2018S1 DATA1001 5exam Solutions

Tom Bishop, Floris Van Ogtop, Di Warren

03/04/2018

Note to markers: Below are only brief answers. We expect students to answer concisely and precisely, in context with evidence. The exam is where we differentiate bewteen students, by testing/expecting rigorous statistical thinking, as the projects marks tend to be high.

Extended Answer Questions

Question 1

1(a)

i. It depends on what our research question is. If we are focussing just on July 2016, then the data is the population. If we are considering this data as a sample of broader data, then it is a sample.

ii. The data came from observations (collected through tap-on/off machines), so it is an observational study. A controlled trial would allow more detailed driling down on the difference between variables.

iii. This is an ethical issue. Is the data owned by collection company, Transport for NSW, Research company or the personal commuter?

1(b)

i. 215630

ii. Qualitative

iii. Missing data?

iv. No tap-ons/offs in that period/location/mode.

1(c) As the data is skewed, the median is more approrpriate. This is due to the massive outliers for peak hour central locations.

1(d)

Examples:

- Redfern train station is busier than the bus stops.

- However, there are outliers for buses indicating some very busy periods.
- The median count for trains across all periods is about 200 in a 15 minute period!

Part	**Significant Error**	**Outcome**
1(a)(i)	1	Assessment with reason
1(a)(ii)	1	2 sensible comments
1(a)(iii)	1	Assessment with reason
(b)	1	4 Correct answers
(c)	1	Assessment with reason
(d)	3	3 sensible comments
Total	8	

Question 2

1(a) 16%, showing reasoning on sketch.

1(b)

i. Tap-ons for trains is always higher than buses, across the time period.

ii. No, as there is a quadratic or perhaps non-linear trend evident for both modes.

iii. There would be a quadratic trend, rather than a random scatter, indicating the linear model is not appropriate.

iv. 7am = 120 minutes.

prediction = 22515.736 + 61.633 x 120 + 6.955 x 120^2

Part	**Significant Error**	**Outcome**
2(a)	2	Answer plus working on curve
2(b)(i)	1	Sensible comment
2(b)(ii)	2	Assessment with reason

Part	Significant Error	Outcome
2(b)(iii)	1	Answer mentions quadratic trend
2(b)(iv)	2	Correct expression, correct value of time
Total	8	

Question 3

3(a) The Central Limit Theorem

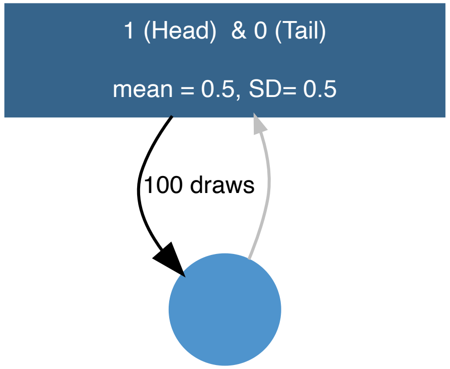
- In terms of box model: When drawing at random with replacement from a box, the probability histogram for the sum (or average) will follow the normal curve even if the contents of the box does not.
- Or more formally: The limiting distribution of the sum of a series of variables approaches the Normal distribution, for sufficiently large sample size.

3(b)

i.

- Mean and SD of Box = 0.5

```
## Warning: package 'DiagrammeR' was built under R version 3.4.3
```

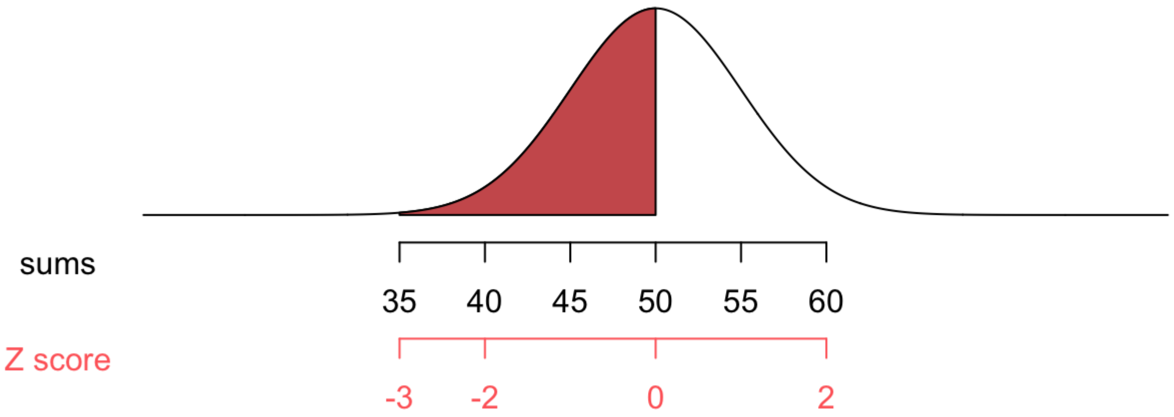


ii.

- Expected Value = $100 \times 0.5 = 50$

- Standard Error = $\sqrt{100} \times 0.5 = 5$.

iii.



As the area between 35 and 50 is exactly 3 SEs from the mean, we know it is approximately 50% (or 99.7%/2) of the area (property of Normal).

iv. The Central Limit Theorem requires a sufficiently large sample size. Here we have 100 draws and the box is already symmetric, making it very appropriate.

3(c)

i. We want a representative sample of Australians who commute, hence if possible we would select a random sample across all of Australia. Eg Randomly choose mailing addresses and contact residents. Or we could do a survey at public transport hubs.

ii. A phone poll could be biased towards those with are home and don't commute. Finding those who are late could be problematic, especially as they probably wouldn't be keen to stop and be surveyed!

iii. Eg Are Sydney trains usually late?

Part	Significant Error	Outcome
3(a)	1	Correct reason
3(b)(i)	1	Correct drawing showing box, sample and draws
3(b)(ii)	1	Correct working

Part	Significant Error	Outcome
3(b)(iii)	1	Correct reasoning from sketch
3(b)(iv)	1	Correct reasoing
2(c)(i)	1	1 propsed method
2(c)(ii)	1	2 suggestions
2(c)(iii)	1	1 question with bias
Total	8	

Question 4

4(a)

i. Skewed data, so need for transformation to symmetrise.

ii.

H: H_0 : There is no difference between tap-ons for ferry and light rail (in log units) vs H_1 : There is a difference.
A: Are the 2 samples independent? Are the 2 populations Normal?
T: $\mathrm{t} = 8.3386$
P: p-value < 2.2e-16
C: Assuming $\alpha = 0.05$, we would strongly reject H_0 in favour of H_1. Hence we conclude that there appears to be significant difference between the 2 modes of transport in terms of tap-ons.

4(b)

H: H_0 : bus:ferry:lightrail:train tap-offs are in the ratio 40:5:5:50.vs H_1 : Not H_0
A:
T: $\mathrm{chi-squared} = 5.7886$
P: p-value = 0.1224
C: Assuming $\alpha = 0.05$, we would retain H_0. Hence we conclude that the model seems to fit.

4(c) More people forget to tap-off on ferry and train.

Part	Significant Error	Outcome
4(a)(i)	1	Correct suggestion
4(a)(ii)	3	Correct hypothesis, p-value and conclusion
4(b)	3	Correct hypothesis, chisquared-value and conclusion
4(c)	1	Insightful observation, like suggestion
Total	8	

Question 5

Part	Significant Error	Outcome
	1	Client: must be relevant to data
	1	Purpose: must be relevant to data and client
	1	Limitations: 2 sensible suggestions eg possible missing data, would be nice to have more demographics on commuters etc
	2	Insights: 2 discoveries relating to the Opal data, referred to in previous parts of questions, or extension or previous parts
	1	Action: linked to client and 1+ of the insights
	1	Communication: well written: easy to read, correct grammar and spelling
	1	HD standard: exceptional insights (insight is a synthesis of parts of previous questions, rather than just listing a previous result
Total	8	