STATISTICS

Data types — Numerical — Continuous — can have dec points
Numerical — Discrete — fixed values (no fractions)
Categorical — Nominal — no order (hair colour)
Categorical — Ordinal — 1st, 2nd, 3rd Good, VG, G, VB..

Secondary Data collected by someone else you use it
Primary Data — you collected you use it

Univariate data → one item of information collected
Bivariate - 2 items of information - height + weight

Collecting data → survey, questionaire
survey methods — observation, interview, postal survey
Respondant - person who answers questionaire

you need to make sure survey is fair → avoids bias
Both in way it asks questions and way it allows answers to be given

Designed Experiments:
Explanatory variable = Controlled variable
or independent variable
Response variable = dependent variable
Ex 1 p56

Control group: given placebo → identical to group of patients who are getting trial drug except they don't get drug.

Bias in sampling → due to
(i) How you choose sample
(ii) Sampling wrong population
(iii) People not responding
(v) Dishonest answers

SAMPLING METHODS

1.) Simple Random Sampling (e.g. put no's in hat, use calculator) → every possible sample has equal chance of being selected.

2.) Stratified Sampling: If population contains groups that are quite different to each other then the population is split into those groups and the no. taken from each group to be part of the sample is proportional to size of the group. Ex 1 pg 60

3.) Systematic sampling - sample obtained by choosing items at regular intervals from an unordered list eg 4th, 14th, 24th....

4. Quota Sampling: Population is divided into groups in terms of age, class.... Then interviewer is told how many people to interview in each group but interviewer makes choice of who to ask (can lead to bias)

5.) Cluster sampling: Population split into groups or clusters (random) Then clusters are chosen randomly and every item in cluster is looked at. Its best if a large no. of small clusters is formed as this minimises the chances of sample being unrepresentative

6.) Convienience sampling → 1st 40 names on list, stand outside shop & select 1st 40 people
- can lead to high levels of bias → unrepresentative.

Census → when entire population is surveyed, SAMPLE: only sample is

GIVE TIME to answering Stats Q's - think re-read Q... use common sense!

MEASURES OF LOCATION (averages) | MEASURES OF SPREAD/VARIATION

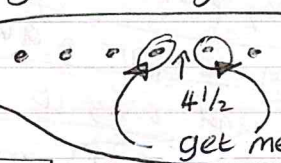
mean, mode, median | Range, Interquartile Range, Standard Deviation

Mode, modal result = most common, most frequent — Q8 pg 68, Q9

Mean of number set = $\frac{\text{sum of numbers}}{\text{how many numbers there are}}$ — Q16 pg 69, Q17?

Median = if data was put in order from lowest to highest, the median is the middle piece of data if n = no of pieces of data then put the in order from lowest to highest and median is $\left(\frac{n+1}{2}\right)^{th}$ piece of data.

If $\left(\frac{n+1}{2}\right)$ is not a whole no, then you must get the average of the two nearest whole numbers to $\frac{n+1}{2}$.



$n = 8$ $\frac{n+1}{2} = 4\frac{1}{2}$

FREQUENCY DISTRIBUTION:

Result/Data (x)							
Frequency (f)							

$\text{mean} = \frac{\Sigma f \cdot x}{\Sigma f}$

mode = data with highest frequency

GROUPED FREQ DISTRIBUTION: Where exact data not given - just intervals.

Result/Data	1-5	6-10	11-15	16-20
Freq	11	12	15	9

How do you find mean? Mid Interval values give approx value for data (1-5) => mid interval = $\frac{1+5}{2} = 3$.

So, use midinterval values to work out mean.

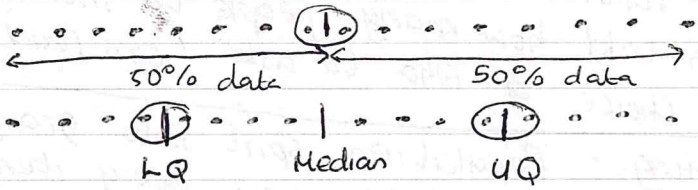
But REM → not 100% accurate.

Mode: Easy to find, not influenced by extreme data values	may not exist Can't use for further analysis
Median: Not affected by extremes, if data is ordered median is easy to calculate	not always a given data value not useful for further analysis
MEAN: - Uses all the data to calculate it - Easy to calculate - very useful for further analysis	- affected/distored by extreme values - not always a given data value '2.3 children'

Range: Largest data value − Smallest data value

$n = 16$ $\frac{n+1}{2} = 8.5$

Interquartile range:



Find median,
Then find median of lower 50% data, median of upper 50% data

$n = 13$ $\frac{n+1}{2} = 7$

DON'T forget to put the data in ORDER!!!

E1 pg 71

STANDARD DEVIATION

Find mean $\bar{x}$, x = data value

$$\sigma = \sqrt{\frac{\Sigma (x-\bar{x})^2}{n}}$$

know how to do this on calc and written!

STD σ of frequency Distribution:

Data (x)										
Freq (f)										

$$\sigma = \sqrt{\frac{\Sigma f (x-\bar{x})^2}{\Sigma f}}$$

E3 p 73

p 78 Q14, 21

know how to do on calc and written !!

Percentiles: Give a measure of your position relative to others in a data set. If you are on 70th percentile ⇒ 70% of data was below yours, 30% was higher.

To find the k^{th} percentile: Find value of no that is k% into the data: Order no's from smallest to highest (n = no. of no.'s)
Then find k% of total no. of data points in set ($\frac{k}{100} \times n$)
If answer is whole no then the k^{th} %ile is the mean of the data at that no and the next one (If $\frac{k}{100} \times n = 5$ then you must average 5th and 6th piece of data)
If answer is not a whole no. then the no is rounded up (6¼ → 7)

Ex 5,6 p75

Stem & leaf plots – don't forget the key!!!

Back to Back Stem & leaf — Ex 2 pg 82

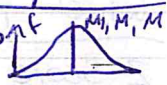
Always do a rough sort unordered!

p 86 Q10

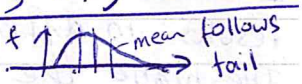
p 89 Q5

Histograms → No GAPS between bars, show continuous data, data always grouped

SYMMETRICAL Distribution: Has axis of Symmetry down middle, also called Normal Distribution (e.g. heights of population, IQ)

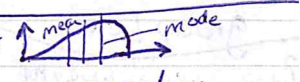


Positively skewed ⇒ tail to right, most data to left



(e.g. the age at which people first learn to ride a bike)

Negatively skewed ⇒ tail to left, most of data to right



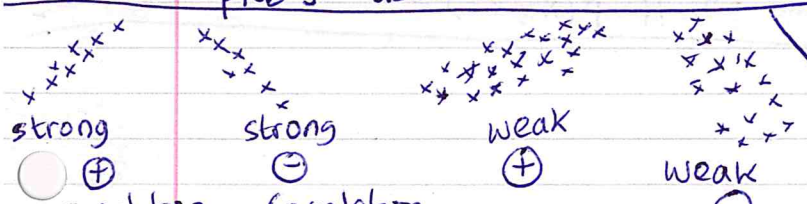
(e.g. the age at which people have to get 1st pair reading glasses)

Uniform distribution ⇒ data evenly spread throughout no mode

BIMODAL distribution ⇒ 2 modes MULTIMODAL ⇒ >3 modes

Distributions and STD: Rem y axis = freq, x axis = data!!!

Scatter plots used with bivariate data



Correlation – measure of the strength of a relationship between 2 variables

e.g. Price of car correlates with age so this makes sense as the age of a car generally causes price to decrease.

e.g. no of pads sold correlates with no of fridges sold – NO CAUSAL relationship There is some other variable at play (lurking variable)

Correlation does NOT imply Causality!!
* direction of causality – which causes which!!

Line of best Fit: (linear regression line) Draw by eye, find m, find eqn

Note if $\bar{x}$ = mean of x's, $\bar{y}$ = mean of y's then $(\bar{x}, \bar{y})$ is always on line of best fit

Use a clear ruler!

Drawing: (ruler on edge)

(i) By eye the distances (dotted) above should = distances below line

(ii) Should be roughly same no. of points above & below line (ignore outliers)

(iii) Line of best fit does not have to contain any of the points OR (0,0)

Slope: Take 2 points that are reasonably far apart use $\frac{y_2 - y_1}{x_2 - x_1}$

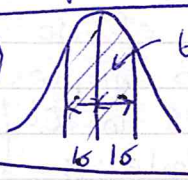
$y = mx + c$

$y - y_1 = \checkmark$

We only deal with linear correlation on our course @11 p157

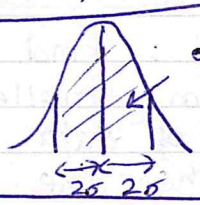
100% correlation ⇒ all points on line of best fit ⇒ correlation = ± 1

EMPIRICAL RULE:



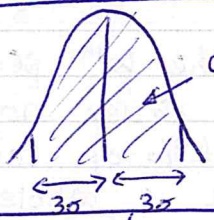
68% of data

σ σ



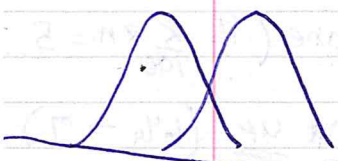
95% of data

2σ 2σ

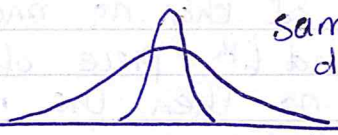


99.7% of data

3σ 3σ



Same σ different $\bar{x}$



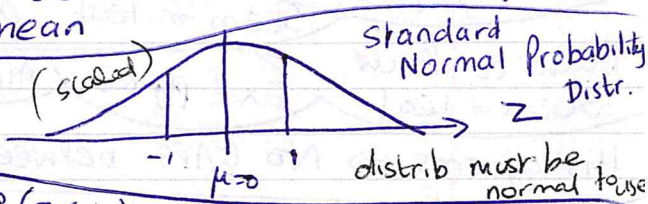
Same $\bar{x}$ different σ's

pg 163 Q 7, 8, 16, 18 p165

z score / standard score = no of std. deviations a value is from the mean

Margin of error = $\frac{1}{\sqrt{n}}$ n = sample size

Z = random variable



distrib must be normal to use

Find value of $k \in \mathbb{R}$ for which $P(-k \leq z \leq k) = 0.95$

$P(-k \leq z \leq k) = P(z \leq k) - P(z \leq -k)$

$= P(z \leq k) - [1 - P(z \geq k)]$

$= P(z \leq k) - [1 - P(z \leq k)] = 2P(z \leq k) - 1$

so $0.95 = 2 \times P(z \leq k) - 1 \Rightarrow P(z \leq k) = 0.975$

$= k = 1.96$

Z score used to compare results in different tests. The higher the z score the better the result relative to the group / class

If data is normally distributed then z score can be used to determine the %ile of a result

In a university exam, marks are Normally distributed mean = 65% σ = 8%. The university decides to give 1st class hons to top 10% of its students. What is least % required to achieve a 1st class hons?

In 93rd % means 93% of results are below yours

want 90% in here → what z score corresponds to 90% → look up tables $z = 1.28$ corresponds to 0.8997

$1.28 = \frac{x - \bar{x}}{\sigma}$

$1.28 = \frac{x - 65}{8}$

$x = 75.24$

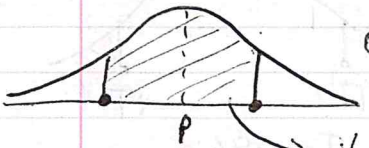
so a student needs 75.24% to achieve a 1st class hons

Q: X is a normal variable with mean = 9. If $P(X \geq 12) = 0.0062$, calculate σ.

Ans: 1.2

INFERENTIAL STAT's

If we took all samples of size n and looked at the proportion within each sample then we would have the Sampling distribution of proportion. The mean $= p$ & it is normally dist



$\sigma_{\hat{p}}$ = Standard Error of proportion $= \sqrt{\frac{p(1-p)}{n}}$

→ if we want to be 95% confident that p lies within an interval then we want $P(z \leq Z \leq z) = 0.95 \Rightarrow P(Z \leq z) = 0.95 + 0.025 = 0.975$

From tables this means we need an interval 1.96σ's either side of $\hat{P}$ so $\hat{p} - 1.96\sqrt{\frac{p(1-p)}{n}} < p < \hat{p} + 1.96\sqrt{\frac{p(1-p)}{n}}$ ← margin of error $= 1.96\sigma_{\hat{p}}$ for 95% confidence

A less accurate Margin of Error is $\frac{1}{\sqrt{n}}$ which is an approximation based on Empirical rule and uses $p = \frac{1}{2}$.

CENTRAL LIMIT THEOREM:

i) Sampling distribution of the mean → always normal if $n \geq 30$

ii) $\mu_{\bar{x}} = \mu$ mean of samples distribution = mean of population

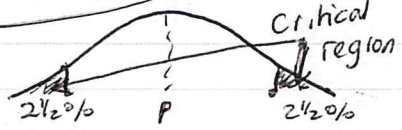
(iii) $\sigma_{\bar{x}} = \sigma/\sqrt{n}$ (σ = std of actual population BUT use sample σ if Population σ is not given.)
Standard Error

Q: If a sample of 100 bags of flour are chosen at random, what is the probability that their mean weight is less than 1.99 kg given that bags of flour are known to have a mean weight of 2 kg, $\sigma = 50$g
Ans: 0.0228

Confidence interval for population mean: $\bar{x} \pm 1.96\,\sigma_{\bar{x}}$ where $\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}}$

$1.96\,\sigma_{\bar{x}}$ = Margin of Error / use σ sample if σ population not known

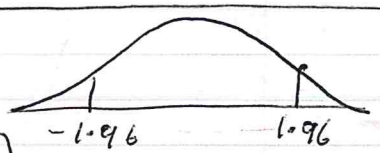
Hypothesis testing for Population Proportion:



(i) State H_0 null Hypothesis {things stay same}
(ii) State H_1 alternative Hypothesis {things not same}
(iii) Obtain test statistic (sample statistic) → this is $1/\sqrt{n}$ (margin of error for 95% of sample proportion)
If $\hat{p}$ is outside $p \pm \frac{1}{\sqrt{n}}$ then it is in the 5% level of significance (critical region and we will reject H_0

If $\hat{p}$ is within $p \pm \frac{1}{\sqrt{n}}$ we fail to reject H_0

Hypothesis testing for Population Mean



(i) State H_0 (things stay as they were)
(ii) State H_1 (things have changed - not same)
(iii) Test statistic: $\frac{\bar{x} - \mu}{\sigma_{\bar{x}}} = \frac{\bar{x} - \mu}{\sigma/\sqrt{n}}$ (how many σ's the sample mean is from the actual mean)

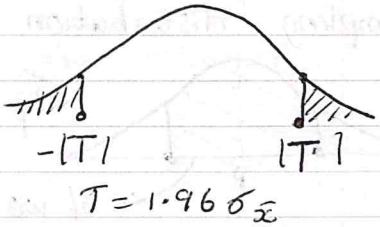
(iv) If test statistic is not within $\pm 1.96\sigma$'s then this means it is in the 5% level/region of significance - its inside the critical region, and we reject H_0. Otherwise, if it is within the $\pm 1.96\sigma$ region, we fail to reject H_0 as the test statistic is NOT in the critical region

P values: An alternative approach to hypothesis testing:

Once H_0 stated

H_1 stated, Test statistic calculated:

Then p = area outside $\pm$ test statistic



= probability of getting a result outside the test statistic.

$$p \text{ value} = P(Z \leq -|T|) + P(Z \geq |T|)$$

$$= 2P(Z \geq |T|) = 2[1 - P(Z < |T|)]$$