A Level Further Maths: Statistics Notes

Contents

Chapter 1

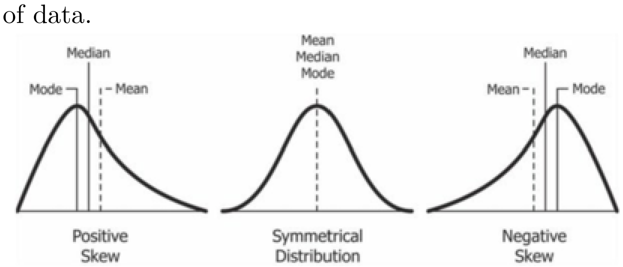
Conditional Probability

Sometimes it is more useful to find the probability of an event A given than an event B has occurred. We write this $p(A \mid B)$.
The conditional probability is defined as:
$p(A \mid B) = \frac{p(A \cap B)}{p(B)}$
If events A and B are independent $p(A \mid B) = p(A)$
Using tables is one of the easiest ways to work out the intersection.

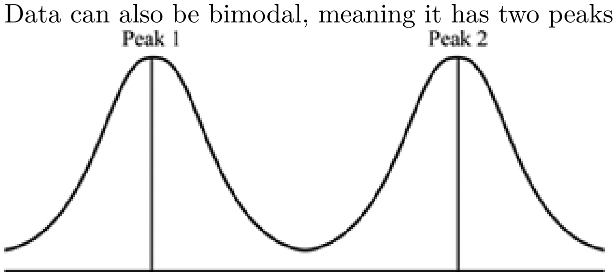
Chapter 2

Data Distribution

Data can be distributed due to its appearance to help us interpret the spread of data.



Data can also be bimodal, meaning it has two peaks



Measures of Central Tendency

There are four measures of central tendency that we should be aware of:

- Mode - The mode is the value that occurs most frequently. The distribution is uni-modal if there is only one mode. If two non-adjacent values occur mode frequently than the rest, the distribution is said to be bimodal, even if the frequencies aren't the same.
- Medial - It is the middle value. It is a good measure as it isn't skewed as much by outliers but most representative for symmetrical data.

- Mean - Found by adding all the values together and dividing by the number of values. We use the symbol $\bar{x}$ to denote the mean. It is the best measure for skewed data but also representative for symmetrical data. It is affected by outliers.
- Midrange - This is the average of the highest and lowest values. It is easy to calculate but only useful when the data is symmetrical and contains no outliers.

Measures of Spread

There are three measures of spread we need to know about which help us talk about how varied the data is.

- Range - This is the highest subtract the lowest. Again this is not useful if we have any outliers.
- Interquartile Range - As seen before. We can also find the semi-interquartile range which is half the interquartile range.
- Standard Deviation - This we will learn about separately but is the most widely used measure of spread.

Standard Deviation

Why is it useful?

- Approximately 68% of values lie within 1 standard deviation of the mean.
- 95% lie within 2

If you have a particular values which is more than two standard deviations from the mean it should be investigated as a possible outlier.

$$\begin{aligned}\text{Sum of Squares } &= S_{xx}\\ &= \sum x^2 - n\bar{x}^2\\ &= \sum x^2 - \frac{(\sum x)^2}{n}\end{aligned}$$

The variance is found by $var = s^2 = \frac{s_{xx}}{n-1}$
The standard deviation is found by the square root of the variance.
$s = \sqrt{\frac{s_{xx}}{n-1}} = \sqrt{var}$

Example 1

Find the standard deviation of 0, 1, 0, 3, 0, 2

$$\begin{aligned}
\bar{x} &= \frac{1+2+3}{6} = 1 \\
s_{xx} &= (0^2+1^2+0^2+3^2+0^2+2^2) - 6\times 1^2 \qquad = 14-6=8 \\
s &= \sqrt{\frac{s_{xx}}{n-1}} \\
&= \sqrt{\frac{8}{5}} = 1.26
\end{aligned}$$

Frequency tables

If you have a frequency table, we can still work out the standard deviation. The only thing that changes is our formula for sum of squares:
$s_{xx} = \sum x^2 f - n\bar{x}^2$

Binomial Distribution

- Running as a set number of trials, n
- Only two possible outcomes
- Probability of success is p
- Trials are independent

$X \sim B(n, p)$ means X is binomial distributed with n trials and a probability of success p.
$P(X = r) = \binom{n}{r} p^r (1-p)^{n-r}$ for r successes out of n

Probability / Cumulative

Binomial PD refers to the probability distribution at a single value. e.g. $P(X = a)$
Binomial CD refers to the cumulative distribution e.g. $P(X \leqslant a)$

Normal Distribution

All normal curves are governed by 2 parameters: The mean (μ) and the standard deviation (σ). They are symmetrical about the mean and there is a point of inflection 1 s.d. away from the mean.

If X has a normal distribution with mean μ and variance σ^2 we write $X \sim N(\mu, \sigma^2)$. The standard normal distribution has mean 0 and variance 1. It is written Z, so $Z \sim N(0, 1)$.

As normal distributions represent continuous data, it only makes sense to find the probability that X takes a value in a particular interval. The probability corresponds to the area under the curve in the interval. There is no simple formula that can be used to find the probability for an interval so we use a calculator.

All normal distributions can be mapped onto each other using a stretch and a shift. It is possible and easy to transform any normal distribution to a standard normal. If $X \sim N(\mu, \sigma^2)$, $Z = \frac{X-\mu}{\sigma} \sim N(0, 1)$

Working Backwards

If we know μ and σ and the probability, we can use the inverse normal function on a calculator to find the interval. We can also use it to form simultaneous equations which can be solved to find one or more of the variables.

Example 1

A machine is designed to fill jars so that mass, X follows a normal distribution with mean μg and s.g σg. If:

- $p(X > 210) = 0.025$
- $p(X < 108) = 0.04$

Find μ and σ to 3.s.f.

First we use the inverse normal function to go from the probabilities to a Z value. With these two values we can form two equations of the form $Z = \frac{x-\mu}{\sigma}$ with the given X values and calculated Z values. This means that we have two equations and two unknowns and they can be solved for μ and σ to give $\mu = 204, \sigma = 3.23$

Approximating a Discrete Distribution Using Normal

A continuity correction must be applied when approximating a discrete random variable (such as the binomial distribution) to a continuous distribution (such as the normal distribution).

We apply a continuity correction by treating each discrete bar as an interval of all the values that would round to it (e.g. 0.5 above and below). So if we want to look at $p(X = 5)$ this is corrected to $p(4.5 < X < 5.5)$.

Approximating the Binomial Using a Normal Distribution

To approximate the binomial using normal we must have:

- n large
- $0 \ll np \ll n$

If the conditions are met then X can be reasonably approximated using $X \sim N(np, npq)$ where $q = 1 - p$. We need to remember to apply a continuity correction as we are going from a discrete distribution to a continuous one.

This is useful where n is very large as the binomial distribution can become quite slow to calculate.

Chapter 3

Data Processing

Different Types of Data

- Categorical - Generally not numerical.
- Discrete - Only certain values, normally integers.
- Continuous - Numerical, real values.
- Ranked - Numerical, ordered.

Categorical data

For categorical data, the most common summary measure of our data is the modal class. This is the class with the highest frequency. Diagrams:

- Bar chart
- Pie chart
- Pictogram
- Pot chart

Ranked Data

If our data is ranked, we normally use stem and leaf diagrams or box plots to represent the data.

Stem and Leaf Diagrams

Example:

Key: 3 | 1 means 31

Stem	Leaf
1	9 9
2	0 4 7 8
3	1 2 2 2 6
4	0 5 5
5	5

Note: includes repeats, only a single digit on the right.

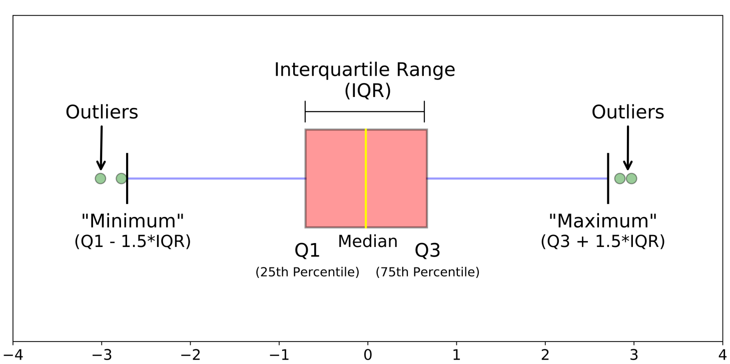
For ranked data, we often use the median, lower and upper quartiles as our summary measures.

- The median value (Q_2) - The middle number
- The lower quartile (Q_1) - The middle number of the lower half
- The upper quartile (Q_3) - The middle number of the upper half

If the number of data-points is off, you just take the middle number. If it is even, take the average between the two middle numbers and when calculating the quartiles, include the middle.

Box Plots

The five key numbers can be shown on a simple diagram known as a box-and-whisker plot.



Outliers

We can say that a data-point is an outlier if the data-point is more than $1.5\times$ IQR beyond or below the lower or upper quartiles.

Product-Moment Correlation Coefficient

The PMCC measures hot close the data is to a straight line.
$r = \frac{S_{xy}}{\sqrt{S_{xx}S_{yy}}}$ where

$$\text{Always} > 0 \begin{cases} S_{xx} = \Sigma x^2 - n\bar{x}^2 \\ S_{yy} = \Sigma y^2 - n\bar{y}^2 \end{cases}$$

$$\text{Positive or negative} \begin{cases} S_{xy} = \Sigma xy - n\bar{x}\bar{y} \end{cases}$$

We are assuming that the underlying population has a bivariate normal distribution. If we show the data on a scatter graph it should form an ellipse.

Chapter 4

Hypothesis Tests

One-Tailed Tests

We have two possible outcomes, H_0, H_1 which represent the null and alternative hypothesis. We initially assume that the null hypothesis is true until proven otherwise.
p is the probability of some event happening and our hypothesise are related to whether we think that the probability is higher or lower than it should be.
The significance level represents the probability of us rejecting the null hypothesis when it is true.

Steps for a One-Tail test

1. Define p and write out H_0 and H_1 in terms of p.
2. State the significance level - if none is mentioned in the question, assume it is 5%.
3. State the distribution, assuming the null hypothesis to be true.
4. Calculate the probability (under H_0) of obtaining result as or more extreme than those collected.
5. Compare the probability with the significance level and make conclusions - can H_0 be rejected or not? Interpret your results in context.

If the significance level is less than the probability, then we say: "There is insufficient evidence to reject the null hypothesis so the result is not significant. The evidence doesn't seem to suggest that..."

If the significance level is more than the probability, then we say: "There is sufficient evidence to reject the null hypothesis so the result is significant. The evidence seems to suggest that..."

Example 1

I have a coin which I suspect is more likely to show heads than tails. Run a hypothesis test at a 5% significance level to test than the claim if I took a sample of 20 coin flips and got 18 heads.

> p is the probability I get a head
>
> $H_0 : p = \frac{1}{2}$
>
> $H_1 : p > \frac{1}{2}$
>
> 5% significance level, one tailed test
>
> $X \sim B(20, \frac{1}{2})$
>
> $P(X \geqslant 18) = 0.0201\%(4dp)$
>
> $0.0201\% < 5\%$

There is sufficient evidence to reject the null hypothesis so the result is significant. The evidence seems to suggest that the coin is biased towards heads.

Two-Tailed Tests

Two tailed tests are similar to one tailed ones except the significance level should be split over the two tails. We end up only testing for one tail.
Our new hypothesises are $H_0 : p = a$ and $H_1 : p \neq a$.
We then check whether the actual value was higher or lower than the expected value ($E(x)$) and then continue as normal.

- $n > E(x) \to P(X \geqslant n)$
- $n < E(x) \to P(X \leqslant n)$

Example 2

I believe than 10% of people are left handed. Run a hypothesis test to test than the claim if I took a sample of 10 people and got 0 left handed people.

> p is the probability a person is left handed
>
> $H_0 : p = 0.1$
>
> $H_1 : p \neq 0.1$
>
> 5% significance level, two tailed test therefore 2.5% each tail
>
> $X \sim B(10, 0.1)$
>
> $E(X) = 1$ $\qquad P(X \leqslant 0) = 0.3487(4dp)$
>
> $0.3487 > 2.5\%$

There is insufficient evidence to reject the null hypothesis so the result is not significant. The evidence doesn't seem to suggest that the proportion of left handed people is not 10%.

Critical values and Critical Regions

Sometimes is is more useful to find the value for which we change from not rejecting the null hypothesis to rejecting it. This value is called the critical value.
The range of values for which you reject the null hypothesis is called the critical region. The range of values for which you can't reject the null hypothesis is called the acceptance region.

Example 3

Throw a coin 10 times. What is the critical region at 5% significance level if $H_1 : p < 0.5$?
Biased against tails? One tailed test at 5% significance level.

$$\begin{aligned} p(X \leqslant 1) &= 0.0107 \\ p(X \leqslant 2) &= 0.0546 \\ X &\leqslant 1 \end{aligned}$$

Hypothesis Tests of The Sample Mean

Given a population X with a mean of μ and a standard deviation σ i.e. $X \sim N(\mu, \sigma^2)$ and a sample of size n is taken, the distribution of the sample means is given by $\bar{X} \sim N(\mu, \frac{\sigma^2}{n})$. We are assuming that the underlying population follows a normal distribution.

$$\text{The null and alternative hypothesise are } H_0 : \mu = k,\, H_1 : \begin{cases} \mu > k \text{ (one tailed)} \\ \mu < k \text{ (one tailed)} \\ \mu \neq k \text{ (two tailed)} \end{cases}$$

Remember to define μ

Example 1

Test results are normally distributed with a mean of 65 and s.d. of 10. A new teacher has a group of 8 students with a mean test score of 72. Is there evidence that the results have significantly improved at a 5% significance level.

$\mu =$ the population mean test score of students.

$$\begin{aligned}
H_0 &: \mu = 65 \\
H_1 &: \mu > 65 \\
X &\sim N(65, 100) \\
\Rightarrow \bar{X} &\sim N(65, \frac{100}{8}) \\
p(\bar{X} &\geqslant 72) = 0.02387 < 0.05
\end{aligned}$$

$\therefore$ There is sufficient evidence to reject the null hypothesis. The evidence suggests that the population mean test score has increased.

Chapter 5

Probability

Estimating Probability

There are two ways to estimate probability:

- Experiment
- Theoretical

Experimental

$p = \frac{n_{\text{events}}}{n_{\text{trials}}}$
Requires data to be collected.

Theoretical

$p = \frac{n_{\text{ways}}}{n_{\text{outcomes}}}$
For example, the probability of getting a 6 on a dice is $\frac{1}{6}$

Modelling Probability

If A is impossible, $P(A) = 0$, if A is certain, $P(A) = 1$
$P(A') = 1 - P(A)$
If events A and B are $P(A \cap B) = P(A) \times P(B)$

Mutually exclusive

Two events are mutually exclusive events if they cannot both happen.
$P(A \cup B) = P(A) + P(B)$

If events are not mutually exclusive $P(A \cup B) = P(A) + P(B) - P(A \cap B)$ We can test for independence as if two events are independent $P(A \cap B) = P(A) \times P(B)$

Discrete Random Variables

A model is a discrete random variable X if it is:

- Discrete
- The actual values of the outcome of the variable can only be predicted with a given probability.

Discrete Random Variables may have a finite or countably infinite number of outcomes.

Notation

The particular values our DRV can take are denoted by r, this $P(X = r)$ means the probability that the DRV X has the outcome r.
The sum of these probabilities equal 1. Formally, $\sum_{r=1}^{n} P(X = r) = p_1 + p_2 + \cdots + p_n = 1$

Expectation

The most useful measure of central tendency is usually the mean (or the expectation). We can apply a similar idea for DRV. We define the expectation as:
$E(x) = \sum rP(X = r)$
Note we often use the Greek symbol μ to represent $E(x)$ as well.
$\bar{x}$ is the mean when it is a sample. μ is the mean when it is a population.