STATS 1000 / STATS 1004 / STATS 1504
Statistical Practice 1
Assignment 2
2020

DEADLINE:

- Wednesday 25^{th} March 2020 (Week 4) 5:00pm.

CHECKLIST

☐: Have you shown all of your working, including probability notation where necessary?

☐: Have you given all numbers to **3 decimal** places.

☐: Have you included all R output and plots to support your answers where necessary.

☐: Have you made sure that all plots and tables each have a caption.

☐: If before the deadline, have you submitted your assignment via the online submission on MyUni?

☐: Is your submission a single word document or pdf file - correctly orientated, easy to read? If not, penalties apply.

☐: Penalties for more than one document - 10% of final mark for each extra document. Note that you may resubmit and your final version is marked, but the final document should be a single file.

☐: If after the deadline, but within 24 hours, have you contacted us via the enquiry page on MyUni and then submitted your assignment online via the online submission on MyUni?

☐: Penalties for late submission - within 24 hours 40% of final mark. After 24 hours, assignment is not marked and you get zero.

☐: Assignments emailed instead of submitted by the online submission on MyUni will not be marked and will recieve zero.

☐: Have you checked that the assignment submitted if the correct one, as we cannot accept other submissions after the due date.

☐: Do not write directly on the question sheet.

1. **Two-way tables in R**

```
## -- Attaching packages ---------------------------------------------------
tidyverse 1.2.1 --

## v ggplot2 3.1.0     v purrr   0.2.5
## v tibble  2.0.1     v dplyr   0.7.8
## v tidyr   0.8.2     v stringr 1.3.1
## v readr   1.3.1     v forcats 0.3.0

## Warning:  package 'tibble' was built under R version 3.5.2

## -- Conflicts ------------------------------------------------------------
tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()

## Warning:  The `printer` argument is deprecated as of rlang 0.3.0.
## This warning is displayed once per session.
```

For full marks, this question must be typed in Word, with all figures and tables included and captioned.

The data in `titanic.xlsx` gives the fate of 1309 passengers on the HMS Titanic. Each row in the dataset corresponds to a passenger and we have four variables:

- pclass - the passenger class with levels 1,2,3.
- survived - a number indicating if the passenger survived (0 - did not survive, 1 - survived).
- sex - the gender of the passenger - levels: male, female.
- age - the age of the passenger when known.

We are interested in whether there is a difference in survival rates for different classes of passengers.

(a) Produce crosstab tables in R to answer the following questions. To obtain full marks, please included appropriate R output to support your answers. You will need at least one table and some calculations, or you can do all of it with the correct two tables.

Be careful to read the question correctly, expecially with regards to percentage and proportion.

[8 marks]

i. What is the total number of first class passengers who survived?

   200 first class passengers survived (Table 1).

ii. What is the total number of third class passengers who survived?

   181 third class passengers survived (Table 1).

iii. What proportion of all the passengers survived?

   0.382 of all passengers survived (Table 1).

iv. What proportion of all the passengers were in third class?

   0.542 of all the passengers were in third class (Table 2).

v. What proportion of the first class passengers survived?

   0.619 of the first class passengers survived (Table 1).

vi. What proportion of the third class passengers survived?

   0.255 of the third class passengers survived (Table 1).

```
##        survived
## pclass   0   1
##      1 123 200
##      2 158 119
##      3 528 181
```

Table 1: Cross tabs table of class (rows) and survival (columns) for passengers in the titanic dataset.

```
##          pclass
## survived   1   2   3
##        0 123 158 528
##        1 200 119 181
```

Table 2: Cross tabs table of survival (rows) and class (columns) for passengers in the titanic dataset.

(b) We would like to assess which class has the lowest survival rate. To do this perform the following:

i. Create a bar-chart of the data with class on the x-axis, number of people who survived and did not survive on the y-axis and bars for

both survival and non-survival. You must include the plot in your assignment for the marks.

[2 marks]

The correct plot is in Figure 1.

ii. From all the results you have so far, which class appears to have the highest survival proportion, and which class appears to have the lowest survival proportion? By considering the context of the data, why do you think this is the case?

[3 marks]

First class has the highest survival proportion and third class has the lowest survival proportion.

At the time the titanic sank there was a strong class system in the UK and so it was probably first class who got first chance in the lifeboats.

[Total: 13]

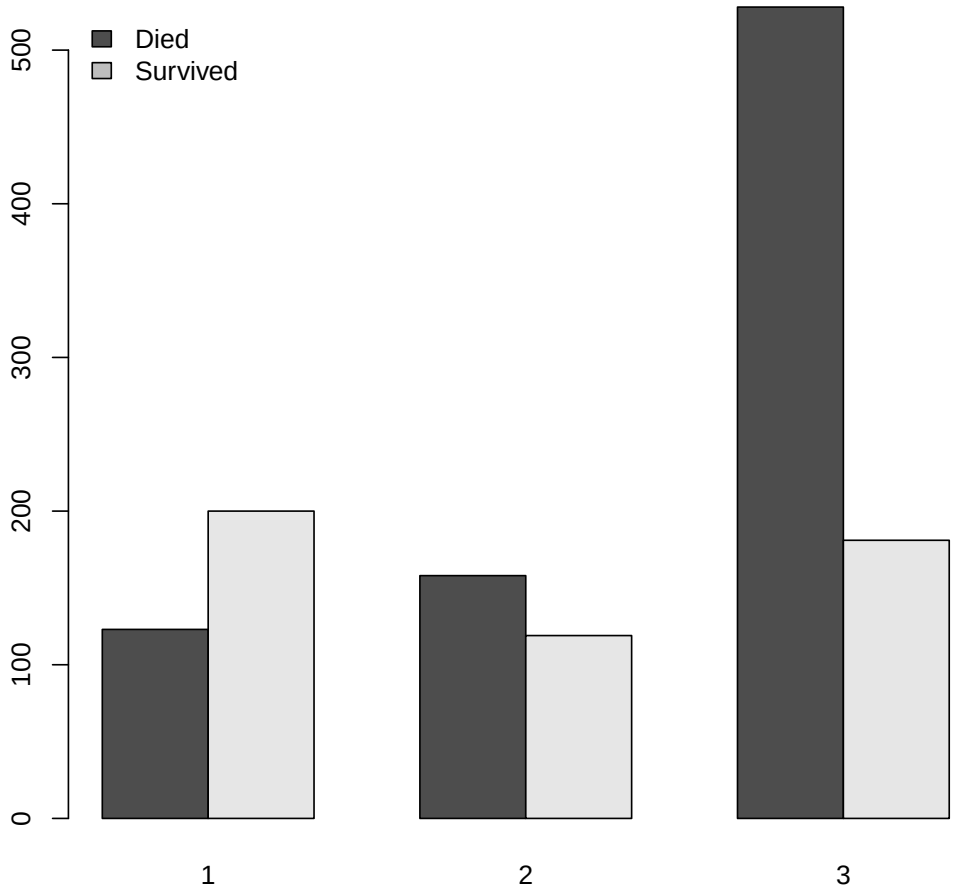

Figure 1: Bar-chart of the number of people within each class that survived or did not survive for passengers in the titanic dataset.

2. **Scatterplots and least-squares line in R**

For full marks, this question must be typed in Word; with all figures and tables included and captioned; and units given

Osteoporosis is a condition where bones become weak. It affects more than 200 million people worldwide. Exercise is one way to produce strong bones and to prevent osteoporosis. Since we use our dominant arm (the right arm for most people) more than our non-dominant arm, we expect the bone in our dominant arm to be stronger than the bone in the non-dominant arm. By comparing the strengths, we can get an idea of the effect that exercise can have on bone strength.

The data in the excel file `bone.xlsx` gives the bone strength ($cm^4/1000$) for the arms of 15 young men (control) and 15 baseball players.

(a) Obtain a scatter plot of the data (Non-dominant on y-axis, dominant on x-axis) and comment on the relationship between `non-dominant` and `dominant` bone strength. *Remember you need to give direction, form (linear versus non-linear), and strength*

[5 marks]

The scatter-plot is given in Figure 2. There is a moderate positive linear relationship between non-dominant and dominant bone strength.

(b) Using R, find the intercept and slope of the least squares line (response variable non-dominant bone strength, predictor variable dominant bone strength) and interpret these parameters in context. For full marks, include the appropriate R table.

[6 marks]

The intercept is 7.331 $cm^4/1000$ and the slope is 0.448 $cm^4/1000$ (Table 3).

If the dominant bone strength is 0 $cm^4/1000$, then we expect the non-dominant bone strength to be 7.331 $cm^4/1000$.

If the dominant bone strength increases by 1 $cm^4/1000$, then we expect the non-dominant bone strength to increase by 0.448 $cm^4/1000$.

(c) Use the least squares line to estimate the mean non-dominant bone strength for a subject with a dominant bone strength of 20 $cm^4/1000$. *Remember to show your working.*

[2 marks]

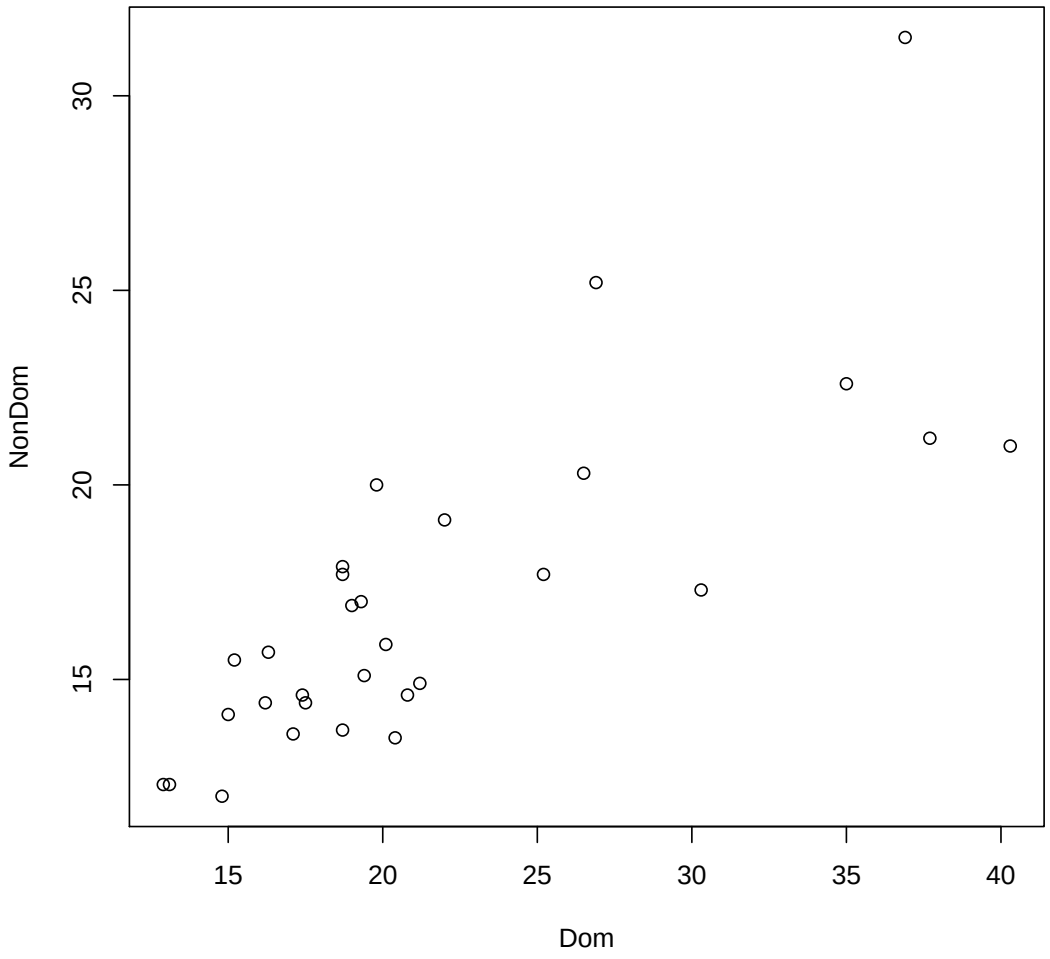

Figure 2: Scatterplot of non-dominant against dominant bone strength ($cm^4/1000$) for the subjects in the bone dataset.

```
##
## Call:
## lm(formula = NonDom ~ Dom, data = bone)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -4.3723 -1.7881 -0.4752  1.0970  7.6497
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  7.33146    1.52264   4.815 4.60e-05 ***
## Dom          0.44766    0.06637   6.745 2.53e-07 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.657 on 28 degrees of freedom
## Multiple R-squared:  0.619,Adjusted R-squared:  0.6054
## F-statistic:  45.5 on 1 and 28 DF,  p-value: 2.526e-07
```

Table 3: Coefficients table for the linear regression of non-dominant on dominant bone strength for the subjects in the bone dataset.

The least squares line is

$$\text{Non-dominant} = 7.331 + 0.448 \times \text{Dominant}.$$

So substituting Dominant $= 20$ into the equation gives

$$\text{Non-dominant} = 7.331 + 0.448 \times 20 = 16.291 \text{ cm}^4/1000.$$

(d) Why should we not use this model to predict for a subject with a dominant bone strength of 60 $\text{cm}^4/1000$.

[2 marks]

The data that we used to fit our model goes from 12.9 to 40.3 and so predicting for a value of 60 would be extrapolation and should not be done.

[Total: 15]

3. **Experimental design**

The following question may be hand-written. Remember to scan and attach to the rest of the assignment in a single document.

(a) What are the three elements of good experimental design?

[3 marks]

- Control.
- Replication.
- Randomisation.

(b) Give two reasons why randomisation is useful in experiments.

[2 marks]

- Averages out lurking variables.
- Prevents selection bias.

(c) What is a placebo?

[1 mark]

A treatment that looks like, smells like, and tastes like the other treatments, but has no active ingredient.

(d) What is a double-blind experiment?

[1 mark]

An experiment where both the participants and the researchers do not know who receives which treatment.

[Total: 7]

[[Assignment total: 35]]