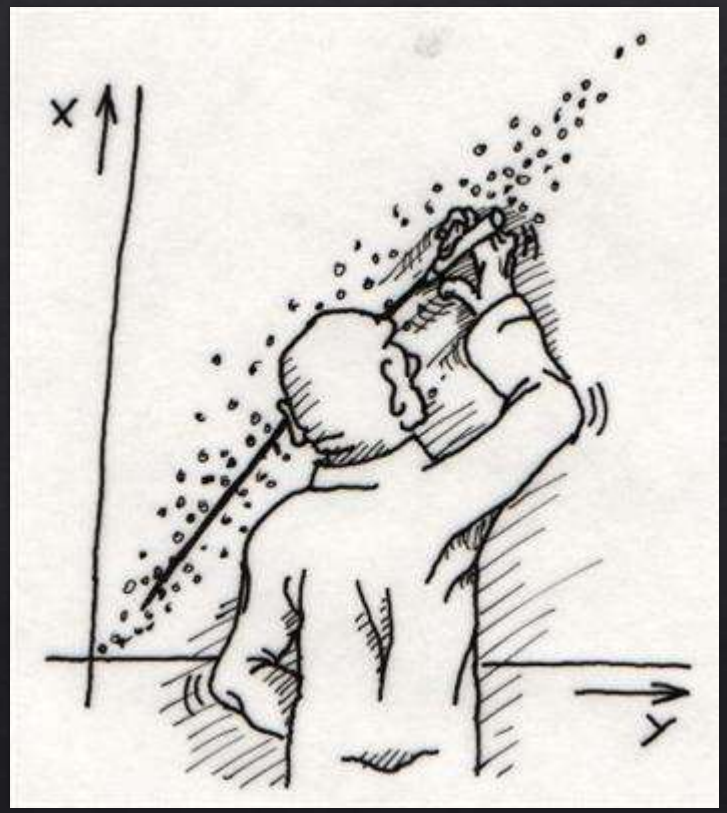
The Practical: Part 1

- https://github.com/HakkinenH/RStatsCourse2026



- Code -> download ZIP -> Unzip and put somewhere useful
- Run 0Packages.R first (if you haven't already) to install packages
- *Optional:* if you want some practice with basic R, then work through
- Work through 2BasicTests.R as your first example

PART 5: Linear Models



Data example!

- A research team has found several previously undiscovered populations of flying whales in Antarctica.
- The team have gathered data on their length, age, location, population and local intensity of background magic (i.e. an environmental variable).
- They wish to investigate what variables might explain their enormous size.

Source: Sampo Jumisko

Check list

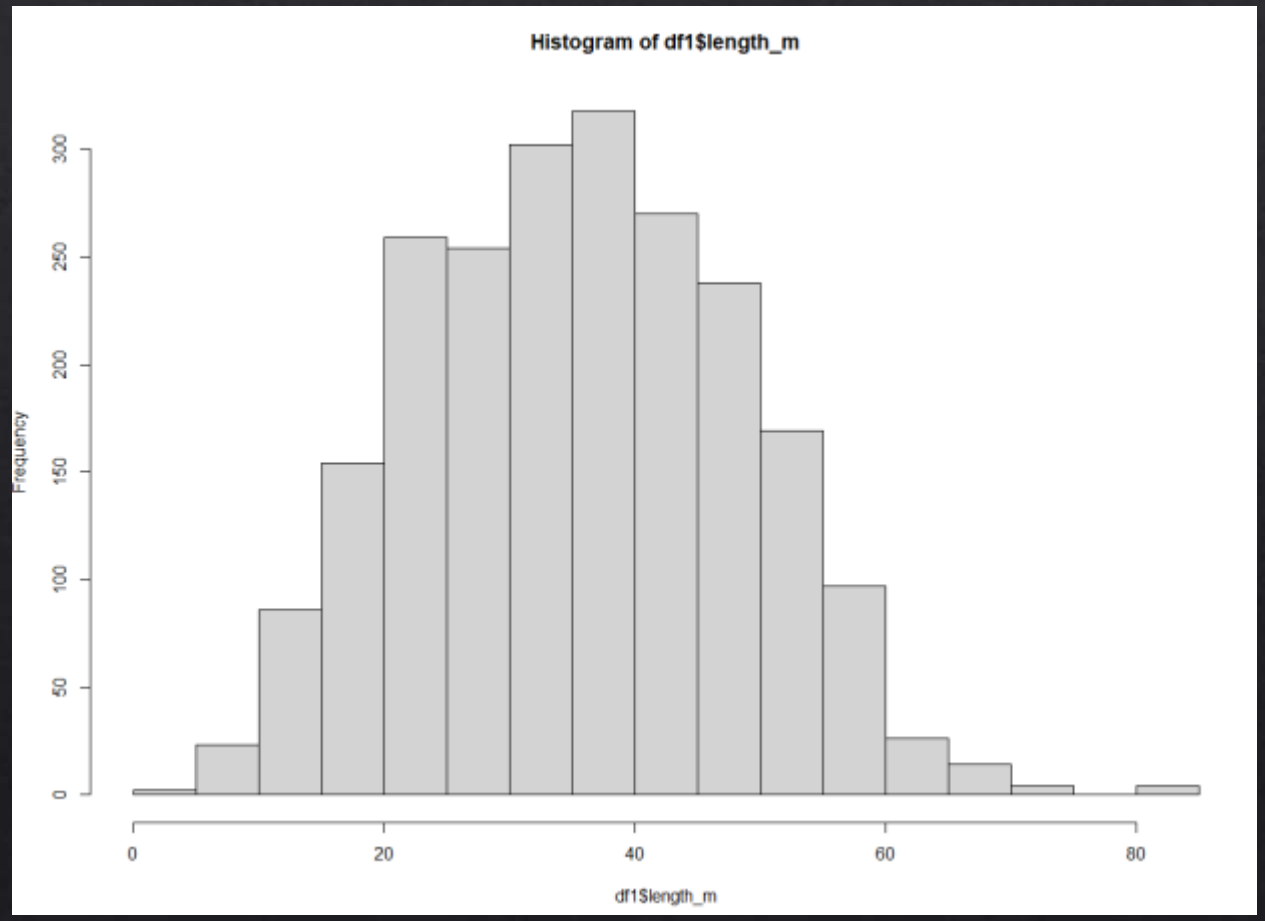
- Steps to making a statistical test:
  1. Think about our study system, identify our response and our predictors and make a formula
  2. Make plots of our data (especially our response variable), and select a distribution.
     i. Check our data fulfils the assumptions of the distribution
  3. Run chosen statistical test
  4. Check the diagnostics and residuals
     i. Consider alternative distributions if necessary
  5. Interpret our output

Data example!

- Step 1: Think about our study system, identify our response and our predictors and make a formula

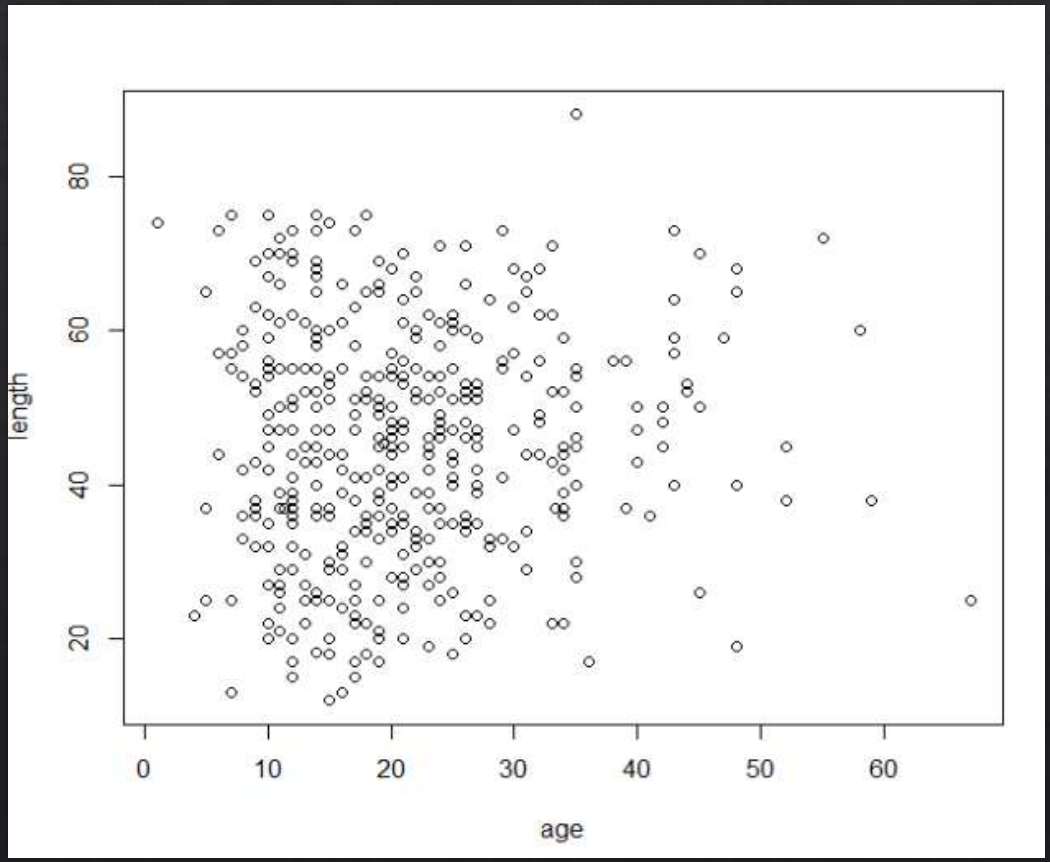
- Response: length
- Predictors: magic, age, location (east or west)

- Let's make some other plots for our data



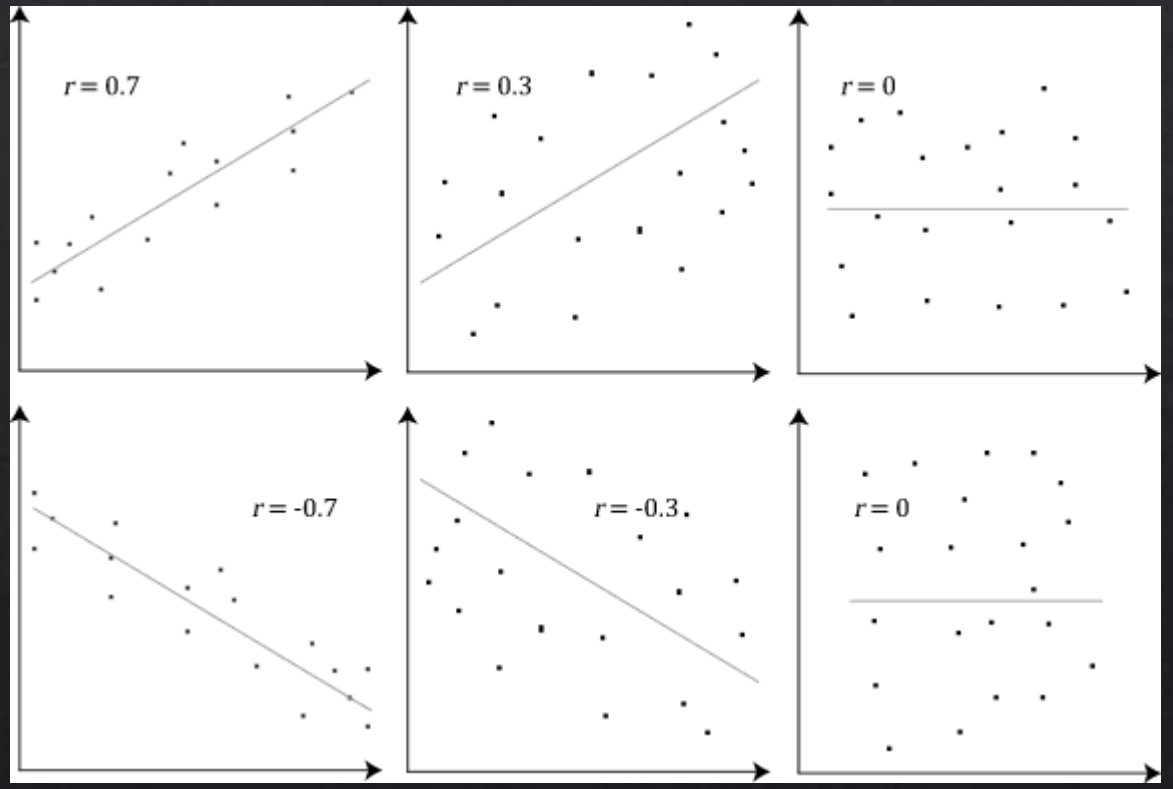
Correlation

- For categories we used mean, median, SD etc. as descriptive statistics
- For pairs of continuous variables we can also use correlation coefficients
- NOT a statistical test, it's descriptive
- Several variants, most common are Pearson's and Spearman's



Correlation

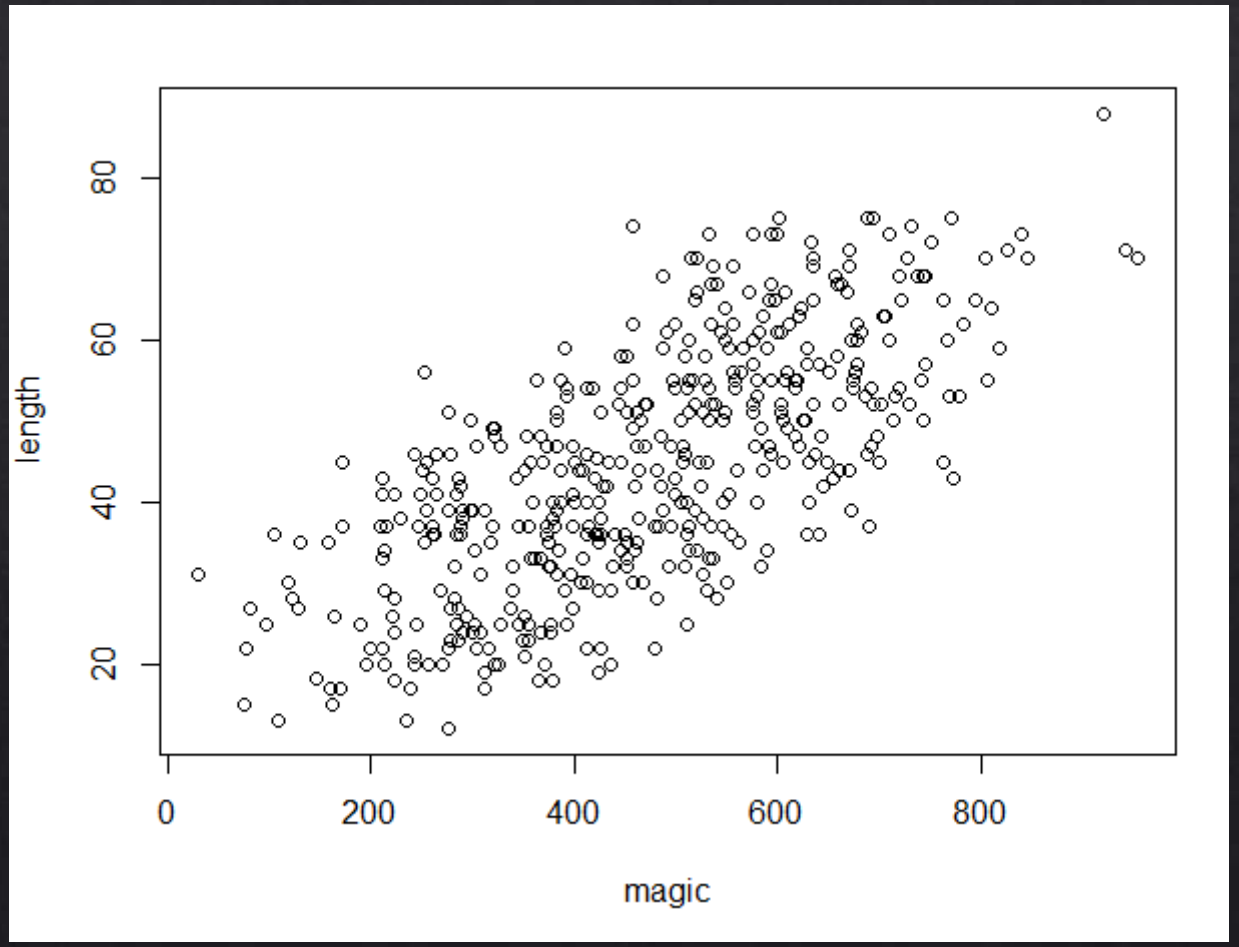
- Most standard correlation coefficients are from -1 to 1
  - 0 means no correlation
  - -1 means perfect negative correlation
  - +1 means perfect positive correlation
- R-squared is just correlation squared, commonly reported in linear models



Correlation

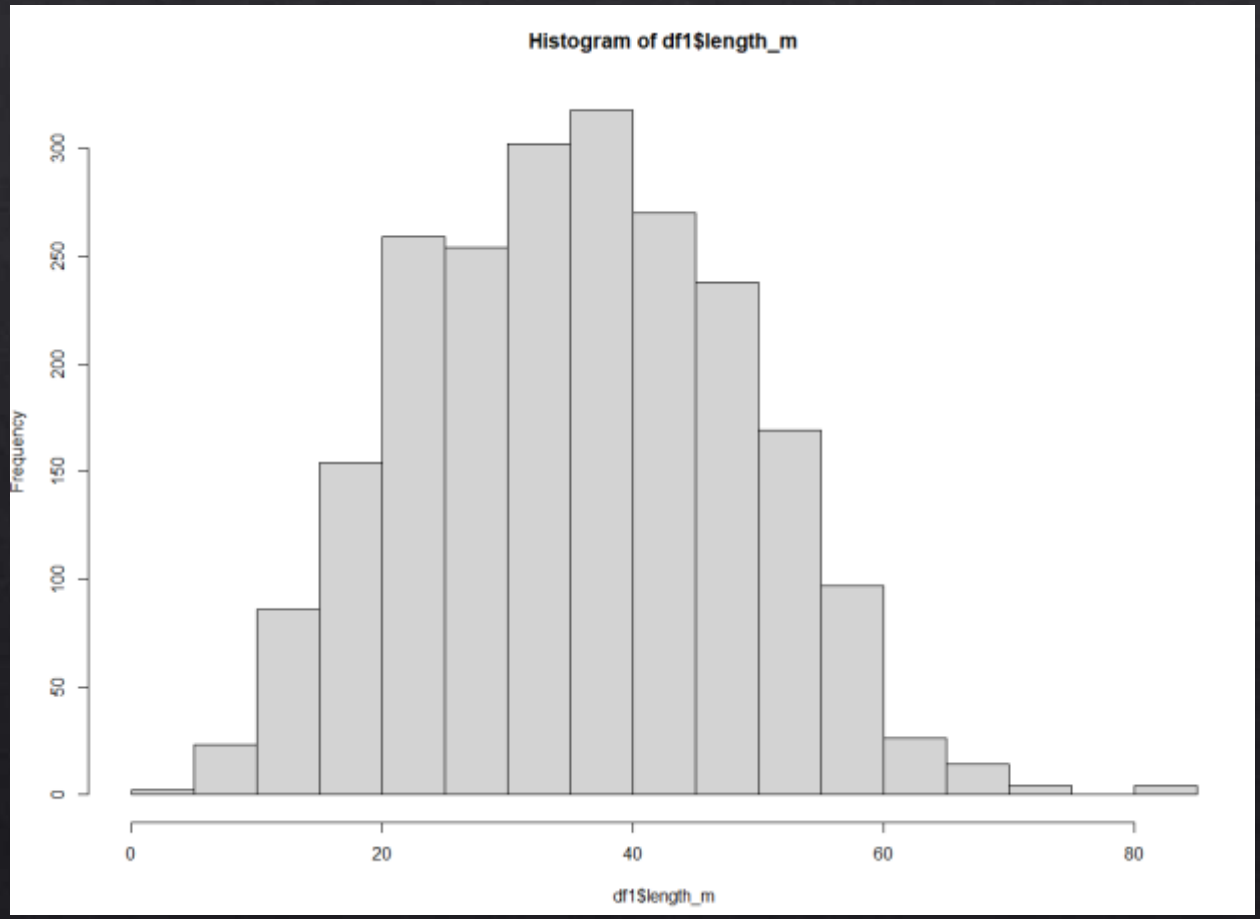
- Age vs Length: Pearson's r =0.07
- Magic vs Length: Pearson's r =0.71

- So our initial exploration suggests there might be a relationship between magic and length, but not age and length



Data example!

- Step 2: Make plots of our data (especially our response variable), and select a distribution
  - Continuous response: normal distribution
  - Continuous predictor: [distribution doesn't matter]
  - Categorical predictor: [distribution doesn't matter]
  - Chosen model: linear model with a normal distribution



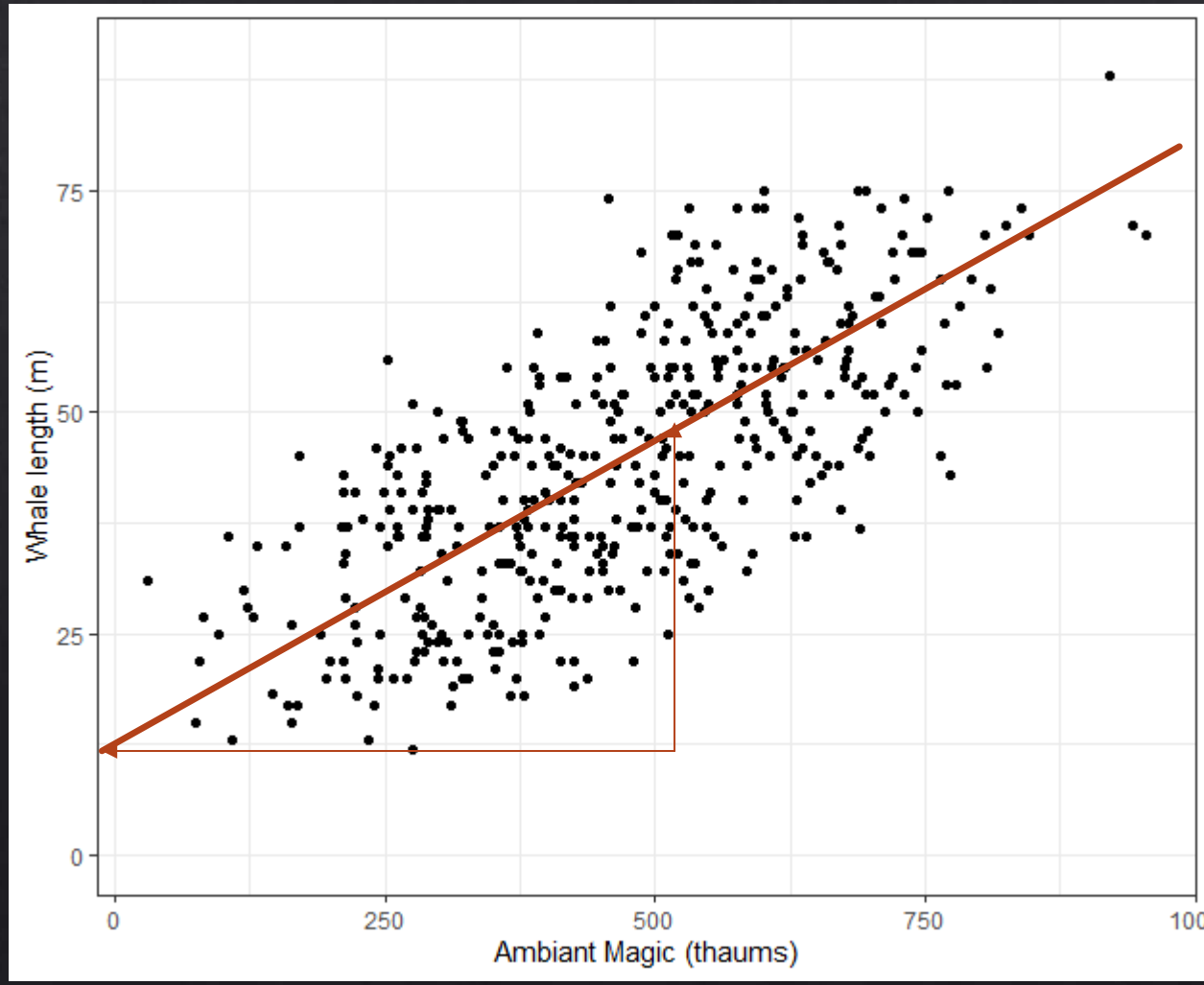
Check model assumptions

- Check our data fulfils the assumptions of the distribution. (I looked them up online)
  - The sample is representative of the population at large.
  - The response variable is normally distributed
  - There are no substantial outliers
  - The predictors are all linearly independent (they do not substantially correlate with each other)
  - There is sufficient data to carry out the test (rule of thumb: 10-20x more data than variables)
  - There is a linear relationship between the independent variable(s) and the dependent variable.
  - The independent variables are measured with no error (i.e. they are fixed effects).
  - Mean of the distribution of errors are 0.
  - The variance of the residuals is constant across observations (homoscedasticity).
  - The residuals are uncorrelated with one another.

Data example!

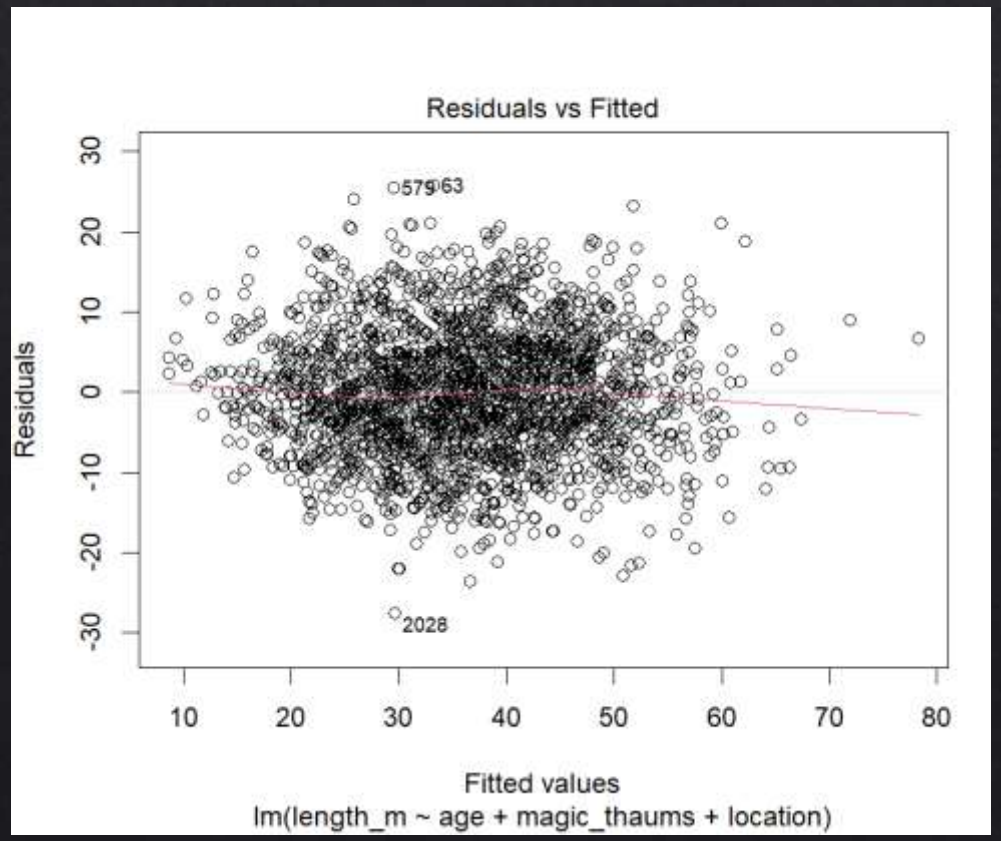
Step 3: Run chosen statistical test (simple version)

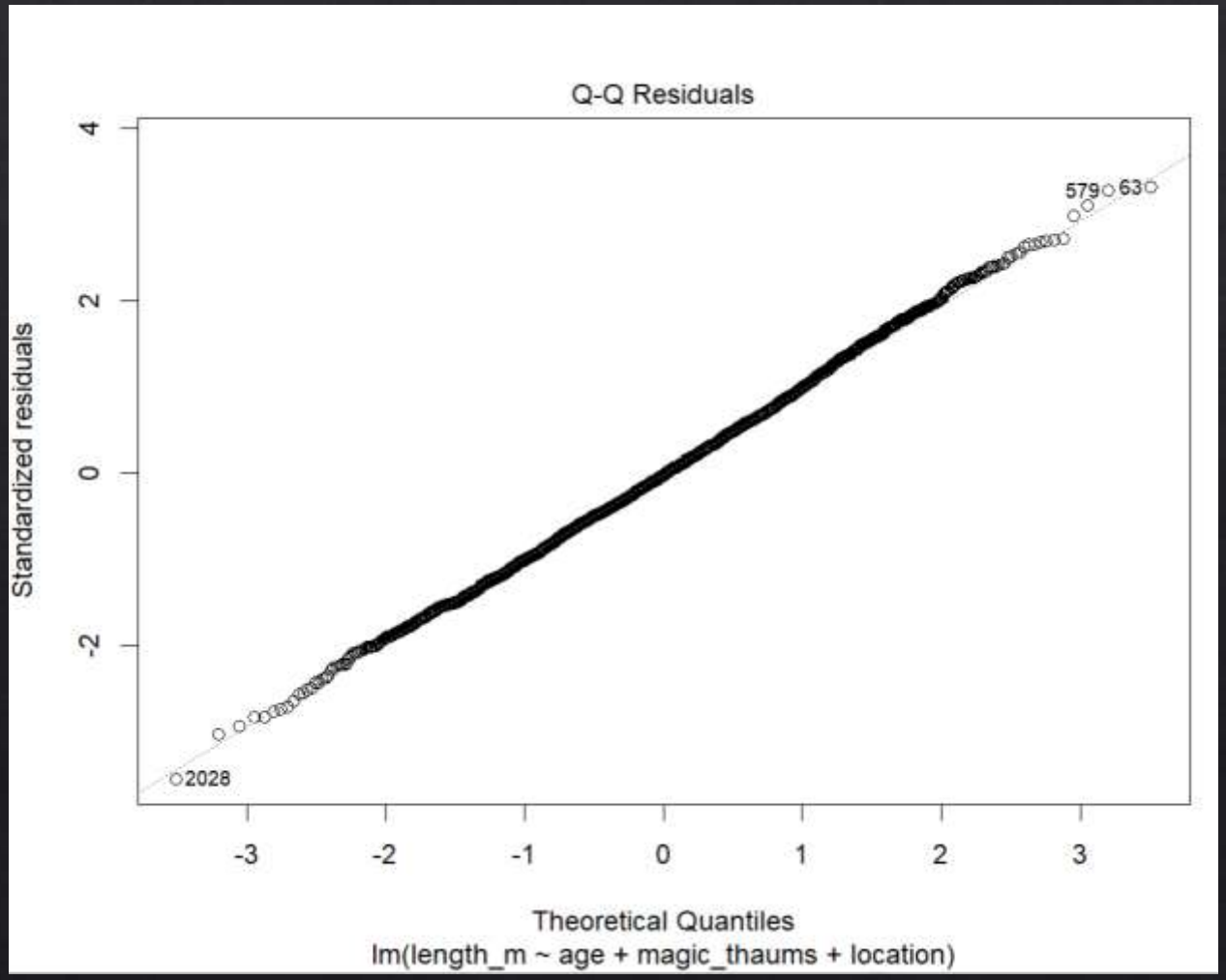
```
Call:
lm(formula = length_m ~ magic_thaums, data = df1)

Residuals:
     Min       1Q   Median       3Q      Max
-23.1046  -8.3972   0.1768   7.8501  30.2264

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)  15.737262   1.462982   10.76   <2e-16 ***
magic_thaums  0.061293   0.002927   20.94   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 10.56 on 442 degrees of freedom
Multiple R-squared:  0.498,     Adjusted R-squared:  0.4969
F-statistic: 438.5 on 1 and 442 DF,  p-value: < 2.2e-16

>
```



Data example!

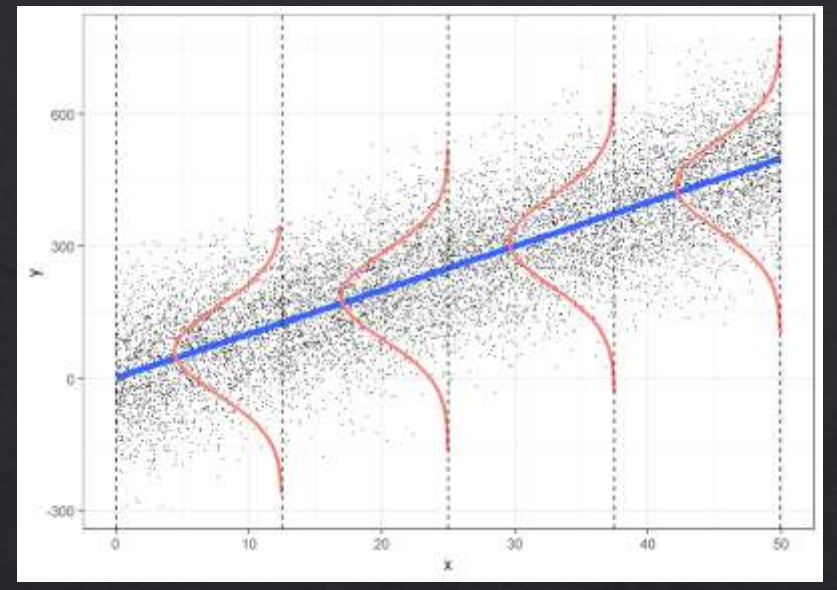
- Step 4: Check the diagnostics and residuals

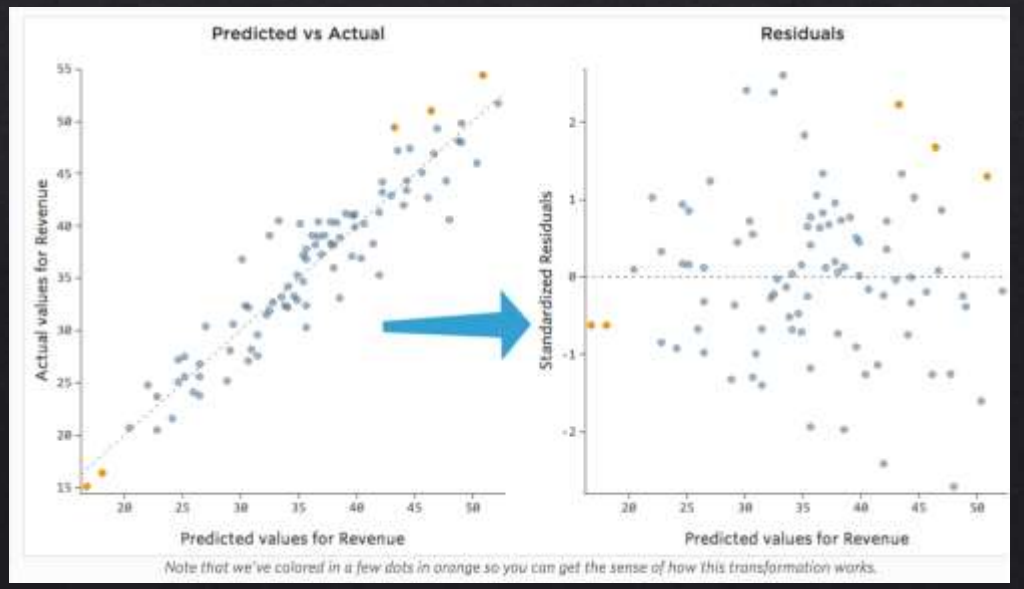




Reminder: residuals

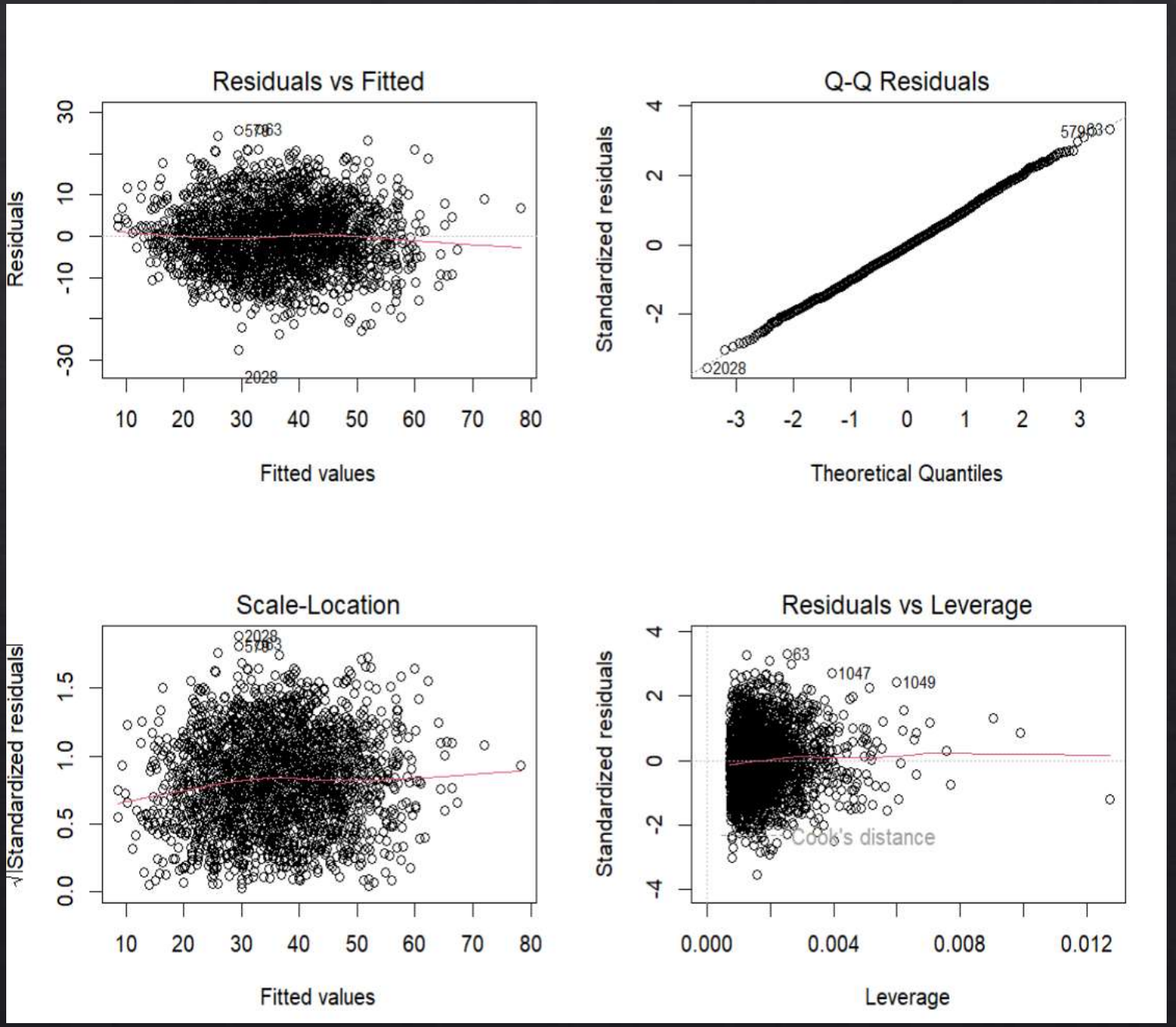
- Residuals measure the difference between “what the model predicts” and “what does the data says”
- No model is perfect so you will always have some residuals
- Residuals should be *random* and (typically) are *normally distributed.*
- Any systemic patterns in your residuals indicate there is a mismatch between your model design and your data





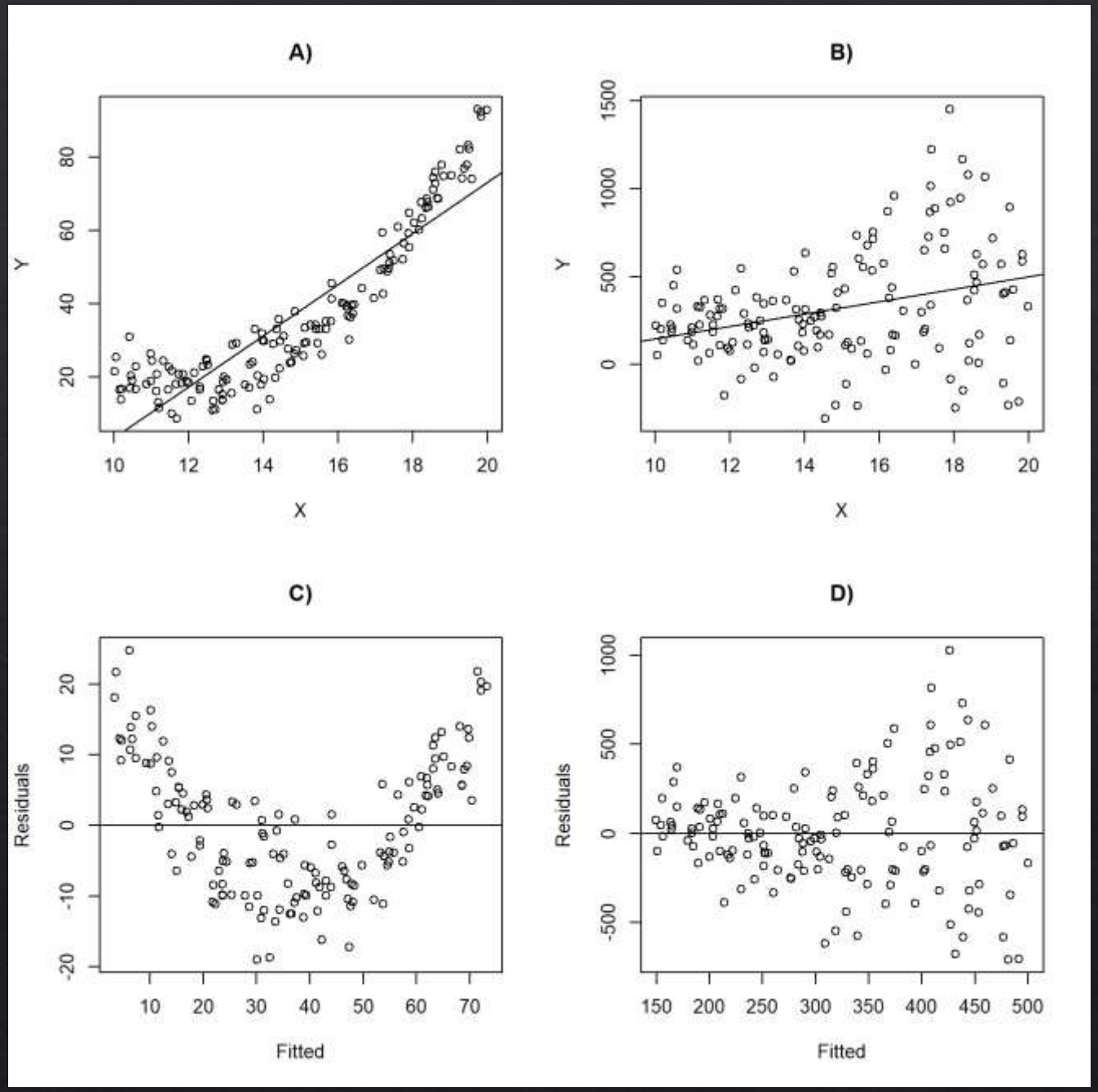
https://statistics4ecologists-v2a.netlify.app/linreg#lmassumptions

Data example!



Bad residuals

- What is wrong with these?



Data example!

Step 3: Run chosen statistical test (full version)

```
lm(formula = length_m ~ age + magic_thaums + location, data = df1)

Residuals:
     Min       1Q   Median       3Q      Max
-23.4968  -8.3161   0.2885   7.9424  30.7625

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)  15.703413   1.890789   8.305 1.24e-15 ***
age           0.018675   0.049870   0.374    0.708
magic_thaums  0.061091   0.002956  20.669  < 2e-16 ***
locationw    -0.428526   1.040472  -0.412    0.681
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 10.58 on 440 degrees of freedom
Multiple R-squared:  0.4984,    Adjusted R-squared:  0.495
F-statistic: 145.7 on 3 and 440 DF,  p-value: < 2.2e-16
```

Data Example!

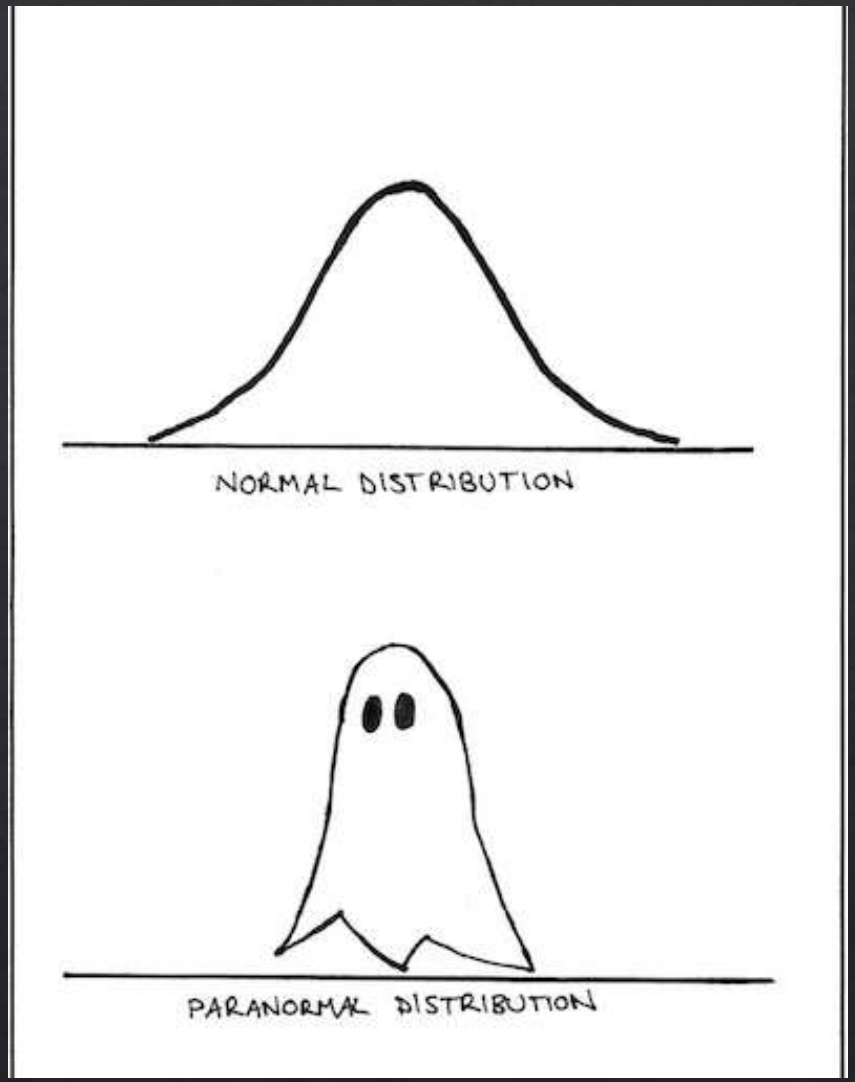
- Step 5: Interpret and report out findings

- *"Ambient background magic significantly, and positively, correlated with whale length (linear regression: $F(3, 440) = 145.5$, $p<0.001$); on average, length increased by 0.06m per unit of magic. There was no significant correlation between age and length ($p=0.71$), nor was there a significant difference between Eastern and Western sites ($p=0.68$).*

- Alternatively: provide a table with estimates, SE, t-values and p-values! For models with lots of variables, this can be much easier.

Congratulations!

- EVERY frequentist model in R (based on a GLM) works in a similar manner to what we just went through. They are all extensions of least squared regression
- The only difference is:
  - They are designed to work with different types of response data (e.g. categorical, binomial etc.)
  - They are designed to work with different distributions
  - They have different assumptions



Response variables is…	Predictor	Response variable distribution	Possible test
Binary (0/1)	1 or more categories	binomial	Binomial test
Count	>=2 categories	Chi-squared	Chi-squared test
Continuous	2 categories	Normal	Linear model with 2 categories
Continuous	2 categories	Student's t-distribution	t-test
Continuous	>2 categories	Normal	Linear model with >2 categories (variant: ANOVA)
Continuous	Continuous	Normal	Linear Regression
Continuous	Continuous + Categorical	Normal	Linear Regression (aka multiple linear regression
Count	Continuous + Categorical	?????	?????
Binary (0/1)	Continuous + Categorical	?????	?????
Proportion	Continuous + Categorical	?????	?????
…	…	…	…

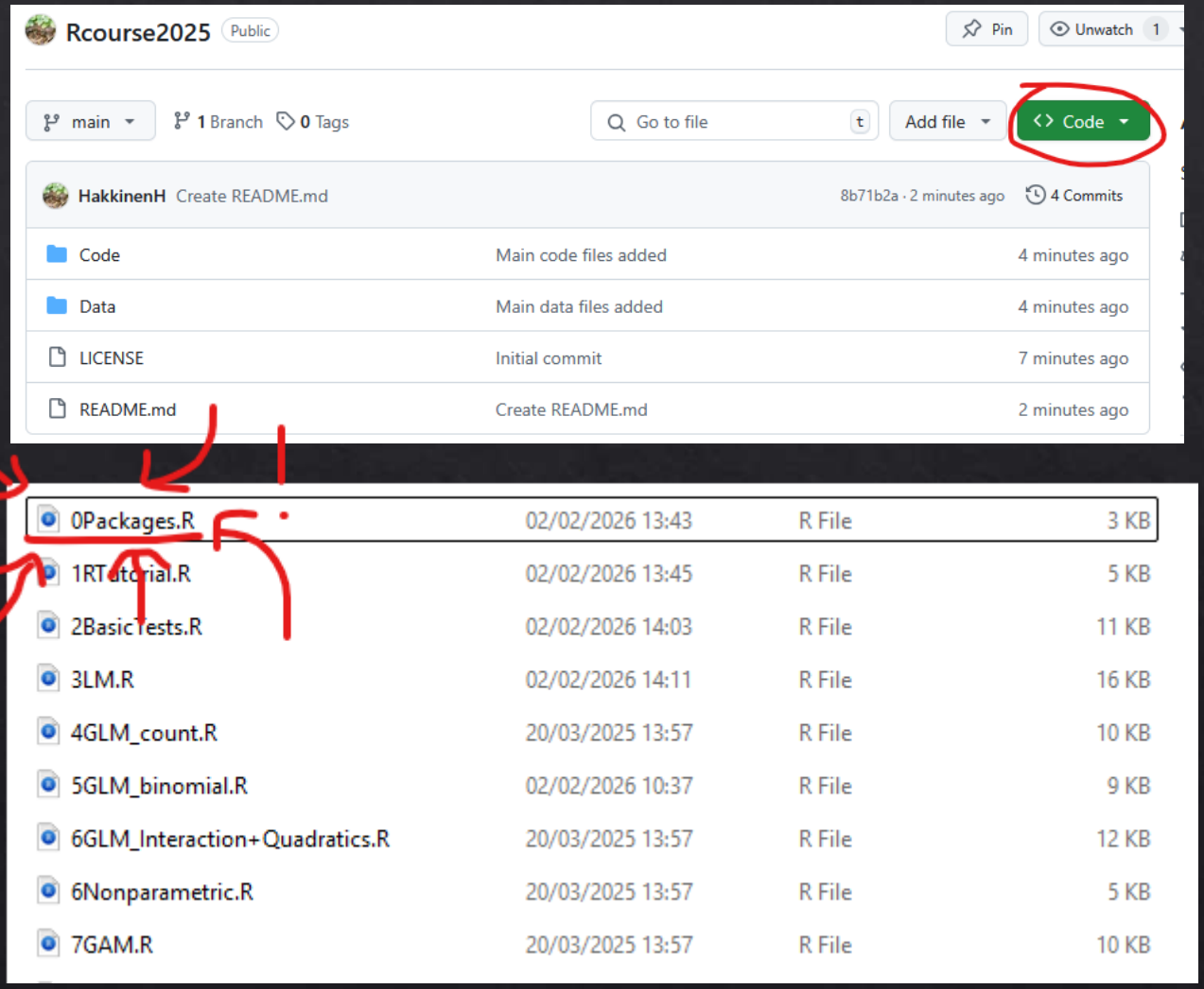
Questions

Next time…

- Learning more distributions!
  - GLMs!
- What to do if there is no distribution that fits
  - Non-parametric tests
- Adding bells and whistles
  - Random effects
- Model comparisons
  - AIC and so on…

The Practical: Part 2

- https://github.com/HakkinenH/Rcourse2026
- Code -> download ZIP -> Unzip and put somewhere useful
- Run 0Packages.R first ((if you haven't already) to install packages
- Work through 3LM.R
- Then pick whatever topic you like from there!



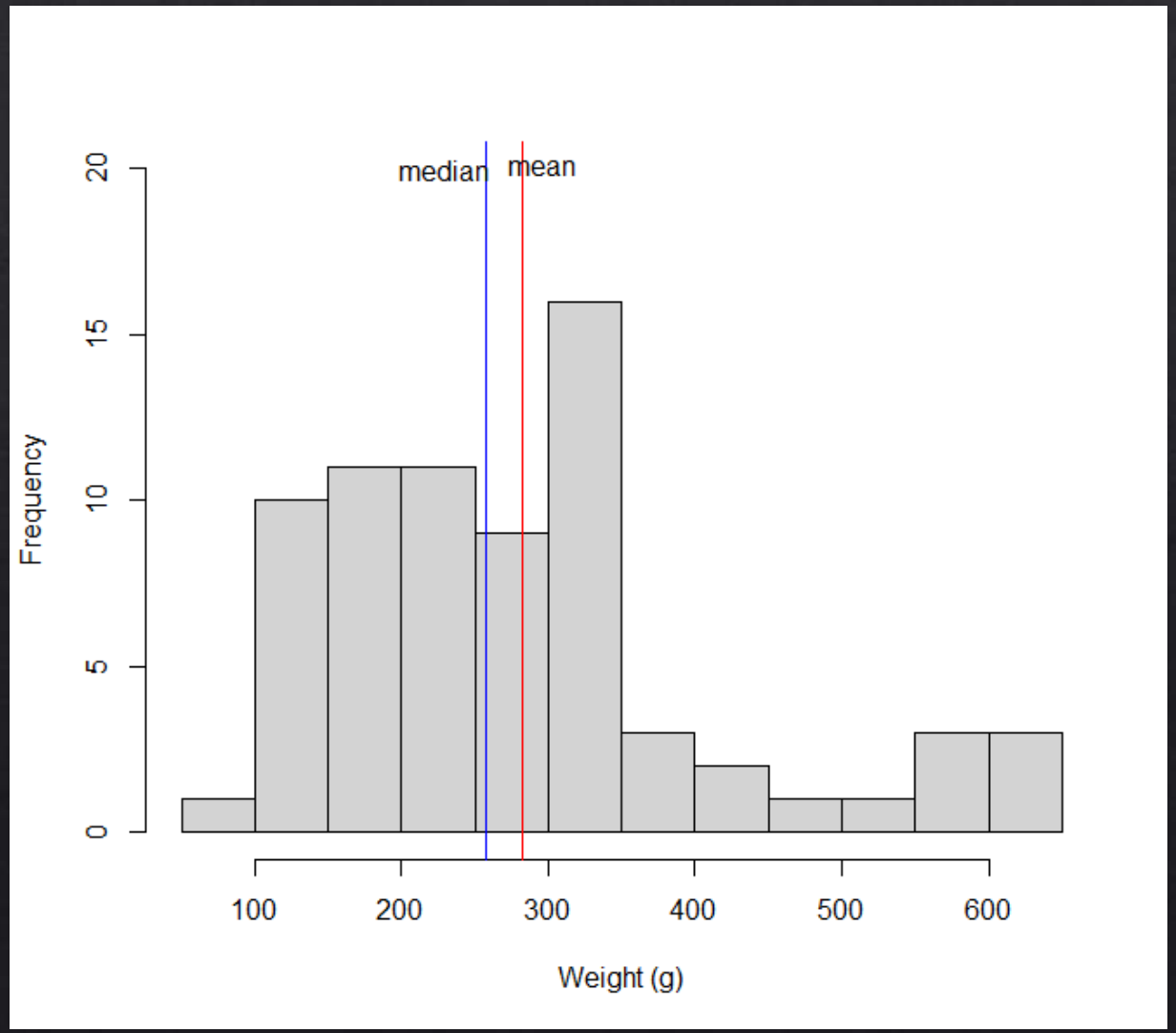
Further resources

- I used these heavily for examples and scripts in this course:
- https://statistics4ecologists-v3.netlify.app
- https://bookdown.org/ndphillips/YaRrr/
- https://r.qcbs.ca/workshop06/book-en/reviewing-linear-models.html
- https://saestatsteaching.tech/nonparametric-methods
- https://r.qcbs.ca/workshop07/book-en/mixed-model-protocol.html
- https://cran.r-project.org/web/packages/DHARMa/vignettes/DHARMa.html
- https://www.flutterbys.com.au/stats/tut/tut8.4a.html
- https://noamross.github.io/gams-in-r-course

Extra Slides

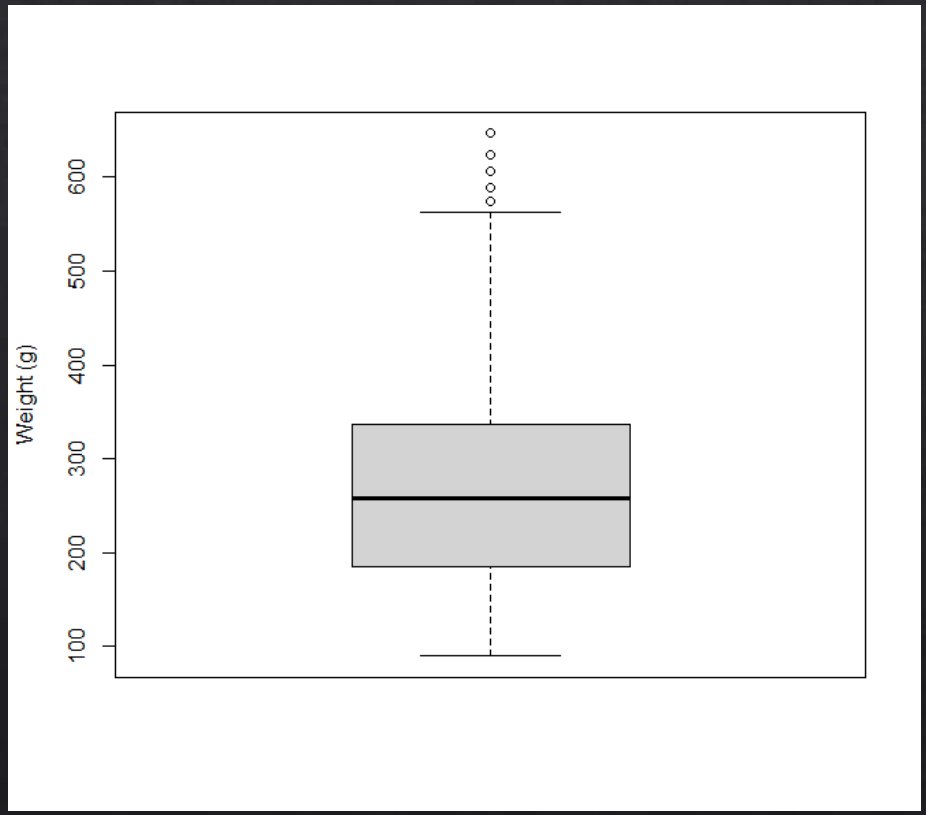
Descriptive Statistics

- Mean and median:
  - Tells us the “average” value, but little else
  - Very few individuals are actually average, so we need other ways to describe data



Descriptive Statistics

- Range:
  - Minimum and maximum values
  - Useful, but outliers can be misleading!
- Interquartile range
  - 25th and 75th percentile
  - Also useful!



Descriptive Statistics

- In statistics we talk a lot about “variance around the mean”, or “on average, how close are individuals to being average”?
- The Mean Absolute Deviance is the average of how far each data point is from the mean -> 68g.
  - On average a chick will be 68g lighter or heavier than the mean.

weight (g)	Mean	Deviance
179	282.6	-103.6
160	282.6	-122.6
136	282.6	-146.6
227	282.6	-55.6
217	282.6	-65.6
168	282.6	-114.6
108	282.6	-174.6
124	282.6	-158.6
143	282.6	-139.6
140	282.6	-142.6
309	282.6	26.4
229	282.6	-53.6
181	282.6	-101.6
141	282.6	-141.6
260	282.6	-22.6
203	282.6	-79.6
148	282.6	-134.6
169	282.6	-113.6
213	282.6	-69.6
257	282.6	-25.6
244	282.6	-38.6
271	282.6	-11.6
243	282.6	-39.6
230	282.6	-52.6
248	282.6	-34.6
327	282.6	44.4
329	282.6	46.4