REVISION SHEET – STATISTICS 1 (MEI)

THE BINOMIAL DISTRIBUTION & PROBABILITY

The main ideas in this chapter are

- Probabilities based on selecting or arranging objects
- Probabilities based on the binomial distribution
- The expected value of a binomial distribution
- Expected frequencies from a series of trials

Before the exam you should know:

- $n!$ is the number of ways of ordering a collection of n objects and ${}^nC_r$ is the number of ways of selecting a group of r objects from a total of n objects.
- when a situation can be modelled by the binomial distribution.
- the formula: $P(X = r) = {}^nC_r p^r q^{n-r}$ and how to use it.
- how to use the binomial distribution tables (in particular that they give cumulative probabilities).
- the mean or expected value of $X \sim B(n,p)$ is np.
- how to calculate expected frequencies when a set of trials is repeated.

Probabilities based on selecting or arranging

- $n! = n \times (n-1) \times (n-2) \ldots \times 2 \times 1$ is the number of ways of ordering a collection of n objects.
- ${}^nC_r = \dfrac{n!}{(n-r)!r!}$ is the number of ways of selecting r objects from n.

Example
Find the number of different 4-digit numbers than can be made using each of the digits 7,8,9,0 once.
Solution
This is the number of ways of ordering the digits 7, 8, 9, 0. For example 7890 and 7809 are two such orderings. This is given by $4! = 4 \times 3 \times 2 \times 1 = 24$.
This can be thought of as: "there are 4 possibilities for the 1st number, then there are 3 possibilities for the 2nd number, then there are 2 possibilities for the 3rd number, leaving only one possibility for the 4th number.

Example
Eddie is cooking a dish that requires 3 different spices and 2 different herbs, but he doesn't remember which ones. In his cupboard he has 10 different jars of spices and 5 different types of herb and he knows from past experience that the ones he needs are there.

(i) How many ways can he choose the 3 spices?
(ii) How many ways can he choose the 2 herbs?
(iii) If he chooses the herbs and spices at random what is the probability that he makes the correct selection?

Solution
(i) ${}^{10}C_3 = 120$ (ii) ${}^5C_2 = 10$ (You can work these out using the ${}^nC_r$ function on a calculator.)
(iii) $1 \div (120 \times 10) = 0.000833$
In part (iii) we multiply the results of (i) & (ii) to get 1200 different possible combinations. Only 1 of these is the correct selection so the probability of making the correct selection is $1 \div 1200$.

Probabilities based on the binomial distribution

The binomial distribution may be used to model situations in which:

1. you are conducting n trials where for each trial there are two possible outcomes, often referred to as success and failure.
2. the outcomes, success and failure, have fixed possibilities, p and q, respectively and $p + q = 1$.
3. the probability of success in any trial is independent of the outcomes of previous trials.

The binomial distribution is then written $X \sim B(n, p)$ where X is the number of successes. The probability that X is r, is given by $P(X = r) = {}^{n}C_{r}\, p^{r} (1-p)^{n-r}$

Example

A card is taken at random from a standard pack of 52 (13 of each suit: Spades, Hearts, Clubs, Diamonds). The suit is noted and the card is returned to the pack. This process is repeated 20 times and the number of Hearts obtained is counted.

(i) State the binomial distribution that can be used to model this situation.
(ii) What is the probability of obtaining exactly 6 Hearts?
(iii) What is the probability of obtaining 6 or less Hearts?
(iv) What is the probability of obtaining less than 4 Hearts?
(v) What is the probability of obtaining 6 or more Hearts?

Solution

(i) $X \sim B(20,0.25)$ (ii) $P(X = 6) = {}^{20}C_{6} \times 0.25^{6} \times (0.75)^{20-6} = 0.1686$

(iii) $P(X \leq 6) = 0.7858$ (This can be read straight from the tables as it is a "≤ probability").

(iv) $P(X < 4) = P(X \leq 3) = 0.2252$

(v) $P(X \geq 6) = 1 - P(X \leq 5) = 1 - 0.6172 = 0.3828$

You need to be very careful with >, < or ≥. These must all be converted to ≤ if you are going to use the tables. In (iv) 'less than 4' is the same as '3 or less'. In (v) the complement of '6 or more' is '5 or less'.

The expected value of a binomial distribution

The expected value (mean) of a binomial distribution $X \sim B(n,p)$ is $E[X] = np$.

Example A die is rolled 120 times. How many 3's would you expect to obtain.

Solution Here success would be defined as getting a 3, and failure not getting a 3. Therefore $n = 120$, $p = 1/6$ and $q = 5/6$. X, the number of 3s obtained is modelled by $X \sim B(120,1/6)$ and so $E[X] = np = 120 \times (1/6) = 20$.

Expected frequencies from a series of trials

If a situation modelled by a binomial distribution is repeated then the expected frequency of a given number of successes is found by multiplying the probability of that number of successes by the number of times the set of trials is repeated.

Example

The probability of an individual egg being broken during packing is known to be 0.01.

(i) What is the probability that a box of 6 eggs will have exactly 1 broken egg in it?
(ii) In a consignment of 100 boxes how many boxes would you expect to contain exactly 1 broken egg?

Solution

(i) Using $X \sim B(6,0.01)$, $P(X = 1) = {}^{6}C_{1} \times 0.01^{1} \times (0.99)^{6-1} = 0.057$.

(ii) $0.057 \times 100 = 5.7$ boxes. (This is an expected value and does not have to be an integer).

REVISION SHEET – STATISTICS 1 (MEI)

DATA PRESENTATION

The main ideas in this chapter are

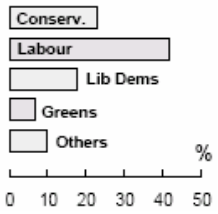
- **Bar charts**
- **Vertical line graphs**
- **Pie charts**
- **Histograms**
- **Cumulative frequency**

Before the exam you should know:

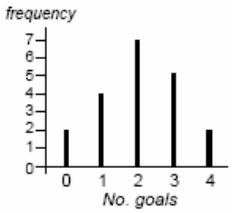
- How to draw, interpret and comment on:
  - Bar charts and pie charts for categorical and discrete data.
  - Vertical line graphs for discrete data.
  - Histograms are used to illustrate grouped, continuous data. The groups can have different width and the area of each column is proportional to the frequency. The vertical axis is frequency density, which is calculated by dividing frequency by class width. There are no gaps between the between the columns.
- About cumulative frequency.
  - That points are plotted at the upper class boundary. The curve is used to find estimates for the median, upper and lower quartiles and the inter-quartile range.
  - The upper and lower quartiles can also be calculated from the data.
- A box and whisker plot is a useful way of showing the median, inter quartile range, range and any outliers

Discrete data

Horizontal bar chart



Vertical line graph

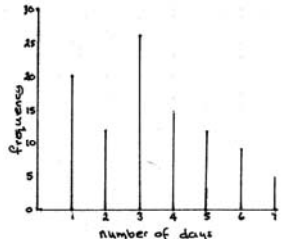


Example: A random sample of cyclists were asked how many days they had used their bicycles in the last week. The results are given in the following table.

Number of days (x)	1	2	3	4	5	6	7
Frequency (f)	20	12	26	15	12	9	6

Illustrate the distribution using a suitable diagram and describe its shape.

Answer

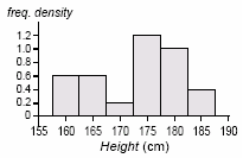


The distribution is bimodal with a slight positive skew

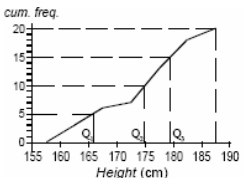
This is a good example of where to use a vertical line graph

Continuous data

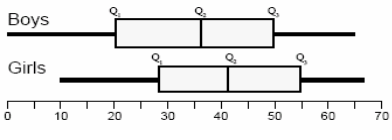
Histogram



Cumulative frequency curve



Box and whisker plot



The box and whisker plot shows the range, median and quartiles. It is a good way of comparing two distributions.

Quartiles and Percentiles

Lower (Q_1) and **upper** (Q_3) **quartiles**: values ¼ way and ¾ way through the distribution.

Percentile: The **nth** percentile is the value **n / 100** way through the distribution.

Inter quartile range (IQR)

A measure of spread calculated by subtracting the lower quartile from the upper quartile: IQR = $Q_3 - Q_1$

An outlier can also be defined as a piece of data at least 1.5 × IQR beyond the nearer quartile
(below $Q1 - 1.5 \times$IQR or above $Q3 + 1.5 \times$IQR)

Example: The numbers of people using the local bus service on 20 weekday mornings were as follows.

184	192	175	171	195	186	178	177	182	165
183	180	186	170	196	171	187	151	186	199

(i) Calculate the median and the inter-quartile range.
(ii) Using the inter-quartile range, show that there is just one outlier. Find the effect of its removal on the median and the inter-quartile range.

Answer: 151, 165, 170, 171, 171, 175, 177, 178, 180, 182, 183, 184, 186, 186, 186, 187, 192, 195, 196, 199
(i) median (Q_2) = (182+183)/2 = 182.5
Lower quartile (Q_1) = (171+175)/2 = 173 Upper Quartile (Q_3) = (186+187)/2 = 186.5
IQR = 186.5 – 173 = 13.5

(ii) 1.5 X IQR = 20.25. Q_1 – 20.25 = 152.75 Q_3 +20.25 = 206.75
So 151 is the outlier.
Remove 151: Q_1 = 173 Q_2 = 182 Q_3 = 189.5 IQR = 16.5

If 151 is removed, the median drops by 0.5 but the IQR increases by 3.

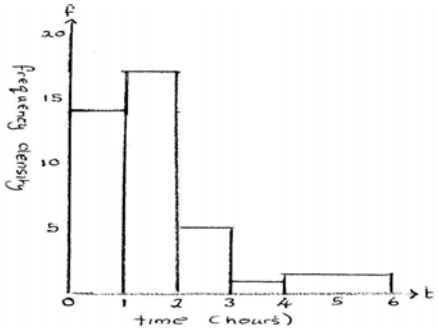
Example: A magazine carried out a survey of the ages of 60 petrol and 60 diesel cars for sale. In the survey, the value of each car was expressed as a percentage of its value when new. The results of the survey are summarised in the table.

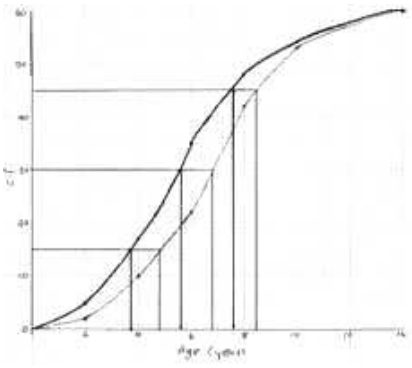
(i) Display the data for petrol cars on a histogram
(ii) Draw a cumulative frequency table for each set of results. On the same axes draw the corresponding cumulative frequency graphs
(iii) Use your curves to estimate the median and inter-quartile range for each type of car
(iv) Comment on the differences in the two distributions

Age in years (x)	petrol cars	diesel cars
$0 \le x < 2$	5	2
$2 \le x < 4$	12	8
$4 \le x < 6$	18	12
$6 \le x < 8$	13	20
$8 \le x < 10$	6	11
$10 \le x < 14$	6	7

Answer

Age in years (x)	**petrol**	Class width	Frequency density	c.f	**diesel**	c.f
$0 \le x < 2$	**5**	2	2.5	5	**2**	2
$2 \le x < 4$	**12**	2	6	17	**8**	10
$4 \le x < 6$	**18**	2	9	35	**12**	22
$6 \le x < 8$	**13**	2	6.5	48	**20**	42
$8 \le x < 10$	**6**	2	3	54	**11**	53
$10 \le x < 14$	**6**	4	1.5	60	**7**	60





	Petrol	Diesel
Median	5.6	6.8
Q_1	3.7	4.8
Q_3	7.6	8.3
IQR	3.9	3.5

The median age of diesel cars is higher, suggesting that diesel cars are generally older than petrol cars.
The inter-quartile range for petrol cars is greater than for diesel cars so there is greater variation in the gases of petrol cars.

REVISION SHEET – STATISTICS 1 (MEI)

DISCRETE RANDOM VARIABLES

The main ideas are

- Discrete random variables
- Expectation (mean) of a discrete random variable
- Variance of a discrete random variable

Before the exam you should know:

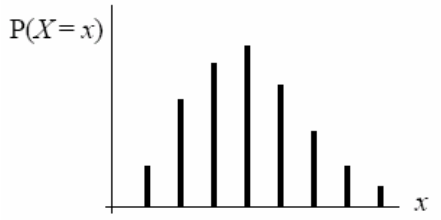
- Discrete random variables are used to create mathematical models to describe and explain data you might find in the real world.
- You must understand the notation that is used.
- You must know that a discrete random variable X takes values r_1, r_2, r_3, r_4, …, r_n with corresponding probabilities: p_1, p_2, p_3, p_4, …, p_n
- Remember that the sum of these probabilities will be 1 so $p_1 + p_2 + p_3 + p_4, \ldots + p_n = \Sigma P(X=r_k) = 1$
- You should understand that the expectation (mean) of a discrete random variable is defined by

  $\Rightarrow E(X) = \mu = \sum rP(X=r_k)$
- You should understand that the variance of a discrete random variable is defined by

  $\Rightarrow \text{Var}(X) = \sigma^2 = E(X-\mu)^2 = \Sigma(r-\mu)^2P(X=r)$

  $\Rightarrow \text{Var}(X) = \sigma^2 = E(X^2) - [E(X)]^2$

Discrete random variables with probabilities p_1, p_2, p_3, p_4, …, p_n can be Illustrated using a vertical line chart:



Notation

- A discrete random variable is usually denoted by a capital letter (X, Y etc).
- Particular values of the variable are denoted by small letters (r, x etc)
- $P(X=r_1)$ means the probability that the discrete random variable X takes the value r_1
- $\Sigma P(X=r_k)$ means the sum of the probabilities for all values of r, in other words $\Sigma P(X=r_k) = 1$

Example: A child throws two fair dice and adds the numbers on the faces. Find the probability that

(i) P(X=4) (the probability that the total is 4)

(ii) P(X<7) (the probability that the total is less than 7)

Answer:

(i) $P(X=4) = \frac{3}{36} = \frac{1}{12}$ (ii) $P(X<7) = \frac{15}{36} = \frac{5}{12}$

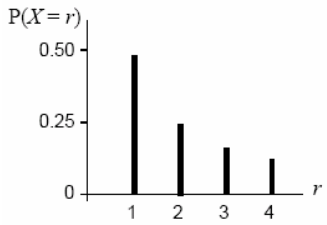
Example: *X* is a discrete random variable given by $P(X = r) = \frac{k}{r}$ for *r* = 1, 2, 3, 4 Find the value of k and illustrate the distribution

Answer: To find the value of *k*, use $\Sigma P(X = x_i) = 1$

$\Sigma P(X = x_i) = \frac{k}{1} + \frac{k}{2} + \frac{k}{3} + \frac{k}{4} = 1$

$\Rightarrow \frac{25}{12}k = 1 \Rightarrow k = \frac{12}{25} = \mathbf{0.48}$

Illustrate with a vertical line chart:



Example Calculate the expectation and variance of the distribution

Answer:

expectation is $E(X) = \mu = \Sigma rP(X = r)$

$= 1\times0.48 + 2\times0.24 + 3\times0.16 + 4\times0.12 = \mathbf{1.92}$

$E(X^2) = \Sigma r^2P(X = r)$

$= 1^2\times0.48 + 2^2\times0.24 + 3^2\times0.16 + 4^2\times0.12 = 4.5$

variance is $\text{Var}(X) = E(X^2) - [E(X)]^2$

$= 4.5 - 1.92^2 = \mathbf{0.8136}$

Using tables: For a small set of values it is often convenient to list the probabilities for each value in a table

r_i	r_1	r_2	r_3	….	r_{n-1}	r_n
$P(X = r_i)$	p_1	p_2	p_3	….	p_{n-1}	p_n

Using formulae: Sometimes it is possible to define the probability function as a formula, as a function of r, $P(X = r) = f(r)$

Calculating probabilities: Sometimes you need to be able to calculate the probability of some compound event, given the values from the table *or* function.

Explanation of probabilities:

Often you need to explain how the probability $P(X = r_k)$, for some value of k, is derived from first principles.

Example: The discrete random variable X has the distribution shown in the table

r	0	1	2	3
$P(X = r)$	0.15	0.2	0.35	0.3

(i) Find $E(X)$.
(ii) Find $E(X^2)$.
(iii) Find $Var(X)$ using (*a*) $E(X^2) - \mu^2$ and (*b*) $E(X - \mu)^2$.
(iv) Hence calculate the standard deviation.

r	0	1	2	3	totals
$P(X = r)$	0.15	0.2	0.35	0.3	**1**
$rP(X = r)$	0	0.2	0.7	0.9	**1.8**
$r^2P(X = r)$	0	0.2	1.4	2.7	**4.3**
$(r - \mu)^2$	3.24	0.64	0.04	1.44	**5.36**
$(r-\mu)^2P(X=r)$	0.486	0.128	0.014	0.432	**1.06**

This is the expectation (μ)

This is $E(X^2)$

This is $Var(X) = \Sigma(r-\mu)^2P(X=r)$

(i) $E(X) = \mu = \Sigma r\, P(X = r) = 0\times0.15 + 1\times0.2 + 2\times0.35 + 3\times0.3$
$= 0 + 0.2 + 0.7 + 0.9$
$= \mathbf{1.8}$

(ii) $E(X^2) = \Sigma r^2\, P(X = r) = 0^2x0.15 + 1^2x0.2 + 2^2x0.35 + 3^2x0.3$
$= 0 + 0.2 + 1.4 + 2.7$
$= \mathbf{4.3}$

(iii) (a) $Var(X) = E(X^2) - \mu^2 = 4.3 - 1.8^2 = \mathbf{1.06}$

(b) $Var(X) = E(X - \mu)^2 = 0.15(0-1.8)^2 + 0.2(1-1.8)^2 + 0.35(2-1.8)^2 + 0.3(3-1.8)^2$
$= 0.486 + 0.128 + 0.014 + 0.432$
$= \mathbf{1.06}$

Notice that the two methods give the same result since the formulae are just rearrangements of each other

(iv) $s = \sqrt{1.06} = 1.02956 = \mathbf{1.030}$ (3d.p.)

standard deviation (s) is the square root of the variance

REVISION SHEET – STATISTICS 1 (MEI)

EXPLORING DATA

The main ideas are

- Types of data
- Stem and leaf
- Measures of central tendency
- Measures of spread
- Coding

Before the exam you should know:

- How to identify whether the data is categorical, discrete or continuous.
- How to describe the shape of a distribution, say whether it is skewed positively or negatively and be able to identify any outliers.
- Be able to draw an ordered stem and leaf and a back to back stem and leaf diagram.
- Be able to calculate and comment on the mean, mode, median and mid-range.
- Be able to calculate the range, variance and standard deviation of the data.

Types of data

Categorical data or qualitative data are data that are listed by their properties e.g. colours of cars

Numerical or quantitative data

Discrete data are data that can only take particular numerical values. e.g. shoe sizes

Continuous data are data that can take any value. It is often gathered by measuring e.g. length, temperature

Frequency Distributions

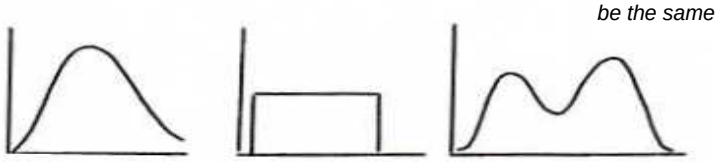
Frequency distributions: data are presented in tables which summarise the data. This allows you to get an idea of the shape of the distribution.

Grouped discrete data can be treated as if it were continuous, e.g. distribution of marks in a test.

Shapes of distributions

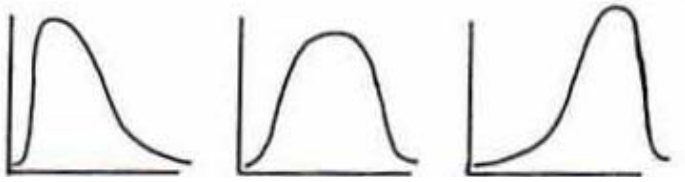
Symmetrical (Unimodal) **Uniform** **Bimodal**

bimodal does not mean that the peaks have to be the same height



Skew

Positive Skew **Symmetrical** **Negative Skew**



Stem and leaf diagrams
A concise way of displaying discrete or continuous data (measured to a given accuracy) whilst retaining the original information. Data usually sorted in ascending order and can be used to find the mode, median and quartiles. You are likely to be asked to comment on the shape of the distribution.

Example
Average daily temperatures in 16 cities are recorded in January and July. The results are
January: 2, 18, 3, 6, -3, 23, -5, 17, 14, 29, 28, -1, 2, -9, 28, 19
July: 21, 2, 16, 25, 5, 25, 19, 24, 28, -1, 8, -4, 18, 13, 14, 21
Draw a back to back stem and leaf diagram and comment on the shape of the distributions.

Answer

Jan					July				
9	5	3	1	**-0**	1	4			
6	3	2	2	**0**	2	5	8		
9	8	7	4	**10**	3	4	6	8	9
9	8	8	3	**20**	1	1	4	5	5	8

The January data is uniform but the July data has a negative skew

Central Tendency (averages)

Mean: $\bar{x} = \frac{\Sigma x}{n}$ (raw data)

$\bar{x} = \frac{\Sigma xf}{\Sigma f}$ (grouped data)

Median: mid-value when the data are placed in rank order

Mode: most common item or class with the highest frequency

Mid-range: (minimum + maximum) value ÷ 2

Outliers

These are pieces of data which are at least two standard deviations from the mean
i.e. beyond $\bar{x} \pm 2s$

Dispersion (spread)

Range: maximum value – minimum value

Sum of squares:

$S_{xx} = \Sigma(x-\bar{x})^2 \equiv \Sigma x^2 - n\bar{x}^2$ (raw data)

$S_{xx} = \Sigma(x-\bar{x})^2 f \equiv \Sigma x^2 f - n\bar{x}^2$ (frequency dist.)

Mean square deviation: msd = $\frac{S_{xx}}{n}$

Root mean squared deviation: *rmsd* = $\sqrt{\frac{S_{xx}}{n}}$

Variance: $s^2 = \frac{S_{xx}}{n-1}$

Standard deviation: $s = \sqrt{\frac{S_{xx}}{n-1}}$

Example: Heights measured to nearest cm:
159, 160, 161, 166, 166, 166, 169, 173, 173, 174, 177, 177, 177, 178, 180, 181, 182, 182, 185, 196.

Modes = 166 and 177 (i.e. data set is *bimodal*), **Midrange** = (159 + 196) ÷ 2 = 177.5 , **Median** = (174 + 177) ÷ 2 = 175.5

Mean: $\bar{x} = \frac{\Sigma x}{n} = \frac{3472}{20}$ = 174.1

Range = 196 – 159 = 37

Sum of squares: $S_{xx} = \Sigma x^2 - n\bar{x}^2$ = 607886 – 20 ×174.1^2 = 1669.8

Root mean square deviation: rmsd = $\sqrt{\frac{S_{xx}}{n}} = \sqrt{\frac{1669.8}{20}}$ = 9.14 (3 s.f.) **Standard deviation:** $s = \sqrt{\frac{S_{xx}}{n-1}} = \sqrt{\frac{1669.8}{19}}$ = 9.37 (3 s.f.)

Outliers (a): 174.1 ± 2× 9.37 = 155.36 *or* 192.84 - *the value* 196 *lies beyond these limits*, *so one outlier*

Example

A survey was carried out to find how much time it took a group of pupils to complete their homework. The results are shown in the table below. Calculate an estimate for the mean and standard deviation of the data.

Time taken (hours), t	0<t≤1	1<t≤2	2<t≤3	3<t≤4	4<t≤6
Number of pupils, f	14	17	5	1	3

Answer

Time taken (hours), t	0<t≤1	1<t≤2	2<t≤3	3<t≤4	4<t≤6
Mid interval, *x*	0.5	1.5	2.5	3.5	5
Number of pupils, f	14	17	5	1	3
fx	7	25.5	12.5	3.5	15
fx^2	3.5	38.25	31.25	12.25	75

$\bar{x}$ = 7+25.5+12.5+3.5+15 / 14+17+5+1+3 = 63/40 = 1.575

S_{xx} = (3.5+38.25+31.25+12.25+75) – (40 X 1.575^2) = 61.025

s = √(61.025/39) = 1.251 (3dp)

Linear coding

If the data are coded as y = ax +b then the mean and standard deviation have the coding are
$\bar{y} = a\bar{x} + b$ (the same coding) and $s_y = as_x$ (multiply by the multiplier of x)

Example

For two sets of data x and y it is found that they are related by the formula *y* = 5*x* – 20:
Given $\bar{x}$ = 24.8 and s_x = 7.3, find the values of $\bar{y}$ and s_y

$\bar{y}$ = (5 × 24.8) – 20 = 102

s_y = 5 × 7.3 = 36.5

REVISION SHEET – STATISTICS 1 (MEI)

HYPOTHESIS TESTING USING THE BINOMIAL DISTRIBUTION

The main ideas are

- Establishing the null and alternative hypotheses
- Conducting the test, doing the necessary calculations
- Interpreting the results

Before the exam you should know:

- The vocabulary associated with hypothesis testing.
- How to write the null and alternative hypotheses.
- How to decide whether the hypothesis test is one or two tailed.
- How to compare a value to the significance level.
- How to find critical values/regions.
- How to decided whether to reject H_0 or not and how to write a conclusion based on the situation.
- How to carry out a 2-tail test.

Vocabulary

You should be familiar with the following terms/notation for binomial hypothesis tests

Probability of success: p **Number of trials:** n **Number of successes:** X

Null Hypothesis (H_0):	The statement that the probability of success is equal to a certain value.
Alternative Hypothesis (H_1):	The statement that the probability of success is actually $<$, $>$ or $\neq$ to the value in given in H_0.
Significance level:	The probability at which you make the decision that an observed outcome hasn't happened by chance, given the probability of success in H_0.
1-tail test:	A test based on the probability in H_0 being either too high or too low (but not both).
2-tail test:	A test based on the probability in H_0 being incorrect (too high or too low).
Critical value:	The maximum (for $<$) or minimum (for $>$) value, X, for the number of successes that would result in rejecting H_0.
Critical region:	The set of values of X for the number of successes that would result in rejecting H_0.
Acceptance region:	The set of values of X for the number of successes that would result in accepting H_0.

Hypothesis Tests

Hypothesis testing is based on assuming that the probability of success, p, takes a certain value, then conducting an experiment to test it. Given this assumption, if the result of the experiment is sufficiently rare (i.e. unlikely to have happened by chance) you can conclude that the probability, p, is likely to be incorrect.

Setting up

The statement of the value of this assumed probability, p, is known as the Null Hypothesis (H_0) (this is what you are testing). You must then decide if the situation leads you to think this value is too high, too low or, in the case of a 2-tailed test, incorrect.

Conducting

The probability of obtaining the value recorded in the experiment, x, or something more extreme is compared to the significance level to see if it is sufficiently rare to reject the null hypothesis. You must use $P(X \leq x)$ or $P(X \geq x)$ as opposed to $P(X = x)$.

Drawing conclusions

If the probability is smaller than the significance level then reject H_0 in favour of H_1, otherwise you accept H_0 at the stated significance level.

Example

The makers of the drink Fizzicola claim that three-quarters of people prefer their drink to any other brand of cola. A rival company suspects that the claim by Fizzicola is exaggerated. They wish to carry out a hypothesis test to test this claim.

(i) State suitable Null and Alternative Hypotheses.

The rival company take a sample of 15 cola drinkers of whom 9 say they prefer Fizzicola to any other brand.

(ii) Using these data, carry out a hypothesis test at the 5% level stating your conclusion carefully.

Solution

(i) H_0: $p = 0.75$; The probability of a person chosen at random preferring Fizzicola is 0.75.
H_1: $p < 0.75$; The claim is exaggerated, the probability of a person chosen at random preferring Fizzicola is less than 0.75.
The alternative hypothesis is based on the rival branding thinking the claim is exaggerated, i.e. the proportion stated is too high.

(ii) From the tables: $P(X \leq 9) = 0.1484$. This value is not significant at the 5% level, therefore we accept H_0. There is not sufficient evidence to suggest Fizzicola are overestimating the proportion.
The probability of 9 or fewer is used, as opposed to exactly 9, as if you would accept 9 as evidence of overestimating then you would have also accepted 8, 7, 6, …
The significance level tells you the value at which a probability is considered so rare that is unlikely to have happened by chance. In this example case 5% is used, so an event with probability smaller than 0.05 is considered rare: 0.1484 is not smaller than 0.05 so the event is not rare.
As the event is not rare, it is likely that it occurred by chance, so there is no evidence to suggest that the makers of Fizzicola were overestimating. Note that you are not saying that they are correct, just that you don't have strong enough evidence to contradict them.

Alternative solution using critical value/critical region

(ii) From the tables: $P(X \leq 7) = 0.0173$, $P(X \leq 8) = 0.0566$. The critical value is 7, (the critical region is 0-7). 9 is not in the critical region (it is in the acceptance region), therefore we do not reject H_0. There is not sufficient evidence to suggest Fizzicola are overestimating the proportion.
The critical value is the largest (because H_1 is <) value of x such that $P(X < x)$ is smaller than the significance level.

This example used an alternative hypothesis of the form H_1: $p < 0.75$ (because the rival firm thought the company was overestimating). This made it easy to read the values for $P(X \leq 7)$, $P(X \leq 8)$ and $P(X \leq 9)$ from the tables. If the alternative hypothesis had been of the form H_1: $p > 0.75$ (e.g. if the firm thought 0.75 was an underestimate), you would need to work with ≥ probabilities, using the complement of the values in the table.

e.g. If the alternative hypothesis had been H_1: $p > 0.75$ you would have calculated $P(X \geq 9)$.
$P(X \geq 9) = 1 - P(X \leq 8) = 1 - 0.0566 = 0.9434$: this is not smaller than 0.05 so you do not reject H_0.

1-tail vs 2-tail tests.

If there is no indication in the situation as whether the probability used in H_0 is too high or too low you use a 2-tailed test, splitting the significance level in half and using half at each end.

Example

A teacher is forming a 12-person committee of students. She does not want the selection system to unfairly favour either boys or girls. Construct a hypothesis test at the 5% level to test this.

Solution

H_0: $p = 0.5$, There is an equal chance of a boy or girls being chosen.
H_1: $p \neq 0.5$, The selection system favours one gender.
You then split the significance level in half forming two critical regions of 2.5% at the top and bottom, totalling 5%.
Critical regions: 0 - 2 and 10 - 12.

REVISION SHEET – STATISTICS 1 (MEI)

PROBABILITY

The main ideas are

- Measuring probability
- Estimating probability
- Expectation
- Combined probability
- Two trials
- Conditional probability

The experimental probability of an event is = $\frac{\text{number of successes}}{\text{number of trials}}$

If the experiment is repeated 100 times, then the *expectation* (expected frequency) of a picture card being chosen = $n \times P(A)$

The sample space for an experiment illustrates the set of all possible outcomes. Any event is a sub-set of the sample space. Probabilities can be calculated from first principles.

Example: If two fair dice are thrown and their scores added the sample space is

+	1	2	3	4	5	6
1	2	3	4	5	6	7
2	3	4	5	6	7	8
3	4	5	6	7	8	9
4	5	6	7	8	9	10
5	6	7	8	9	10	11
6	7	8	9	10	11	12

If event *A* is "the total is 7" then

$P(A) = \frac{6}{36} = \frac{1}{6}$

If event *B* is "the total > 8" then

$P(B) = \frac{10}{36} = \frac{5}{18}$

If the dice are thrown 100 times, the expectation of event B is

100 X P(B) = 100 X $\frac{5}{18}$ = 27.7778

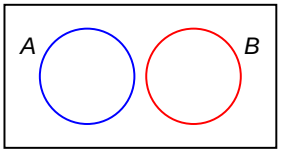
or 28 (to nearest whole number)

Before the exam you should know:

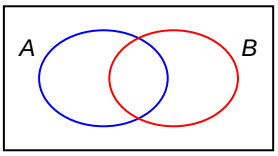
- The theoretical probability of an event A is given by $P(A) = \frac{n(A)}{n(\xi)}$ where A is the set of favourable outcomes and ξ is the set of all possible outcomes.
- The complement of A is written A' and is the set of possible outcomes not in set A. P(A') = 1 – P(A)
- For any two events A and B:
- $P(A \cup B) = P(A) + P(B) – P(A \cap B)$
- [or P(A or B) = P(A) + P(B) – P(A and B)]
- Tree diagrams are a useful way of illustrating probabilities for both independent and dependent events.
- Conditional Probability is the probability that event B occurs if event A has already happened. It is given by $P(B \mid A) = \frac{P(A \cap B)}{P(A)}$

More than one event

Events are **mutually exclusive** if they cannot happen at the same time so P(A **and** B) = $P(A \cap B) = 0$



Addition rule for mutually exclusive events:
P(A **or** B) = $P(A \cup B)$ = P(A) + P(B)



For non-mutually exclusive events
$P(A \cup B) = P(A) + P(B) – P(A \cap B)$

Example: An ordinary pack of cards is shuffled and a card chosen at random.

Event *A* (card chosen is a picture card): $P(A) = \frac{12}{52}$

Event *B* (card chosen is a 'heart'): $P(B) = \frac{13}{52}$

Find the probability that the card is a picture card **and** a heart.

$P(A \cap B) = \frac{12}{52} \text{ X } \frac{13}{52} = \frac{3}{52}$:

Find the probability that the card is a picture card **or** a heart.

$P(A \cup B) = P(A) + P(B) – P(A \cap B)$

$= \frac{12}{52} + \frac{13}{52} – \frac{3}{52} = \frac{22}{52} = \frac{11}{26}$

Tree Diagrams

Remember to multiply probabilities along the branches (*and*) and add probabilities at the ends of branches (*or*)

Independent events

P(*A* and *B*) = P(*A* ∩ *B*) = P(*A*) × P(*B*)

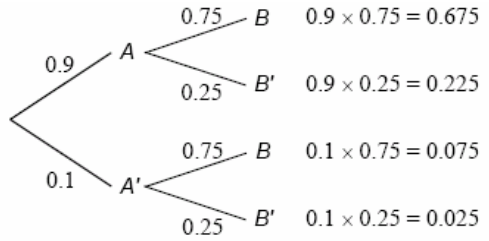
Example1: A food manufacturer is giving away toy cars and planes in packets of cereals. The ratio of cars to planes is 9:1 and 25% of toys are red. Joe would like a car that is not red.

Construct a tree diagram and use it to calculate the probability that Joe gets what he wants.

Answer:

Event *A* (the toy is a car): P(*A*) = 0.9
Event *B* (the toy is not red): P(*B*) = 0.75

The probability of Joe getting a car that is not red is 0.675



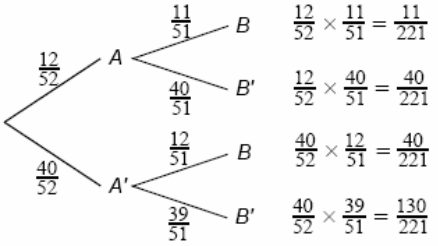
Example 2: dependent events

A pack of cards is shuffled; Liz picks two cards at random without replacement. Find the probability that both of her cards are picture cards

Answer:

Event *A* (1st card is a picture card)
Event *B* (2nd card is a picture card)

The probability of choosing two picture cards is $\frac{11}{221}$



Conditional probability

If A and B are **independent events** then the probability that event B occurs is not affected by whether or not event A has already happened. This can be seen in example 1 above. For independent events P(B/A) = P(B)

If A and B are dependent, as in example 2 above, then P(B/A) = $\frac{P(A \cap B)}{P(A)}$

so that probability of Liz picking a picture card on the second draw card given that she has already picked one picture card is given by P(B/A) = $\frac{P(A \cap B)}{P(A)} = \frac{11/221}{3/13} = \frac{11}{51}$

The multiplication law for dependent probabilities may be rearranged to give P(*A* and *B*) = P(*A* ∩ *B*) = P(*A*) × P(*B*|*A*)

Example: A survey in a particular town shows that 35% of the houses are detached, 45% are semi-detached and 20% are terraced. 30% of the detached and semi-detached properties are rented, whilst 45% of the terraced houses are rented. A property is chosen at random.

(i) Find the probability that the property is rented

(ii) Given that the property is rented, calculate the probability that it is a terraced house.

Answer

Let A be the event (the property is rented)
Let B be the event (the property is terraced)

(i) P(rented) = (0.35 X 0.3) + (0.45 X 0.3) + (0.2 X 0.45) = 0.33

The probability that a house is detached and rented	*The probability that a house is semi-detached and rented*	*The probability that a house is terraced and rented*

(ii) P(A) = 0.33 from part (i)

P(B/A) = $\frac{P(A \cap B)}{P(A)} = \frac{(0.2 \times 0.45)}{(0.33)}$ = 0.27 (2 decimal places)