Statistical Methods

F21RP - Research Methods and Project Planning

(based on slides by Frank Broz, Helen Hastie, Oliver Lemon, Judy Robertson and Tom Methven)

Learning Outcomes

- Making sense of quantitative data:

  A. Types of data

     - Nominal, Ordinal, Interval, Ratio

  B. Descriptive statistics

     - Measure of Central Tendency and Spread

  C. Inferential statistics

     - Correlation study

     - Hypothesis test

Experimental Design

1. Determine

   (a) the research questions of the project

   (b) the hypotheses being tested

2. Design the experiment(s)

3. Run the experiment AND/OR user study [takes a lot of time!]

4. Analyse measured data AND/OR questionnaires using statistical tests

5. Summarise and present the result

Experimental Design

1. Determine

   (a) the research questions of the project

   (b) the hypotheses being tested

2. Design the experiment(s)

3. Run the experiment AND/OR user study [takes a lot of time!]

4. Analyse measured data AND/OR questionnaires using statistical tests

5. Summarise and present the result

Descriptive vs Inferential Stats

Mean time on task: Baseline > System

Statistics

Predicted time on task for Baseline > System for whole population

Descriptive:
Organisation, summarisation & display of data

Inferential:
Using a sample to infer about a Population
Hypothesis testing

A. Quantitative data types

- Nominal
- Ordinal
- Interval
- Ratio
- These distinctions really matter when you are doing statistical tests!

Nominal

- Numbers which represent names or categories:
  - No intrinsic ordering
    - E.g., Male/Female or Red/Yellow/Blue

Category	1	2	3	4	5
(Meaning)	Students	Teaching Staff	Non-teaching Staff	Visitors	Other
	650	34	43	17	2

People using Heriot-Watt's website

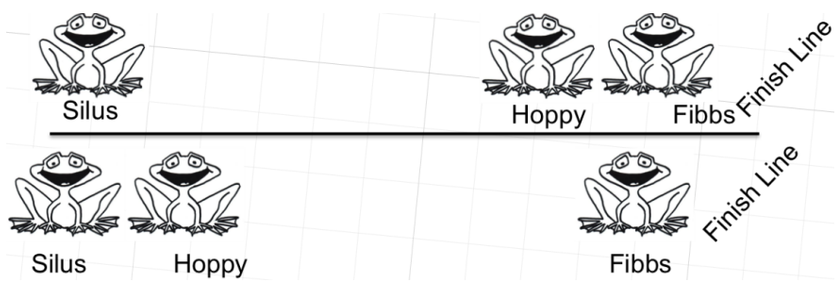
Category	A	B	C	D	E
(Meaning)	Competitive	Solitary	Associative	Parallel	Co-operative
	23	19	8	6	7

Type of play method supported by computer games

Ordinal

- Order - positioning information
  - Relative positions but not distance between scores
  - Usually numbers that rely on human judgement
    - e.g. marking, questionnaire data, Likert scale

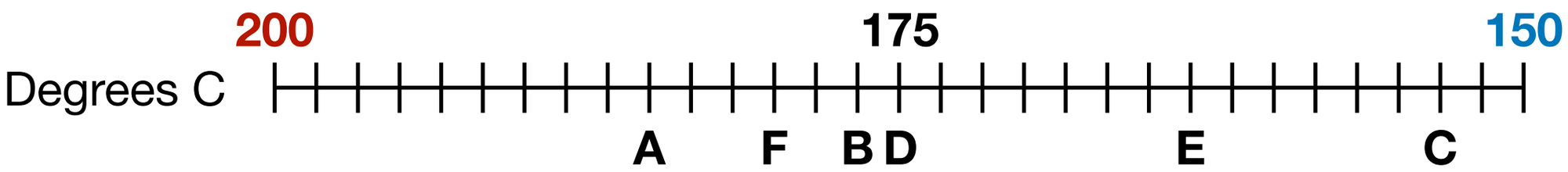
Person	Score	Rank of score
A	18	5.5
B	25	7
C	14	1
D	18	5.5
E	15	3
F	15	3
G	15	3
H	29	8



- Can you say that H's performance is twice as good as E, F or G's?

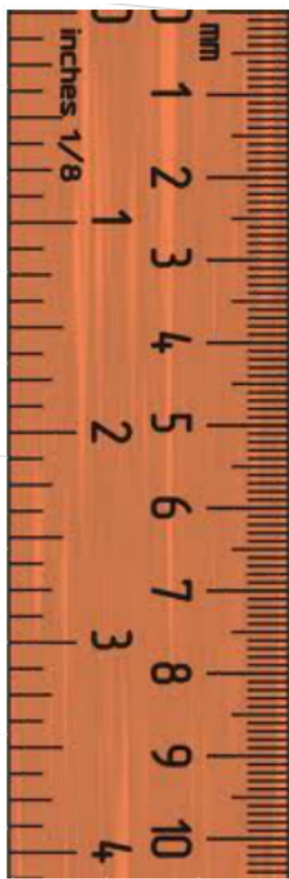
Interval

- Data which has been measured using a scale with equal intervals on it
  - No absolute zero
  - E.g., temperature at which optimum performance is gained



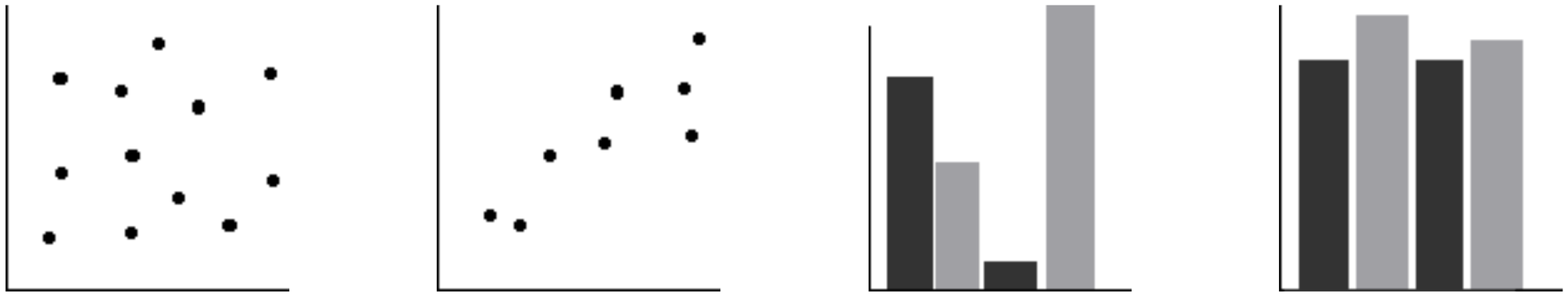
Ratio

- Ratio scales = interval scales with a necessary and absolute zero:
  - Timing
    - There is no negative number of seconds
  - Height
    - There is zero height but not negative height



Making sense of quantitative data

- Draw or chart the data
  - “Eyeball” the data -> interesting patterns
- Gather descriptive statistics
  - summarise a set of data into just a couple of numbers that represent the entire data set
- Helps decide whether to bother carrying out statistical tests



B. Descriptive Statistics

- Measures of Central Tendency
  - Mode
  - Median
  - Mean
- Measures of Spread
  - Range
  - Standard Deviation
  - Variance

Measures of central tendency

- Single number that is used to represent the general magnitude of scores in the data set
- Representative value of the set
  - Mode
  - Median
  - Mean

Mode

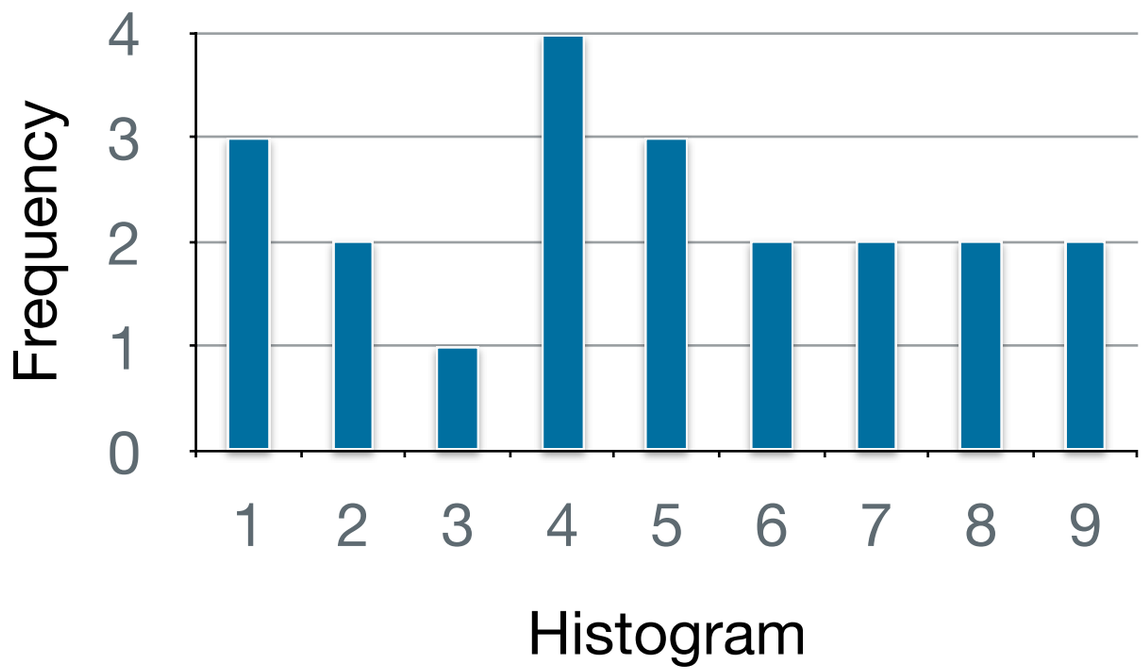
- **The most frequent score**

Example: 5 9 7 4 6 8 2 4 1 3 5 1 4 6 9 8 7 5 2 4 1

Ordered: 1 1 1 2 2 3 4 4 4 4 5 5 5 6 6 7 7 8 8 9 9

Frequency count: 3 2 1 4 3 2 2 2 2

Mode



Mode

- There is no mode when all the scores are different (or there is the same number of many scores)
  - E.g., 2 5 7 9 11 1 8 0
- Sometimes there is more than one mode
  - E.g., 2 5 7 9 9 11 1 1 8 0
- Limitation
  - Does not take into account other scores so comments about distribution may be misleading

Median

- The middle score (of an ordered set)

Example (odd # of scores):

3 8 11 11 12 13 24 35 46

Median

Median

- The middle score (of an ordered set)

Example (even # of scores):

3 8 11 11 12 13 24 35 46 48

Median

(12+13) / 2 = 12.5 = median

Limitation
- does not take into account extreme scores

Mean

- The arithmetic average of scores
  - mathematical centre of the distribution of scores
  - BUT it is not the middle score
    - it is the computed centre

$$Mean : \bar{x} = \frac{1}{N} \sum_{i} x_i$$

Mean

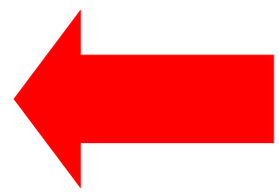
- The arithmetic average of scores

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

Example: 3 8 11 11 12 13 24 35 46 48

$$\bar{x} = \frac{3 + 8 + 11 + 11 + 12 + 13 + 24 + 35 + 46 + 48}{10}$$

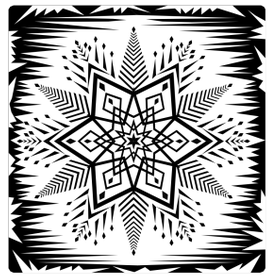
$$\bar{x} = \frac{211}{10} = 21.1$$



Mean

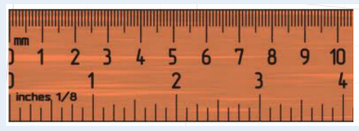
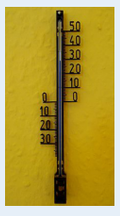
Beware of inappropriate averaging...

With your head in an oven
and your feet in ice
you would feel,
on average,
just fine



The majority of people have more than the average number of legs
(Mean = 1.9999)

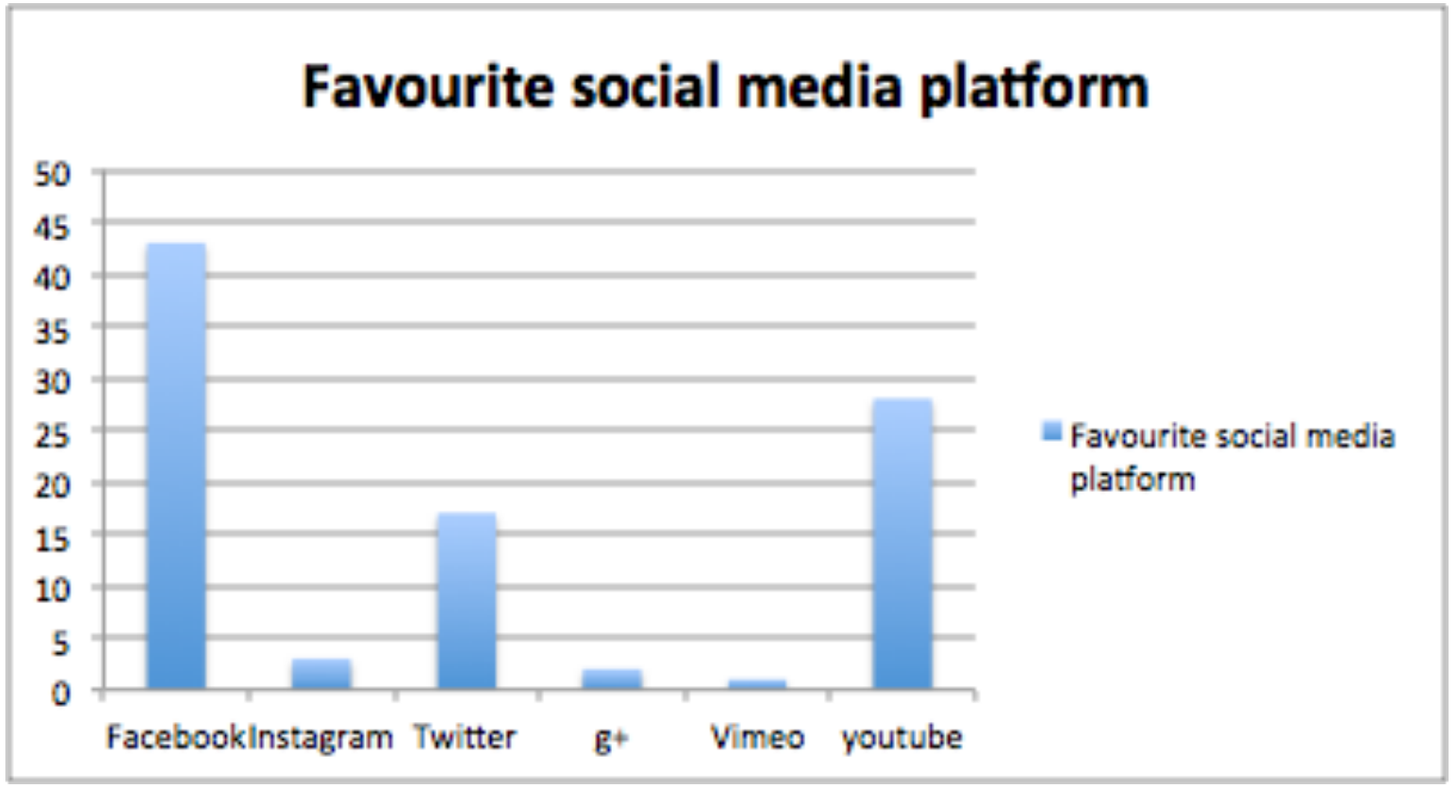
What statistic to use?

Data type	Reporting Central Tendency
Nominal 	Mode
Ordinal 	Mode or Median
Interval and ratio  	Mode, Mean or Median

What statistics to use?
Nominal

- If Nominal: you must use the mode
- Mode represents the most popular

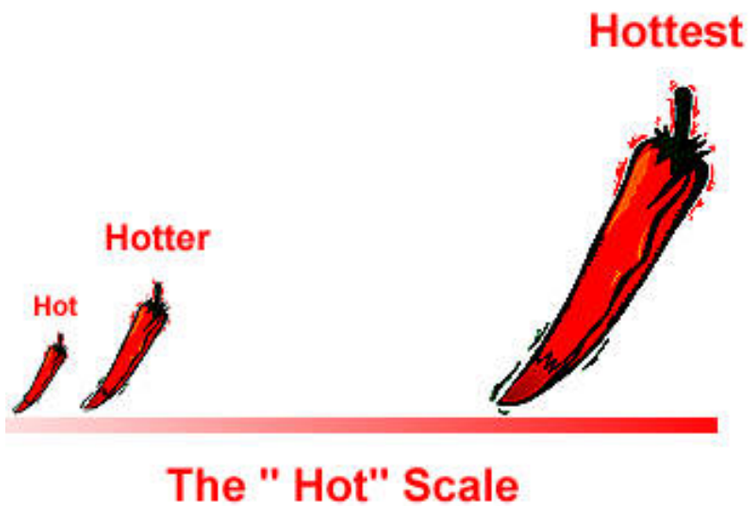
	Favourite social media platform
Facebook	43
Instagram	3
Twitter	17
Google+	2
Vimeo	1
Youtube	28

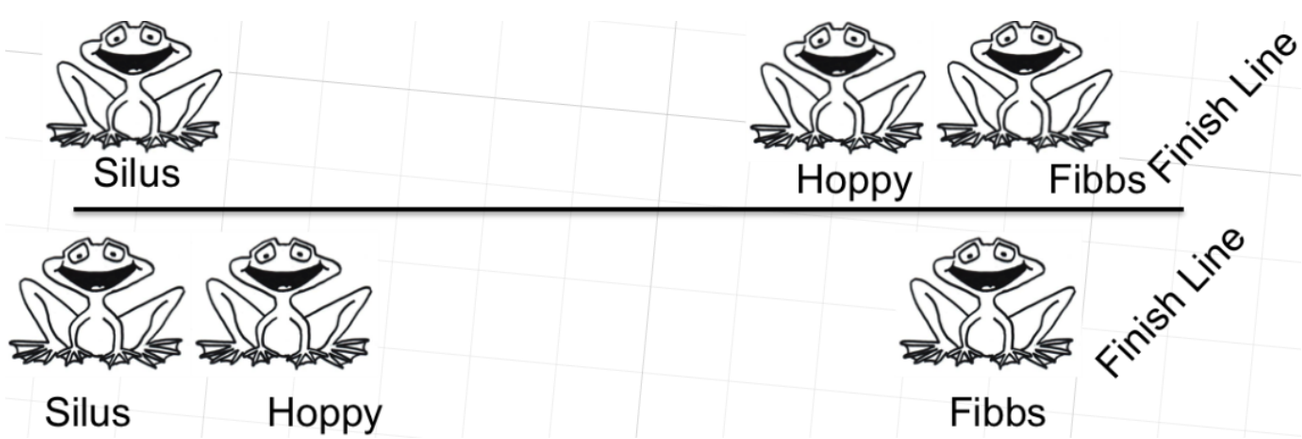


What statistics to use?
Ordinal

- If Ordinal: you can use either the mode or the median

- No averaging of Likert scale responses!

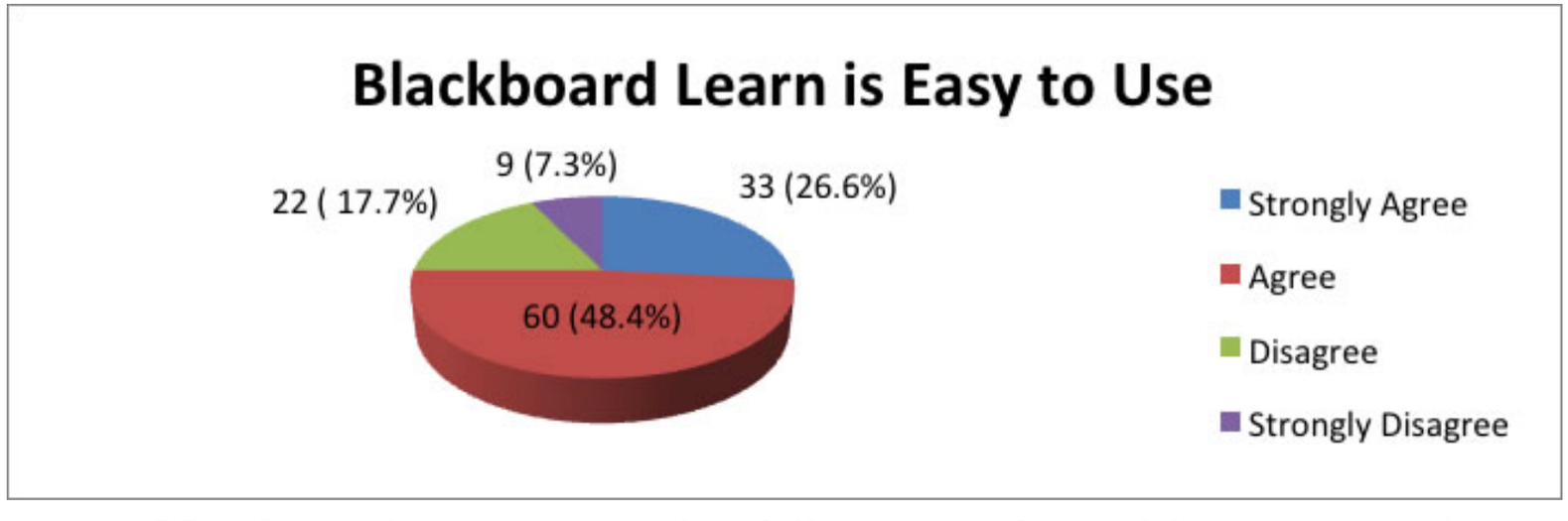




What statistics to use?
Ordinal

- Mode = the most popular response
- Examples
  - mode = 4, i.e., 80% of participants rated the game as 4 on a 5-point rating scale
  - 60% of females compared to 80% of males rated the interface 6 or above on a 7-point rating scale

4-Strongly Agree
3-Agree
2-Disagree
1-Strongly disagree

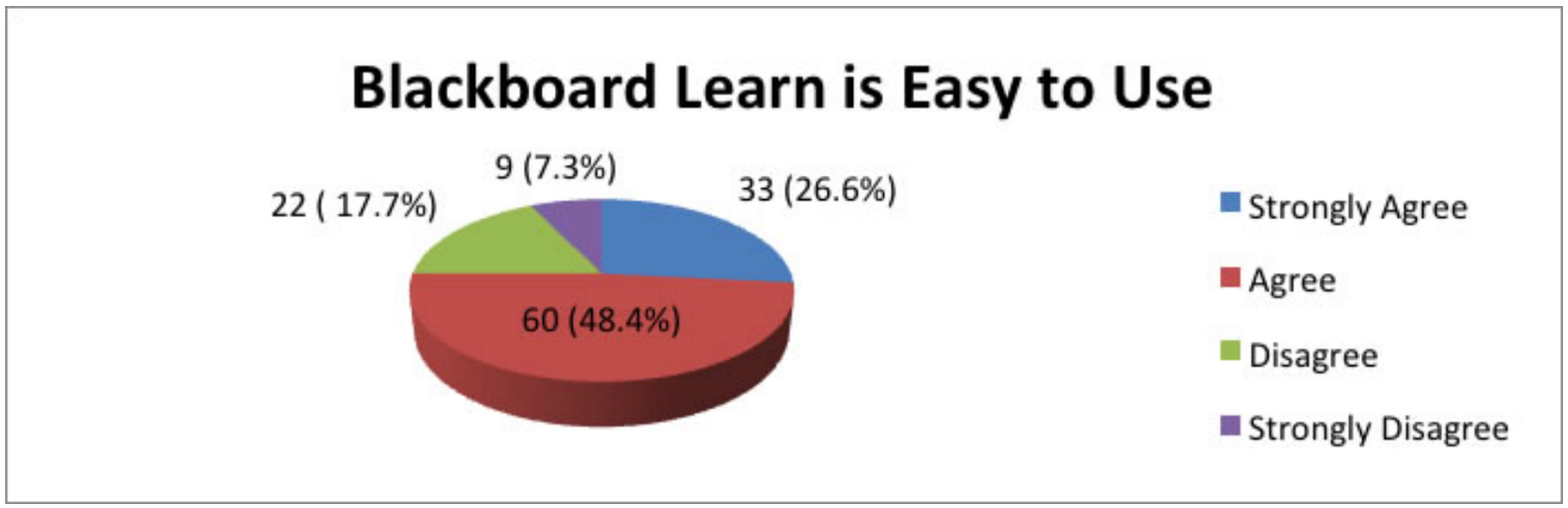


Mode = 3 on a 4-point Likert scale with 48.4% of the population selecting ‘Agree’

What statistics to use? Ordinal

- Median

4-Strongly Agree
3-Agree
2-Disagree
1-Strongly disagree

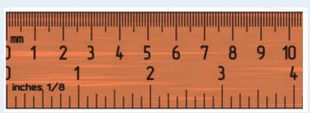
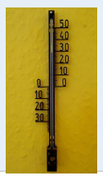


1111111112222222222222222233333333333333333333
333333333333333333333333333333344444444444444
4444444444444444444

Median = 3

What Statistic to Use?

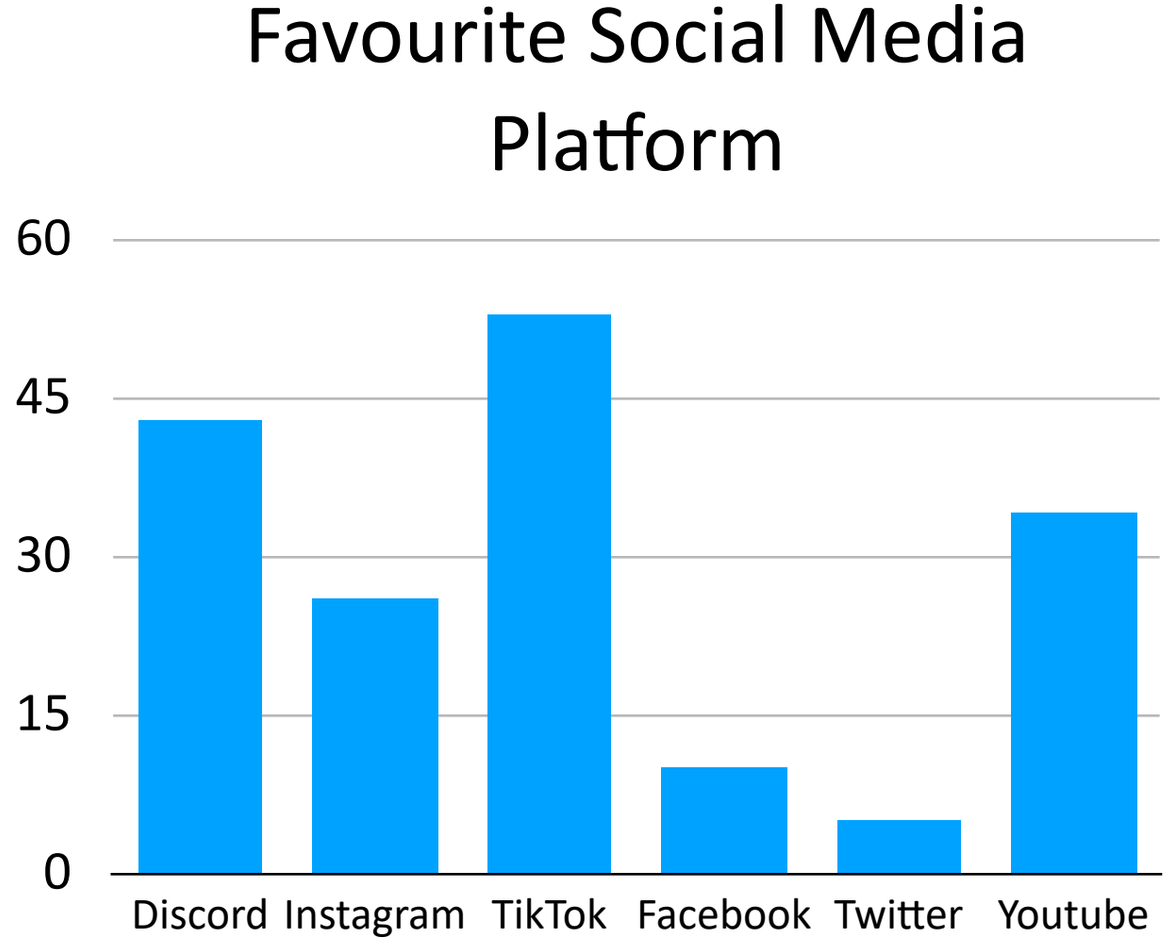
- If Interval or Ratio: use any but the mean is the most typical
- Examples
  - The average time to complete task is 15.5 seconds
  - The average temperature in Edinburgh in November is 8 degrees Celsius :(

Data type	Reporting Central Tendency
Nominal	Mode
Ordinal	Mode or Median
Interval and ratio  	Mode, Mean or Median

What statistics to use? Nominal

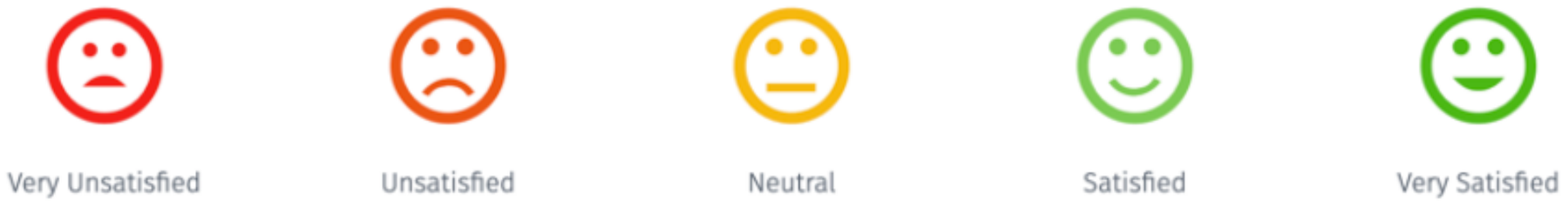
- If Nominal: you must use the mode
- Mode represents the most popular

	Favourite social media platform
Discord	43
Instagram	26
TikTok	53
Facebook	10
Twitter	5
Youtube	34
Total	128



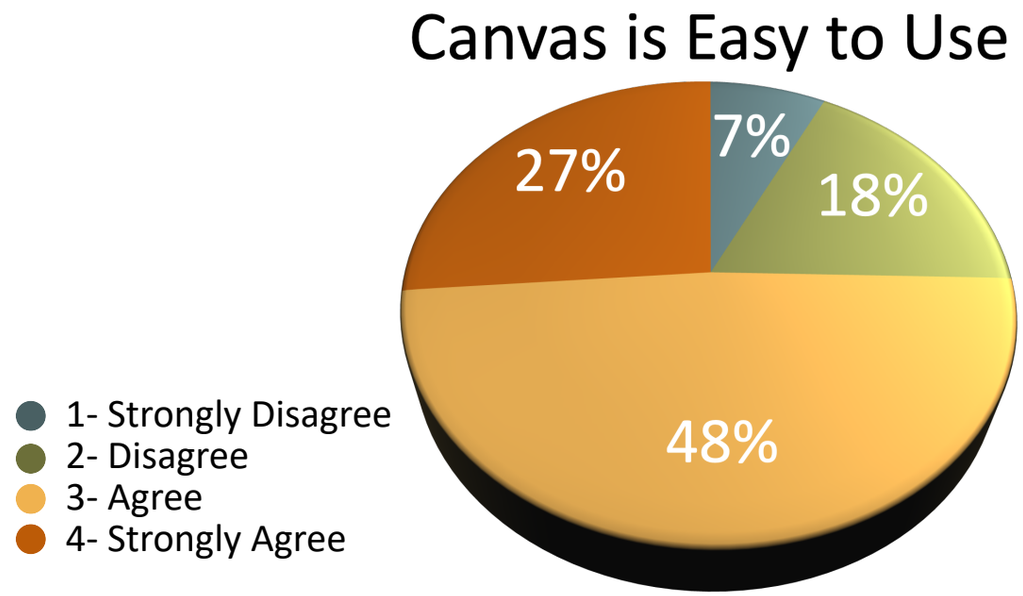
What statistics to use? Ordinal

- If Ordinal: you can use either the mode or the median
- No averaging of Likert scale responses!



What statistics to use? Ordinal

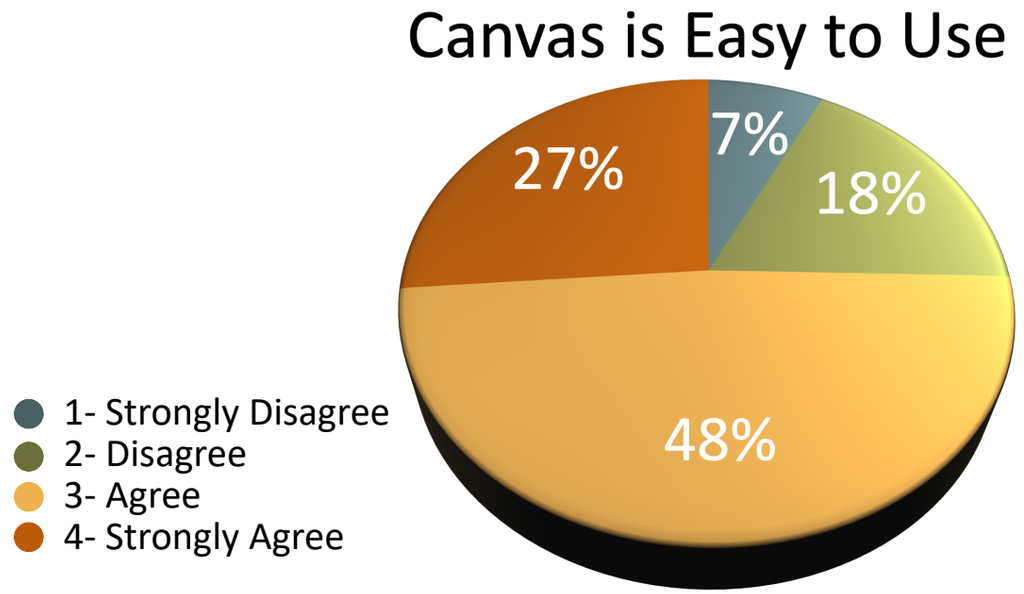
- Mode = the most popular response
- Examples
  - Mode = 4, i.e., 80% of participants rated the game as 4 on a 5-point rating scale
  - 60% of females compared to 80% of males rated the interface 6 or above on a 7-point rating scale



Mode = 3 on a 4-point Likert scale with 48.4% of the population selecting ‘Agree’

What statistics to use? Ordinal

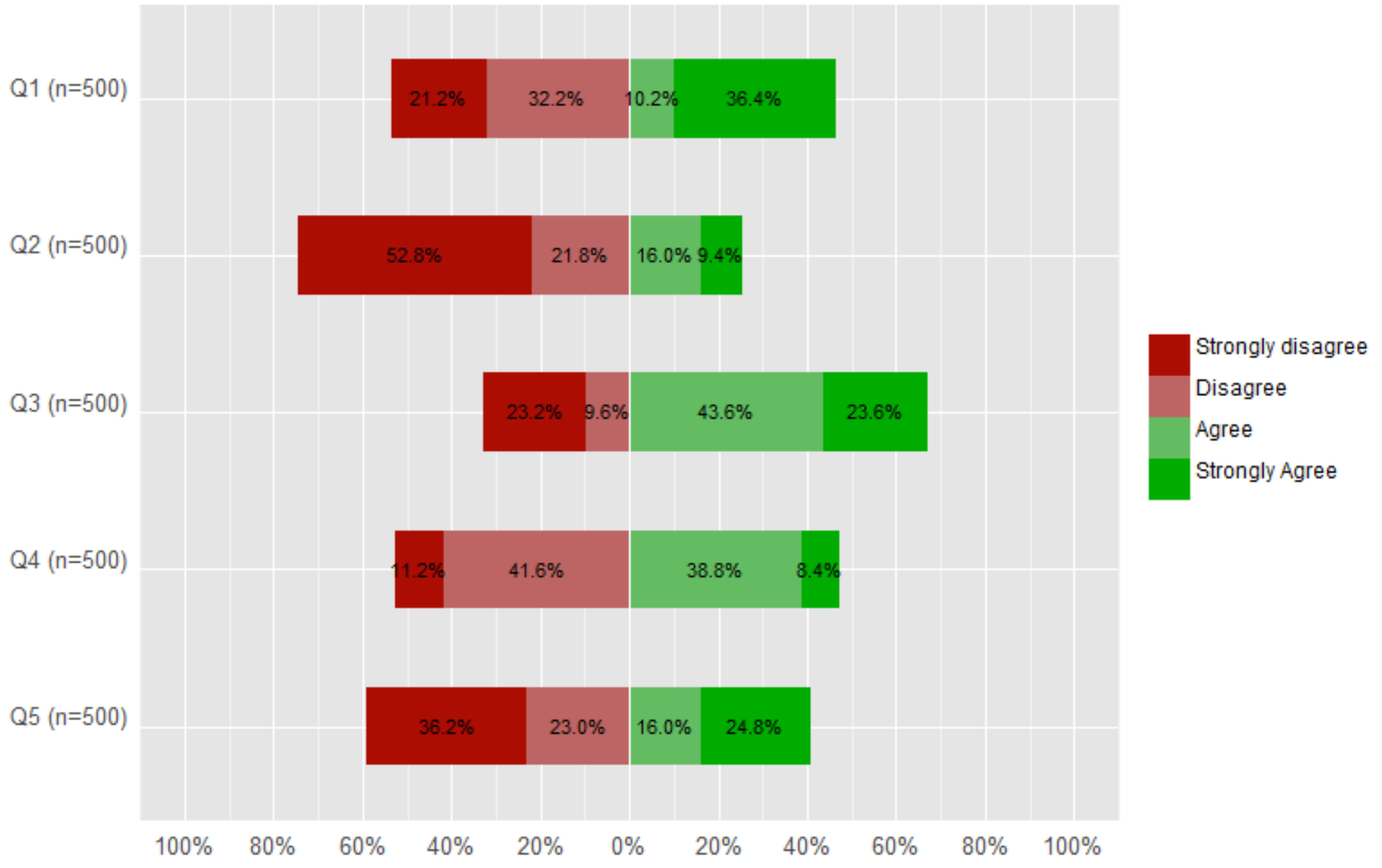
- Median



11111111122222222222222222233333333333333333333
33333333333333333333333333333333333344444444
4444444444444444444444444

Median = 3

Likert scales visualisation



What Statistics to Use? Interval / Ratio

- If Interval or Ratio: use any but the mean is the most typical
- Examples
  - The average time to complete the task is 15.5 seconds
  - The average temperature in Edinburgh in November is 8 degrees Celsius :(

B. Descriptive Statistics

- Measures of Central Tendency
  - Mode
  - Median
  - Mean
- Measures of Spread
  - Range
  - Standard Deviation
  - Variance

Measures of Spread

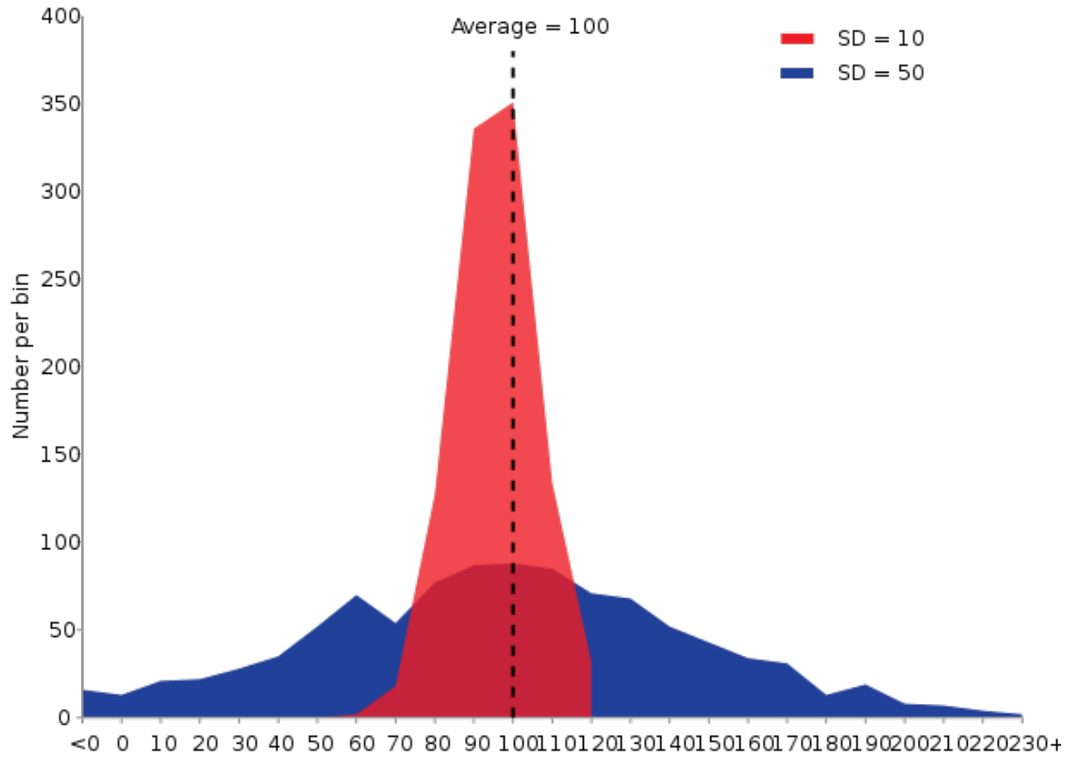
- Magnitude of deviation from central tendency
  - Range
  - Standard Deviation
  - Variance

Dispersion: Range, variance and Standard deviation

- Average time on task:
  blue = men and red = women

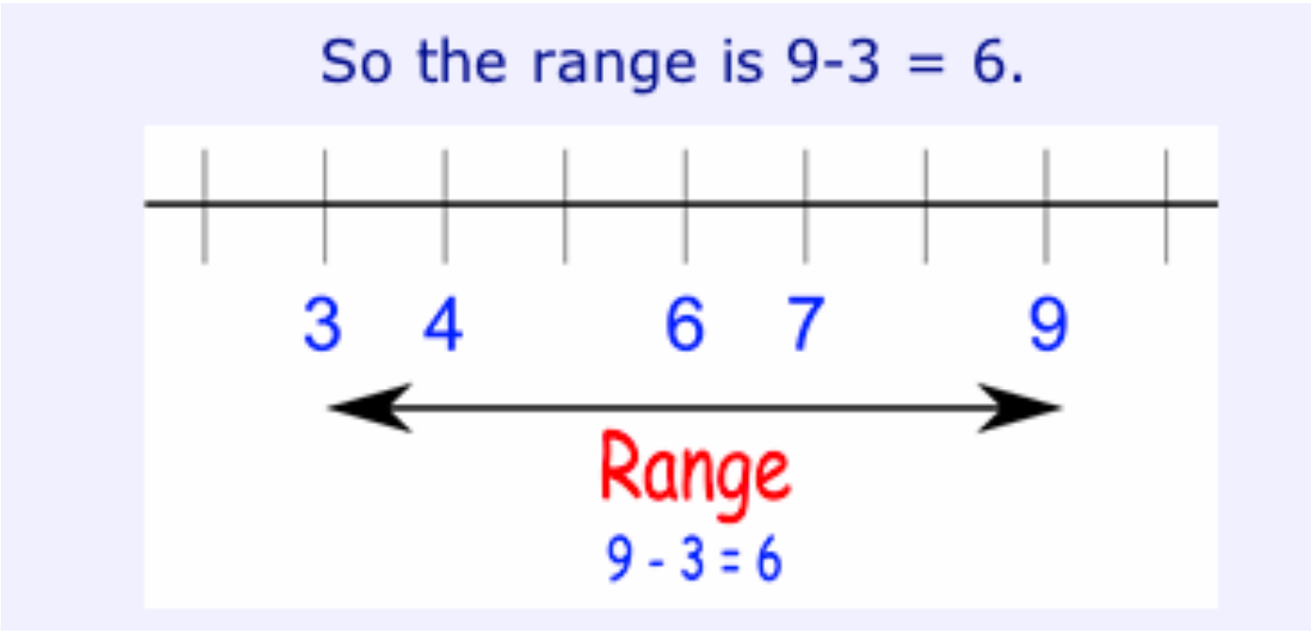
- They have equal mean and different standard deviations

  - What does this tell you?



Dispersion: Range

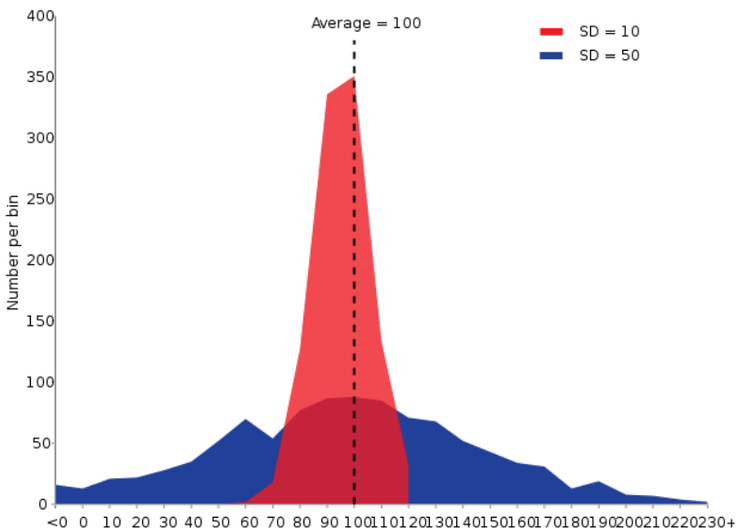
- Range: difference between the lowest and the highest number
- Example: {4,6,9,3,7}
- Range can be misleading. E.g. {8, 11, 5, 9, 7, 6, 3616}
  - Lowest value: 5; highest value: 3616
  - Range is 3615-5=3611



Variance/Standard Deviation

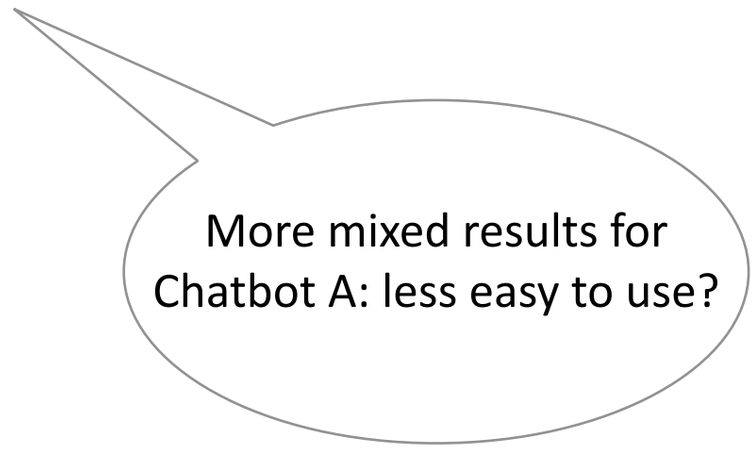
- Variance captures the spread of numbers:
  - How far the numbers are from the mean
- Standard deviation = the square root of variance
  - Preferred over variance as it is expressed in the same units as the data
- Low standard deviation
  - Data points tend to be very close to the mean
- High standard deviation
  - Data points are spread out over a large range of values

Use R, Python, or Excel to calculate

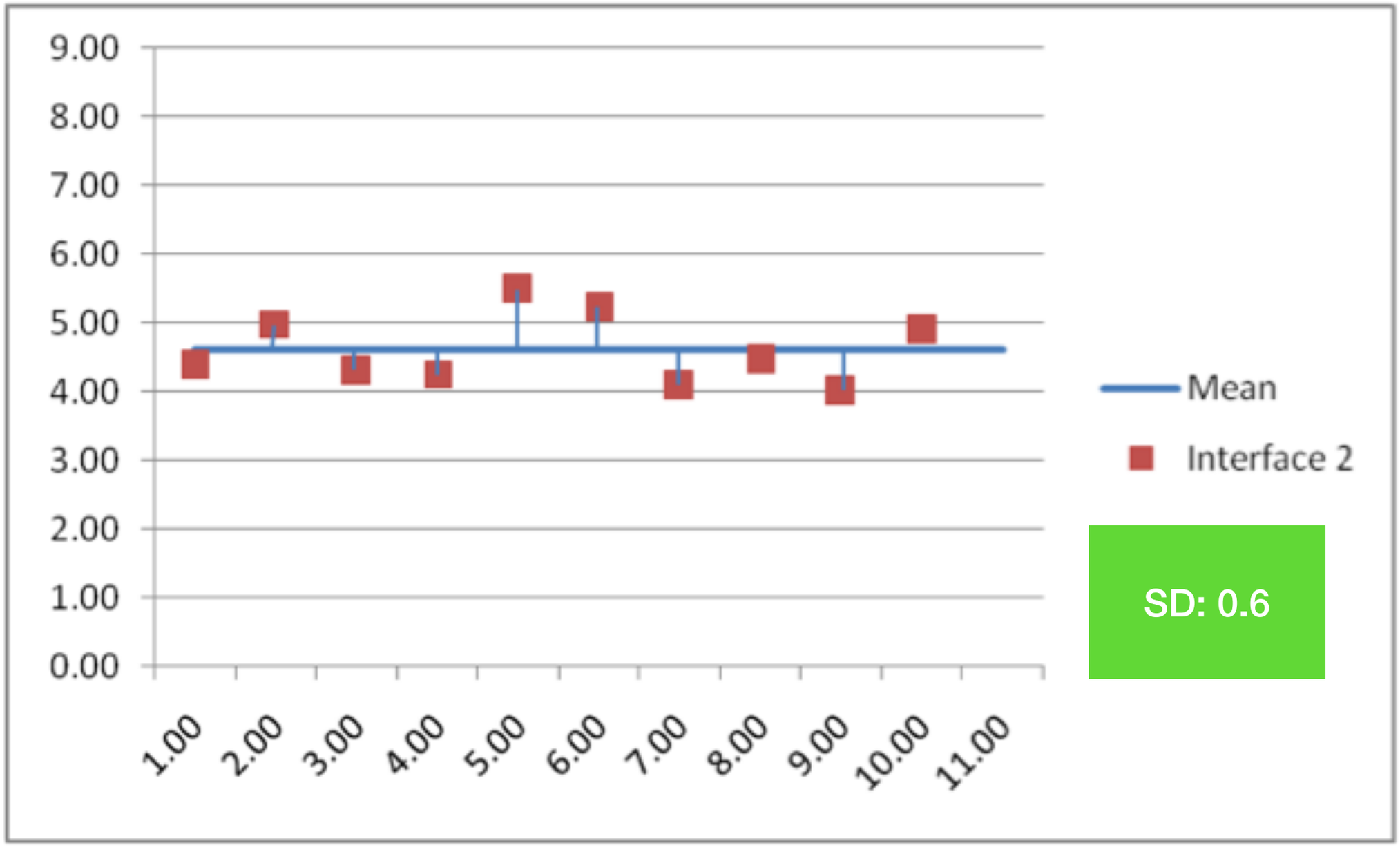


Why is dispersion important?

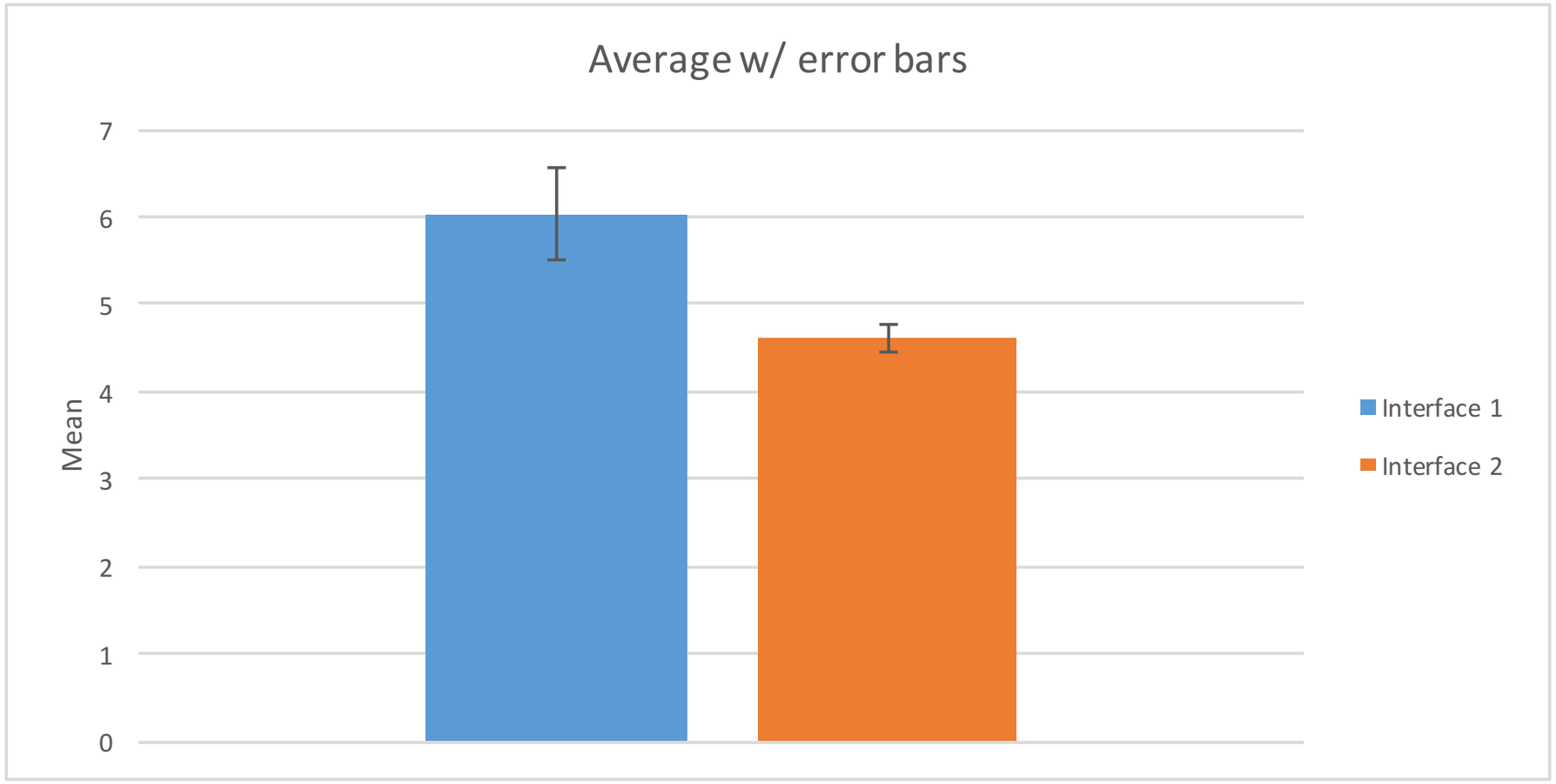
- You can say things like
  - Children only use the range 3-5 on a 5-point rating scale
  - The time to look up a hotel using Chatbot A had a higher variance/standard deviation than Chatbot B



Example -
Putting it all together



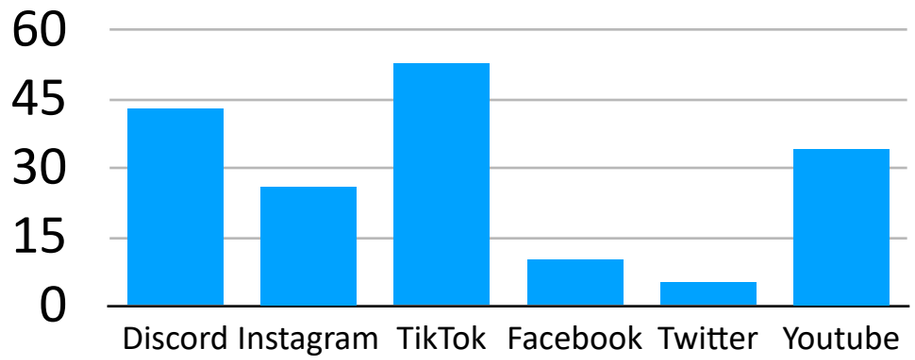
Example - Error Bars

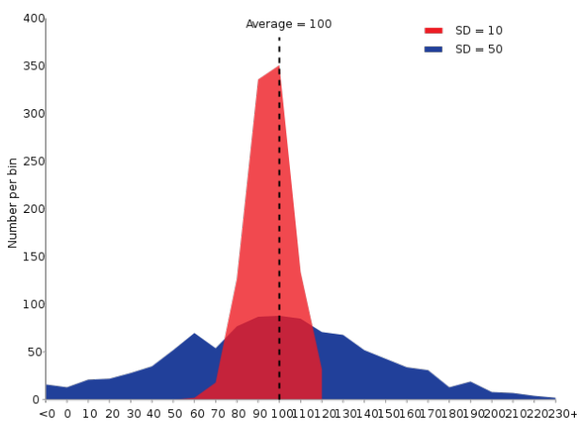


- Error bars: Plot standard error
- Represent variability of the data
- Often represent one standard deviation, or a certain confidence interval (e.g. 95%)

Descriptive Statistics - Summary

- Spend ‘quality time’ investigating your data
- Describe the central tendency:
  - Frequencies, percentages
  - Mode, Median, Mean
- Describe the variability:
  - Min, Max, Range
  - Standard Deviation, Variance





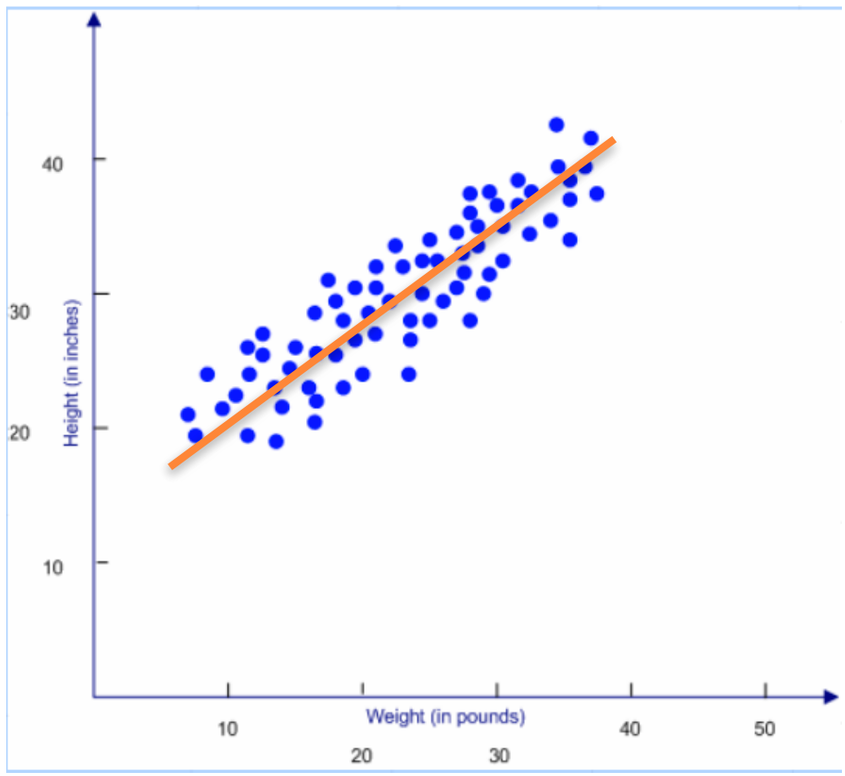
C. Inferential Statistics

- Use experimental sample to infer about population
- Investigate relationships between variables
  - Correlation
  - Hypothesis testing

Measuring Relationships

- We think that there might be a relationship between how tall someone is and how much they weigh.
  - How do we investigate if this is true?
- Correlations don't assume a cause/effect: just some relation
- Scatterplots

Correlation Coefficient r

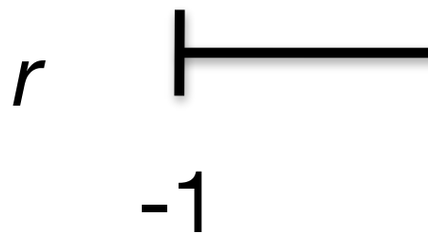


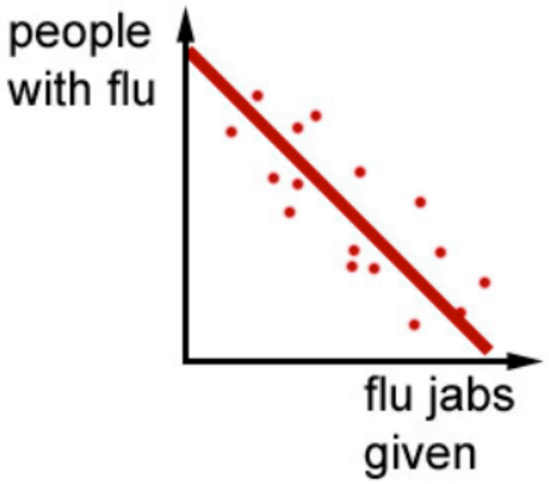
Correlations

Perfect negative relationship

No relationship

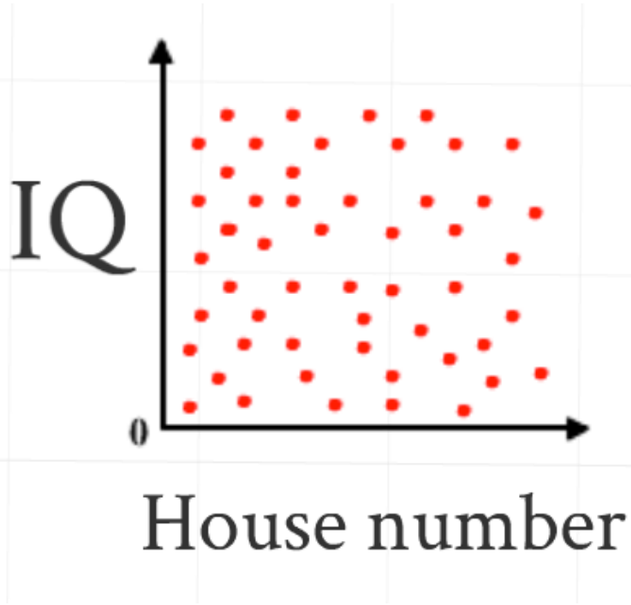
Perfect positive relationship

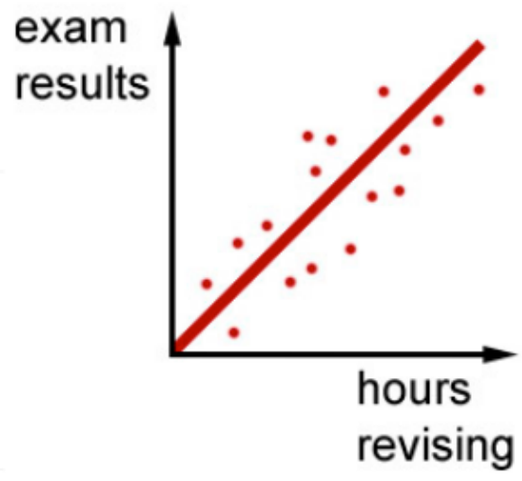




NEGATIVE CORRELATION
- when more jabs are given the number of people with flu falls.
- flu jabs prevent flu.

r=-0.5





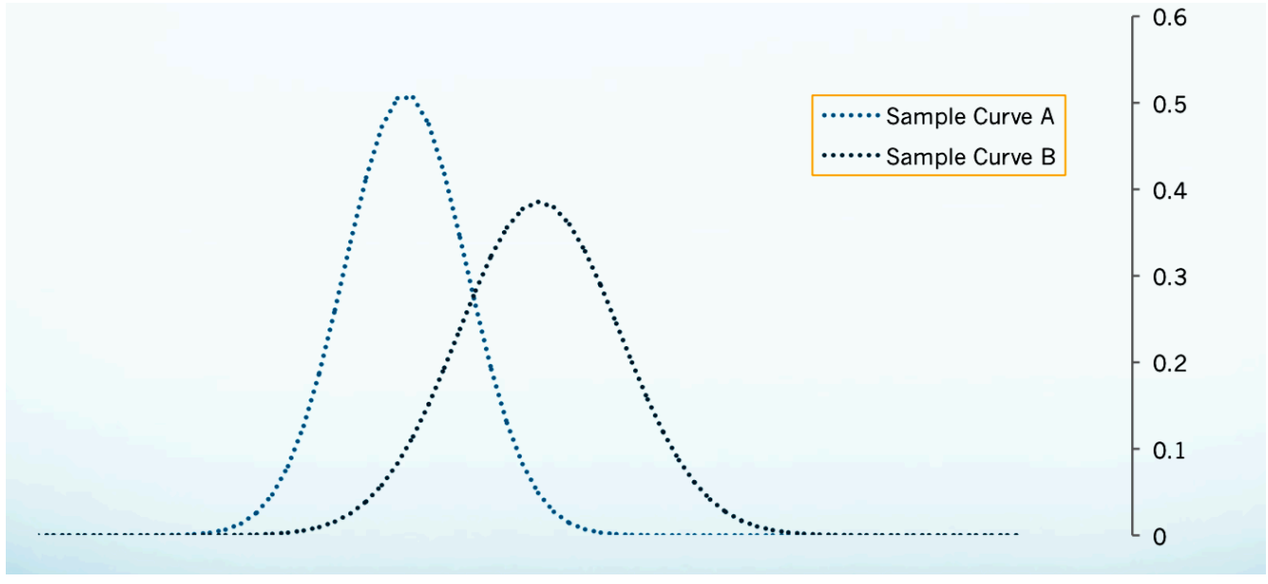
POSITIVE CORRELATION
- people who do more revision get higher exam results.
- revising increases success.

r=+0.5

Correlation doesn't mean causation!

Hypothesis testing

- There are many statistical significance tests
  - for different types of data
  - and different experimental designs
- Tests provide a way of
  - figuring out if a phenomenon is
    - random chance or
    - a real effect
  - measuring the strength of experimental evidence



Hypothesis testing with statistical significance tests

- Step 1: Create a hypothesis
- Step 2: Decide on your experimental design
  - Use paired (within-subjects) or unpaired (between-subjects) test
- Step 3: Run the test
  - Look at the output probability of the test
    - measure of reliability

Step 1: Create a Hypothesis

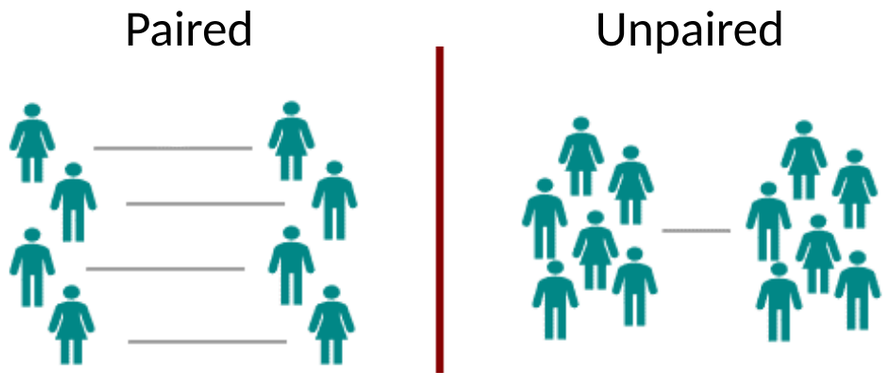
- Every hypothesis has a null hypothesis (H0)
  - H1 = Computers are better at playing chess than humans
  - H0 = Computers are NOT better at playing chess than humans
- The null hypothesis is the state of the world until proven different

Reject H0?

- In hypothesis testing we either reject or accept the null hypothesis
  - Example: Smoking
    - H1=Smoking causes cancer
    - H0=Smoking does *not* cause cancer.
      - There is evidence which allows us to reject this null hypothesis
  - Example: Shirt color
    - H1 = People with blue t-shirts are more intelligent
    - H0= People with blue t-shirts are *not* more intelligent
      - No evidence so we cannot reject the null hypothesis
      - We accept the null-hypothesis
- The null hypothesis has more weighting
  - need significant evidence to reject the null hypothesis

Step 2: Paired or unpaired?

- Depends on your experiment design
- Within-subjects → paired
  - same number of data points
  - same source of data



- Between-subjects → unpaired
  - different source of data (different subjects)
  - Potentially different number of data points
    - different # of people in each condition

Hypothesis testing with statistical significance tests

- Step 1: Create a hypothesis
- Step 2: Decide on your experimental design
  - Use paired or unpaired test
- **Step 3: Run the test**
  - Look at the output probability of the test
  - p-value - the probability that what we are seeing just happened by chance

Step 3: Run the test

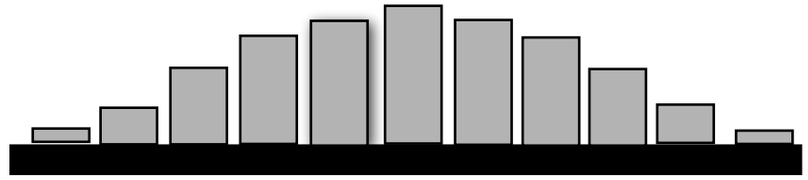
- If the result from your test is
  - $p < 0.05$
- Then we
  - reject the null hypothesis
  - your hypothesis is (probably) right
- Else
  - we accept the null hypothesis
  - better luck next time :(

Step 3: Run the test

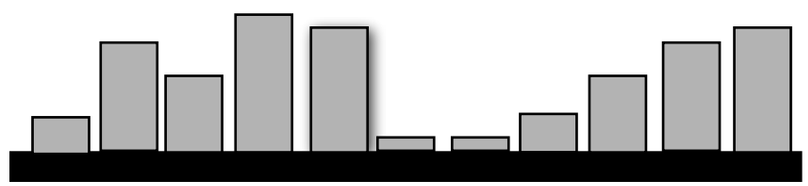
- We can either ‘Reject’ or ‘Accept’ the null hypothesis
  - based on the probability (p) that the evidence is just a random event - that it happened by chance
- if p is small (i.e. $p < 0.05$)
  - there is little chance it’s a random event that happened by chance.
- If $p = 0.05$,
  - we are 95% confident we can reject the null hypothesis
- $p < 0.05$ STANDARD THRESHOLD in CS

Which statistical significance test to use?

- Parametric
  - e.g., t-test



- Non-parametric
  - e.g., Mann-Whitney



T-test (or Student's T-test)

- Test if two groups are different
  - Comparison of means
    - Interval or ratio data
- Assumptions
  - Populations are normally distributed
  - Variances are equal
- Robust
  - For small sample sizes if assumptions hold
  - For large (> 30) sample sizes even if assumptions violated

Example -
T-Test using Excel

- TTEST in Excel will give a 'p-value' directly

Person 1	4.28	4.38
Person 2	2.78	4.99
Person 3	7.63	4.3
Person 4	7.93	4.27
Person 5	7.19	5.5
Person 6	5.73	5.22
Person 7	8.4	4.09
Person 8	5.88	4.46
Person 9	5.6	4
Person 10	4.89	4.9

=TTEST(G2:G11,H2:H11,1,1)

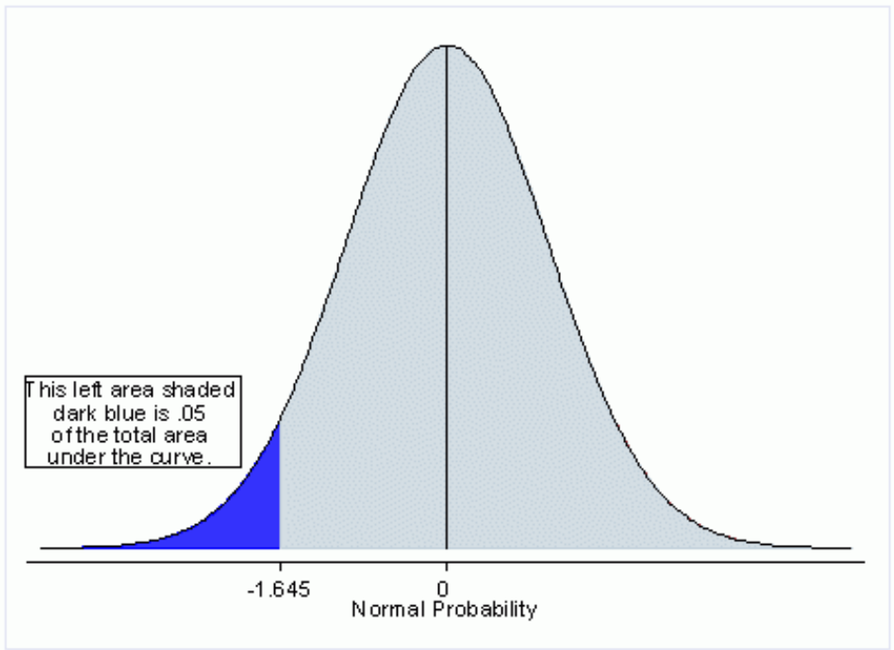
TTEST(**array1**, array2, tails, type)

array1 - series of results
array2 - series of results

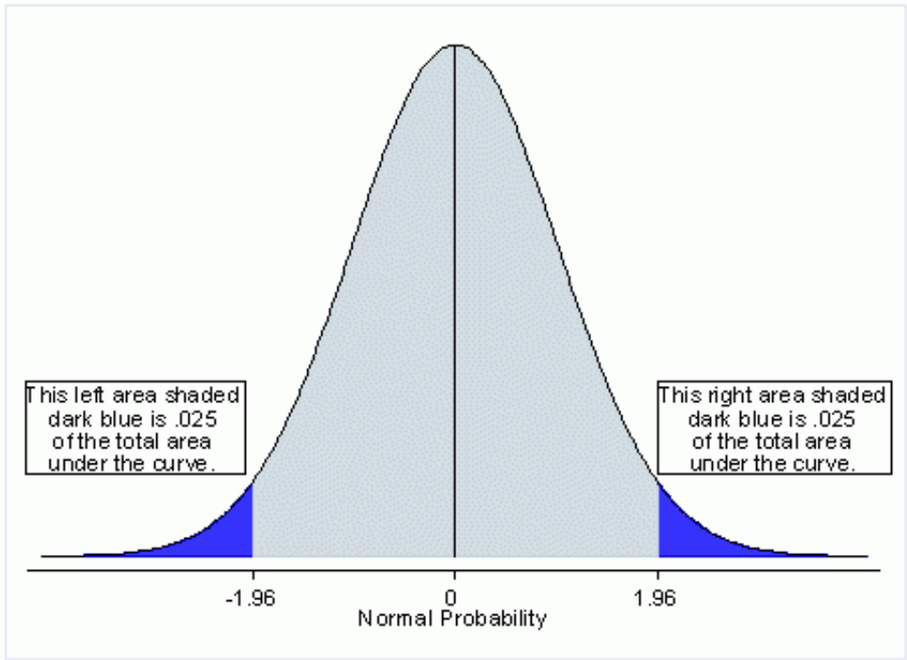
'tails' = 1 or 2
'type' = 1 for dependant
t-test

One or two tailed t-test

- Null hypothesis takes different forms
  - One-tailed - H0 is difference in one direction
  - Two-tailed - H0 is no difference



One-tailed



Two-tailed

One tailed t-test

- Are you only interested in difference in one direction?
  - Large difference in opposite direction is not significant
- Example
  - H1= Our new app will be faster to use than the old app
  - H0 = Our new app will not be faster to use than the old app

H1 = Mean A < Mean B

H0 = Mean A ≥ Mean B

Two tailed t-test

If you're not sure, use 2 tailed t-test

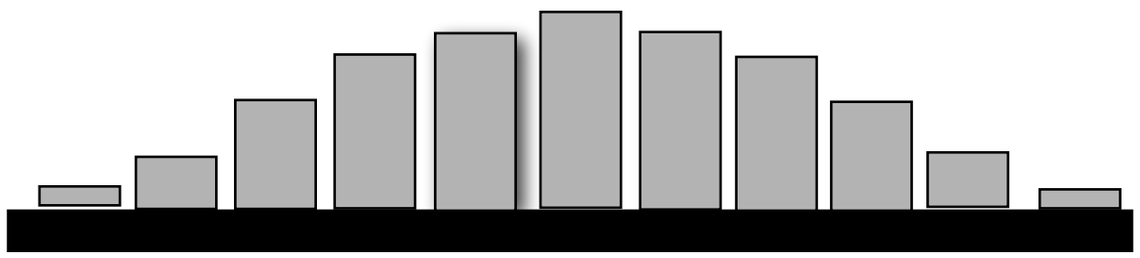
- Example
  - Two new maths teachers with different teaching methods (Mr Clark and Mrs Brown) join a school
  - H1= One class will have higher test scores than the other
  - H0=Both teachers have the same average test scores

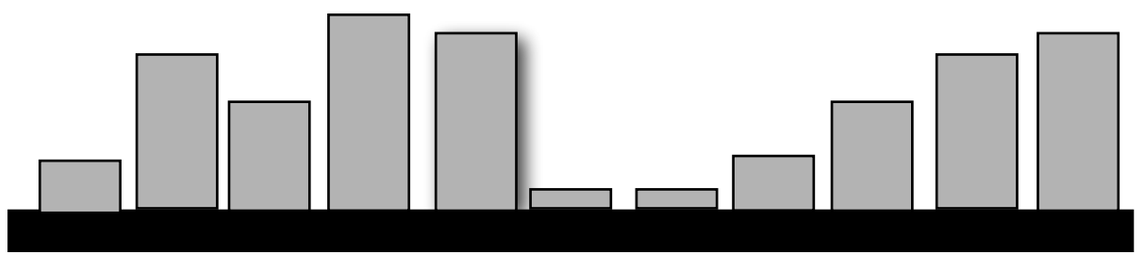
H1 = Means are different

H0 = Means are same

Non-parametric tests

- T-test assumes that your data is normally distributed
  - What if it's not?
  - Also, compares means (interval or ratio data)
- Non-parametric tests don't assume data is from a certain type of distribution



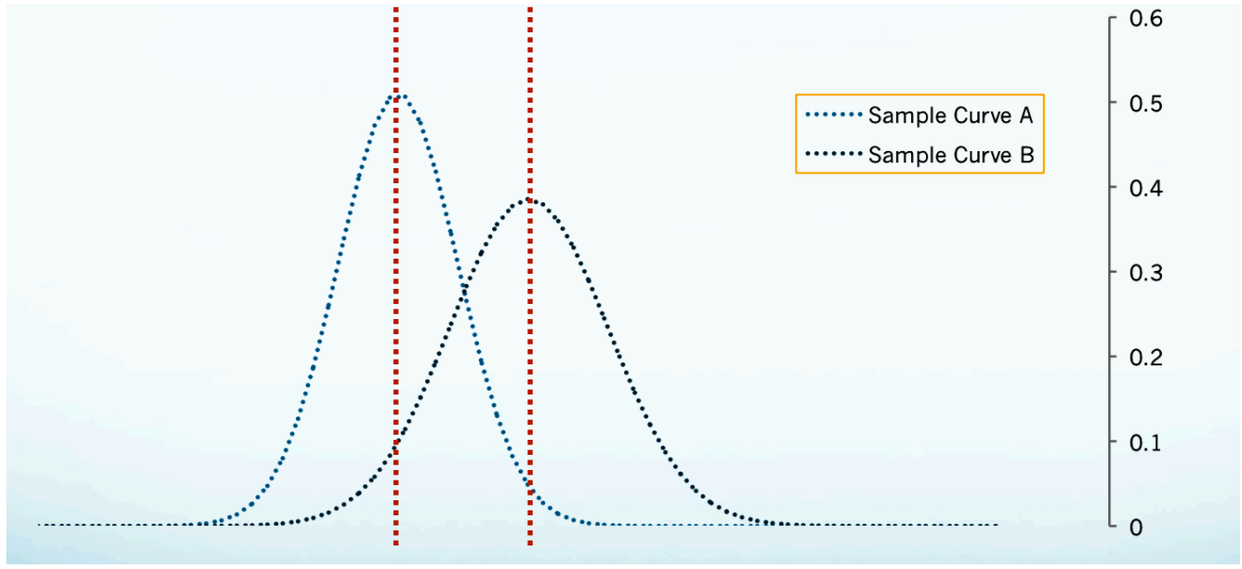


Non-parametric tests

- Hypothesis testing
  - Mann-Whitney (between/unpaired) or
  - Wilcoxon signed rank (within/paired)
- Descriptive statistics: report median (ordinal data)
  - Likert scale responses, ratings

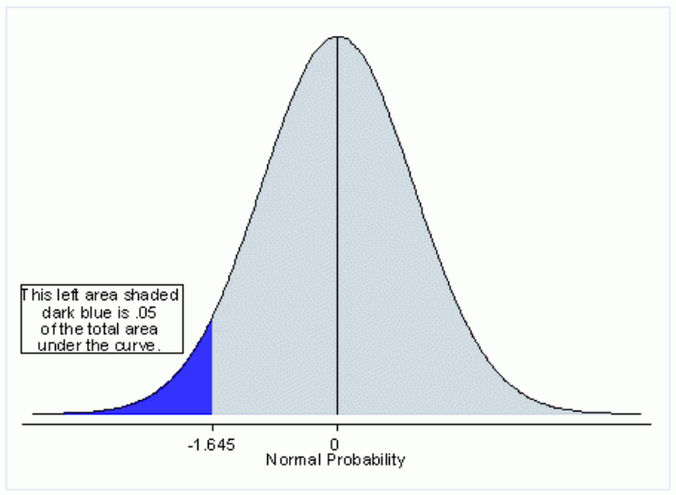
T-test (or Student's T-test)

- Test if two groups are different
  - Comparison of means
    - Interval or ratio data
- Assumptions
  - Populations are normally distributed
  - Variances are equal
- Robust
  - For small sample sizes if assumptions hold
  - For large (> 30) sample sizes even if assumptions violated

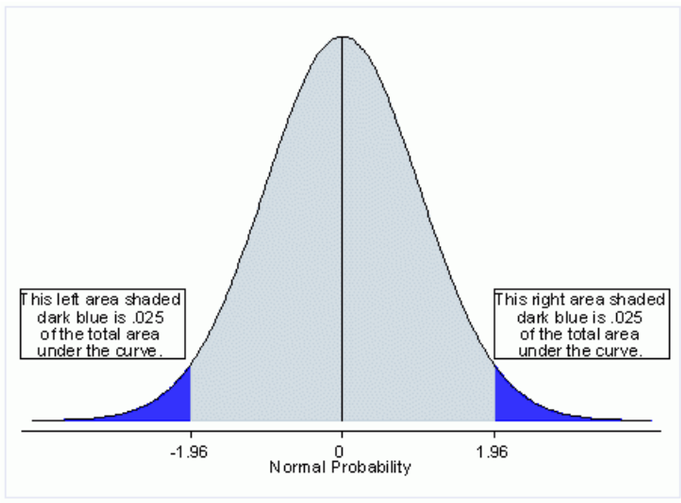


One or two tailed t-test

- Null hypothesis takes different forms
  - One-tailed - H0 is difference in one direction
  - Two-tailed - H0 is no difference



One-tailed



Two-tailed

One tailed t-test

- Are you only interested in difference in one direction?
  - Large difference in opposite direction is not significant
- Example
  - H1 = Our new app will be faster to use than the old app
  - H0 = Our new app will not be faster to use than the old app

H1 = Mean A < Mean B

H0 = Mean A ≥ Mean B

Two tailed t-test

If you're not sure, use 2 tailed t-test

- Example

- Two new maths teachers with different teaching methods (Mr Clark and Mrs Brown) join a school

- H1= One class will have higher test scores than the other

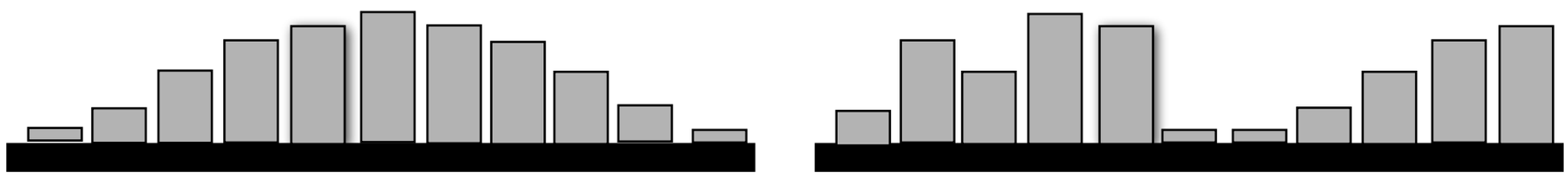
- H0=Both teachers have the same average test scores

H1 = Means are different

H0 = Means are same

Non-parametric tests

- T-test assumes that your data is normally distributed
  - What if it's not?
  - Also, compares means (interval or ratio data)
- Non-parametric tests don't assume data is from a certain type of distribution

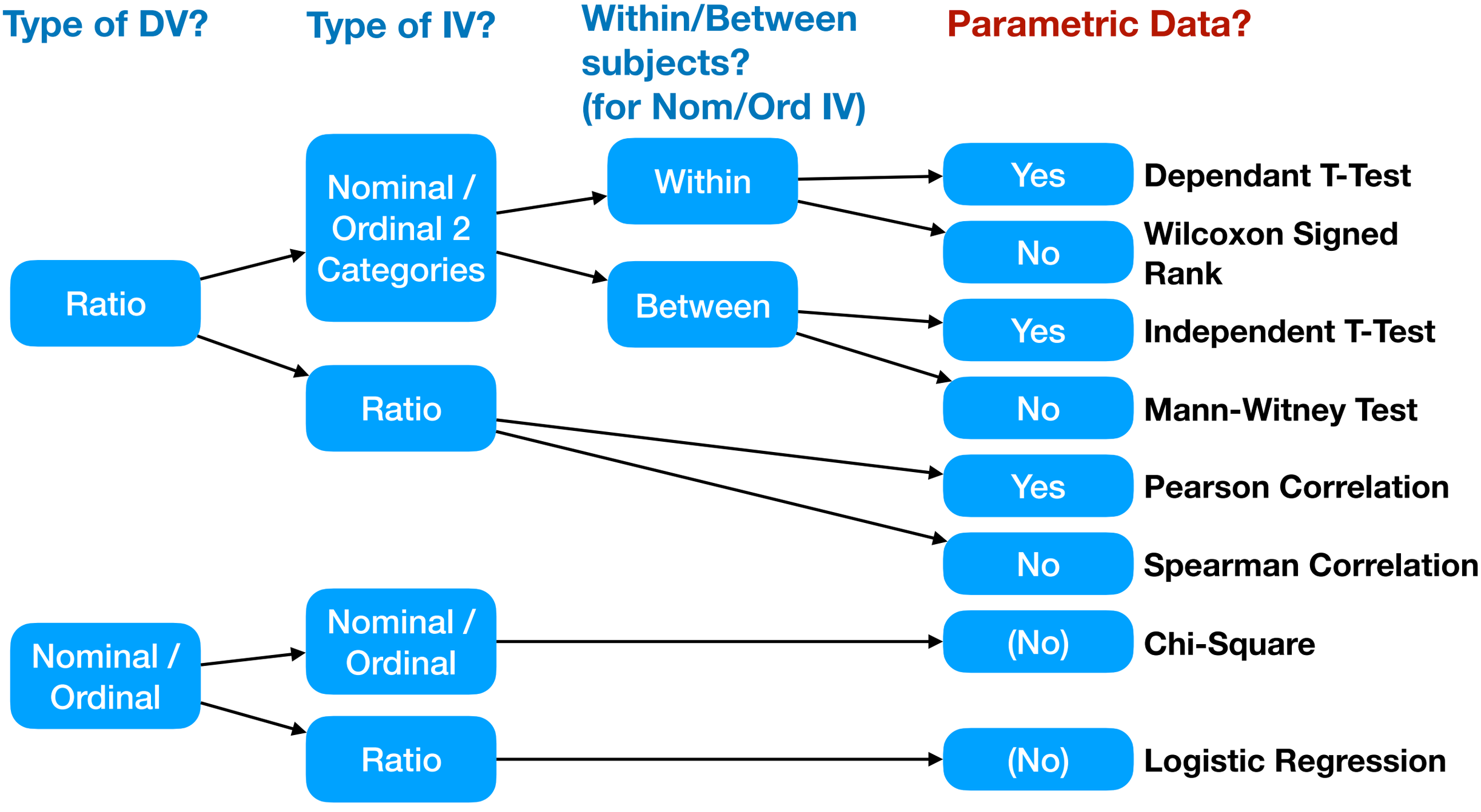


Non-parametric tests

- Hypothesis testing
- Mann-Whitney (unpaired) or Wilcoxon signed rank (paired)
- Report median (ordinal data)
    - Likert scale responses, ratings

Decision Tree - 1 DV and 1 IV

(cf. Field 2005)



Hypothesis testing don’ts

- Don’t use a one-tailed test just because the two-tailed test result wasn’t significant
- Don’t
    - Test lots of variables you didn’t have hypotheses about before collecting data
        - Report whatever you find that’s significant
- That’s bad science!

Attributions

- https://media.4rgos.it/i/Argos/6514064_R_Z001A?Web&$DefaultPDP570$
- https://www.joshuakennon.com/wp-content/uploads/2012/11/Net-Worth-and-Income-By-Education-Level.png
- https://jcebmo.org/wp-content/uploads/on-the-ballot.jpg
- https://www.pymnts.com/wp-content/uploads/2016/09/US-Income-On-The-Rise.jpg
- https://hsastrology.weebly.com/uploads/1/2/6/8/12687052/chinese-years_orig.png
- https://stats.idre.ucla.edu/other/mult-pkg/faq/general/faq-what-are-the-differences-between-one-tailed-and-two-tailed-tests/