Seat No	
Student Name	
Student ID	
Signature	

EXAM COVER SHEET

NOTE: DO NOT REMOVE this exam paper from the exam venue

EXAM DETAILS

Course Code: MATH1324

Course Description: Introduction to Statistics

Date of exam: **Start time of exam:** 5:45 PM **Duration of exam:** 2hr 15min

Total number of pages (incl. this cover sheet) ....

ALLOWABLE MATERIALS AND INSTRUCTIONS TO CANDIDATES

1. Write your full name and student number on each exam booklet together with the number of exam books used.
2. Students must not write, mark in any way any exam materials, read any other text other than the exam paper or do any calculations during reading time.
3. All mobile phones must be switched off and placed under your desk.
4. This is an **OPEN BOOK** Exam.
   Candidates are permitted to bring two double-sided hand written (or typed) A4 notes, a text book and all other lecture materials typed (or hand written).
5. Commence each question on a new page. Carry out the instructions on the front cover of the exam script book and the front of this exam paper.
6. Non text storing calculators are allowed.
7. Electronic dictionaries are allowed.
8. The exam has 50 multiple choice questions worth one mark each and 10 short answer questions worth five marks each (100 marks total).
9. Answer the multiple choice questions using the multiple choice answer sheet provided. Ensure you fill out your student details clearly. Use a pencil just in case you need to change your answers.
10. Write your answers to the short answer questions in the script book provided. Start a new page for each question.

MATH1324 Exam

Sample Exam (with Solutions)

The real exam will be open-book, 50 multiple-choice questions (1 mark each) and 10 short answer questions (5 marks each). Questions are conceptual. No computation or memorisation of R code needed. The exam will cover Modules 1 - 9. Duration of real exam will be 2 hours 15 minutes.

Multiple Choice (15 Questions - 1 Mark each)

Answer the following multiple choice questions using the MCQ answer sheet provided. Choose the one, best response from the alternatives provided. Ensure you fill out your name and student number clearly on the answer sheet.

1. Which of the following is an example of a qualitative variable?
   (a) Hair colour
   (b) Height
   (c) Distance travelled to university
   (d) Exam score

---

2. An ordinal scale is an example of what type of variable?
   (a) Qualitative
   (b) Quantitative
   (c) Ratio
   (d) Either A or B

---

3. The median is best defined by which one of the following?
   (a) The sum of all the scores in a quantitative dataset divided by the number of data points.
   (b) The middle value that splits an ordered quantitative dataset in half.
   (c) The most frequently occurring observation in a dataset.
   (d) The average deviation from the mean.

---

4. The middle 50% of a quantitative variable's distribution is known as which one of the following?
   (a) Median
   (b) Range
   (c) Standard deviation
   (d) Inter-quartile range

---

5. If two events cannot occur together, they are said to be?
   (a) Dependent
   (b) Independent
   (c) Conditional
   (d) Mutually exclusive

---

6. Suppose the average number of hospital admissions due to childhood asthma in Victoria each year is 946. Which probability distribution is best suited to model the annual expected number of childhood asthma admissions in Victoria?
   (a) Normal distribution
   (b) Binomial distribution
   (c) Poisson Distribution
   (d) F distribution

7. As sample size increases, the width of a statistic's 95% confidence interval:
   (a) increases
   (b) decreases
   (c) stays the same
   (d) varies randomly

8. A 95% CI is an example of which one of the following?
   (a) A measure of variation.
   (b) A point estimate.
   (c) A measure of central tendency.
   (d) An interval estimate.

9. An investigator wants to determine if dice produced by a particular manufacturer are fair. They purchase a large quantity of dice and perform many thousands of experiments of rolling the dice and recording the number that comes up. Which statistical test could they use to determine if the dice are fair by comparing the results of their experiment to a fair die that has a 1/6 chance of landing on one of it's six sides?
   (a) A Chi-square test of association.
   (b) A correlation.
   (c) A Chi-square goodness of fit test.
   (d) A paired-samples t-test.

10. A researcher wants to compare the mean hours spent sleeping between males and females. Which of the following tests is most likely to be the appropriate Hypothesis test for this situation?
   (a) The one-sample t-test
   (b) The two-sample t-test
   (c) The paired samples t-test
   (d) The Chi-square test of association

11. If we are told that the results of a hypothesis test was statistically significant this means that the decision of the test was to:
   (a) fail to reject the Alternate hypothesis
   (b) reject the Null hypothesis
   (c) accept the Null hypothesis
   (d) accept the Alternate hypothesis

12. Imagine that there are 100 different researchers each studying the sleeping habits of patients with apnoea. Each researcher takes a random sample of size 50 from the same population. Each researcher is trying to estimate the mean hours of sleep that patients with apnoea get at night, and each one constructs a 95% confidence interval for the mean. With greatest likelihood, how many of these 100 confidence intervals will NOT capture the mean?
   (a) 1 or 2
   (b) About 5
   (c) About half
   (d) 95 to 100

13. As the sample size increases, the standard error of the mean:
   (a) decreases
   (b) increase
   (c) does not change
   (d) varies randomly

14. Investigators want to conduct a university survey pertaining to the number of hours of paid work

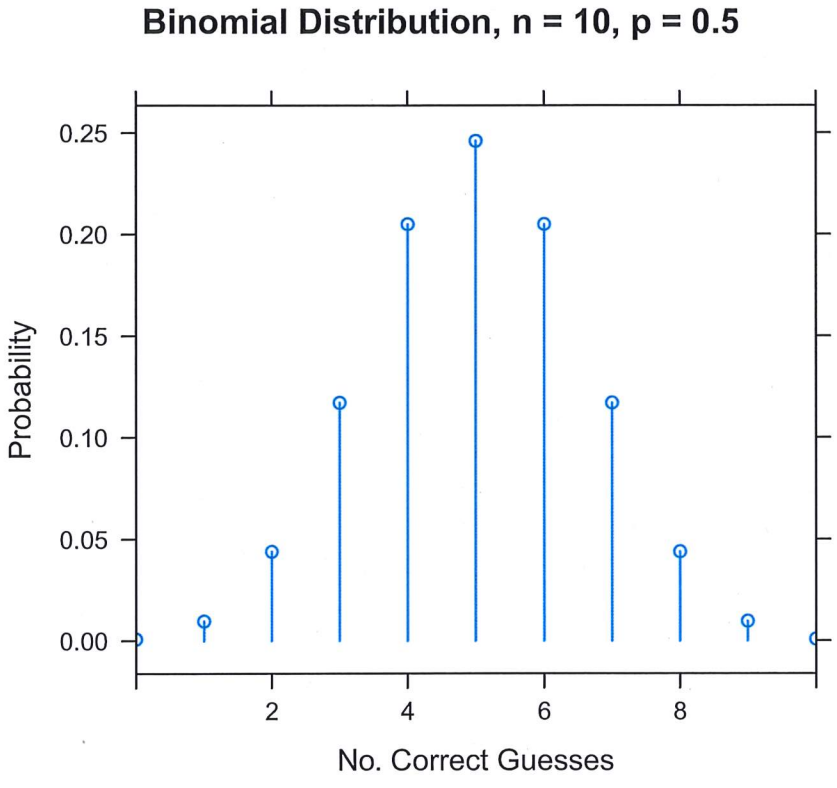
postgraduate students complete each week. Which one of the following sampling methods will most likely result in a randomly representative sample of postgraduate students?

(a) Use a university enrolment list to randomly select postgraduate students from the list and approach these students individually to ask them to fill out the survey.
(b) Survey each postgraduate student at the university.
(c) Ask for postgraduate student volunteers to fill out the survey.
(d) Ask all the students of MATH1324 to fill out the survey.

---

15. The following binomial distribution plot shows the probability of correctly guessing a certain number of questions on a True/False quiz with 10 questions. Which one of the following outcomes is LEAST likely?

(a) $Pr(X = 5)$
(b) $Pr(4 \leq X \leq 5)$
(c) $Pr(X < 5)$
(d) $Pr(X > 7)$



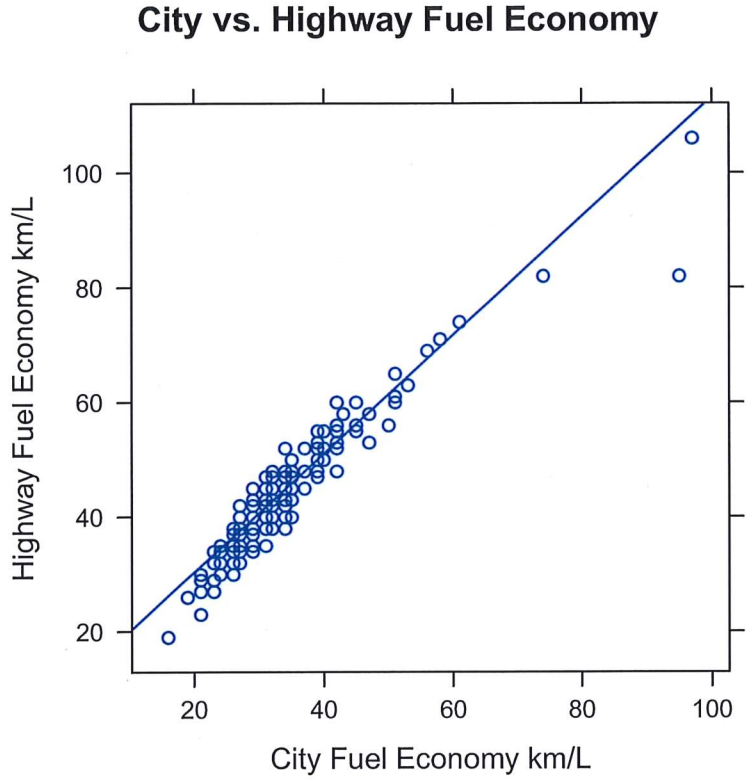
Short Answer (5 Questions - 5 marks each)

Respond to the following short answer questions in the script book provided. Ensure you fill out your name and student number. Start a new page for each question. Number and label each short answer response.

16. A simple linear regression analysis was performed to model the relationship between 428 different cars' city fuel economy (km/L) and their highway fuel economy. The main output from the regression analysis is shown below. Use this output to answer the following questions:

(a) Describe the nature of the relationship between highway and city fuel economy. (1 mark)

(b) Interpret the R-square statistic from the regression model. (1 mark)
(c) Interpret the intercept value from the regression model. (1 mark)
(d) Interpret the slope value from the regression model. (1 mark)
(e) Is there a statistically significant relationship between city and highway fuel economy? Explain. (1 mark)

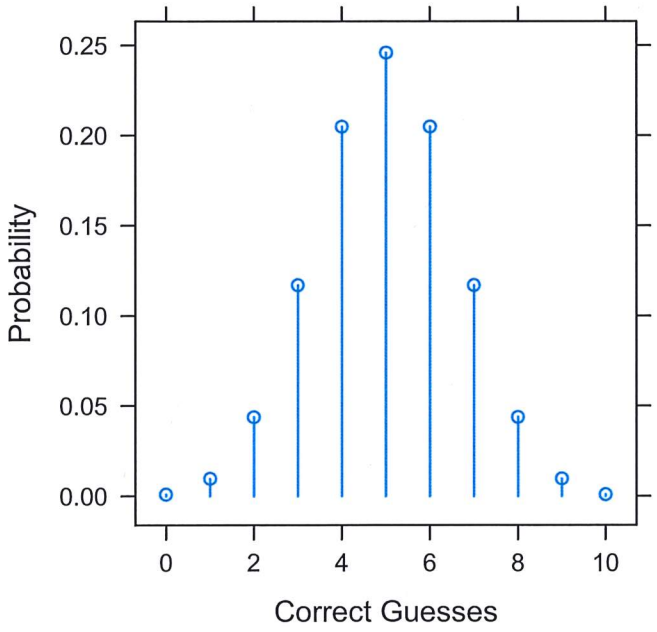



```
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)   9.71625    0.63491   15.30   <2e-16 ***
## Economy_city  1.03519    0.01899   54.51   <2e-16 ***
##
## Residual standard error: 3.238 on 412 degrees of freedom
##   (14 observations deleted due to missingness)
## Multiple R-squared:  0.8782, Adjusted R-squared:  0.8779
## F-statistic:  2971 on 1 and 412 DF,  p-value: < 2.2e-16
```

17. During the 2014 World Cup soccer tournament, the media claimed that Paul, the famous octopus, had correctly predicted the winning team of the match 80% of the time. The table below shows the probability density and cumulative density functions of Paul correctly guessing, just by chance (50%), the outcomes of 10 World Cup matches. Using this scenario and the output provided, answer the following questions:

(a) What probability distribution has been used to calculate the probability of Paul correctly guessing the winners? (1 mark)
(b) What are the parameters of the distribution used to calculate the probability of guessing? (1 mark)
(c) What's the probability of Paul guessing 8/10 correct? (1 mark)
(d) What's the probability of Paul guessing 10/10? (1 mark)
(e) What's the probability of Paul guesting 5 or less winners? (1 mark)

```
##        PDF    CDF
## 0   0.0010 0.0010
## 1   0.0098 0.0107
## 2   0.0439 0.0547
## 3   0.1172 0.1719
## 4   0.2051 0.3770
## 5   0.2461 0.6230
## 6   0.2051 0.8281
## 7   0.1172 0.9453
## 8   0.0439 0.9893
## 9   0.0098 0.9990
## 10 0.0010 1.0000
```



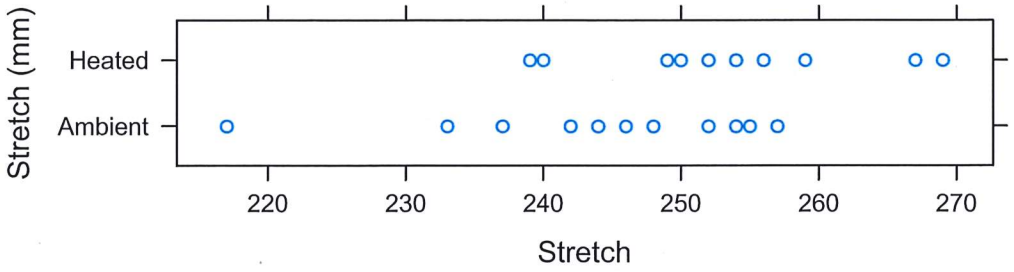
18. Twenty-one elastic bands were randomly divided into two groups. One of the sets was placed in hot water (60-65 degrees C) for four minutes, while the other was left at ambient temperature. After a wait of about ten minutes, the amounts of stretch, under a 1.35 kg weight, were recorded. The R output from a two-sample t-test has been reported below. Stretch was assumed to be normally distributed. Use this output to answer the following questions:
    (a) Explain why the two-sample t-test was selected. (1 mark)
    (b) Explain whether the assumption of equal variance was met. (1 mark)
    (c) Interpret the results of the two-sample t-test and draw a conclusion from the experiment. (3 marks)

```
##   Condition min     Q1 median     Q3 max     mean        sd  n missing
## 1   Ambient 217 239.50    246 253.00 257 244.0909 11.734177 11       0
## 2    Heated 239 249.25    253 258.25 269 253.5000  9.924717 10       0
```

```
## Levene's Test for Homogeneity of Variance (center = median)
##       Df F value Pr(>F)
## group  1  0.0947 0.7616
##       19
```

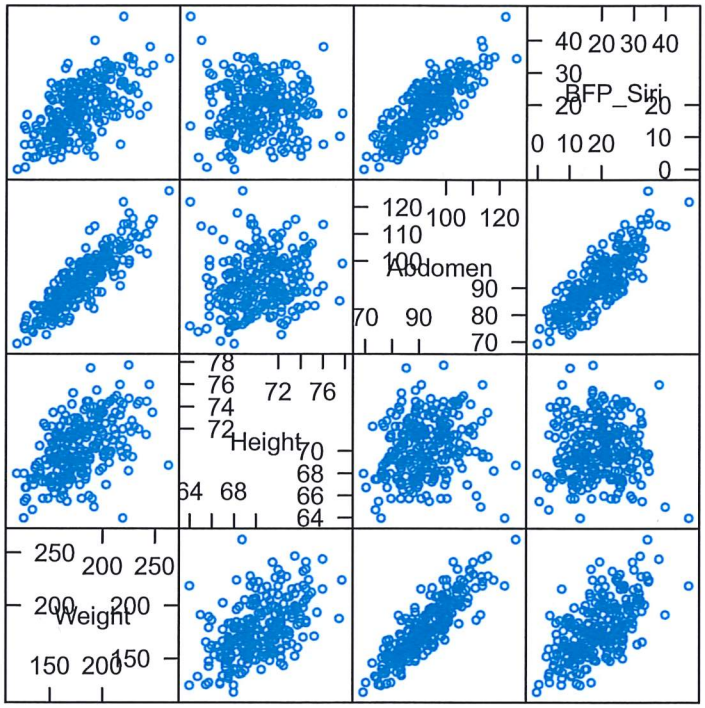
```
##
##  Two Sample t-test
```

```
##
## data:  Stretch by Condition
## t = -1.973, df = 19, p-value = 0.06322
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -19.3905091   0.5723273
## sample estimates:
## mean in group Ambient  mean in group Heated
##              244.0909              253.5000
```



19. The following Pearson correlation matrix and scatter plots show the bivariate relationships among 250 adult males' abdomen, weight, height and body fat % (BFP_Siri) measurements. Using this output, answer the following questions:
    (a) Which two variables share the strongest linear relationship? (1 mark)
    (b) Which two variables share the weakest linear relationship? (1 mark)
    (c) Describe the nature of the relationship between abdomen and weight measurements. (1 mark)
    (d) Explain whether or not the correlation between height and weight was statistically significant. (2 mark)

```
##           Weight Height Abdomen BFP_Siri
## Weight      1.00   0.51    0.87     0.62
## Height      0.51   1.00    0.19    -0.03
## Abdomen     0.87   0.19    1.00     0.82
## BFP_Siri    0.62  -0.03    0.82     1.00
##
## n= 250
##
##
## P
##           Weight Height Abdomen BFP_Siri
## Weight           0.0000 0.0000  0.0000
## Height    0.0000        0.0030  0.6438
## Abdomen   0.0000 0.0030         0.0000
## BFP_Siri 0.0000 0.6438 0.0000
```



Scatter Plot Matrix

20. According to the Australian Bureau of Statistics, the percentage of people aged 0 - 14, 15 - 64 and 65 years and over in Australian are 18.9%, 66.2% and 14.9%, respectively. An investigator gathers a large random sample of 500 Australians and wants to know if their sample can be considered representative of the wider Australian distribution of age. They perform a Chi-square goodness of fit test. The R output is reported below. Use this output to answer the following questions.
    (a) Explain why a Chi-square goodness of fit test was used. (1 mark)
    (b) Explain what the expected values represent. (1 mark)
    (c) Interpret the results of the Chi-square goodness of fit test and draw a conclusion about the representativeness of the investigator's sample. (3 marks)

```
## [1] "Population Proportions - Age Bands"

##  0 - 14 15 - 65     65+
##   0.189   0.662   0.149

## [1] "Obeserved - Age Bands"

##  0 - 14 15 - 65     65+
##     107     325      68

##
##  Chi-squared test for given probabilities
##
## data:  df$Observed
## X-squared = 2.3293, df = 2, p-value = 0.312

## [1] "Expected counts - Age Bands"

##  0 - 14 15 - 65     65+
##    94.5   331.0    74.5
```

MC Answers: 1) a, 2) a, 3) b, 4) d, 5) d, 6) c, 7) b, 8) d, 9) c, 10) b, 11) b, 12) b, 13) a, 14) a, 15) d.
Short Answers:

16. a. Positive linear (maybe a slight curvilinear relationship) (1 mark)
    b. Based on a linear relationship, city fuel economy explained 87.8% of the variation in Highway fuel economy. (1 mark)
    c. When city fuel economy was 0, the estimated highway fuel economy was 9.71. (1 mark)
    d. As city fuel economy increased by 1, highway fuel economy increases on average by 1.04. (1 mark)
    e. Yes, there was a statistically significant linear relationship between city and highway fuel economy. This was because the overall F-test of the linear regression model was statistically significant, $p < .001$ (students might also interpret the hypothesis test of the slope which was also statistically significant). (1 mark).

17. a. Binomial distribution (1 mark)
    b. $p = .5$, $n = 10$ (1 mark)
    c. 0.0439 (1 mark)
    d. 0.001 (1 mark)
    e. 0.6230 (1 mark)

18. a. Because the investigator compared two independent groups of rubber bands that were randomly allocated to different conditions. (1 mark)
    b. The Levene's test of equal variance was not statistically significant, $p = 0.76$, therefore, the assumption of equal variance was not violated. (1 mark)
    c. The results of the two-sample t-test comparing the mean stretch length between the ambient and heated conditions was not statistically significant. This decision was due to the p-value of the test, $p = 0.06$, being greater than the 0.05 level of statistical significance and the 95% CI of the difference between means, (-19.39, 0.57), capturing the Null hypothesised value of 0. The results of the statistical test failed to find evidence that heating a rubber band will increase its mean stretch length. (3 marks)
    d. (1 mark for concluding that the test was not statistically significant. 1 mark for explaining why [p-value or CI] and 1 mark for drawing a conclusion in context)

19. a. Abdomen and weight, $r = 0.87$ (1 mark)
    b. BFP_Siri and Height, $r = -0.3$ (1 mark)
    c. Positive, as abdomen measurements increase, BFP also increases (1 mark)
    d. The correlation between height and weight, $r = .51$, was statistically significant (1 mark). This was because the p-value of the correlation, $p < .001$, was less than the significance level of 0.05 (1 mark).

20. a. Because the investigator compared the distribution of a categorical variable to an assumed population distribution. (1 mark)
    b. The expected values represent the expected frequencies that should result from a sample 500 assuming the population distribution of age categories according to the ABS is true.
    c. The results of the Chi-square test of association were not statistically significant (1 mark). This was because the p-value of the test was greater than the 0.05 significance level (1 mark). The investigator's sample did not significantly deviate from the ABS distribution of age categories (1 mark).