Introduction to Statistics

Worksheets and Worked Examples

Basic Probability Questions Q1a. Two fair dice are thrown. What is the probability of at least one odd number? Q1b. What is the probability of at least one odd number if four fair dice are thrown?

Introduction to Statistics

Example 1 A fair die is thrown. The number shown on the die is the random variable X. Tabulate the possible outcomes. Solution X takes the six possible outcomes 1, 2, 3, 4, 5, 6 which each have probability 1/6 (i.e. one sixth).

k	1	2	3	4	5	6
$\Pr(X = k)$	1/6	1/6	1/6	1/6	1/6	1/6

Introduction to Statistics

Example 2 Two unbiased spinners, one numbered 1, 3, 5, 7 and the other numbered 1, 2, 3 are spun. The random variable X is the sum of the two results. Find the probability distribution for X.

Solution Listing all the possible outcomes is best done in a table.

Introduction to Statistics

Chi Square The table below shows the relationship between gender and party identification in a US state.

DemocratIndependentRepublicanTotal Male27973225577 Female16547191403 Total444120416980

Test for association between gender and party affiliation at two appropriate levels and comment on your results.

Introduction to Statistics

Set out the null hypothesis that there is no association between method of computation and gender against the alternative, that there is. Be careful to get these the correct way round!

H0: There is no association. H1: There is an association.

Work out the expected values. For example, you should work out the expected value for the number of males who use no aids from the following: (95/195) 22 = 10.7.

Structure Structure of Resource

1) Descriptive Statistics 2) Probability Distributions 3) Inference: Confidence Intervals 4) Inference: Hypothesis Testing 5) Simple Linear Regression 6) Chi Square Tests

Part 1 Statistics Gamblers Ruin Monte Carlo Revision of normal distribution statistical process control and Six Sigma histograms density curves and boxplots some remarks on binomial distribution Part 2 Describing,Exploring and Comparing Data Part 3 Revision of Normal Distribution 3.1 Overview 3.2 The standard normal distributions 3.3 Applications of normal distributions 3.4 Sampling distributions and estimators 3.5 The central limit theorem 3.6 Determining normality

Part 4A Estimates and Samples Sizes Part 4B Sampling distributions significance alpha central limit theorem p-values Type I and Type II errors inference procedures Testing normality Part 5 Hypothesis testing and inference procedures 5.1 Overiew 5.2 Basics of Hypothesis Testing 5.3 Testing a claim about a proportion 5.4 Testing a claim about a mean sigma known 5.5 Testing a claim about a mean sigma unknown confidence intervals margin of error standard error

Chapter 6A Hypothesis testing about two samples 6.1 Overview 6.2 Inferences about two proportions 6.3 Inference about two means (Independent Samples) 6.4 Inferences about Matched Pairs (The Paired t-test) Part 6B More Inference Procedures Outliers Grubbs test shapiro wilk test Non parametric tests Wilcoxon test kolmogorov smirnov test chi square goodness of fit table Part 7 Correlation and Simple Linear Regression Slope and Intercept estimates prediction interpolation extrapolation Correlation and Causation Spurious Correlation Definitions Part 7B Multiple Linear Regression Akaike information criterion Adjusted r squared overfitting multicollinearity Part 8 Multinomial Experiments and Contingency Tables 8.1 Overview 8.2 Multinomial Experiments: Goodness of fit 8.3 contingency tables

For this component of the module, students will need to familiarise themselves with the Murdoch Barnes statistical tables.

The necessary sections are provided in handbook.

Section 1 Data and Sampling

1.2 Introduction

The aim of statistics is to provide insight by means of numbers.

Difference between and experiment and an observational study

Experiment: Researcher has ability to control important variables.

Observational Study: Researcher does not have ability to control important variables

Example 1: Health Insurance [pg 4] No it is an observational study. The researcher has no influence on the how they respond.

Example 2: advertising [pg4] Yes this was an experiment. The researcher was able to control how much TV advertising each student watched.

0.1 1.4 Sampling

We often read headlines in newspapers saying things like 80% of the population are satisfied with the governments performance. How can the newspaper make such a statement when they havent asked everyone in the country their opinion of the government?

The newspaper has taken a representative subset of the population and assumed that what happens for that subset is what happens for the whole population. Is this assumption valid?

- How do you select a representative subset? What mistakes can you make selecting this subset and what can be done to correct these mistakes?
- Without understanding the concepts behind selecting a subset of the population i.e. sampling, we can make serious errors in our conclusions about the population.

0.1.1 1.4.1 Definitions

First, we need to define some terms. These terms will be illustrated using the example of a pre-election poll on which political party is going to win the election.

Population: the entire group of objects/subjects about which information is wanted. For our example, the population is all adults on the electoral register.

Sample: any subset of a population e.g. a representative subset of individuals from the electoral register.

Unit: any individual member of the population e.g. an individual on the electoral register.

Sampling frame: a list of the individuals in the population e.g. the electoral register.

Variable: we can measure its value for each person and its value will change from person to person e.g. the political party the individual will vote for.

Parameter: this represents some value (e.g. an average value or a percentage) that we are interested in calculating for the population for example the percentage of adults on the electoral register who will vote for a particular political party or the average age of the voters.

Measures of Dispersion

Examples of Quantiles

Quartiles are just one type of quantile.

- Percentiles
- Deciles

The Binomial Distribution

In the last class, we looked at how to compute the mean, variance and standard deviation.
As these are key outcomes of this part of the course, we shall briefly go over this material again
The mean (i.e. average) value is denoted with a bar over the set name i.e. " ".
(pronounced x bar) is the sample mean.

The Binomial Distribution

Example
A sample data set comprised of five values.
What is the sample mean value of data set " " (i.e. What is ?)
The sample mean is 44

The Binomial Distribution

Variance How do we calculate the variance? We can use scientific calculators or we can calculate it by hand using the following formula :

We are calculating the difference between each observation x and the mean . Remark : The mean is used in the calculation. Some of the differences will be positive and some will be negative so we square the differences to make them all positive.

The Binomial Distribution

- An easier formula to use if you are calculating the sample standard deviation by hand is
- The population variance (which is rarely know) is denoted by the Greek letter (sigma squared).
- Important :The standard deviation σ is the square root of the variance σ^2.

The Binomial Distribution

- The standard deviation for the sample is called s and the standard deviation for the population is called σ.
- The standard deviation is often preferred to the variance as a descriptive measure because it is in the same units as the raw data e.g. if your data is measured in years, the standard deviation will also be in years whereas the variance will be in years squared.

Measures of Centrality and Dispersion

2.2.3 Measures of Centrality and Dispersion for Grouped data [Page 27]
: Frequency of class i : Midpoint of class i
Sample variance for grouped data
Sample variance for grouped data
Median
Lower limit of median class
: Sample Size

: cumulative frequency to class m-1
: frequency of median class
: width of frequency class
MidpointFrequencyCmlt Freq 420 - 439429.588 440 - 459449.51725 460 - 479469.51237 (median in here) 480 - 499489.5845 500 - 519509.5752 520 - 539529.5456 540 - 559549.5258 560 - 579569.5462 580 - 599589.5264 600 - 619609.5670 Total70
The median is in the third class interval
The lower limit of the third class is 460
The width is 20 (not 19)
There is a frequency of 12 in that interval
The sample size n = 70
[Rest on Overhead]
Mean 493.21 Variance 3017.888 Std.Dev 54.935
[Page 29]
Week 4 (from Lecture 4A)

- The first midterm is to take place Monday of Week 5 at 4pm.
- The first midterm will cover:
  - Basic Probability
  - Descriptive statistics (mean, median variance etc)
  - Discrete probability distributions (binomial and Poisson)
  - The exponential distribution
  - Some of the normal distribution will be included.

Overview of Current Part of Course Probability Distributions (Question 2 for End Of Year Exam)

- Discrete Probability Distributions
  - Binomial Probability Distribution (Week 3)
  - Geometric Probability Distribution (Week 3)
  - Poisson Probability Distribution (Week 3/4)
- Continuous Probability Distributions
  - Exponential Probability Distribution (Week 4)
  - Uniform Probability Distribution (Week 4)
  - Normal Probability Distribution (Week 4/5)

0.2 Continuous Distributions

- (The Continuous Uniform Distribution, Not examinable)
- The Exponential Distribution (Examinable for midterm)

- The Normal Distribution
- The Standard Normal (Z) Distribution.
- Applications of Normal Distribution

Confidence Interval for a Mean (Small Sample)

- The mean operating life for a random sample of $n = 10$ light bulbs is $\bar{x} = 4,000$ hours, with the sample standard deviation $s = 200$ hours.
- The operating life of bulbs in general is assumed to be approximately normally distributed.
- We estimate the mean operating life for the population of bulbs from which this sample was taken, using a 95 percent confidence interval as follows:

$$4,000 \pm (2.262)(63.3) = (3857, 4143)$$

Confidence Interval for a Mean (Small Sample)

- The point estimate is 4,000 hours. The sample standard deviation is 200 hours, and the sample size is 10. Hence

$$S.E(\bar{x}) = \frac{200}{\sqrt{10}} = 63.3$$

- From last slide, the t quantile with $df = 9$ is 2.262.
- The normal distribution
- The Central Limit Theorem
- Sampling Distributions
- Standard error
- $p-$ values
- Testing for Normality
- **Standard Error** Hypothesis Testing

$$\sqrt{\frac{\pi(1-\pi}{n}}$$

- **Standard Error** Confidence Intervals

$$\sqrt{\frac{\bar{p}(1-\bar{p}}{n}}$$

0.2.1 Move to Scratchpad

- In a group of 24 adults, only 6 people watch televised pay to view sports events.
- A visitor successively asks two people chosen at random from the community whether they watch these TV programmes.
- What distribution should be used in this instance?
- What is the probability that the second person chosen does watch them?

0.3 The Variance

- Suppose a continuous uniform distribution has following parameters
  - minimum $a = 1$
  - maximum $b = 9$
- What is the probability of X being between 5 and 8?
  - The lower bound L is 5
  - The upper bound U is 8
- What is $P(5 \leq X \leq 8)$?

$$P(5 \leq X \leq 8) = \frac{\mathbf{8-5}}{\mathbf{9-1}} = \frac{\mathbf{3}}{\mathbf{8}} = \mathbf{0.375}$$

Question 5 - Sampling

A lot contains 13 items of which 4 are defective. Three items are drawn at random from the lot one after the other. Find the probability p that all three are non-defective.

0.3.1 Pooled Standard Deviations

Sample sizes and degrees of freedom

Suppose one has two independent samples,

x1, ..., xm and y1, ..., yn, and wishes to test the hypothesis that the mean of the x population is equal to the mean of the y population:

- H_0 : x = y.
-

Let

X and Y denote the sample means of the xs and ys and let sx and sy denote the respective standard deviations.

The standard test of this hypothesis H0 is based on the t statistic

$$TS = \frac{X - Y}{Sp(1/m) + (1/n)}$$

S_p is the pooled standard deviation

Sp=(m -1)sx2+ (n -1)sy2m+n-2

Under the hypothesis

H_0, the test statistic T has a t distribution with $m + n2$ degrees of freedom when both the xs and ys are independent random samples from normal distributions the standard deviations of the x and y populations, σ_x and σ_y, are equal

Suppose the level of significance of the test is set at α. Then one will reject H_o

when$|T|t_{(n+m2,\alpha/2)}$,where tdf, is the (1) quantile of a t random variable with df degrees of freedom.

0.3.2 Other Stuff

It is a one tailed test

$H_o : \mu = 80$ $H_a : \mu \neq 80$

The significance level is 5

what is the column to use?

what is the degrees of freedom Is it a large sample or a small sample?

$$\sqrt{3}1.09 \times 1.08 \times 1.07]$$

$$\sqrt{\frac{\hat{p}1 - \hat{p}}{n}}$$

$$\sqrt{\frac{\hat{p_1}1 - \hat{p_1}}{n_1} + \frac{\hat{p_2}1 - \hat{p_2}}{n_2}}$$

Binomial Distribution

There are n independent trials The probability of a success is

0.3.3 Grouped Data

The table below gives the number of thunderstorms reported in a particular summer month by 100 meteorological stations.

Number of thunderstorms:	0	1	2	3	4	5
Number of stations:	22	37	20	13	6	2

(a) Calculate the sample mean number of thunderstorms.

(b) Calculate the sample median number of thunderstorms.

(c) Comment briefly on the comparison of the mean and the median.

In a certain large population 45% of people have blood group A. A random sample of 300 individuals is chosen from this population. Calculate an approximate value for the probability that more than 115 of the sample have blood group A.

If X is the number in the sample with group A, then X has a binomial (300, 0.45) distribution, so

$$E[X] = 300 \times 0.45 = 135$$

and

$$Var[X] = 300 \times 0.45 \times 0.55 = 74.25$$

. Then, using the continuity correction,

$$P(X > 115) = P(X > 115.5)$$

$$1 - \frac{115.5 - 135}{\sqrt{74.25}}$$

$$P(X > 115.5) = 1 - \theta(-2.26) = \theta(2.26) = 0.99$$

The mean of a sample of 30 claims is \$5,200. Six have mean of \$ 8000 (i.e. group 1) Ten have mean of \$ 3100 (i.e. group 2)

Compute the mean for the remaining claims

$$\text{Total Costs} = (Cost for Group 1) + (Cost for Group 2) + (Cost from Group 3)$$

- Total Cost for all three groups : $\$5200 \times 30 = \$156000$
- Cost for Group 1 : $\$8000 \times 6 = \$48000$
- Cost for Group 2 : $\$3100 \times 10 = \$31000$

Necessarily the cost for group 3 is \$77000
The mean claim for group 3 is therefore

$$\frac{\$77000}{14} = \$5500$$

Poisson/Binomial/Exponential

- Poisson

  Find P(X=0) for Poisson Mean (m=0.5)

$$P(X = 0) = \frac{e^{-0.5}}{0!} = 0.606$$

- Binomial
- Exponential

  No Claim in the next two years

$$= (0.606)^2 = 0.368$$

- Time Until Next Claim

  $\mu = 0.5$

  $T \approx exp(0.5)$

- $P(XT > 2) = exp(-1) = 0.368$

0.3.4 Dice Roll Example

Suppose someone asks you to play the following game. Roll a die. If the die shows a 1, 2, 3, or 4, you lose \$10. If the die shows a 5, you win \$15. If the die shows a 6, you win \$30. What is your expected value?

The mean and standard deviation of the following

We are told the following piece of information $\bar{x} = 44$ So what is the coefficient of determination?

0.4 Poisson Approximation

n = 25 p = 0.1, 0.2

Poisson approximation of Binomial (letting $m = np$

- $m_1 = 2.5$
- $m_2 = 5$

Find $P(X \geq 5)$

$P(X \leq 4) = 1 - P(X \geq 5)$

From Tables 0.89118 0.44049

(Rest : Compare to Real Answers)

A random sample of 10 is taken from a normal distribution of $\mu = 20$ and $\sigma^2 = 1$. Let s^2 be the sample variance.

Find $P(S^2 > 1)$

Solution

$$\frac{(n-1)S^2}{\sigma^2} \sim \chi^2_{n-1}$$

$$9S^2 \sim \chi^2_9$$

$$P(S^2 > 1) = P(\chi^2_9 > 9) = 1.05627 = 0.437$$

PMS Spring 2008

$$f(x,y) = \frac{4}{3}(1 - xy)\ 0 < x < 1, 0 < y < 1$$

The Marginal PDF of X and Y is given by

$$f(x) = \frac{2}{3}(2 - x)\ 0 < x < 1$$

$$f(x,y) = \frac{2}{3}(2-y)\ 0 < y < 1$$

Show that the conditional expectation of Y given X is given by

$$f(y|x) = \frac{2(1-xy}{2-7}\ 0 < y < 1$$

Solution

$$f(y|x) = \frac{f(x,y)}{f(x)} = \frac{\frac{4}{3}(1-xy)}{\frac{2}{3}(2-x)} = \frac{2(1-xy)}{2-x}$$

0.4.1 Question 1 : Probability Distribution

Introduction

Consider playing a game in which you are winning when a ***fair die*** is showing 'six' and losing otherwise.

0.4.2 Part 1

If you play three such games in a row, find the probability mass function (pmf) of the number X of times you have won.

- Firstly: what type of probability distribution is this?
- Is this the distribution ***discrete*** or ***continuous***?
- The outcomes are whole numbers - so the answer is discrete.
- So which type of discrete distribution? (We have two to choose from. See first page of formulae)
- **Binomial:** characterizing the number of ***successes*** in a series of n ***independent trials***, with the ***probability of a success*** in each trial being p.
- **Poisson:** characterizing the ***number of occurrences*** in a ***unit space*** (i.e. a unit length, unit area or unit volume, or a unit period in time), where λ is the the number of occurrences per unit space.

Example

In the above example where the die is thrown repeatedly, lets work out $P(X \leq t)$ for some values of t.
P(X $\leq$ 1) is the probability that the number of throws until we get a 6 is less than or equal to 1. So it is either 0 or 1.

- P(X = 0) = 0
- $P(X = 1) = 1/6$.
- Hence $P(X \leq 1) = 1/6$

Similarly, $P(X \leq 2) = P(X = 0) + P(X = 1) + P(X = 2)$
= 0 + 1/6 + 5/36 = 11/36

Question 10.

a. (1 Mark) The observation of the air pressure at a volume of 5 cubic metres was 19.87 bars. Calculate the residual from the regression model corresponding to this observation.

b. (3 marks) Using the table above to justify your conclusion, test the null hypothesis that there is no monotonic (systematic) relationship between volume and pressure. State the null and alternative hypotheses clearly.

c. (2 marks) Briefly explain why the use of linear regression to describe pressure as a function of volume is inappropriate.

0.4.3 Standardisation Formula

$$Z = \frac{X - \mu}{\sigma} \tag{1}$$

0.4.4 Pooled Standard Deviations

Sample sizes and degrees of freedom

Suppose one has two independent samples,

x1, ..., xm and y1, ..., yn, and wishes to test the hypothesis that the mean of the x population is equal to the mean of the y population:

- H_0 : x = y.
-

Let

X and Y denote the sample means of the xs and ys and let sx and sy denote the respective standard deviations.

The standard test of this hypothesis H0 is based on the t statistic

$$TS = \frac{X - Y}{Sp(1/m) + (1/n)}$$

S_p is the pooled standard deviation

Sp=(m -1)sx2+ (n -1)sy2m+n-2

Under the hypothesis

H_0, the test statistic T has a t distribution with $m + n2$ degrees of freedom when both the xs and ys are independent random samples from normal distributions the standard deviations of the x and y populations, σ_x and σ_y, are equal

Suppose the level of significance of the test is set at α. Then one will reject H_o

when$|T|t_{(n+m2,\alpha/2)}$,where tdf, is the (1) quantile of a t random variable with df degrees of freedom.

Type 1	Type 2
$n_1 = 100$	$n_2 = 100$
$\bar{x}_1 = 25$ hours	$\bar{x}_1 = 23$ hours
$s_1 = 4$ hours	$s_1 = 3$ hours

$$S.E.(\bar{x}_1 - \bar{x}_2) = \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}$$

$$S.E.(\bar{x}_1 - \bar{x}_2) = \sqrt{\frac{4^2}{100} + \frac{3^2}{100}}$$

$$S.E.(\bar{x}_1 - \bar{x}_2) = \sqrt{25/100}$$

Two Sample Tests *Test Statistic*

$$TS = \frac{(25 - 23) - 0}{?} = 2.?$$

Critical Value
The crtical value is 1.645.

- Large aggregate sample
- One-tailed procedure
- Significance level $\alpha = 0.05$

Chapter 1

Probability Distributions

Review Question 5 : Choose Operator Evaluat the following

- ${}^{10}C_0$
- ${}^{10}C_1$
- ${}^{6}C_3$

Solutions

- ${}^{10}C_0 = 10!/(10! \times 0!) = 1$
- ${}^{10}C_1 = 10!/(9! \times 1!) == 1$
- ${}^{6}C_3$

Review Question 5 : Probability If A and B are events such that $P(A|B^c) = 2P(A|B)$ and $P(B^c) = 2P(B)$. (The event B^c is the complement of event B.) Compute the probability: $P(B^c|A)$

- Complement events: $P(B^c) = 1 - P(B)$. Also we are told $P(B^c) = 1 - P(B) = 2P(B)$
- Necessarily $1 = 3P(B)$. Therefore $P(B) = 1/3$ and $P(B^c) = 2/3$
- Total Probability: $P(A) = P(A \text{ and } B) + P(A \text{ and } B^c)$
-

2009/GD5/Q8

8. Monthly magazines are often directed at special interests of their readers, such as the countryside, wildlife, sport, motor cars. The marketing manager for one of these magazines is trying to encourage advertisers to place advertisements in her magazine. She therefore wishes to collect information on the number and types of people who read the magazine, regularly or occasionally, and their interests.

(i) Paying particular attention to the problems of constructing a sampling frame, undercoverage, non-response and likely requirements of advertisers, recommend a method for selecting a sample from the readership. (10)

(ii) Comment on the types of information the manager should aim to collect, in addition to standard questions about name, address, age, occupation, etc. (5)

(iii) How could the frame be used, and how could the results of the exercise be kept up to date, to satisfy the likely needs of the advertisers? (5)

Q1

A multinational corporation (MNC) has two manufacturing facilities one in Limerick, the other in Texas, both producing identical products and employing roughly the same number of direct employees, i.e. directly involved in manufacturing. The Operations Director randomly selected 9 weeks from the previous years production figures. Each of the weeks chosen, employees worked 5 full days and no overtime. The production figures for each facility is as follows (units in thousands)

IrelandTexas 13.83.7 23.93.8 34.03.6 43.73.3 53.63.4 64.13.5 73.93.6 84.03.7 94.13.4

(a)You are required to construct a box plot for each factorys output and to comment on the salient features of each plot. Is there evidence from the box plots to reject the hypothesis that there is no difference in the production figures between the two facilities? Use the box plot to justify your comments. (8 marks)

(b)Compare and contrast the Mean and the Median as measures of location (2 marks)

(c)Two independent resistors are located in a simple circuit as shown below. The probability that A and B works is 0.8 and 0.7 respectively. You are required to calculate the probability that the signal will travel from x to y. If the two resistors were arranged in series, what difference would this make to the probability of the signal reaching y? (10 marks)

(c)A random sample of 100 parts (from (a) above) revealed that for 4 of those parts, the critical diameter measurements were outside spec. You are requested to construct a 95

If the company wished to estimate the defective rate to within + or - 2(8 marks)

Q3 The data from Question 1 (reproduced below) should be used to answer the following question.

IrelandTexas 13.83.7 23.93.8 34.03.6 43.73.3 53.63.4 64.13.5 73.93.6 84.03.7 94.13.4

IrelandTexas Sample Mean3.903.56 Standard Deviation0.173x

(a)Fill in the missing section from the above table(2 marks) (b)Does the data provide sufficient evidence to support the hypothesis that the populations are of equal variances? Use a 5(c)You are required to test the hypothesis that there is no statistical difference between the number of units produced at each plant. Use a test with a 5

.

Q1

A telecommunications company has two servers from different suppliers. The Operations Director randomly selected 9 months from the previous years figures. The down time figures for each server is as follows (down time per month in hours)

Server 1Server 2 15.35.4 25.45.5 35.56.1 45.25.0 55.15.1 65.65.2 75.45.3 85.55.4 95.65.9

(a)You are required to construct a box plot for each servers output and to comment on the salient features of each plot. Is there evidence from the box plots to reject the hypothesis that there is no difference in the down time between the two servers? Use the box plot to justify your comments. (8 marks)

(b)Compare and contrast the Mean and the Median as measures of location (2 marks)

(c)Two independent resistors are located in a simple circuit as shown below. You are required to calculate the probability that the signal will travel from x to y. Resistor A has a mean time of 20 hours to failure and Resistor B a mean time of 40 hours. Assume that the respective failure rates follow an exponential distribution. What is the probability that signal integrity will be maintained between x and y for 30 or more hours? (10 marks)

Q3 Two facilities, one in Limerick and the other in Galway produce identical products on similar lines. The production manager randomly selected 9 weeks from the previous years production. The number of units

produced in each facility was recorded in thousands as follows:

LimerickGalway 13.83.7 23.93.8 34.03.6 43.73.3 53.63.4 64.13.5 73.93.6 84.03.7 94.13.4

LimerickGalway Sample Mean3.903.56 Standard Deviation0.173x

(a)Fill in the missing section from the above table(2 marks) (b)Does the data provide sufficient evidence to support the hypothesis that the populations are of equal variances? Use a 5(c)You are required to test the hypothesis that there is no statistical difference between the number of units produced at each plant. Use a test with a 5

MA4704 Technological Maths 4

Q1 A multinational company has two sub suppliers, one in Taiwan and the other in mainland China. Both plants have similar production facilities and are dedicated exclusively to supplying a company based in Ireland with components. The number of defects per shipment is recorded as follows:

SubS 116181922222528282831332723 SubS 22223252727283032333536383 8

(a)You are required to construct a box plot for each plants output and to comment on the salient features of each plot. Is there evidence from the box plots to reject the hypothesis that there is no difference in the defect figures between the two facilities? Use the box plot to justify your comments. (8 marks)

(b)Compare and contrast the Mean and the Median as measures of location (2 marks)

(c)The mean of three numbers: w, y, z is x

(i) Express x in terms of w, y, z (ii) If a and b are constants express the means of aw + b; ay + b; az + b in terms of (10 marks)

1.1 Question 5

The National Roads Authority is studying the relationship between the number of bidders on a Motorway project and the winning (lowest) bid for the project. Of particular interest is whether the number of bidders increases or decreases with the amount of the winning bid.

Observation number Number of Bidders (x)Winning Bids (y) 154.0 286.9 343.9 497.8 532.7 666.1 744.4 876.2

(a) You are required (i)To draw a Scatter Gram and comment on its features (ii) Find the regression equation and plot the regression equation on the scattergram (iii) Use a 5(12 marks)

(b) The above data was entered into Minitab and the following output was generated: You are requested to fill in the blanks in the following table: Analysis of Variance

Source DF SS MS F P Regression x 20.160 20.160 x 0.000 Residual Error x xx xxx Total x 21.460

Observation Number of BiddersWinning BidFitted valueResidual 28 6.9 x x

(8 marks)

Question 3 (a)Suppose 40% of employees in a large company favour unionisation. A poll of 10 employees in this company is taken.

(i)What is the probability that 4 or more employees polled favour unionisation? (ii)What is the probability that less than 2 employees polled favour unionisation? (iii)What is the probability that exactly 5 employees polled favour unionisation? (iv)What is the mean and variance for this distribution? (9 marks)

(c)Telephone calls coming in to a busy switchboard follow a Poisson distribution with 3 calls expected in a one minute period. The switchboard operator can answer at most 3 calls in a one minute period; the fourth and succeeding calls receive a busy signal.

(i) Find the probability of receiving a busy signal. (ii) The switchboard operator leaves the switchboard unattended for 2 minutes. What is the probability that exactly 1 call will be missed during that 2 minute period? (6 marks)

(d)In what circumstances can the Poisson distribution be used to approximate the Binomial distribution? (2 marks)

Q4

(a) Four per cent of PCB boards purchased over the Internet from the Far East have some defect. From a large consignment of boards, 50 are chosen at random. What is the probability that: (i)3 or more boards have some defect? (ii)Exactly 2 boards are defective? (iii)Less than 3 boards are defective? (iv)If more than 5 boards from the 50 were defective what action would you take? (justify)

(b) Flaws occur in a hard wood timbers at the rate of 1.5 per linear metre section. Calculate the probability that: (i)3 or more flaws will occur in a 3 metre length (ii)Exactly 4 flaws will occur in a 10 metre length (iii)8 or less flaws will occur in a 6 metre length

(c)There is a constant probability of 0.05 that the power supply in telecoms network will not start. You are requested to calculate the probability that the power supply will fail the 5th time it is activated.

1.2 Random Samples

- Consider two random samples drawn from X and Y respectively.
- When these observations are plotted on a scatterplot, it may be the case that some sort of relationship **appears** to exist (when in fact it doesn't).
- The smaller the number of observations, the more likely this erroneous conclusion will occur.

1.2.1 Introduction to Statistics

2.2.3 Grouped Data Median [Pg 27]
Lower limit of median class
: Sample Size
: Cumulative frequency to class m-1
: frequency of median class
: width of frequency class

1.2.2 Introduction to Statistics

```
MidpointFrequencyCmlt Freq
420 - 439429.58 4398
440 - 459449.517 45925
66469.512 47937
480 - 499489.58 49945
500 - 519509.57 51952
520 - 539529.54 53956
540 - 559549.52 55958
560 - 579569.54 57962
580 - 599589.52 59964
600 - 619609.56 61970
Total70
```

1.2.3 Introduction to Statistics

Remark : The 26th to 37th items are in the third class. (m = 3)
Lower limit of median class460
: Sample Size70
: Cumulative frequency to class m-125
: frequency of median class 12
: width of frequency class20
The sample size n = 70
Therefore
Median is estimated as