Maths Statistics Notes

Data collection

Population = The whole set of items of interest

Census = Taking observations / measurements of every member of a population

Sample = A selection of observations taken from a subset of the population which is used to find out information about the whole population.

Sampling unit = An individual unit of a population

Sampling frame = when sampling units are numbered / named to form a list

A simple random sample is one where every possible sample of size n has an equal chance of being selected.

In systematic sampling the required elements are chosen at regular intervals from a sampling frame.

In stratified sampling the population is divided into mutually exclusive strata (eg. males and females) and a random sample is taken from each.

In quota sampling an interviewer or researcher selects a sample that reflects the characteristics of the whole population.

Opportunity sampling consists of taking the sample from people who are available at the time the study is carried out and who fit the criteria.

Variables or data can be either qualitative or quantitative.

Continuous variable = can take any value in a given range

Discrete variable = can only take specific values in a given range (typically integers)

When data is in a grouped frequency table, the specific data values are not shown. The groups are also known as classes

- Class boundaries tell you the maximum and minimum values that belong in each class
- The midpoint is the average of the class boundaries
- The class width is the difference between the upper and lower class boundaries

The Large data set

You need to know:

- The types and ranges of data
- The characteristics of each location
- May need to recall trends from within the data set, or identify a location based on given data

Measures of location and spread

Mode / modal class = value / class that occurs the most often

Median = the middle value when the data values are put in order

Mean = $\bar{x} = \frac{\sum x}{n}$ Frequency Table mean = $\bar{x} = \frac{\sum xf}{\sum f}$

Lower Quartile : divide n by 4. → whole number: LQ is halfway between this data point and the one above; not a whole number: round up and LQ is this data point

Upper Quartile : times n by $^3/_4$

Range : difference between largest and smallest values

Interquartile Range (IQR) : difference between upper and lower quartile

Interpercentile Range : difference between the values for two given percentages

Variance = $\frac{\sum(x-\bar{x})^2}{n} = \frac{\sum x^2}{n} - \left(\frac{\sum x}{n}\right)^2 = \frac{S_{xx}}{n}$

$S_{xx} = \sum(x-\bar{x})^2 = \sum x^2 - \frac{(\sum x)^2}{n} \Rightarrow S_{xx} = \text{Variance} \times n$

Standard deviation = $\sigma = \sqrt{\text{variance}}$

$\sigma^2 = \frac{\sum f(x-\bar{x})^2}{\sum f} = \frac{\sum fx^2}{\sum f} - \left(\frac{\sum fx}{\sum f}\right)^2$

If data is coded using the formula $y = \frac{x-a}{b}$

- The mean of the coded data is given by $\bar{y} = \frac{\bar{x}-a}{b}$
- The standard deviation of the coded data is given by $\sigma_y = \frac{\sigma_x}{b}$ where σ_x is the standard deviation of the original data.

Representations of Data

A common definition of an outlier is any value that is:

greater than $Q_3 + k(Q_3 - Q_1)$

less than $Q_1 - k(Q_3 - Q_1)$

The process of removing anomalies from a data set is known as cleaning the data

On a histogram, to calculate the height of each bar (the frequency density) use area of bar = k × frequency

Joining the middle of the top of each bar in a histogram with equal class widths forms a frequency polygon.

When comparing data sets, you can comment on:

- a measure of location
- a measure of spread

Correlation

Bivariate data = data with pairs of values for two variables.

Correlation describes the nature of the linear relationship between two variables.

When two variables are correlated, you need to consider the context of the question and use your common sense to determine whether they have a causal relationship.

The regression line of y on x is written in the form $y = a + bx$.

The coefficient b tells you the change in y for each unit change in x.

- If the data is positively correlated, b will be positive
- If the data is negatively correlated, b will be negative

You should only use the regression line to make predictions for values of the dependent variable that are within the range of the given data.

Statistical Distributions

A probability distribution fully describes the probability of any outcome in the sample space.

The sum of the probabilities of all outcomes of an event add up to 1. For a random variable, X, you can write $\Sigma P(X = x) = 1$

You can model X with a binomial distribution $B(n, p)$ if:

- There are a fixed number of trials, n
- There are two possible outcomes (success or failure)
- There is a fixed probability of success, p
- The trials are independent of each other

If a random variable X has a binomial distribution $B(n, p)$:

$$P(X = r) = \binom{n}{r} p^r q^{n-r} \qquad \text{mean} = np \qquad \text{variance} = npq$$

Hypothesis testing

Null hypothesis H_0: The hypothesis that you assume to be correct

Alternative hypothesis H_1: tells us about the parameter if your assumption is shown to be wrong

When H_1 is of the form $p < \ldots$ or $p > \ldots$ it's called a one-tailed test.

Critical region: A region of the probability distribution which, if the test statistic falls within it, would cause you to reject the null hypothesis

Critical value: The first value to fall inside the critical region

Actual significance: The probability of incorrectly rejecting the null hypothesis

For a two-tailed test the critical region is split at either end of the distribution

For a two-tailed test, either double the p-value for your observation, or halve the significance level at the end you are testing

Regression, correlation, and hypothesis testing

The product moment correlation coefficient describes the linear correlation between two variables. It can take values between -1 and 1.

ρ = PMCC for a whole population

r = PMCC for a sample

↳ For one-tailed use: $H_0: \rho = 0$ $H_1: \rho > 0$ or < 0

For two-tailed use: $H_0: \rho = 0$ $H_1: \rho \neq 0$

When doing correlation tests you'll either use the tables to find the critical value for r, or they'll give you the probability of achieving the r value randomly (p-value) and you just compare it to the significance level

Conditional Probability

The event of A and B can be written as $A \cap B$ $\cap$ = intersection

The event of A or B can be written as $A \cup B$ $\cup$ = union

The event not A can be written as A' $'$ = compliment

The probability that B occurs given that A has already occurred is written as $P(B|A)$.

For independent: $P(A) = P(A|B) = P(A|B')$

$P(A \cap B) = P(A) \times P(B)$

$P(A \cup B) = P(A) + P(B) - P(A \cap B)$ $\quad P(B|A) = \frac{P(B \cap A)}{P(A)}$

$P(B \cap A) = P(B|A) \times P(A)$

The Normal Distribution

The area under a continuous probability distribution is equal to 1. If X is a normally distributed random variable, you write $X \sim N(\mu, \sigma^2)$ where μ is the population mean and σ^2 is the population variance.

The normal distribution

- has parameters μ, the population mean, and σ^2, the population variance
- is symmetrical (mean = median = mode)
- has a bell-shaped curve with asymptotes at each end
- has a total area under the curve equal to 1
- has points of inflection at $\mu + \sigma$ and $\mu - \sigma$

$\# P(Z < a) = \Phi(a)$

The standard normal distribution has mean 0 and standard deviation 1.

The standard normal variable is written as $Z \sim N(0, 1^2)$

If $X \sim N(\mu, \sigma^2)$, you can code X using $Z = \frac{X - \mu}{\sigma}$. The resulting z-values have $\mu = 0$, $\sigma =$

If n is large and p is close to 0.5 then the binomial distribution $X \sim B(n, p)$ can be approximated by the normal distribution $N(\mu, \sigma^2)$

where: $\mu = np$ and $\sigma = \sqrt{npq}$, (variance $= npq$)

When using this approximation you need to apply a continuity correction.

For a random sample of size n taken from a random variable $X \sim N(\mu, \sigma$

the sample mean is normally distributed with $\bar{X} \sim N\left(\mu, \frac{\sigma^2}{n}\right)$.

For the sample mean of a normally distributed random variable, $\bar{X} \sim N\left(\mu, \frac{\sigma^2}{n}\right)$

$Z = \dfrac{\bar{X} - \mu}{\frac{\sigma}{\sqrt{n}}}$ is a normally distributed random variable with $Z \sim N(0, 1)$