STATISTICS 2 Summary Notes

1. *Discrete Random Variables*

- discrete if a list could be made of all of the possible values the variable could take

- **Probability Distribution** – a list or tables showing the probability of each value occurring
  - tree diagram may be needed to help you calculate the probabilities – remember to multiply along the branches

  The sum of the probabilities = 1

- **Probability Function** (sometimes easier than making a list)

  e.g. X is the result when a fair tetrahedral die is rolled $P(X=x)$ $\begin{cases} \frac{1}{4} & x = 1,2,3,4 \\ 0 & \text{otherwise} \end{cases}$

- **Cumulative Distribution Function** – shows $P(X \le x)$ for all x

x	<1	1-	2-	3-	≥ 4
$P(X \le x)$	0	¼	½	¾	1

MEAN = $\mu = E(X) = \sum x_i p_i$ (Expectation of X)

VARIANCE = $\sigma^2 = E(X^2) - (E(X))^2$

$\sum x_i^2 p_i$ - mean^2

Standard Deviation = $\sqrt{\text{variance}}$

x	0	1	2
P(X=x)	0.5	0.2	0.3

E(X) = 0×0.5 + 1×0.2 + 2×0.3
= 0.8

Var(X) = 0^2×0.5 + 1^2×0.2 + 2^2×0.3 – 0.8^2
= 0.76

Expectation of a function of a r.v.

$E(g(X)) = \sum g(x_i)p_i$

$E(4x^3) = \sum 4x^3 p$ $\quad E\left(\frac{1}{x}\right) = \sum \frac{1}{x} p$

$E(4X^3) = 4\times0^3\times0.5 + 4\times1^3\times0.2 + 4\times2^3\times0.3$
$= 10.4$

- **Mean and Variance of functions of a r.v**

E(aX) = aE(X) **E(X+b) = E(X) + b** **E(aX+b) = aE(X) + b**

Var(aX) = a^2Var(X) **Var(X+b) = Var(X)** **VAR(aX+b) = a^2Var(X)**

2. *The Poisson Distribution*

- number of events occurring in a fixed interval of time or space

Conditions
- time of each event (or position) is **independent** of previous events
- probability of each event occurring in a given interval of time(space) is fixed
- two evens **cannot** occur at exactly the same time (or position)

$$P(X=x) = \begin{cases} e^{-\lambda}\dfrac{\lambda^x}{x!} & x = 0,1,2\ldots\ldots \\ 0 & \text{otherwise} \end{cases}$$

λ is the rate at which events occur (on average) in the required interval of time or space

Example
The rate at which calls are received by a call centre is 2 calls per minute
Work out the probability that exactly 6 calls are received in 4 minutes.

Average 2 calls per minute $\Rightarrow$ Average 8 calls per 4 minutes

$$P(X=6) = e^{-8}\frac{8^6}{6!}$$

USING TABLES – tables give the $P(X \leq x)$ – remember to use the correct λ

USING THE RECURRENCE FORMULA – if you need to calculate a succession of values of x : P(X=1) P(X=2) P(X=3)…..

$$P(X = x_n) = \frac{\lambda}{n} \times P(X = x_{n-1})$$

eg If we know P(X=1) then

$$P(X=2) = \frac{\lambda}{2}P(X = 1) \qquad P(X=3) = \frac{\lambda}{3}P(X = 2)$$

Sum of INDEPENDENT random variables
Make sure both variables averages for the same unit of time(or space) before adding.

Example – At a checkpoint on average 300 cars pass per hour and the mean time between lorries is 5 minutes. Find the probability that exactly 6 vehicles pass the point in a 1 minute period

Cars 300 per hour $\Rightarrow$ 5 cars per minute
Lorries 12 per hour $\Rightarrow$ 0.2 lorries per minute

Vehicles = 5.2 per minute $P(x=6) = e^{-5\cdot2}\dfrac{5\cdot2^6}{6!} = 0.1515$

Binomial – questions on the Poisson distribution can include use of the binomial theorem – look for Probability when multiples of the time interval are needed

Example –
*What is the probability that **exactly 6** cars pass the checkpoint in at least 3 or the next 4 minutes ?*

Probability of success = 0.1515 n = 4
$P(X \geq 3) = P(X=3) + P(X=4)$

$P(X=3) + P(X=4) = {}_4C_3(0{\cdot}1515)^3(1 - 0{\cdot}1515) + {}_4C_4(0{\cdot}1515)^4$

$= 0.0123$

Easy Marks : Mean = Variance = λ
- occasionally a question will ask you to give a reason why the variable may follow a Poisson Distribution – simple answer the mean is approximately equal to the variance – make sure you use an unbiased estimate of the population variance $(\sigma_{n-1})^2$ for your comparison

3. *Continuous Random Variables*

- the variable can take an infinite number of possible values
- $P(X=0) = 0$ and $P(X < t) = P(X \leq t)$

- **Probability Density Functions f(x)**
  - area under the graph represents the probability

  - **sketching** the graph is always a good idea – evaluate at the interval bounds to plot the key points - check for **quadratic** or **linear** to get the general shape of each section

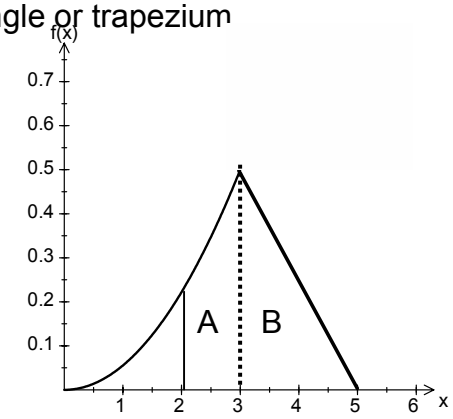
  - **TOTAL area under the curve in the required interval = 1**
  - If linear graph then you can use formulae for the area of triangle or trapezium
  - If quadratic – **integrate** to calculate probabilities

Example A

$$f(x) = \begin{cases} \frac{1}{18}t^2 & 0 \leq x \leq 3 \\ \frac{1}{4}(5-t) & 3 \leq x \leq 5 \\ 0 & \text{otherwise} \end{cases}$$

Find P(2 < x < 5)

Sketching the graph



AREA A $= \int_2^3 \frac{1}{18}x^2 dx = \frac{19}{54}$

AREA B = ½ × 2 × 0.5

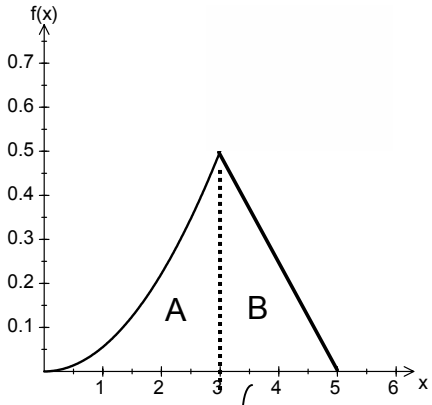
P(2 < X < 5) $= \frac{19}{54} + \frac{1}{2} = \frac{23}{27}$

If asked to find an unknown within the function (often denoted k) – sketch the graph – set the total area = 1 to and solve

- **(Cumulative) Distribution Function F(x)**
  - gives the probability that the value is less than x - P(X<x) or P(X≤x)
  - integral of f(x)
  - useful when finding medians F(x) = 0.5, Lower Quartile F(x) = 0.25 etc…..

Example :Find F(x) for the probability density function defined in example A

Consider each section of the graph !



SECTION A – quadratic
If 0< c< 3 then

$$P(X<c) = \int_0^c \frac{1}{18}x^2 dx = \frac{c^3}{54}$$ using this P(X <3) = ½

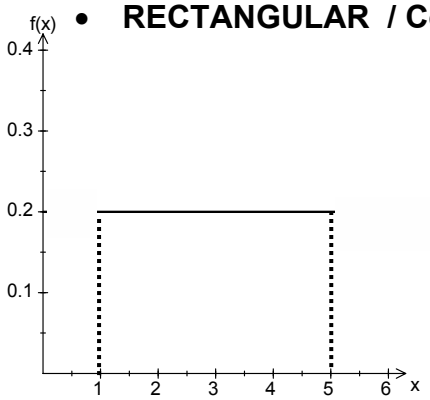
SECTION B – linear
If 3 < c < 5 then

$$P(X<c) = ½ + \int_3^5 \frac{1}{4}(5-x)dx = \frac{1}{8}(10c - c^2 - 17)$$

$$F(x) = \begin{cases} \frac{c^3}{54} & 0 \le x \le 3 \\ \frac{1}{8}(10c - c^2 - 17) & 3 \le x \le 5 \\ 1 & x \ge 5 \end{cases}$$

Don't forget this part!!!

- **RECTANGULAR / Continuous UNIFORM distribution**



$$f(x) = \begin{cases} \frac{1}{b-a} & a < x < b \\ 0 & \text{otherwise} \end{cases}$$

Probability found by working out the area of a rectangle

$$F(x) = \begin{cases} 0 & x \le a \\ \frac{x-a}{b-a} & a \le x \le b \\ 1 & x \ge b \end{cases}$$

E(X) = mean = ½ (a+b)

Var(X) = σ^2 $= \frac{1}{12}(b-a)^2$

If you are given the mean and the variance

solve simultaneously to find the values of a and b

4. *Estimation*

- Useful formulae to learn

$\bar{x}$	$(\sigma_n)^2$	$(\sigma_{n-1})^2$
To calculate the **mean**	the **sample variance**	the **Unbiased Estimate of the population variance**
$\bar{x} = \frac{\sum x}{n}$	Sample Var = $\frac{\sum (x-\bar{x})^2}{n}$	$\frac{n}{n-1} \times \text{SampleVariance}$
$\bar{x} = \frac{\sum fx}{\sum f}$	Sample Var = $\frac{\sum x^2}{n} - \text{mean}^2$	
	Sample Var = $\frac{\sum fx^2}{\sum f} - \text{mean}^2$	

- **CONFIDENCE INTERVALS**

- **Interpretation of a 95% CI** – different samples of size n lead to different values of $\bar{x}$ and hence to different 95% confidence Intervals. On average 95% of these intervals will contain the true population mean**.**

-Check the degree of accuracy required e.g. 3 d.p., 3sf ……

- Write confidence intervals as
(Lower limit , Upper limit)

Population Variance Given – Any sample size – Use Z tables (Standard Normal)
- Population Normally distributed

$$\bar{x} \pm Z_{\alpha} \times \sqrt{\frac{\text{popn var}}{n}}$$

If looking for a 95% look up 0.975 in the percentage tables

Population Variance unknown – Large sample size >30 - Use Z tables
- because of the large sample size – can use Z values due to the **Central Limit Theorem** population does not need to be normally distributed.

Example
A firm offers free bottled water to all 135 employees who work the night shift. The amounts they consume on the first night have a mean of 960 ml with a standard deviation of 240 ml.

Calculate a 90% confidence interval for the mean stating any assumptions you have made.

Good idea to list all values

n = 135
mean = 960
sample variance = 57600
population variance = 58029.9
90% CI Z=1.6449

$$960 \pm 1{\cdot}6449 \times \sqrt{\frac{58029{\cdot}9}{135}}$$

(925.9 , 994.1)

Assumptions
- the data can be regarded as a random sample
- the large sample size – Central Limit theorem – no restrictions in the distribution of the population

Population Variance unknown – Small sample size < 30 - Use t tables (d.f = n-1)

$$\bar{x} \pm t_{\alpha} \times \sqrt{\frac{\text{popn var}}{n}}$$

Use t tables and **n-1** degrees of freedom (γ)

Example

20 bottles are selected from a production line, The contents of each is recorded (x ml)

$$\sum x = 1518{\cdot}9 \qquad \sum (x - \bar{x})^2 = 7{\cdot}2895$$

Stating any assumptions you make calculate a 95% confidence interval for the mean

n = 20
mean = 75.945
sample variance = 0.364475
population variance = 0.38366
degrees of freedom = 19
95% CI T= 2.093

$$75{\cdot}945 \pm 2{\cdot}093 \times \sqrt{\frac{0{\cdot}38366}{20}}$$

(75.66, 76.23)

Assumptions – contents are normally distributed – sample selected randomly

5. *Hypothesis Testing*

STEP 1
State the null hypothesis- **always in terms of $H_0 : \mu = a$** – never $\bar{x}$

STEP 2
State the alternative **$H_1 : \mu \neq a$** 2 tail test – divide significance level by 2 before using tables to find the critical value

$H_1 : \mu > a$ 1 tail test – **positive** critical value
$H_1 : \mu < a$ 1 tail test – **negative** critical value

STEP 3 - Test statistic **Z** or **T**

Variance Known or n > 30

$$z = \frac{\bar{x} - \mu}{\sqrt{\frac{\text{popn var}}{n}}}$$

Unknown Variance and n < 30

$$t = \frac{\bar{x} - \mu}{\sqrt{\frac{\text{popn var}}{n}}}$$

- List the variables
- calculate the statistic

STEP 4
- Use tables to find the critical value - n-1 degrees of freedom if using t
  - check for 1 or 2 tail

-Sketch a graph – mark the critical value
– shade the critical/ rejection region
- mark the position of your test statistic

STEP 5
As … > …. There is no sufficient evidence at the …% significance level that the mean differs from **a** – Accept H_0

As … > …. There is sufficient evidence at the …% significance level that the mean differs from **a** – Therefore accept H_1 and conclude that ….

Significance level : - If the value of the test statistic falls in the critical region then the outcome is said to be significant at the ……% level

TYPE 1 ERROR - *The probability of obtaining a value of a test statistic in the critical region even when the null hypothesis is correct* - **Rejecting H_0 and accepting H_1 when H_0 is actually correct**

TYPE 2 ERROR - *The probability is not fixed, since it depends upon the extent to which the value of μ deviates from the value given in H_o. If the μ is close to this value then the probability of Type 2 error is large* **H_0 is accepted even though it is incorrect**

6. Chi-squared Goodness of fit Test

- can be used for testing whether a die is biased or whether variables are independent

- test statistic involves squares – only interested in **upper limit critical values**

- always **state H_0 before starting** – The variables are **independent**

- to calculate expected frequencies $\frac{\text{row total} \times \text{column total}}{\text{TOTAL}}$

- **check totals** for expected frequencies - for calculation errors

Test Statistic

Special case 2x2 table – 1 degree of freedom

$$X^2 = \sum \frac{(|O - E| - 0{\cdot}5)^2}{E}$$

General

$$X^2 = \sum \frac{(O - E)^2}{E}$$

O : Observed Frequency
E : Expected Frequency under H_0

- **EXPECTED FREQUENCY MUST BE GREATER THAN 5** - group appropriately – remember to adjust degrees of freedom –number of groups used – 1

- DO NOT use percentages – always frequencies

CHI-SQUARED TABLES – n-1 degrees of freedom – n is the number of groups used in the calculation

Example : The table shows the fate of the passengers on the titanic grouped according to class. Test at the 1% level if there is a relationship between class and the chance of survival

	Survived	Died	Total
1st Class	200	123	323
2nd Class	119	158	277
3rd Class	181	528	709
Total	500	809	1309

HYPOTHESIS

H_0 : The chance of survival is independent of the class of travel

H_1 : There is an association between the class of travel and the chance of survival.

EXPECTED RESULTS if independent

	Survived	Died
1st Class	323*500/1309=**123.4**	323*809/1309=**199.6**
2nd Class	277*500/1309=**105.8**	277*809/1309=**171.2**
3rd Class	709*500/1309=**270.8**	709*809/1309=**438.2**

All expected frequencies greater than 5 so no need to group
6 groups used so 5 degrees of freedom needed in tables

TEST STATISTIC

$$\mathbf{X^2} = \frac{(200 - 123{\cdot}4)^2}{123{\cdot}4} + \frac{(119 - 105{\cdot}8)^2}{105{\cdot}8} + \ldots\ldots\ldots\ldots + \frac{(709 - 438{\cdot}2)^2}{438{\cdot}2}$$

$= \mathbf{127.8}$

CRITICAL VALUE - $\chi^2 = 15{\cdot}086$ (1% - 5 degrees of freedom)