Semester 1 Main, 2018

The University of Sydney
School of Mathematics and Statistics

DATA1001/ENVX1002
Foundations of Data Science/Introduction to Statistical Methods

June 2018 Lecturers: T Bishop, F Van Ogtrop, D Warren

Time Allowed: **Reading - 10 minutes; Writing - 3 hours**

Exam Conditions: You may bring in 1 double-sided A4 page of notes. Writing is not permitted at all during reading time. Calculators are not permitted.

Family Name: .... SID: ....

Other Names: .... Seat Number: ....

Please check that your examination paper is complete (21 pages) and indicate by signing below.
I have checked the examination paper and affirm it is complete.

Signature: .... Date: ....

This examination has two sections: Multiple Choice and Extended Answer.

The Multiple Choice Section is worth 33% of the total examination.
There are 30 questions. The questions are of equal value.
All questions may be attempted.

Answers to the Multiple Choice questions must be entered on
the Multiple Choice Answer Sheet before the end of the examination.

The Extended Answer Section is worth 67% of the total examination.
There are 5 questions. The questions are of equal value.
All questions may be attempted. Working must be shown.

THE QUESTION PAPER & 1 PAGE OF NOTES MUST NOT BE REMOVED FROM THE EXAMINATION ROOM.

Marker's use only	

Extended Answer Section

Answer these questions in the spaces provided.
All questions are based on the Opal Card data story.
Source: https://opendata.transport.nsw.gov.au/

1. In late 2012, Transport for NSW introduced the Opal Card, which is a credit-card sized smartcard for paying for public transport in Sydney and nearby areas. A passenger taps "on" when they begin their journey, and taps "off" when they reach their destination.

Data was collected over a week in July 2016 for the following 6 variables:

- mode of public transport - bus, train, light rail and ferry.
- date - in yyyymmdd format, eg 20160730 is 30/07/2016.
- tap type - "on" and "off".
- time - in 24hr time collected in 15 minute intervals.
- location - denoted by postcode and names of train stations, ferry wharves and light rail stops.
- count - the number of taps "on" or "off".

```
opal = read.csv("data/opal.csv")
dim(opal)

## [1] 215630      6

head(opal,4)

##   mode     date tap  time  loc count
## 1  bus 20160730  on 02:30 2000   415
## 2  bus 20160730  on 02:30 2135    18
## 3  bus 20160730  on 02:30   -1    24
## 4  bus 20160730  on 02:30 2010    31
```

The Extended Answer Section begins on the next page

(a) (*i*) Is this data a population or sample? Explain.

(*ii*) Did the data collection come from a controlled experiment or an observational study? What difference does that make to any conclusions drawn?

(*iii*) Who owns this data? Suggest one possible ethical issue.

(b) (*i*) How many observations/records are there? __________

(*ii*) What type of variable is `mode` ? ____________________

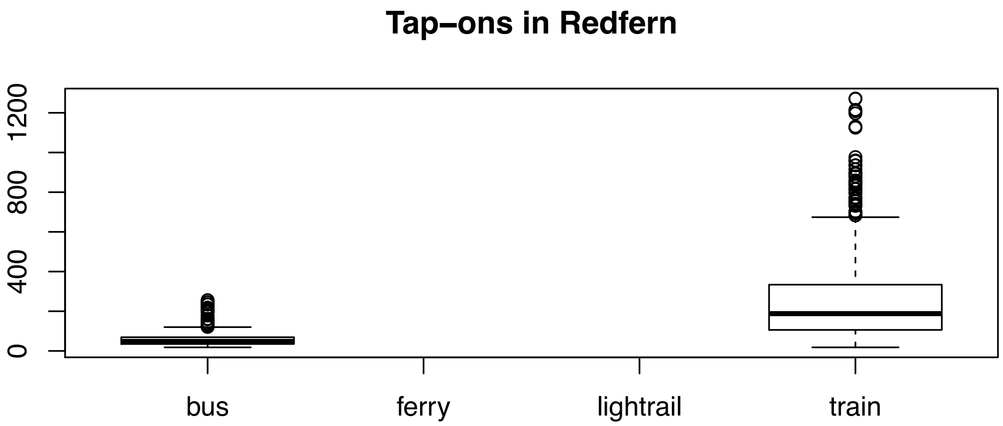
(*iii*) Why might the location of the 3rd subject be recorded as `-1` ?

(*iv*) What would a value of "0" represent in the `count` column?

(c) Here we focus on the counts for the tap-ons over the week. Would the mean or median be a better summary of centre? Explain.

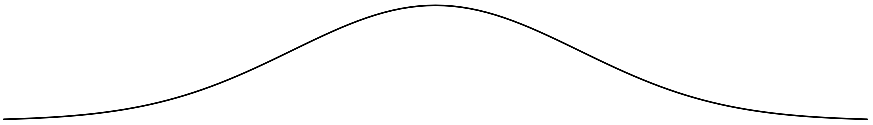
```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    18.0    29.0    50.0   109.4   104.0 14396.0
```

(d) Here we focus on the tap-ons in Redfern over the week. Make 3 comments, in context.

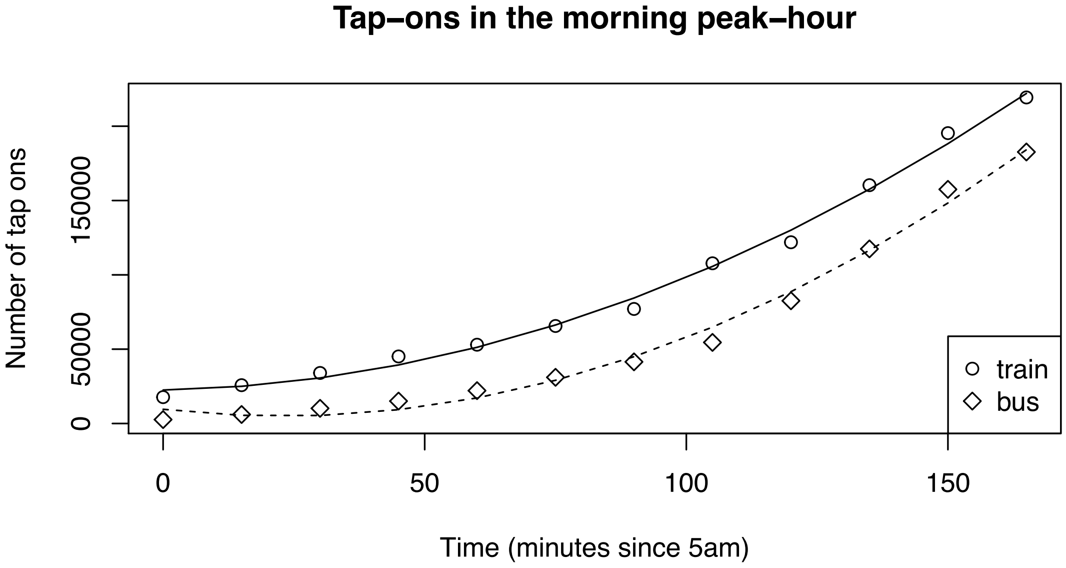


2.(a) Suppose the number of people caught not having an Opal card at Redfern station each weekday can be modelled by a Normal curve with mean = 10 and SD = 3.

Find the chance that the number of people caught on a certain day is more than 13, by showing clear working on the curve below.



(b) Transport for NSW wants to model the total number of tap-ons for buses and trains in the morning peak-hour between 5am-7:45am (a 165 minute period).



(*i*) What do you notice?

(*ii*) Would a linear regression model be appropriate? Explain.

(*iii*) If you fitted a linear regression model to the tap-ons for trains, what would the residual plot look like?

(*iv*) Suppose we fit a quadratic model to the tap-ons for trains. Predict how many taps occcur at 7am.
(Note: Leave your answer as an expression. Don't evaluate.)

```
time2 = time^2
lm(opaltime_m$train~time + time2)

##
## Call:
## lm(formula = opaltime_m$train ~ time + time2)
##
## Coefficients:
## (Intercept)         time        time2
##   22515.736       61.633        6.955
```

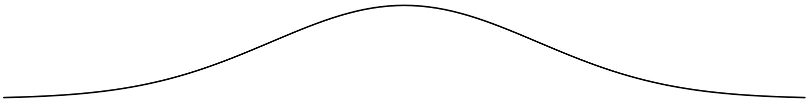
3.(a) In your own words, explain why the Central Limit Theorem is important.

(b) A fair coin is tossed 100 times.

(*i*) Draw a box model to represent the number of heads.

(*ii*) Show that the expected value and standard error of the number of heads are 50 and 5 respectively.

(*iii*) By annotating the Normal curve, show that the chance of getting between 35 and 50 heads is approximately 50%.



(*iv*) Is it valid to use a Normal curve here? Explain.

(c) Choice magazine wants to investigate how often Sydney commuters are late to work.

(i) Describe a possible survey method.

(ii) Discuss 2 possible limitations or sources of bias.

(iii) Propose a biased question.

4. Transport for NSW is interested in comparing the tap-ons for all 4 modes of travel.

(a) Consider the difference between tap-ons for ferry and light rail.
Note: `ferryon$count` is the tap-ons for ferry.

```
##
##  Welch Two Sample t-test
##
## data:  log(ferryon$count) and log(raillighton$count)
## t = 8.3386, df = 5636.1, p-value < 2.2e-16
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  0.1138784 0.1838814
## sample estimates:
## mean of x mean of y
##  3.886913  3.738033
```

(i) Why might the test have been performed on the `log` of the data?

(ii) Perform an appropriate hypothesis test.

H:

A:

T:

P:

C:

(b) It is claimed that the proportions of tap-ons for bus:ferry:lightrail:train is 40:5:5:50. Respond to this claim by using a hypothesis test.

```
##       bus     ferry lightrail     train 
## 41.462751  1.645960  1.229094 55.662195 
## 
##  Chi-squared test for given probabilities
## 
## data:  total_tapon_prop
## X-squared = 5.7886, df = 3, p-value = 0.1224
```

H:

A:

T:

P:

C:

(c) Consider the difference between tap-ons and tap-offs. What story does this output tell us?

```
total_tapon-total_tapoff
      bus     ferry lightrail     train 
    32879     -7335      2249    -29138 
```

5. You are a data scientist reporting to a client on the Opal card data. Choose your client and define the purpose of the report. Discuss the limitations of the data and 2 interesting insights, using evidence from previous questions. Suggest an action for the client.

Client: ______________________________

Purpose of Report: ____

Limitations of data:

Interesting Insights:
1.

2.

Suggested Action:

This page will not be marked - it is for your working.

End of Extended Answer Section

End of Examination