



Prifysgol Cymru
Y Drindod Dewi Sant
University of Wales
Trinity Saint David

www.uwtsd.ac.uk

Research Methods and Professional Development ACFA7004

Statistics for Research Methods

Dr Seena Joseph

Quantitative analysis

- **Quantitative Data:** quantitative information – i.e. can be expressed in numbers
- **Quantitative analysis** - The broad process of collecting, measuring, and interpreting numerical data to answer research questions or solve problems.
 - It's the big picture—like figuring out "How accurate is the developed ML model?"
- **Statistical analysis** – the process of collecting, organising, interpreting, and drawing conclusions from data using statistical methods and techniques.
- A specific part of quantitative analysis where you apply statistical methods to draw conclusions or make predictions from data.
 - It involves analysing numerical data to uncover patterns, relationships, and trends, as well as making inferences and predictions based on the data.

Quantitative analysis

- **Variables**

- **Measurable quantities**

- **Example:**

- Number of students, Test score, Time to complete a task, study time, temperature, ...
 - Accuracy, training time, precision, recall, customer churn rate, cluster count, algorithm execution time, bug severity, programming language, number of attacks, threat level, attack type, response time, ...
 - Variables can be **discrete** (countable/whole numbers) or **continuous** (measurable with decimals)

Variables

Discrete variables

- Countable, distinct values—often whole numbers or categories
 - Can only take on particular values,
 - e.g. predicted class labels = (positive , negative);
 - Number of students in a class (45, 60, ...)
 - Satisfaction rate(1, 2, 3, 4, 5)
 - Number of cyber attacks
 - Attack types (“Phishing”, “DoS”, “DDoS”)

Continuous variables

- Infinite values within a range—measured, not counted.
 - Can take of any value,
 - e.g. Height (measured in inches or centimeters),
 - Weight (measured in pounds or kilograms)
 - Prediction accuracy (85.58), training rate (0.34),
 - Execution time (45.5 ms), churn probability (.67), transaction value (135.50), data transfer rate (15.6 MB/s)

Scales of measurement

- Scales of measurement (also called **levels of measurement**) describe how **numerical and categorical data** are defined and categorised.
- It defines the nature of numerical data utilized in quantitative analysis and determines the permissible statistical operations that can be applied to it.
- Different types of variables have different types of rules that determine what values can be assigned to them.
- Variables are measured on one of four different scales.
 - **Nominal**
 - **Ordinal**
 - **Interval**
 - **Ratio**

Scales of measurement

- **Nominal** - Categorical data and numbers that are simply used as identifiers or names or labels
 - Just categories; these are categories with identity only, no order
 - E.g. programming language (Python, C++);
 - Gender (Male, Female)
 - Type of attack(phishing, DoS, DDoS)
 - User access role (admin, guest, user)
 - ML model (SVM, LR, DT)
 - Authentication method (Password, Biometric, OTP)
 - These categories have no inherent order

Scales of measurement

- **Nominal** - Survey questions example

Examples of nominal variables	
Variable	Categories
Zip code	• 2138 • 90210 • 1007
Political preferences	• Republican • Democrat • Independent
Employment status	• Employed • Unemployed
Literary genre	• Comedy • Drama • Satire • Epic • Tragedy

Scales of measurement

Ordinal - represents an ordered series of relationships or rank order.

- Fundamentally this does not represent a measurable quantity.
 - Things are ranked according to magnitude, e.g. Bad→Good, Rich→Poor,
 - Satisfaction rate (very dissatisfied → very satisfied),
 - Error severity level (Low → critical) ; Likert scale response (disagree → agree);
 - Language ability (beginner, intermediate, Fluent)
 - Rating a cyber security threat from 1 (low) to 6 (critical) shows order—6 is worse than 3, it is not exactly ‘twice’ than 3. It’s order, not a ruler.

Scales of measurement

Ordinal – Survey questions example

Examples of ordinal scale survey questions

Question	Options
What is your age?	• 0-18 • 19-34 • 35-49 • 50+
What is your education level?	• Primary school • High school • Bachelor's degree • Master's degree • PhD
In the past three months, how many times did you buy groceries online?	• None • 1-4 times • 5-9 times • 10-14 times • 15 or more times

Scales of measurement

Interval

- A scale which represents quantity and has equal units but for which zero represents simply an additional point of measurement.
- Things with no meaningful/true zero point ; difference between values are consistent
 - (Temperature in Celsius—like 10°C , 20°C , 30°C . The difference between 10°C and 20°C is the same as 20°C to 30°C (10 degrees), but 0°C isn't "no temperature"—it's just freezing, and you can go below it (e.g., -5°C).

Scales of measurement

Interval

Survey questions example

What is your preferred temperature range when working in a lab or office?

- ☐ Below 18°C
- ☐ 18°C – 21°C
- ☐ 22°C – 25°C
- ☐ 26°C – 29°C
- ☐ Above 29°C

On a scale of 0 to 100, how satisfied are you with the accuracy of your ML models?

- ☐ 0 – 20 (Not satisfied at all)
- ☐ 21 – 40 (Slightly satisfied)
- ☐ 41 – 60 (Neutral)
- ☐ 61 – 80 (Satisfied)
- ☐ 81 – 100 (Highly satisfied)

Scales of measurement

Ratio

- The ratio scale of measurement is similar to the interval scale in that it also represents quantity and has equality of units.
 - Things have a meaningful/true zero point and are measured from there; It has an absolute zero (no numbers exist below the zero).
 - e.g. height, weight, storage capacity, network bandwidth, Accuracy.

Scales of measurement

Ratio

Survey questions example

How many ML models have you built in the past year?

- ☐ 0
- ☐ 1 – 5
- ☐ 6 – 10
- ☐ 11 – 20
- ☐ More than 20

How many hours per week do you spend on ML-related tasks?

- ☐ 0 – 5 hours
- ☐ 6 – 10 hours
- ☐ 11 – 20 hours
- ☐ 21 – 30 hours
- ☐ More than 30 hours



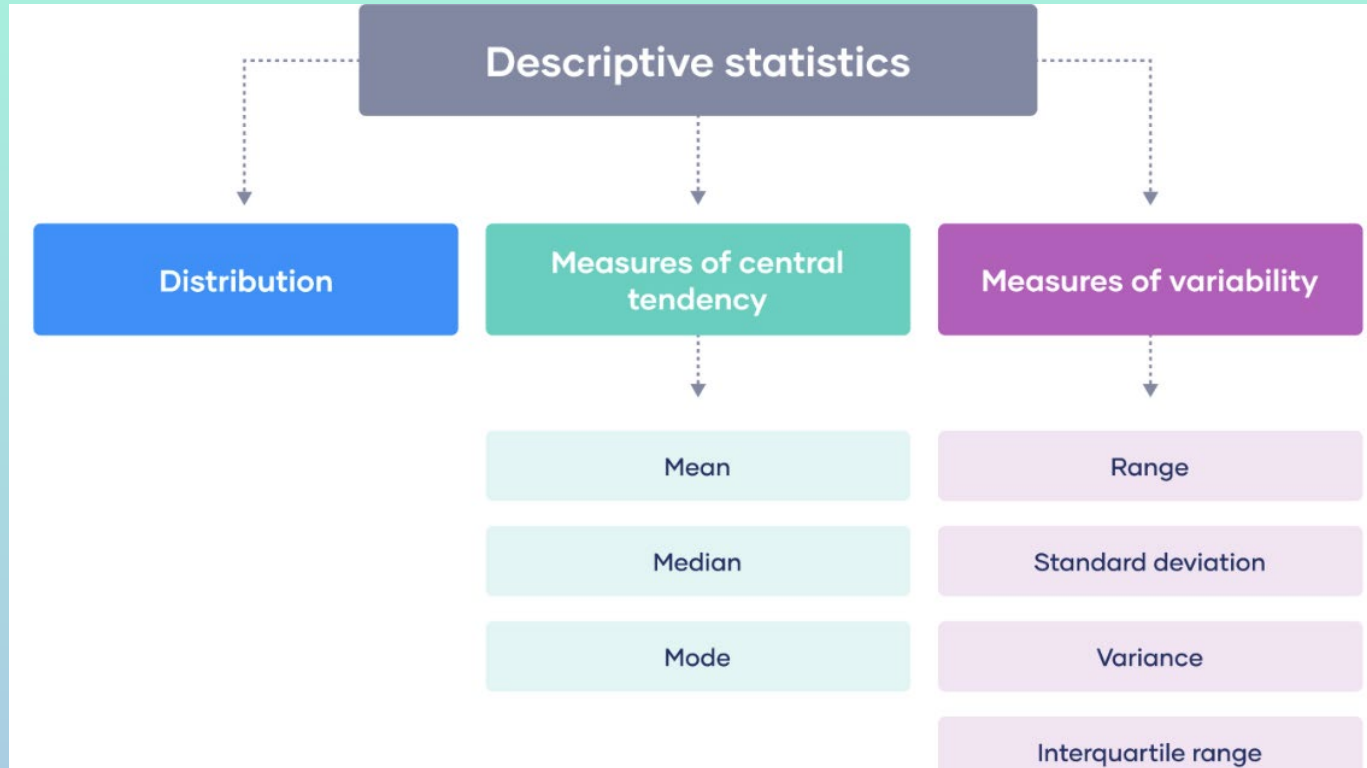
Scales of measurement - Summary

Scale	Definition	Order	Equal Gaps	True Zero	What You Can Do	Everyday Example	ML Example	Data Sci Example	Soft Eng Example	Cyber Example
Nominal	Labels, no order—just names	No	No	No	Count how many (mode)	Eye color (blue)	Labels (cat, dog)	Customer (new)	Language (Python)	Attack (phishing)
Ordinal	Order, but gaps aren't equal	Yes	No	No	Rank, median	Movie stars (1, 2, 3)	Perf (poor, good)	Rating (1-5)	Priority (low, high)	Threat (low, high)
Interval	Equal gaps, zero isn't "nothing"	Yes	Yes	No	Add, subtract, average	Temp (20°C)	Time diff (10 ms)	IQ score (100)	Perf score (50)	Time since (10 min)
Ratio	Equal gaps, zero means "none"	Yes	Yes	Yes	All math (add, multiply)	Height (50 cm)	Accuracy (90%)	Time on site (60 s)	Memory (50 MB)	Data stolen (1 GB)

Statistics

- Two primary types of statistics
- **Descriptive Statistics**
 - Used to describe a set of quantitative data.
 - Summarisation of data
- **Inferential Statistics**
 - Used to make predictions about large groups (i.e. populations) based upon the analysis of data gathered from a smaller group (i.e. samples)

Statistics



Frequency Distributions

- When data is first gathered it is generally in the form of an unordered set of scores
- The first thing to do with it is to organise and display it using a frequency distribution
- Suppose you have an EQ test in which the students gained the following scores:

Emotional Intelligence Test Scores												
46	50	48	47	48	47	49	43	47	46	50	48	49
46	46	45	46	46	47	46	46	46	48	47	46	47
44	49	47	48	49	48	48	49	45	48	46	46	51
48	44	45	44	46	49	50	48	43	48	46	48	46
48	46	47	47	47	47	49	49	46	47	47	44	45
45	48	50	48	47	47	49	47	45	48	49	45	47
47	44	48	47	47	51	47	46	47	46	45	47	45
45	47	48	48	46	48	45	50	47				

- What can we say about the data? Not a lot...

Frequency distributions

- The first thing to do is to create a *frequency distribution*
 1. Locate the lowest and highest scores – 43 and 51
 2. Record all possible scores, even if they're not present in the dataset – 43, 44, 45, 46, 47, 48, 49, 50, 51
 3. Go through the list of scores and tally each score one by one. At the end of this process our data looks like this:

Score	Tally
51	
50	
49	
48	
47	
46	
45	
44	
43	

Frequency distributions

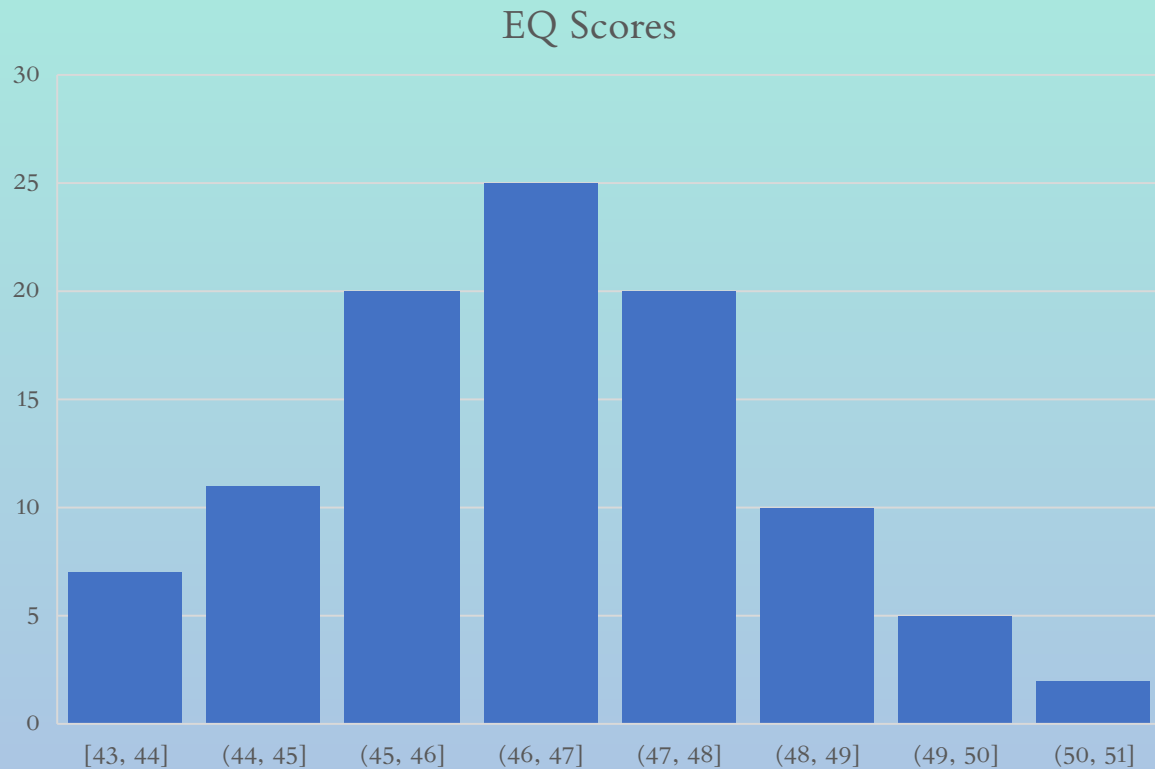
4. Now count the number of tallies to find the frequency

Score	Tally	<i>f</i>
51		2
50		5
49		10
48		20
47		25
46		20
45		11
44		5
43		2

- This diagram provides a lot more information than the raw data
 - The data is grouped around 47
 - Most data is in the 45-49 range
 - The most common score was 47, with 25 students

Frequency Distributions

- These can also be created using Excel with the histogram tool from the Analysis ToolPak



Exercise

- Using the data for a midterm exam (RM 7 Midterms (Excel)), create the frequency distribution for the dataset:

Midterm Examination Scores									
44	35	20	40	38	52	29	36	38	38
38	38	41	35	42	50	31	43	30	37
32	47	43	41	47	32	38	29	41	26
41	51	48	49	37	26	34	48	23	29
38	47	41	33	39	48	38	20	48	34
29	44	29	33	35	50	41	38	35	35
32	26	24	38	38	56	56	48	41	29
26	26	38	37	44	24	44	47	32	41

Grouped frequency distributions

- If scores vary more than in the previous example it can be difficult to see patterns in the frequency distribution
- It may be easier to see them if there are fewer scores, so grouping the scores becomes an option
- Suppose 80 students take a midterm exam

Midterm Examination Scores									
44	35	20	40	38	52	29	36	38	38
38	38	41	35	42	50	31	43	30	37
32	47	43	41	47	32	38	29	41	26
41	51	48	49	37	26	34	48	23	29
38	47	41	33	39	48	38	20	48	34
29	44	29	33	35	50	41	38	35	35
32	26	24	38	38	56	56	48	41	29
26	26	38	37	44	24	44	47	32	41

Grouped frequency distributions

To create a grouped frequency distribution

1. Find the lowest and highest scores : 20 and 56
2. Subtract lowest from highest to find the range : $56 - 20 = 36$
3. Determine the width of each class interval so that there are ten to twenty intervals. Use an integer width whenever possible – in this case we'll use a class interval of 3
4. Determine the starting point of the lowest interval. In this case it could be 18-20, 19-21, or 20-22
5. Record the score limits for all possible class intervals
6. Go through the list of scores and tally each score one by one.
7. Convert the number of checks for each interval into a frequency

Grouped frequency distributions

- Using an interval width of 3 and a starting point of 20 gives a grouped frequency distribution that looks like this:

Score Limits	Tally	<i>f</i>
56-58		2
53-55		0
50-52		4
47-49		10
44-46		4
41-43		11
38-40		14
35-37		9
32-34		8
29-31		8
26-28		5
23-25		3
20-22		2

- Once the scores are placed in a group you lose information about the individual scores or use the granularity.
- Decisions about interval size and start point will alter the distribution

Exercise

- Using the midterm exam data, create the following grouped frequency distributions:
 - One with an interval width of 4 and a start point of 20
 - One with an interval width of 4 and a start point of 18
- What differences can you see between these distributions?

Descriptive Statistics

- Frequency distributions are good for getting a graphical view of data, sometimes an actual number is what is required
 - Average MP's salary
 - Average number of minutes spent on FB each day
- Single numbers can be misleading, as we also need to know how the scores vary within a dataset
- To do this we need to look at
 - Measures of central tendency
 - Selecting a measure of central tendency
 - Measures of variability

Measures of Central Tendency

- These provide an average or typical score that describes the level of a set of scores
 - Useful when comparing individual scores with the group or an external standard
- The most commonly-used measure of centre is the *mean*
 - Add up all of the scores and divide by the number of scores
$$\bar{X} = \frac{\sum X}{N}$$
where X is the set of scores and N is the number of scores
- $\bar{X}$ is used when talking about the mean of a sample
- μ is used when talking about mean of the population from which a set is drawn (this difference is important later on)

Measures of Central Tendency

- The *median* is the score that divides the frequency distribution in half, i.e. the middle score
 - The easiest way to find this is to order the scores from lowest to highest and to find the middle score
- Given an odd set of scores: 113, 95, 94, 133, 112
 - Rearrange: 94, 95, 112, 113, 133
 - Take the middle value: 112
- Given an even set of scores: 113, 95, 94, 133, 112, 102
 - Rearrange: 94, 95, 102, 112, 113, 133
 - Take the mean of the middle two values:
 $(102 + 112)/2 = 107$

Measures of Central Tendency

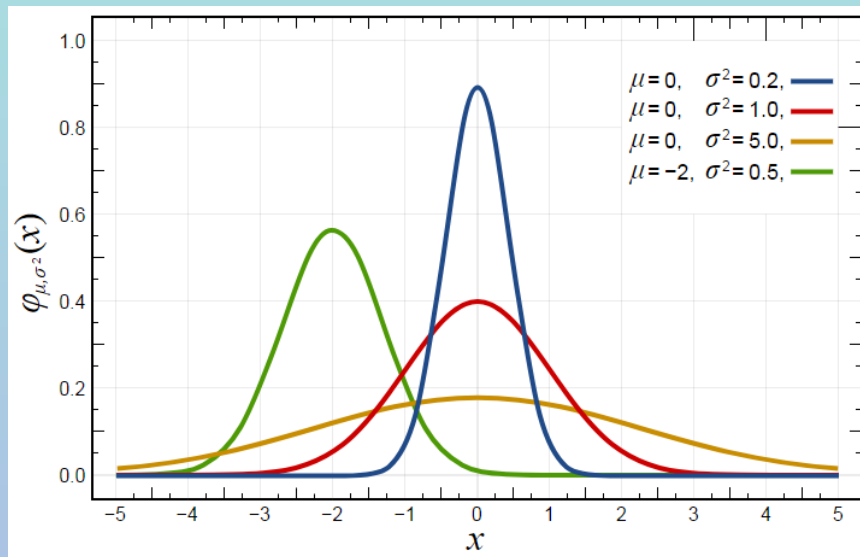
- The *mode* is the most frequently occurring score in a set of scores
 - This is the largest column on the frequency distribution
- Given a set of scores 110, 105, 100, 100, 100, 100, 99, 98
 - 100 appears 4 times, more than any other, so is the mode
- Given a set of scores 110, 105, 105, 103, 100, 100, 99, 98
 - 105 and 100 both appear twice so we have two modes
 - Bimodal and multimodal distributions are possible

Selecting a measure of central tendency

- In a symmetrical unimodal distribution the mean, median and mode will have the same values.
- The mean is the ‘balance point’ of a distribution
 - It reflects the position of every score in a dataset and is the most affected by any skewness in the dataset
- If a dataset is skewed then it is better to report the median than the mean
- A skewed dataset is one that is unbalanced compared to a ‘Normal’ dataset

The Normal curve

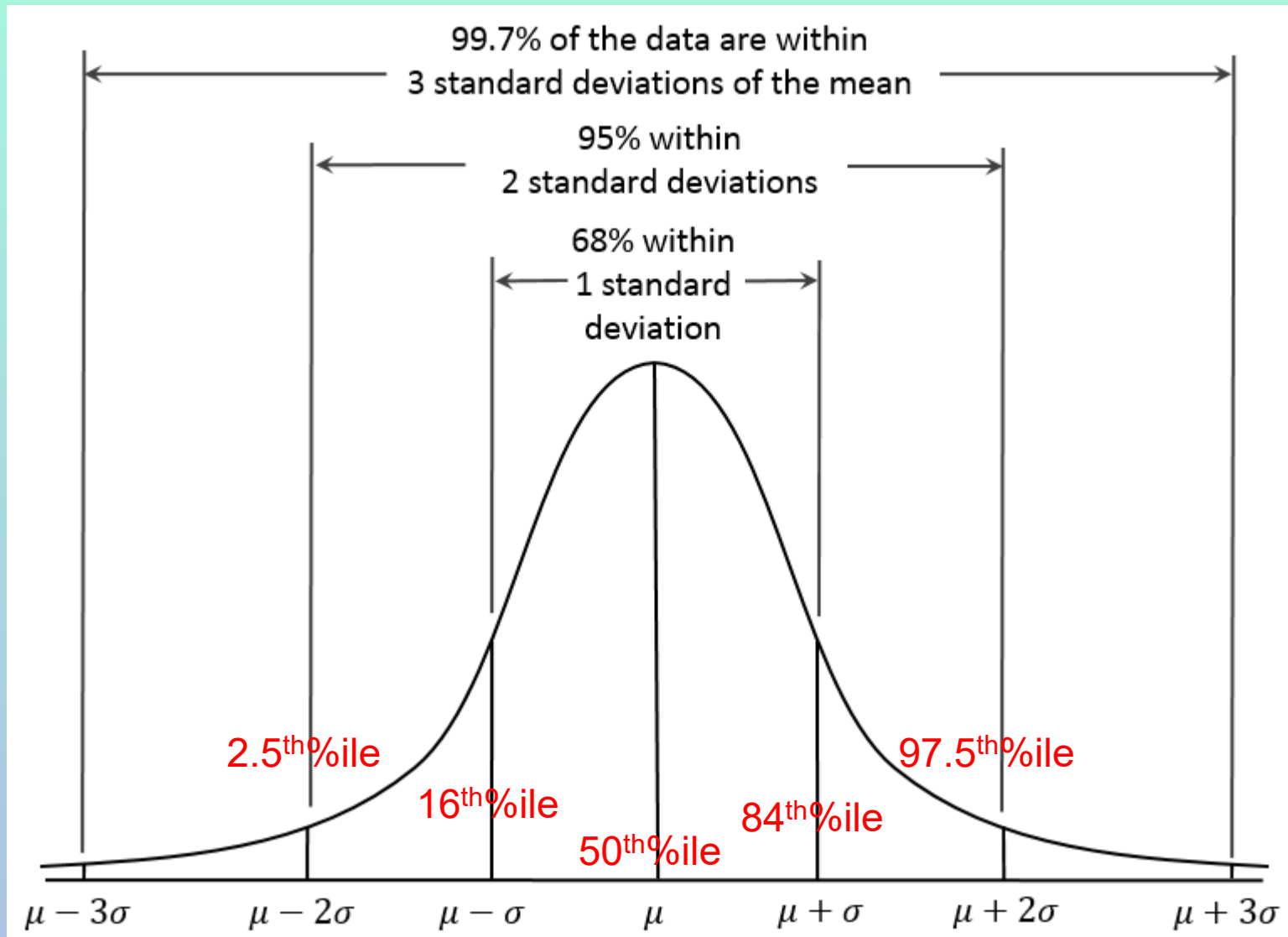
- The Normal curve is a mathematical idealisation of a particular type of symmetrical distribution.
- The **normal curve**, also known as the **normal distribution** or **Gaussian distribution**, is a bell-shaped curve that represents the distribution of data where most values cluster around the mean, with fewer values appearing as you move further away in either direction.
- Many actual data distributions approximate the normal curve so we can use it to help interpret a wide range of datasets
 - Data that approximates the Normal curve is known as *normal data*



The Normal curve

- All normal curves share the following features but differ in their means and standard deviations
 1. All normal curves are symmetrical about the mean
 2. All normal curves are unimodal
 3. Since the curves are symmetrical and unimodal, mean=median=mode
 4. All normal curves are asymptotic to the horizontal axis of the distribution
 - Scores are continuous and mathematically describe an infinite set
 5. All normal curves have the same proportions of scores under the curve relative to particular locations on the horizontal axis when the scores are expressed in a similar basis

The Normal curve



The Normal curve

Example

Let's use an example with a normal distribution of student test scores:

- Mean score: 70
- Standard deviation: 10

1. Within One Standard Deviation:

- $\text{Mean} \pm 1\sigma = 70 \pm 10 = 60 \text{ to } 80$
- About 68% of students scored between 60 and 80.

2. Within Two Standard Deviations:

- $\text{Mean} \pm 2\sigma = 70 \pm 2(10) = 50 \text{ to } 90$
- About 95% of students scored between 50 and 90.

3. Within Three Standard Deviations:

- $\text{Mean} \pm 3\sigma = 70 \pm 3(10) = 40 \text{ to } 100$
- About 99.7% of students scored between 40 and 100.

Skewed distributions

- The Normal distribution is symmetrical, but some frequency curves are not
 - If a test is very easy most people get high scores
 - If a test is very difficult most people get low scores
- Depending on which way the bulk of the scores have shifted we can say if a curve is positively skewed or negatively skewed

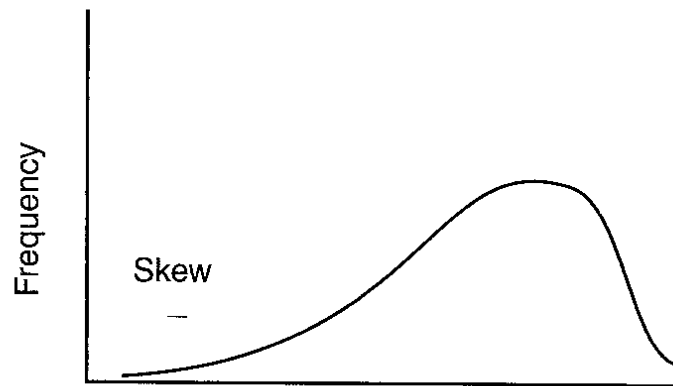


FIGURE 4-4 A negatively skewed distribution.

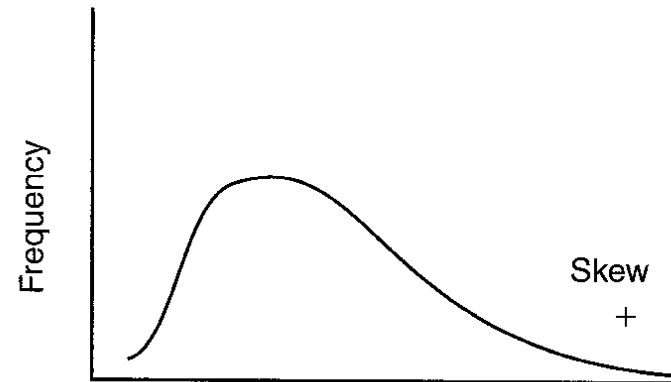


FIGURE 4-5 A positively skewed distribution.

Variation with Skewness

Dataset X: 1, 2, 2, 2, 3

Mean $\bar{X} = 2$

Median = 2

Mode = 2

Dataset Y: 1, 2, 2, 2, 3, 6

Mean $\bar{Y} = 2.667$

Median = 2

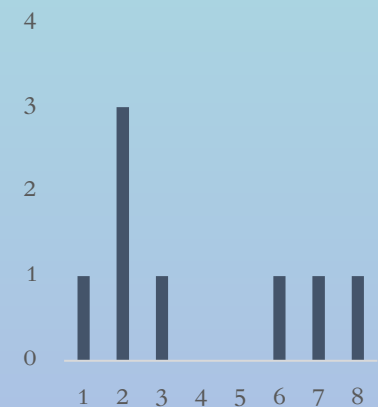
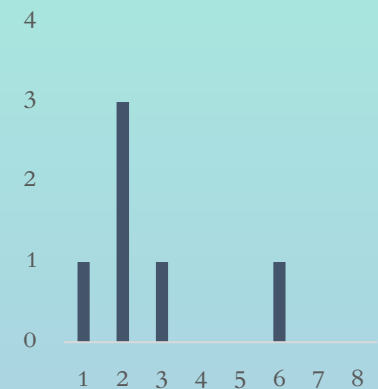
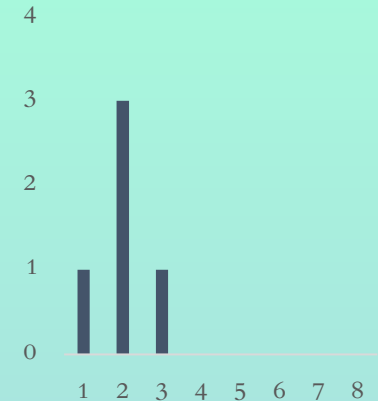
Mode = 2

Dataset Z: 1, 2, 2, 2, 3, 6, 7, 8

Mean $\bar{Z} = 3.875$

Median = 2.5

Mode = 2



Range

- The simplest measure of variability is the *range*
- The range is simply the highest score (H) minus the lowest score (L)
 - $\text{Range} = H - L$
- In classroom X
 - $\text{Range} = H - L = 160 - 40 = 120$
- In classroom Y
 - $\text{Range} = H - L = 102 - 96 = 4$
- This gives a very basic measure of overall variability as it doesn't take any skewness into account
- Better measures are the *variance* and the *standard deviation*

Variance

- The variance is the most commonly used measure of variability
- It is defined as the mean of the sum of the squares of the deviation scores

$$SD^2 = \frac{\sum(X - \bar{X})^2}{N - 1}$$

- $(X - \bar{X})$ is the *deviation score*
 - Positive scores are above the mean
 - Negative scores are below the mean
 - If we calculate the mean deviation these would cancel out, so we square them to change all negatives to positives (common statistical technique) then take the mean value of the squared scores

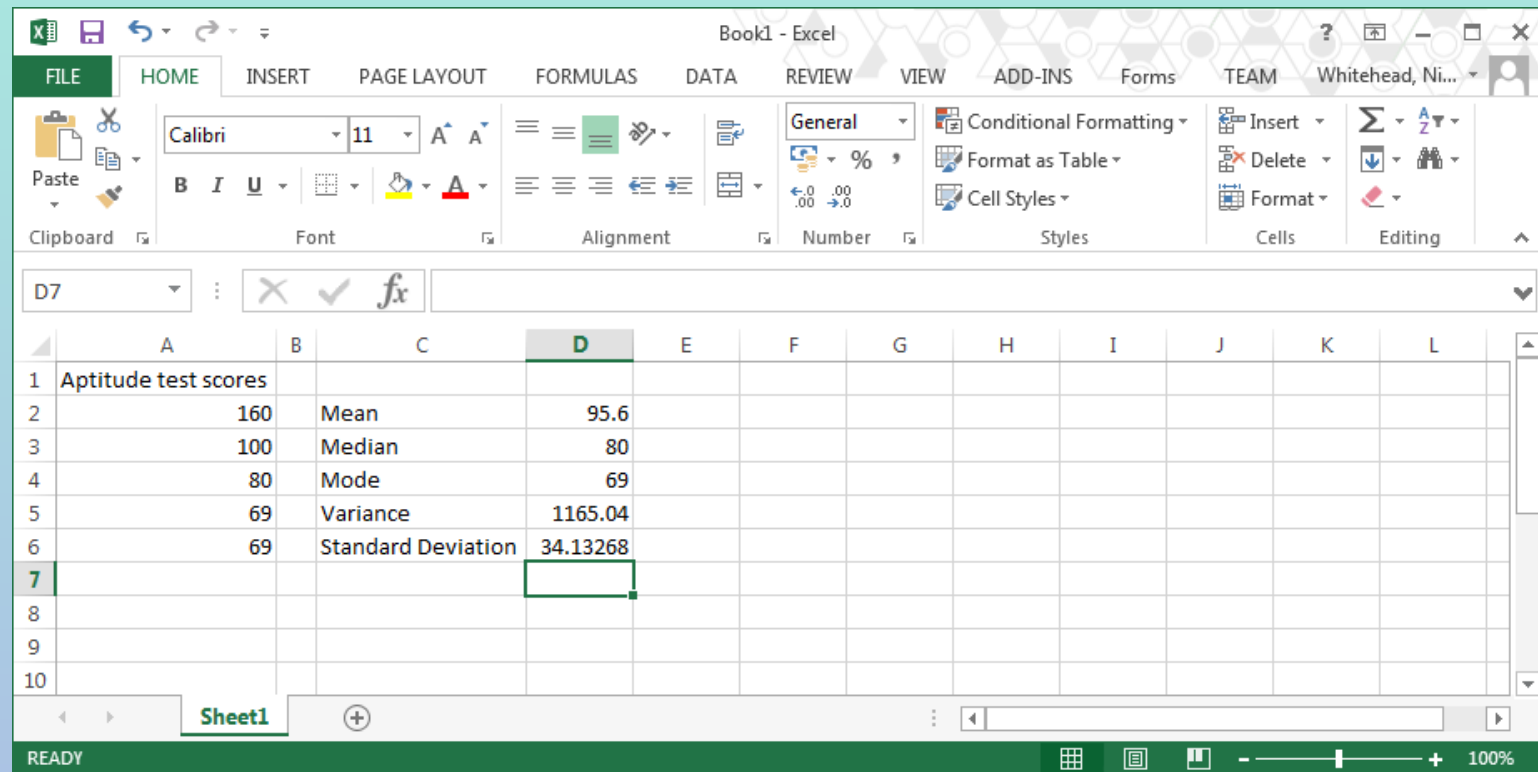
Standard Deviation (S.D.)

- Another common measure of variation is the *standard deviation*
- This is easy – take the square root of the variance

Scores (X)	
9	Mean $\bar{X} = 5$
7	Variance $SD^2 = 10$
5	Standard Deviation $SD = 3.16$
3	
1	

Using Excel

- Calculating the means, variances and standard deviations can be done by hand but is better done with a spreadsheet or a statistical package



The screenshot shows the Microsoft Excel interface with the 'Book1 - Excel' title bar. The ribbon is set to 'HOME' with tabs for FILE, HOME, INSERT, PAGE LAYOUT, FORMULAS, DATA, REVIEW, VIEW, ADD-INS, Forms, TEAM, and Whitehead, Ni... The ribbon includes groups for Clipboard, Font, Alignment, Number, Styles, Cells, and Editing. The formula bar shows 'D7' and the formula 'fx'. The spreadsheet has columns A through L and rows 1 through 10. The data is as follows:

	A	B	C	D	E	F	G	H	I	J	K	L
1	Aptitude test scores											
2		160	Mean	95.6								
3		100	Median	80								
4		80	Mode	69								
5		69	Variance	1165.04								
6		69	Standard Deviation	34.13268								
7												
8												
9												
10												

The status bar at the bottom shows 'READY', 'Sheet1', and a zoom level of 100%.

Using Excel

- The five main Excel functions are:
 - Mean: AVERAGE(range)
 - Median: MEDIAN(range)
 - Mode: MODE(range)
 - If there is no mode then it will return the value #N/A
 - Variance: VAR.S(range)
 - This will ignore any non-numerical cells in the range
 - Standard Deviation: STDEV.S(range)
 - This will ignore any non-numerical cells in the range

	A	B	C	D	E
1	Aptitude test scores				
2	160		Mean	=AVERAGE(A2:A6)	
3	100		Median	=MEDIAN(A2:A6)	
4	80		Mode	=MODE(A2:A6)	
5	69		Variance	=VAR.S(A2:A6)	
6	69		Standard Deviation	=STDEV.S(A2:A6)	

Exercise

- Download the file `Midterms.xlsx` and calculate the mean, median, mode, variance and standard deviation of the data therein

Summary

- All of today's material looked at descriptive statistics of univariate data
 - Frequency distributions
 - Measures of central tendency
 - Selecting a measure of central tendency
 - Measures of variability
 - Using Excel for statistics
- Next session we'll look correlation and how this can be used to make predictions using linear regression.

Reference

- *Burr, V., Statistics for the Terrified*, 3rd ed. New York, NY, Pearson, 2011.
- [The Beginner's Guide to Statistical Analysis | 5 Steps & Examples](#)