STATISTICS REVISION 2

Discrete probability distributions (Binomial and Poisson)

1. At the hot drinks counter in a cafeteria both tea and coffee are sold. The number of cups of coffee sold per minute may be assumed to be a Poisson variable with mean 1.5 and the number of cups of tea sold per minute may be assumed to be an independent Poisson variable with mean 0.5.

(i) Calculate the probability that in a given one-minute period exactly one cup of tea and one cup of coffee are sold.

(ii) Calculate the probability that in a given three-minute period fewer than 5 drinks altogether are sold.

(iii) In a given one minute period exactly 3 drinks are sold. Calculate the probability that these are all cups of coffee.

2. 80% of the inhabitants of a certain African village are known to have a particular eye disorder. 12 people are waiting to see the doctor. What is the most probable value of X if X is the number of people who have the eye disorder?

3. If X is a Poisson variable with mean 3.4, find the most probable value of X.

4. Eggs are packed in boxes of 500. On average, 0.8% of the eggs are found to be broken when the eggs are unpacked.

(i) Find the probability that in a box of 500 eggs, exactly 3 will be broken.

(ii) A supermart unpacks 100 boxes of eggs. What is the probability that there will be exactly 4 boxes containing no broken eggs?

Normal Distribution

5. Sugar is sold in bags labelled as containing 1kg and in bags labeled as containing 0.5kg. The mass of sugar in a 1kg bag may be assumed to have a normal distribution with mean 1005g and standard deviation 2g, and the mass of sugar in a 0.5kg bag may be assumed to have a normal distribution with mean 505g and standard deviation 2g. The mass of sugar in any bag may be assumed to be independent of the mass in any other bag. Bags of sugar are selected at random for testing. Find the probability that

(i) a 1kg bag will contain less than 1000g of sugar;

(ii) two 0.5kg bags will together contain less than 1000g of sugar;

(iii) two 0.5kg bags will together contain more than 1kg bag.

The total mass of sugar in ten 1kg bags is denoted by X_g. Find the value of m such that P(X_g > m) =0.75.

6. In a certain subject, the examination marks of the boys are normally distributed with mean 55 and standard deviation 11 and the marks of the girls are also normally distributed with mean 58 and standard deviation 8. If one boy and one girl are chosen at random from the complete list of candidates, calculate the probability that

(i) the girl's mark exceeds the boy's;

(ii) the girl's mark is at least 20 more than the boy's;

(iii) the difference between the two marks is more than 20;

(iv) their average mark is over 70;

(v) the girl's mark is at least twice as high as the boy's.

7. Weights of persons using a certain lift are normally distributed with mean 70kg and standard deviation 10kg. The lift has a maximum permissible load of 300kg.

(i) If 4 persons from this population are in the lift, determine the probability that the maximum load is exceeded.

(ii) If one person from this population is in the lift and he has luggage weighing 3 times his own weight, determine the probability that the maximum load is exceeded.

8. A pair of dice is tossed 100 times and the total observed on each occasion. What is the probability of getting more than 25 sevens? How many tosses would be required in order that the probability of getting at least one seven is 0.9 or more?

9. A certain office has a stock of identical printed forms which are used independently. On any working day at most one of the forms is used, and the probability that one form is used is 1/3. There are 250 working days in the year. Using a suitable approximation, calculate the number of forms that must be in stock at the beginning of the year if there is to be a 95% probability that they will not all be used before the end of the year.

If one form in 100 is unusable through faulty printing, and these faults occur at random, calculate the probability that in a batch of 250 forms there will be not more than one that is unusable.

10. A radioactive disintegration gives counts that follow a Poisson distribution. If the mean number of particles recorded in a one second interval is 69, calculate the probability of

(i) less than 60 particles for a one second interval;

(ii) more than 150 particles in a two second interval;

(iii) more than 700 particles in a ten second interval.

Sampling Theory

11. A random sample of size 15 is taken from a **normal** distribution with mean 60 and standard deviation 4. Find the probability that the mean of the sample is less than 58. Of 100 samples, each of size 15, how many samples would have their means lying between 58 and 62?

12. The weight of a chocolate produced is known to be **normally** distributed with mean 10g and variance 4g. The chocolates are packed at random into boxes consisting of 25 chocolates each. Find the probability that

(i) a chocolate chose at random weights between 9.5g and 10.5g,

(ii) the contents of a box weight between 247g and 253g,

(iii) the average weight of the chocolates in the boxes lies between 9.9g and 10.1g.

13. A large number of random samples of size n is taken from a **normal** distribution with mean 74 and variance 36. Determine the value of n if $P(\overline{X}>72)=0.854$, where $\overline{X}$ is the random variable denoting the sample mean.

Central Limit Theorem

14. If a random sample of size 30 is taken from each of the following distributions, find for each case, the probability that the sample mean exceeds 5.

(i) $X \rightarrow P_o\ (4.5)$ (ii) $X \rightarrow B(9,0.5)$

15. Packets of a particular drink are found to have a mean of 200ml and a standard deviation of 15ml. In a random sample of size 36, find the probability that the total amount of drink exceeds 7 litres. Another type of drink is such that the mean is 200ml and the standard deviation is 20ml. If 50 packets of each type were taken, what is the probability that the difference in the total amounts is more than 100ml?

16. In an experiment, 10 dice are thrown and the number of sixes are recorded. If the 10 dice are thrown 50 times, find the approximate probability that the average number of sixes obtained is less than 2.

* 17. A random variable X has the following probability distribution:

(a) P(X=0)=0.1, P(X=2)=0.3, P(X=5)=0.6

Based on this distribution, X also has expectation value of 3.6 and variance of 3.24.

X_1, X_2, X_3, ... X_n are n independent observations of X.

(i) Find $P(\sum_{i=1}^{n} X_i > 10)$ for n=3.

(ii) Find $P(\sum_{i=1}^{n} X_i > 350)$ for n=100.

(b) 100 independent observations of X are taken and the mean of these observations denoted by $\overline{X}_{100}$ where $\overline{X}_{100} = \frac{1}{100}(\overline{X}_1 + \overline{X}_2 + ... + \overline{X}_{100})$

Another 200 independent observations of X are taken and the mean of these observations is denoted by $\overline{X}_{200}$. Find $P(2\overline{X}_{100} < \overline{X}_{200} + 4.6)$

Estimation Theory

18. A random sample drawn from a large population contains 20 observations $x_1, x_2, ... x_n$ such that $\sum x = 22.8$ and $\sum x^2 = 27.55$. Obtain the unbiased estimates of the population mean and the population variance.

19. A random sample of 10 bulbs is taken and their lifetimes x_i are obtained. The results

are summarized by

$\sum(x_i - 1000) = 1890$; $\sum(x_i - 1000)^2 = 362050.2$

Find the unbiased estimates of the mean and and variance of X.

Hypothesis Testing

20. A machine packs flour into bags. The mass of a filled bag has a normal distribution with mean 1506.5g and variance 0.16. A sample of 11 bags was tested and found to have a mean mass of 1506.8g. Test whether the sample provides significant evidence, at the 5% level, that the machine is providing overweight bags.

21. It is claimed that the masses of components produced at a particular workshop are normally distributed with a mean mass of 6g and a standard deviation of 0.8g. If this claim is accepted, at the 5 % level, on the basis of the mean mass obtained from a random sample of 50 components, between what values must the mean mass of the 50 components in the sample lie?

22. A manufacturer claims that the average life of his electric light bulbs is 2000 hours. A random sample of 64 bulbs is tested and the life, x in hours, recorded. The results o obtained are as follows:

    $\sum(x - 2000) = -208$, $\sum(x - 2000)^2 = 746$

    Is there sufficient evidence, at the 2% level, that the manufacturer is over-estimating the average length of life of his light bulbs?

23. A sample of size 100 is taken from a normal population with unknown mean and known variance of value 36. An investigator wishes to test: $H_0 : \mu = 65$ against $H_0 : \mu > 65$. He decides on the following criteria: reject H_0 if $\overline{X} > 66.5$ and do not reject H_0 if $\overline{X} \leq 66.5$. Find the level of significance, α .

Correlation and Regression

24. The growth of children from early childhood through adolescence generally follows a linear pattern. Data on the heights of female Americans from 4 to 9 years old, were compiled and the least square regression line was obtained as $y = 80 + 6x$, where y is in cm and x is in years.

(a) Interpret the value of the estimated slop, $b = 6$.
(b) Would interpretation of the value of the estimated y-intercept, $a = 80$, make sense here?
(c) Predict the height of a female American at 8 and 25 years old. Give a reason if any of the predicted value is inaccurate.

25. An electric fire was switched on in a cold room and the temperature of the room was noted at 5 minute intervals.

Time, minutes, from switching on fire, x	0	5	10	15	20	25	30	35	40
Temperature, °C,y	0.4	1.5	3.4	5.5	7.7	9.7	11.7	13.5	15.4

You may assume that $\sum x = 180$, $\sum y = 68.8$, $\sum xy = 1960$, $\sum x^2 = 5100$

(a) Plot the data on a scatter diagram.

(b) Calculate the regression line of y on x. Predict the temperature 60 minutes from switching on the fire.

(c) Starting with the equation of the regression of y on x, derive the equation of the regression line of

(i) y on t where y is the temperature in °C and t is time in hours.
(ii) z on x where z is temperature in K and x is time in minutes.
(A temperature in °C is converted to K by adding 273.)

(d) Explain why in (b), the regression line of y on x was calculated rather than the regression line of x on y. If instead of the temperature being measured at 5-minute intervals, the time for the room to reach predetermined temperatures (eg 1, 4, 7, 10, 13°C) had been observed, what would appropriate calculation have been? Explain your answer.

26. When a car is driven under specific conditions of load, tyre pressure and surrounding temperature, the temperature, T °C, generated in the shoulder of the tyre varies with the speed, V kmh^{-1}, according to the linear model T=a+bV, where a and b are constants. Measurements of T were made at eight different values of V with the following results:

v	20	30	40	50	60	70	80	90
T	45	52	64	66	91	86	98	104

[$\sum v = 440$, $\sum v^2 = 28400$, $\sum t = 606$, $\sum t^2 = 49278$, $\sum vt = 37000$]

(i) Calculate the equation of the estimated regression line of T on V.

(ii) Estimate the expected value of T when V=60.

(iii) It is given that for each value of V, the measured value of T contains a random error which is normally distributed, with zero mean and variance 16. Calculate the probability that, when V=60, the measured value of T exceeds 91.

27. The following summary data refer to concentration of carbon dioxide in the atmosphere (y) in parts per million, for the 8 years 1971, 1973, 1975, …… 1985 (x).
$\sum(x-1971) = 56$, $\sum(x-1971)^2 = 560$, $\sum(y-325) = 69$, $\sum(y-325)^2 = 887$,
$\sum(x-1971)(y-325) = 704$

(i) Let u=x-1971 and v=y-325. Calculate the equation of the least squares regression line of v on u. Hence find the equation of the least squares regression line of y on x.

(ii) Calculate the linear (product moment) correlation coefficient for x and y. Comment on what its value implies about the regression line.

(iii) Estimate the concentration of carbon dioxide in the atmosphere in (a) 1974 and (b) 1988.

28. A research worker Dr Lin Guistic gave each of eight children a list of words of varying difficulty and asked them to define the meaning of each word. The table shows age in years and number of correctly defined words for each child.

	A	B	C	D	E	F	G	H
Age(x)	2.5	3.1	4.3	5.0	5.9	7.1	8.1	9.4
No. of correct words (y)	9	13	18	25	35	53	81	132

(i) Comment on the suitability of finding the linear regression line of y on x.

(ii) Calculate the product moment correlation coefficient between lny and lnx. Comment on the suitability of using $\ln y = a + b \ln x$

(iii) Find the least squares regression line of $\ln y$ on $\ln x$. Sketch this line on the scatter diagram of $\ln y$ against $\ln x$.

29. The data shows the result of an experiment to investigate the relationship between stress x and the time taken to rupture, t of a brass material.

x	22.5	25.0	28.0	30.5	38.0	40.5	42.5	48.0	54.5	55.0	70.0
t	44.0	42.0	33.5	28.0	18.0	13.6	15.0	10.3	9.0	6.3	4.0

(i) Obtain the scatter diagram and comment on any relationship between x and t.

(ii) State with a reason which of the following models is more appropriate to fit the data points:

(a) $x = at^b$, where $a > 0$ and $b < 0$

(b) $x = a + bt^2$, where $a > 0$ and $b < 0$