STATISTICAL METHODS

CIS4027-N-BF1-2018

THOMAS RUDDOCK

Q5114161

8TH JANUARY 2018

Acknowledgments

I would like to thank Claudio Angione and Elisabeth Yaneske. For helping me learn a new module and concepts within that module. They have been very helpful throughout the course and helped me make sure I understand the topics which have been taught. It is understandable that I wouldn't be able to understand all aspects of the topic and which I don't. But will still read up and complete examples in my own time after the semester is over.

Table of Contents

Acknowledgments 2
Qualitative statistics: Thematic analysis 3
1 Table for comments (20), key themes, number of times they occur. 3
2 Description of the themes identified 5
Probability and statistics fundamentals (module power point) 6
3 Define the two statements 6
4 Derive the final formula of Bayes' theorem and describe an example where it can be used. 6
5 Describe the different scales of measurements provide two examples for each scale 7
6 Game, three small boxes on the table 7
Central tendency and variability 8
7 The 2010 Salaries of the white house staff 8
Statistical tests 10
8 A variable X follows a Normal distribution with mean 1 and standard deviation 2. What is the probability P(X<0)? 10
9 You would like to test whether an herb works for the treatment of insomnia. 100 people volunteered to take part in the study (gather information on after herd dosage) 10
10 Use hint (t-test, r studio) 10
Regression (power point E) 11
11 Changing from box office to budget 11
12 What if we want to use linear regression? 12
13 And does it perform well 12

Qualitative statistics: Thematic analysis

1 Table for comments (20), key themes, number of times they occur.

In this sub-section the table below will display comments that are randomly picked from the URL: **https://www.theguardian.com/commentisfree/2018/dec/07/social-media-teenagers-problems-banning-phones-children-support.** It will also display the owner of the comment, themes that come from the comment. Then a second table below will show the key themes from the previous table and the amount of times they are displayed.

No.	Comment	Owner of Comment	Themes
1	Social media, like drink, drugs and a fried breakfast, is fine in moderation	Axel Seaton	Moderation, 18+ topics, social media, correct use
2	The cycle of consumerism.	vammyp	Capitalist
3	Look here, headline-writer, I don't think I was going to demonise social media anyway, but I object to your telling me not to. Don't be so bossy.	MichaelBulley	Social media, demonise, object to telling me what to do, oppose headline writer
4	Absolutely. The antidote to misuse is not disuse but correct use	SignificantOther	Antidote, correct use, misuse, certainty
5	That's a fair point which is true of almost any technology, but I think the problem with a smartphone is that it's always with us. Add social media to that and I can see how some people find that they can't escape.	JohnI	Truth, fair point, problem, smartphone, social media, escape
6	Posting here is no different to social media	Bopstar	Social media, no difference
7	I thought it exacerbated problems? I'm 30 and glad it wasn't around when I was growing up (13-18 years old).	Salcombe	Age
8	There are far too many people walking around like Zombies with their phones in hand. It's time to develop social skills and less social media.	bigsands	Zombies with their phones, social skills, less social media
9	Correct. Unfortunately, those of us that can moderate have to suffer the consequences of those that can't.	Right89	Correct, can moderate, suffer the consequences
10	So, the cause is the solution?	Initalyperora	The solution, the cause
11	I think that's what Homer Simpson says about beer.	Bluejay2011	Homer Simpson, beer
12	I'm not sure there's a world of difference. It's still about sharing your view of the world.	Bopstar	Wold, difference, your view
13	To paraphrase Homer Simpson (he was talking about alcohol) Social media - the cause of and answer to all the world's problems.	Colonelhackney	Homer Simpson, alcohol, worlds problems, social media
14	There's an article here today about the "Sharp rise in number of young people seeking help for anxiety". I don't believe there's any help "teens with problems" need that can't be better provided in the real world rather than through "social media".	taninfan	Age, anxiety, help, teens with problems, social media, real world
15	Ridiculous generalisation of not one but two generations. Bravo. The link between using	KrisPWales	Generalisation, technology, crumbling under pressure,

	technology and "crumbling under pressure" was particularly baffling.		
16	Never trust any business that presents itself as a 'community'	geoffhoppy	Never trust, community, presents
17	Ha, and I thought social media was the cause not the remedy	1Waffle12	social media
18	The pros of the entire internet are far outweighed, in my opinion, by its cons. Unplug it.	Choller21	unplug, entire internet,
19	Ah, social media. The environment to convince yourself you are important in a world where, you are just another exploitable dot on the landscape. A global disconnect.	nihilist	Social media, environment, exploitable dot, landscape, global disconnect
20	The cause and not the solution to most of a teenager's problems.	Therebelalliance	The cause, the solution, teen with problems

Theme	**Count**
Social media	8
Moderation	1
18+ topics (beer, drink, drugs)	3
Correct use	3
Capitalist	1
Antidote	1
Misuse	1
Certainty	1
truth	1
Fair point	1
Problem	4
Smartphone	1
escape	1
No difference	2
Age	2
glad it wasn't around	1
Zombies with their phones	1
Social skills	1
Less social media	1
can moderate	1
suffer the consequences	1
The solution	2
The cause	3
Homer Simpson	2
Worlds problems	1
anxiety	1
Help	1
Teen with problems	2
Real world	1
Generalisation	1
Technology	1
crumbling under pressure	1
Never trust	1
presents	1
community	1
unplug	1
entire internet	1
environment	1

exploitable dot	1
landscape	1
global disconnect	1

2 Description of the themes identified

In this sub-section the table below will display the themes that where identified and have key features that will make the theme more recognisable by using describing words to fit the theme. Some themes will be in the same box section as they talk about the same theme.

Theme	Key Features
Social media, less social media	Online communications
Moderation, can moderate	Online regulations
18 + topics (beer, drink, drugs, alcohol)	Adult talking points
Correct use, the solution,	Solutions for the issue
Capitalist	Economy structure
Antidote, help	Fix issues
Misuse, the cause, suffer the consequences, problem	Acknowledging the intent
Certainty, truth	100% Certain
Fair point	Know the other side
Smartphone, technology	Physical Hardware
Escape, unplug	Trapped
No difference, generalisation	Uninformed
Age, glad it wasn't around	Smugness
Zombie with their phones, social skills, anxiety, teen with problems, crumbling under pressure	No conference
Homer Simpson	Cartoon
World problems, real world, environment, landscape, global disconnect	Physical connection
Never trust, exploitable bot	Trust issues
Community, entire internet	All connected
Presents	Wall to cover oneself

Probability and statistics fundamentals

3 Define the two statements

Given the event E: "Tonight it will start raining at 11pm", define:

- **An event F_1 such the E and F_1 are independent**
  P (E and F_1) = P (E) * P (F_1)
  P (E ∩ F_1) = P (E) * P (F_1)

- **An event F_2 such the E and F_2 are dependent**
  P (E and F_2) = P (E) * P (F_2|E)
  P (E ∩ F_2) = P (E) * P (F_2|E)

4 Derive the final formula of Bayes' theorem and describe an example where it can be used.

Final Formula of Bayes' theorem

$$P(A|B) = \frac{P(B|A)\,P(A)}{P(B)}$$

Bayes' Theorem where it can be used example

We can use the Bayes' theorem to find out a patient's probability of having liver disease if they are alcoholic.

- A will be the event meaning "patient has liver disease" we will use past data that tells us that 10% of people entering a clinic have liver disease. Which will look like this: P (A) = 0.10.
- B will mean that the litmus test that "patient is an "alcoholic." 5% of all the clinics patients are alcoholics, will look like this: P (B) = 0.05.
- We also know that among these patients diagnosed with liver disease, 7% are alcoholics. This is the B|A. Which is the probability that the patient is alcoholic, given that they have liver disease, which is 7%.

From this the Bayes' Theorem will tell use:

$$P(A|B) = \frac{(0.07*0.1)}{0.05} = 0.14$$

Which translates to, if the patient is an alcoholic, their chances of having liver disease is 0.14 (14%). This is an increase from the 10% suggested by the past data. But it's still unlikely that any of the particular patients has liver disease.

5 Describe the different scales of measurements provide two examples for each scale

There are 4 different scales of measurements, they will be listed below with a description of each one and they will all have 2 example of each scale.

- Nominal Scale
  Nominal scale are used to label variables, which don't have any quantitative value and have no particular relationship between the variables for example:
  1. What is your eye colour?
     (Blue, green, brown, hazel, etc)
  2. What is your gender?
     (Male, female)
- Ordinal Scale
  With Ordinal scales, the order of the values which are significant and important. By ordering the observations from low to high which have any ties to attributes to the lack of measurement sensitivity. Examples below:
  1. How satisfied are you with our teaching?
     (Very unsatisfied, unsatisfied, neutral, satisfied and very satisfied)
  2. Have you finished the assignment?
     (Haven't started, started, half way through, almost finished, completed)
- Interval Scale
  The Interval scales are a numeric scale, where we both know the order and the exact differences between the values. The variables doesn't have a "natural" zero value. Examples below:
  1. Celsius temperature every day for the month of December?
     (2.0, 4.0,-1.2, 1.0, 3.0, etc)

  2. What is your family income?
     (Less than 20K, 21k to 30K, 31K to 40k, 41K plus)
- Ratio Scale
  The ratio scale will tell us about the order, they tell us the exact value between units, and they have an absolute zero which allows for the wide range of descriptive and inferential statistics to be applied. These variables can be meaningfully added, subtracted, multiplied, and divided. Examples below:
  1. What is your current height?
     (less than 5 feet, 5,1 to 5,5, 5,6 to 6, more than 6 feet)
  2. What is your current weight in kilograms?
     (Less than 50 KG, 51 to 70 KG, 71 to 90 KG, 91 to 110 KG, more than 110 KG)

6 Game, three small boxes on the table

$$P(A|C) = \frac{P(C|A)\,P(A)}{P(C)} \qquad P(1/3|2/3) = \frac{P(2/3|1/3)\,P(1/3)}{P(2/3)}$$

No, I switch to box C because if I didn't switch the chance of winning is 1/3 and switch gives me a chance of 2/3. Meaning always switch gives the player better odds of winning.

Central tendency and variability

7 The 2010 Salaries of the white house staff

Preforming a pipeline of the data in the excel sheet in RStudio first bit was the central tendency. This includes the mean, mode and median. I started off with importing the files into RStudio then set a new directory in my ICA folder. And used WHS to be the 2010 white house staff document.

Step one.

```
> WHS <- read.csv("whiteHouseStaff.csv")
```

This will read the excel file which was saved as a .csv and use WHS as a quicker way to type stuff out.

Step 2.

```
> WHS
```

This will show all the data in RStudio at a table.

Step 3.

```
> WHS[WHS$Employee.Status == "Employee",c("Employee.Name", "Salary"")]
```

This will display the employees name and salary.

Step 3.

```
> mean(WHS$Salary)
[1] 82721.34
```

This line will show the mean of the salary, which is 82721.34. It adds all the salaries and divided them by how many they are, which got the result.

Step 4.

```
> median(WHS$Salary)
[1] 66300
```

This line above will show the median for all the salaries on the excel file, which is 66300. RStudio gets this by sorting out the number in order and finds the middle number, which got the result.

Step 5.

```
temp <- table(as.vector(whs))
```

This creates a new table and will allow me to focus on the salaries line. Then in console type “temp”.

```
> temp

     0  21000  37826  37983  42000  42738  42840  43656  43860  44402  45000  45594  45900  46745  47500  47532  48095
     3      1      1      1     64      1      1      1      2      3     31      1     15      1      1      1      1
 48450  49000  49069  50000  51000  51630  53550  55000  55080  56092  56100  56791  57000  58511  59160  60000  60232
     1      1      1     16      8      1      1      9      2      2     10      1      1      4      1     12      5
 61200  62000  62500  62544  63240  63673  64439  64548  65000  65393  66000  66300  66335  68230  70000  70126  71400
     6      1      1      1      2      1      2      1      3      1      1     12      1      1      4      1      9
 72000  72876  73917  74958  75000  75480  76500  78000  79560  79864  80000  81600  84855  85000  85680  89033  89846
     1      1      1      2      4      1      5      7      1      1      7      1      2      2      2      4      1
 90000  91800  92001  92341  93840  94969  96900  97936  99000 100000 100904 102000 102829 105211 106839 107770 110000
     4      4      3      1      3      2      2      1      3     10      2      9      1      1      1      1      2
110500 112774 113000 113605 114000 115000 115731 120000 122744 123758 126251 129758 130000 130500 132009 136134 139500
     1      1      5      1      3      1      1      8      1      1      1      1      6     27      1      1      1
140000 140259 144868 145000 147500 148510 149000 150000 153300 153500 155500 158500 162500 162900 165000 172000 172200
     2      1      1      1      4      2      2      3      1      2      5      9      1      1      1      1     23
179700
     2
```

It will display all salaries and the number of times they are displayed. Then type shown below.

```
names (temp)[temp == max (temp)]
```

This will then spit out the mode in the console which is 42000 which is displayed 64 times

```
> names (temp)[temp == max (temp)]
[1] "42000"
```

Step 6.

```
> range(WHS$Salary)
[1]      0 179700
```

Range is to find the difference between the lowest and highest in this instance is 0.179700 this because some staff have no salaries.

Step 7.

```
> quantile(WHS$Salary, 0.10)
  10%
42000
```

The quantile of the first 10% is 42000.

Step 8.

To find the IQR is to find the 25% and the 75% and the median between them to, the code used is below,

```
> quantile(WHS$Salary, 0.25, 0.75)
  25%
45900
```

```
> quantile(WHS$Salary, 0.75)
   75%
113000
```

We then did the function for the Interquartile range (IQR). Which is 67100. Code below.

```
> IQR (WHS$Salary)
[1] 67100
```

Statistical tests

8 A variable X follows a Normal distribution with mean 1 and standard deviation 2. What is the probability P(X<0)?

The probability is P (X<0.3156) 31.56%

And in the standard normal probabilities table

P (X<-0.48) 48%

9 You would like to test whether an herb works for the treatment of insomnia. 100 people volunteered to take part in the study (gather information on after herd dosage)

- Design the experiment and define what could be a null and alternative hypothesis in this case.

Since we have 100 people as a solid number we can split that in half and have two test group which one will have a correct dosage called group A and the other half group B won't have any herbs just to go to bed as normal to get a solid base line of what insomnia is. We will ask the volunteers to record their sleep patterns and hours either on a mobile phone app or notepad, whichever is better for them. This test will go on for a month to have 30 days of evidence. We then gather the evidence in the two groups and plot their sleeping patterns in groups, then check to see if it has improved over the 30 days test.

The null hypothesis for this is if group A has nothing happening to improve the treatment of insomnia (H_0).

The alternative hypothesis for this test will be that Group A has better data than Group B on improving insomnia (H_1).

- Describe in general what could be an error of type I or type II.

We than can reject the null hypothesis when the herbs have no effect on Group A, this is a false positive result (type I error). Next, we fail to reject the null hypothesis when there is a genuine effect like Group B is having a better sleep than the Group A, this a false negative result (type II error).

10 Design and perform a statistical test to check if this is a statistically significant reduction, or the reduction in performance is just due to chance. (One-sample T-test)

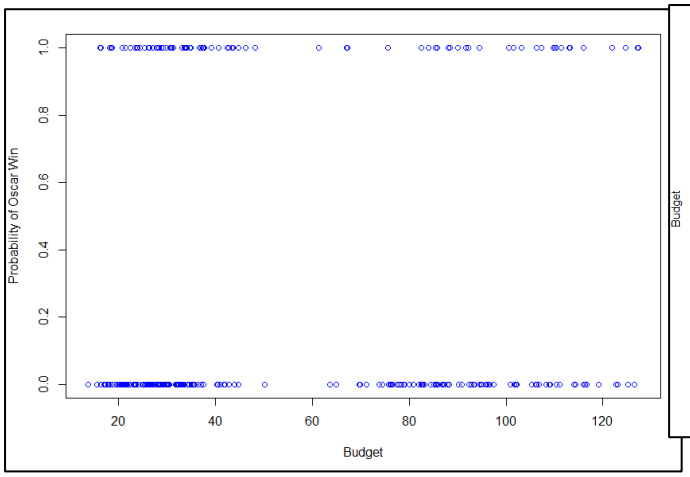
(I forgot to do it, must have missed it or thought I did it.)

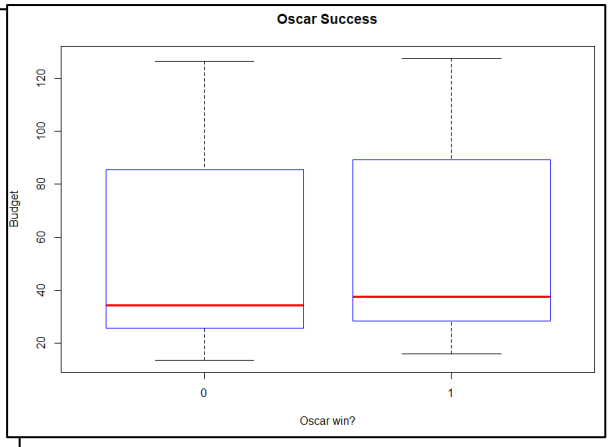
The assumptions that would be needed to use a Z-text is if the main population is unknown for example the manufacture don't know how many CPU's have been created, but since they are all accounted for the Z-test is not needed.

Regression

11 Changing from box office to budget

From the box plot and a probity plot image it shows that there is no reason to think having a larger budget will more likely get the movie an Oscar, a larger number of movies are on the lower even of the budget, which happens to be under 40 million.





Using the logit scores which gets the z and value

```
> summary(fullboxofficemodel)

Call:
glm(formula = boxOffice$Oscar ~ ., family = binomial, data = boxOffice)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.9831  -0.8079  -0.5245   0.9991   2.3983

Coefficients:
              Estimate Std. Error z value Pr(>|z|)
(Intercept)  -6.981440   1.481505  -4.712 2.45e-06 ***
BoxOffice     0.016751   0.003449   4.857 1.19e-06 ***
Budget        0.017038   0.015759   1.081   0.2796
CountryEurope 0.914720   1.388378   0.659   0.5100
CountryIndia -0.004290   1.831527  -0.002   0.9981
CountryOther  1.803408   1.563432   1.153   0.2487
CountryUK     2.523901   1.143281   2.208   0.0273 *
Critics       0.005410   0.007346   0.737   0.4614
Length        0.025874   0.014031   1.844   0.0652 .
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 374.60  on 299  degrees of freedom
Residual deviance: 316.97  on 291  degrees of freedom
AIC: 334.97

Number of Fisher Scoring iterations: 4
```

The Z-value is 1.081

```
Budget          0.017038    0.015759    1.081    0.2796
```

And the P-value is 0.2796

We accept the null hypothesis as it shows a low budget movie (40 mill) is more likely to win an Oscar over a high budget movie. This shows that there is connection between low budget and winning an Oscar.

12 What if we want to use linear regression?

The Linear regression will show this to be more true as when the scatter plots are added it will show more plots under the 40 million zone and stat to thin out when more money is spent, and yes they will be out liars which have very high budgets, but that doesn't not skew the results too much.

13 And does it perform well

Yes it will perform well by backing up the result that lower budgets movies ae more likely to get an Oscar than high budgets.