INDIVIDUAL ASSESSMENT COVER SHEET

Faculty of Design and Creative Technologies

AUT
TE WĀNANGA ARONUI O TĀMAKI MAKAU RAU

First Name	**Nikisha**	**Family Name**	**Chhima**	**Student ID No**	22168889
Course Name	**Statistics of Data Science**	**Course Code**	**COMP616**	**Assignment Due Date**	2nd June
Lecturer	**Ryan Ip**	**Tutorial Day**	**Wednesday**	**Date Submitted**	27th May
Tutor		**Tutorial Time**	**12pm-2pm**	**No.Words/Pages**	**9**

In order to ensure fair and honest assessment results for all students, it is a requirement that the work that you hand in for assessment is your own work. If you are uncertain about any of these matters, then please discuss them with your lecturer.

Plagiarism and Dishonesty are methods of cheating for the purposes of General Academic Regulations (GAR) http://www.aut.ac.nz/calendar

Assignments will not be accepted if this section is not completed and signed.

Please read the following and **tick** ✓ to indicate your understanding:

1.	I understand it is my responsibility to keep a copy of my assignment.	☒ **Yes**	☐ **No**
2.	I have signed and read the **Student's Statement below**.	☒ **Yes**	☐ **No**
3.	I understand that a software programme (Turnitin) that detects plagiarism and copying may be used on my assignment.	☒ **Yes**	☐ **No**

Student's Statement:

This assessment is entirely my own work and has not been submitted in any other course of study. I have submitted a copy of this assessment to Turnitin, if required. In this assessment I have acknowledged, to the best of my ability:

- The source of direct quotes from the work of others.
- The ideas of others (includes work from private or professional services, past assessments, other students, books, journals, cut/paste from internet sites and/or other materials).
- The source of diagrams or visual images.

Student's Signature:  **Date: 27/05/2024**

The information on this form is collected for the primary purpose of submitting your assignment for assessment. Other purposes of collection include receiving your acknowledgement of plagiarism polices and attending to administrative matters. If you choose not to complete all questions on this form, it may not be possible for the Faculty of Design and Creative Technologies to accept your assignment.

Assignment 2 - ANSWER BOOKLET

COMP616 / STAT604

Nikisha Chhima, Student ID: 22168889

INSTRUCTIONS: Use this file to write your answer to Assignment 2. You will need to knit it to PDF and then submit this PDF to Canvas.

Insert as many R code chunks as needed, using:

```
# copy and paste
```

If mathematical notation/ formulae are needed, please copy-paste and modify those in the .Rmd files used in the labs.

ANSWERS FOR QUESTION 1

1(a)

```
#Spearman's Correlation test
weight <- c(19.0, 17.2, 16.7, 17.0, 17.4, 12.4, 10.1, 14.6, 13.4, 14.4)
height <- c(101.0, 113.2, 102.9, 80.4, 96.9, 108.7, 97.2, 98.4, 104.0, 81.9)
cor.test(weight,height,method="spearman")
```

```
##
##  Spearman's rank correlation rho
##
## data:  weight and height
## S = 180, p-value = 0.8114
## alternative hypothesis: true rho is not equal to 0
## sample estimates:
##         rho
## -0.09090909
```

```
#Kendall's tau test
weight<- c(19.0, 17.2, 16.7, 17.0, 17.4, 12.4, 10.1, 14.6, 13.4, 14.4)
height <- c(101.0, 113.2, 102.9, 80.4, 96.9, 108.7, 97.2, 98.4, 104.0, 81.9)
cor(weight,height,method="kendall")
```

```
## [1] -0.06666667
```

1(b)

$$P(X = 10) = C_{10}^{12}\ (0.5)^{10}\ (1 - 0.5)^{12-10} = 0.01611328$$

```
p_10 <- (factorial(12) / (factorial(10) * factorial(12 - 10))) * (0.5^10) * ((1 - 0.5)^(12 - 10))
p_10
```

```
## [1] 0.01611328
```

$$P(X = 11) = C_{11}^{12}\ (0.5)^{11}\ (1 - 0.5)^{12-11} = 0.002929688$$

```
p_11 <- (factorial(12) / (factorial(11) * factorial(12 - 11))) * (0.5^11) * ((1 - 0.5)^(12 - 11))
p_11
```

```
## [1] 0.002929688
```

$$P(X = 12) = C_{12}^{12}\ (0.5)^{12}\ (1 - 0.5)^{12-12} = 0.0002441406$$

```
p_12 <- (factorial(12) / (factorial(12) * factorial(12 - 12))) * (0.5^12) * ((1 - 0.5)^(12 - 12))
p_12
```

```
## [1] 0.0002441406
```

$$P(X >= 10) = 0.01611328 + 0.002929688 + 0.0002441406 = 0.01928711$$

So the manually calculated p-value is around 0.01928711.

```
pvalue <- 1 - pbinom(9, 12, 0.5)
pvalue
```

```
## [1] 0.01928711
```

In conclusion, as the p-value (0.0192) is smaller than the significance level of 0.05, there is enough evidence to reject the null hypothesis so we favour the alternative hypothesis. Suggesting that the median of the population is not 55.

1(c)

A suitable hypothesis test would be the Runs Test (Exact Method) as this type of test tests for sequence randomness.

The minimum values of the test statistic concerning the test selected is 2 and the maximum value is 10

1(d)

I've used 'exact=False' as there are a presence of ties.

```r
locA <- c(6,5,4,3,7,8)
locB <- c(3,2,7,4,6,4)
wilcox.test(locA, locB, alternative="greater", exact=FALSE, correct=FALSE)
```

```
##
##  Wilcoxon rank sum test
##
## data:  locA and locB
## W = 24.5, p-value = 0.146
## alternative hypothesis: true location shift is greater than 0
```

In conclusion as the p-value is greater than the 5% significance level ($0.146 > 0.05$). We do not have enough evidence against the null hypothesis. So we can not confirm that the UV index at location A is higher than at location B at 5% significance level.

1(e)

```r
#probability of testing positive:
prob_positive <- (0.90 * 0.10) + ((1 - 0.95) * (1-0.10))

#Bayes' theorem - probability that you test positive and you have COVID
prob_covid_given_positive <- (0.90 * 0.10) / prob_positive
prob_covid_given_positive
```

```
## [1] 0.6666667
```

ANSWERs FOR QUESTION 2

2(a)

The Kurskal-Wallis Test would be a suitable non-parametric method for this purpose. We use this when we have more than two independent samples of measurements and want to test if all the medians are the same or not.

2(b)

Let M denote the median. The hypotheses is written as:

$H_0 : Martanddesign = Meducation = Mmidwifery = Mlaw$:

H_1 :: Not all medians are equal

2(c)

```
results <-c(41, 52, 32, 44, 45, 49, 43, 60, 55, 53, 48, 26, 27, 46, 51, 17, 30, 23, 39, 37, 13, 29)
departments <- c(rep("Art and Design",6), rep("Education",5), rep("Midwifery",5), rep("Law",6))
print(rbind(results, departments))
```

```
##             [,1]             [,2]             [,3]             [,4]
## results     "41"             "52"             "32"             "44"
## departments "Art and Design" "Art and Design" "Art and Design" "Art and Design"
##             [,5]             [,6]             [,7]        [,8]
## results     "45"             "49"             "43"        "60"
## departments "Art and Design" "Art and Design" "Education" "Education"
##             [,9]        [,10]       [,11]       [,12]       [,13]
## results     "55"        "53"        "48"        "26"        "27"
## departments "Education" "Education" "Education" "Midwifery" "Midwifery"
##             [,14]       [,15]       [,16]       [,17] [,18] [,19] [,20] [,21]
## results     "46"        "51"        "17"        "30"  "23"  "39"  "37"  "13"
## departments "Midwifery" "Midwifery" "Midwifery" "Law" "Law" "Law" "Law" "Law"
##             [,22]
## results     "29"
## departments "Law"
```

2(d)

```
midwifery_results <- c(26, 27, 46, 51, 17)
sum_midwifery_ranks <- sum(rank(midwifery_results))
print(sum_midwifery_ranks)
```

```
## [1] 15
```

2(e)

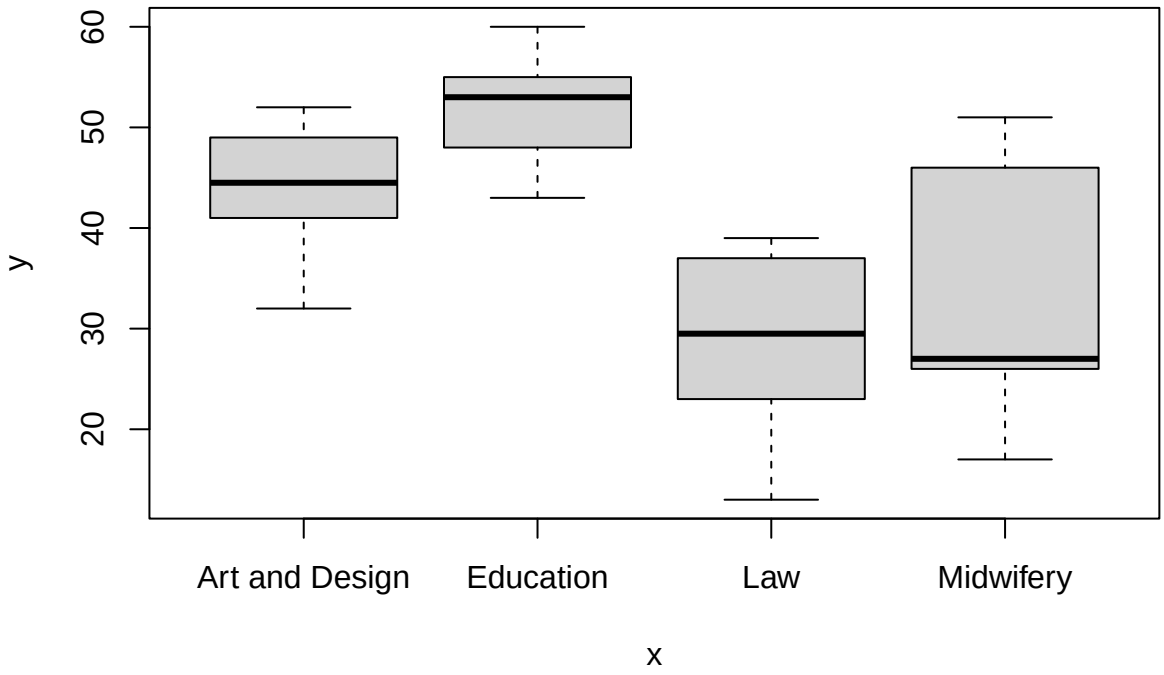
The p-value is 0.01089 which is less than the 5% level of significance. Therefore, we have enough evidence to reject the null hypothesis and favour the alternative hypothesis. This indicates that some of the median test results are different from each other.

```
kruskal.test(results, departments)
```

```
##
##  Kruskal-Wallis rank sum test
##
## data:  results and departments
## Kruskal-Wallis chi-squared = 11.16, df = 3, p-value = 0.01089
```

2(f)

```
plot(as.factor(departments), results)
```




```
pairwise.wilcox.test(results,departments,p.adjust.method="bonferroni")
```

```
##
##  Pairwise comparisons using Wilcoxon rank sum exact test
```

```
##
## data:  results and departments
##
##           Art and Design Education Law
## Education 0.753          -         -
## Law       0.052          0.026     -
## Midwifery 1.000          0.333     1.000
##
## P value adjustment method: bonferroni
```

The findings show that there are significant differences between 'Education' and 'Art and Design', 'Law' and 'Art and Design', 'Law' and 'Education', and 'Midwifery' and 'Education'. We can determine that the response test results were highest in the 'Education' department and lowest in the'Law' department by analysing the observation from the boxplots created as well.

ANSWERS FOR QUESTION 3

3(a)

A Chi-sq test of independence would be a suitable non-parametric method for this purpose. We use this when we have one sample of counts where the sample can be categorised into two ways according to two factors, we want to test if the two factors are independent or not.

3(b)

H0: Size and Area are not associated H1: Size and Area are associated

3(c)

```
table <- as.table(rbind(c(150,57,14),c(96, 45, 38),c(52, 25, 23)))
dimnames(table) <- list(Size = c("Large", "Medium", "Small"),
                        Areas = c("Area A","Area B","Area C"))
print(table)
```

```
##         Areas
## Size     Area A Area B Area C
##   Large     150     57     14
##   Medium     96     45     38
##   Small      52     25     23
```

3(d)

```
chisq.test(table)$expected
```

```
##         Areas
## Size      Area A Area B Area C
##   Large  131.716 56.134  33.15
##   Medium 106.684 45.466  26.85
##   Small   59.600 25.400  15.00
```

```
prop.table(table,2)
```

```
##         Areas
## Size        Area A    Area B    Area C
##   Large  0.5033557 0.4488189 0.1866667
##   Medium 0.3221477 0.3543307 0.5066667
##   Small  0.1744966 0.1968504 0.3066667
```

According to the data above, I can see that Area C has a larger proportion of medium-sized abalones than Area A and B, while Area A has a significantly higher proportion of huge abalones than Area C. These observations suggests a potential association between size of abalones and the areas at which they are found. From this information my expected outcome for this Chi-sq test of independence would be to reject the null hypothesis and favour the alternative hypothesis. That there is an association between size and area.

3(e)

```
exp_count_large_areaC <- (221 * 75) / 500
exp_count_large_areaC
```

```
## [1] 33.15
```

```
exp_count_medium_areaC <- (179 * 75) / 500
exp_count_medium_areaC
```

```
## [1] 26.85
```

3(f)

```
chisq.test(table)
```

```
##
##  Pearson's Chi-squared test
##
## data:  table
## X-squared = 24.561, df = 4, p-value = 6.164e-05
```

As the p-value (0.0000616) is less than 5% significance level, we have enough evidence to reject the null hypothesis and favour the alternative hypothesis which was predicted in question 3(d). There is evidence suggesting that size and area of the abalones are not independent, therefore size and area of abalones being associated with each other.

END OF ASSIGNMENT 2