Module 1 report

470518795

13/08/2019

1: Setup

```
library(tidyverse)
library(janitor)
```

```
# import data and identify missing values
mydata=readr::read_csv("data/mydata.csv",na=c("NA","","  ","n/a"))
```

```
#cleans variable names
mydata=mydata %>% janitor::clean_names()
```

```
# NZ government collects data using census
glimpse(mydata)
```

```
## Observations: 200
## Variables: 78
## $ year
<dbl> …
## $ region
<chr> …
## $ gender
<chr> …
## $ age
<dbl> …
## $ country
<chr> …
## $ new_zealand_european
<chr> …
## $ maori
<chr> …
## $ samoan
<chr> …
## $ cook_islands_maori
<chr> …
## $ tongan
<chr> …
## $ niuean
<chr> …
## $ chinese
<chr> …
## $ indian
<chr> …
## $ other_ethnicity
<chr> …
## $ languages_spoken
<dbl> …
## $ eye_colour
<chr> …
## $ handedness
<chr> …
## $ height
<dbl> …
## $ right_foot_length
<dbl> …
## $ wrist_circumference
<dbl> …
## $ left_thumb_circumference
<dbl> …
## $ travel_method_to_school
<chr> …
## $ travel_time_to_school
<dbl> …
## $ bag_weight
<dbl> …
## $ litter_in_lunch
<chr> …
## $ fruit_vegetables_in_lunch
<dbl> …
## $ memory_time
<dbl> …
## $ reaction_time
```

```
<dbl> …
## $ time_standing_on_left_leg
<dbl> …
## $ physical_activity_before_school
<chr> …
## $ physical_activity_at_school
<chr> …
## $ physical_activity_after_school
<chr> …
## $ physical_activity_on_the_weekend
<dbl> …
## $ scheduled_activities_in_last_week
<dbl> …
## $ screen_time_after_school
<dbl> …
## $ favourite_video_game
<chr> …
## $ own_cell_phone
<chr> …
## $ facebook_account
<chr> …
## $ instagram_account
<chr> …
## $ snapchat_account
<chr> …
## $ reddit_account
<chr> …
## $ you_tube_channel
<chr> …
## $ technology_none_of_these
<chr> …
## $ check_messages_as_soon_as_you_wake_up
<chr> …
## $ respond_to_messages_immediately
<chr> …
## $ take_phone_to_school
<chr> …
## $ lose_focus_as_school_due_to_phone
<chr> …
## $ feeling_without_phone_angry
<chr> …
## $ feeling_without_phone_anxious
<chr> …
## $ feeling_without_phone_frustrated
<chr> …
## $ feeling_without_phone_happy
<chr> …
## $ feeling_without_phone_lonely
<chr> …
## $ feeling_without_phone_relieved
<chr> …
## $ feeling_without_phone_sad
<chr> …
## $ feeling_without_phone_neutral
<chr> …
## $ screen_time_opinion_on_your_phone
<chr> …
## $ screen_time_opinion_on_social_media
```

```
<chr> …
## $ screen_time_opinion_playing_video_games
<chr> …
## $ bed_time
<drtn> …
## $ wake_time
<drtn> …
## $ sleep_time
<dbl> …
## $ time_you_get_home_from_school
<drtn> …
## $ time_you_ate_dinner
<drtn> …
## $ climate_change_opinion
<chr> …
## $ how_true_i_get_carried_away_by_my_feelings
<dbl> …
## $ how_true_i_say_the_first_thing_that_comes_into_my_mind_without_thinking_enough_a
bout_it <dbl> …
## $ how_true_i_cant_stop_myself_from_doing_something_even_if_i_know_it_is_wrong
<dbl> …
## $ how_true_i_try_to_talk_out_a_problem_instead_of_fighting
<dbl> …
## $ how_true_it_is_easy_for_me_to_make_friends
<dbl> …
## $ how_true_i_know_how_to_stand_up_for_myself_without_being_mean
<dbl> …
## $ how_wrong_drink_alcohol
<dbl> …
## $ how_wrong_smoke_tobacco_cigarettes
<dbl> …
## $ how_wrong_smoke_e_cigarettes
<dbl> …
## $ how_wrong_smoke_marijuana
<dbl> …
## $ how_wrong_caregivers_parents_drink_alcohol
<dbl> …
## $ how_wrong_caregivers_parents_smoke_tobacco_cigarettes
<dbl> …
## $ how_wrong_caregivers_parents_smoke_e_cigarettes
<dbl> …
## $ how_wrong_caregivers_parents_smoke_marijuana
<dbl> …
```

2: Guiding questions

> **1: Who collects this data and how is it reported?**

CensusAtSchool Project which is hosted by the Departemnt of Statitics at the University of Auckland association with Stats NZ and the Ministry of Education collects data.

The data is reported on the website: https://new.censusatschool.org.nz/ (https://new.censusatschool.org.nz/), which can be used by individuals for multiple purpose: including teaching. The data is available in csv file, which can be directly opened in Excel.

2: How are missing values recorded, and why might they occur? In the analysis below you will need to think about how to handle missing values.

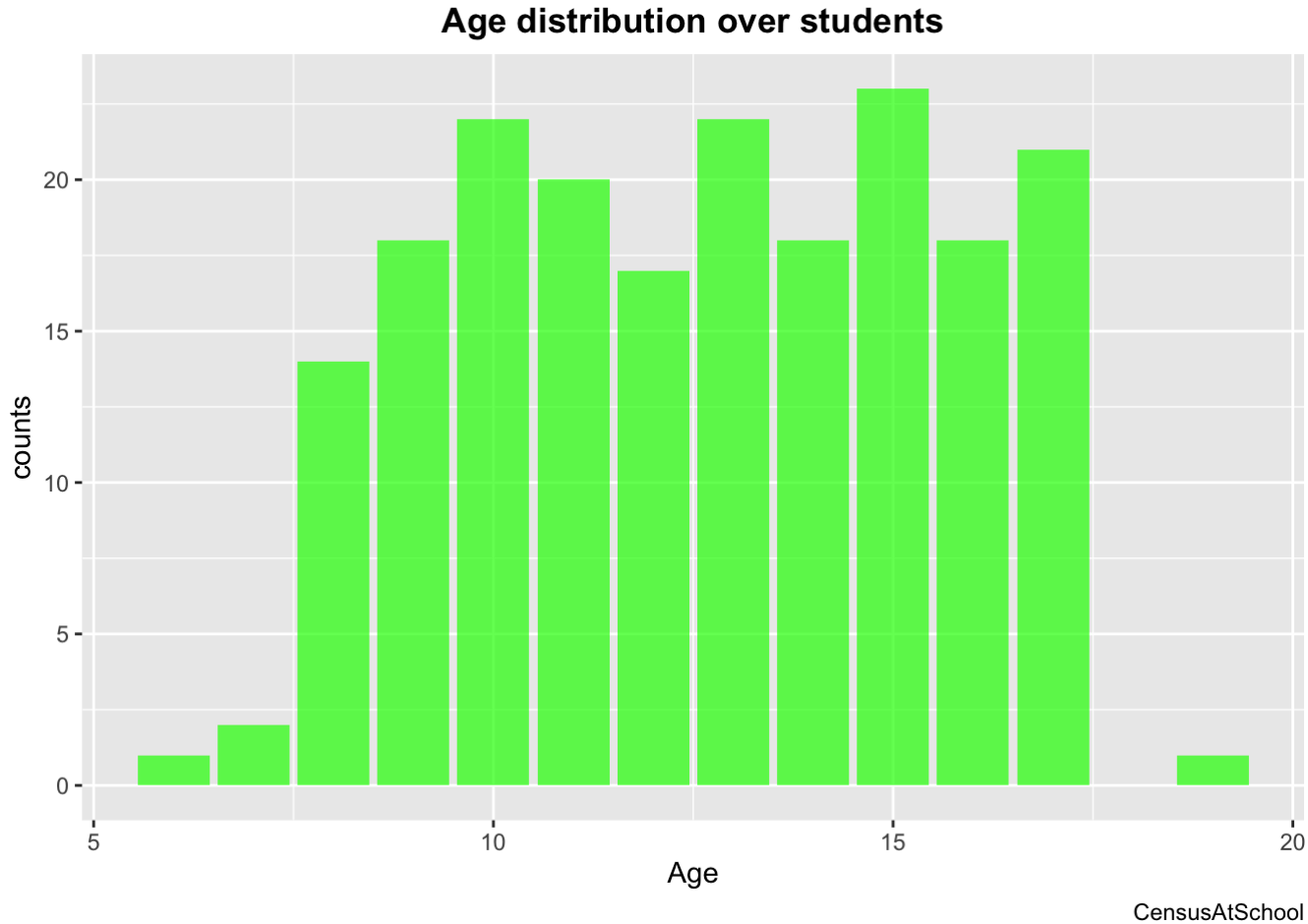
All missing values are being recorded as “NA” or "" or " " or “n/a”.

There are some numeric variables which recorded as missing values because it is likely that the student accidentally ignore the value due to carelessness.

Most missing values are appears in descriptive questions, for example, how wrong…, which is also likely because of ignorance. Moreover, it might because there is no standard answer, so researcher do not need or cannot use this type of data to make their analysis.

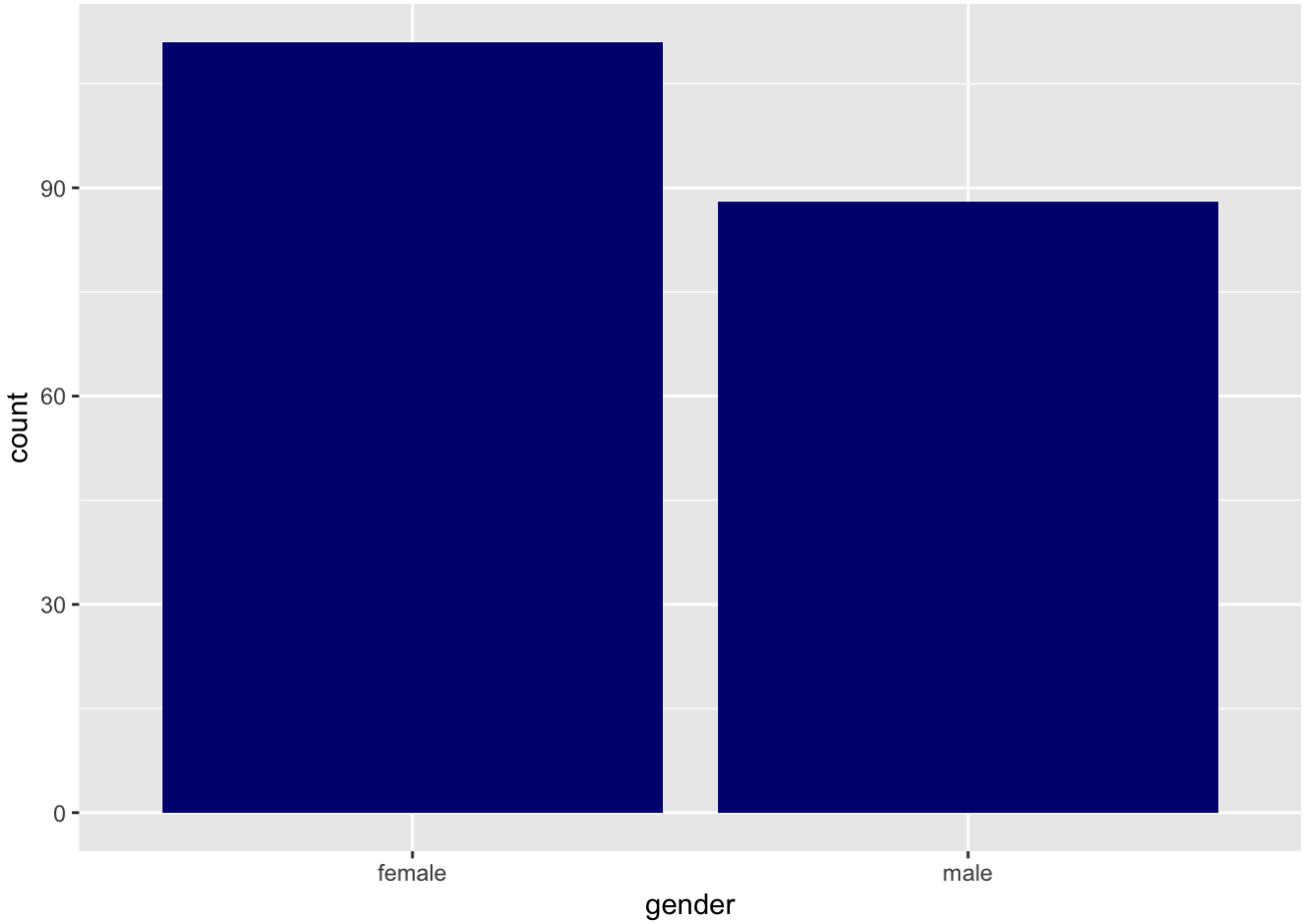
3: Provide some general demographic information about your sample.

```
# age distribution of data
ggplot(data=mydata,aes(x=age)) +
  geom_bar(fill="green",alpha=0.7)+
  labs(x="Age",y="counts",
       title = "Age distribution over students",
       caption = "CensusAtSchool")+
  theme(plot.title = element_text(hjust = 0.5, face = 'bold'))
```



By looking at the age distribution, we can find that most samples from survey are aged between 7 and 18 years old. Merely few outside this range.

```
gender<-mydata %>% drop_na(gender)
ggplot(data=gender,aes(x=gender))+
  geom_bar(fill="navy")
```




```
#sample size
n=count(mydata)
n
```

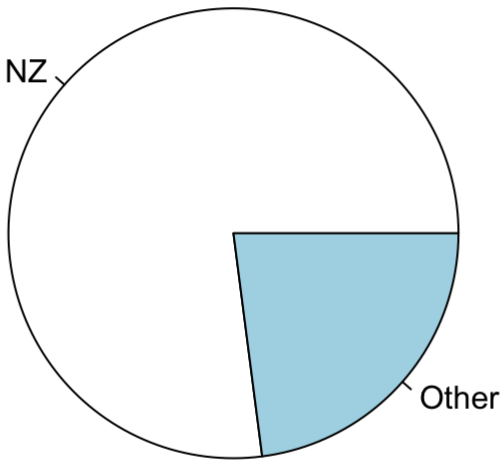
```
## # A tibble: 1 x 1
##       n
##   <int>
## 1   200
```

The sample size is 200. We can see that percentage of female is greater than that of male. There is one sample missing.

```
country<-mydata %>% drop_na(country)
country<-ifelse(mydata$country=="New Zealand","NZ","Other")
table(country)
```

```
## country
##    NZ Other
##   151    45
```

```
pie(table(country))
```



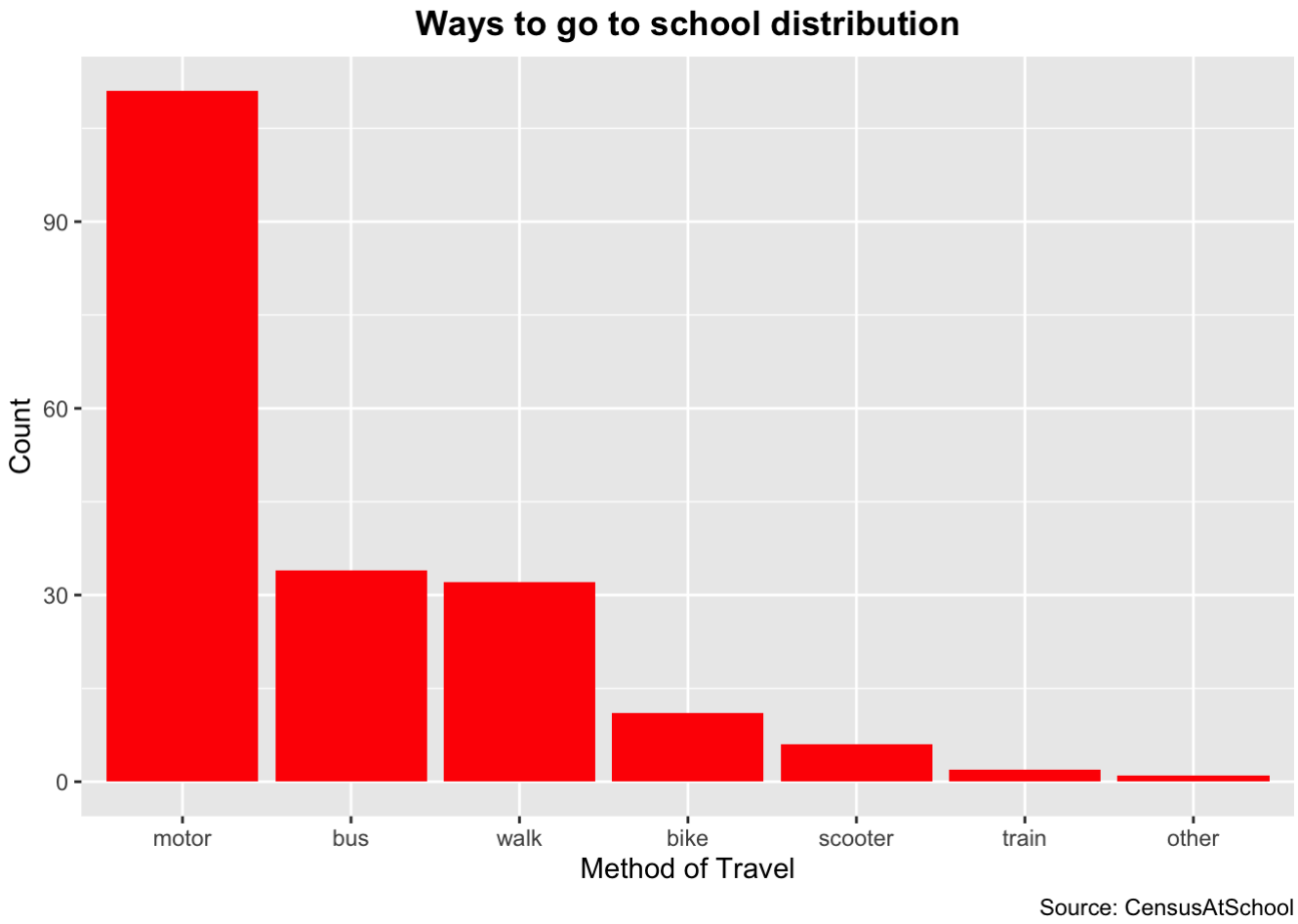
Over the pie chart, we can see that most students are NewZealand citizens, there are about less than 1/4 students from other countries.

4: What is the most common method of travel to school.

```
travel_method<-mydata %>% count(travel_method_to_school,sort=TRUE) %>% drop_na()
travel_method
```

```
## # A tibble: 7 x 2
##   travel_method_to_school     n
##   <chr>                   <int>
## 1 motor                     111
## 2 bus                        34
## 3 walk                       32
## 4 bike                       11
## 5 scooter                     6
## 6 train                       2
## 7 other                       1
```

```
#visualisation
ggplot(travel_method,aes(x=reorder(travel_method_to_school,-n),y=n))+
  geom_bar(fill="red",stat="identity")+
  labs(x="Method of Travel",y="Count",
       title="Ways to go to school distribution",
       caption = "Source: CensusAtSchool")+
  theme(plot.title = element_text(hjust = 0.5, face = 'bold'))
```



It is obvious that the most common way to travel to school is using motor through data and graph.

5:What are the most common favourite video games? You may want to use forcats::fct_lump() to group the least common games together.

```
video_games<-mydata %>% group_by(favourite_video_game)%>% summarise(count=n())%>% arr
ange(desc(count)) %>% drop_na()
video_games
```

```
## # A tibble: 67 x 2
##    favourite_video_game          count
##    <chr>                         <int>
##  1 Don't Have One                   62
##  2 Fortnite                         26
##  3 Roblox                           18
##  4 Minecraft                        13
##  5 PlayerUnknown's Battlegrounds     3
##  6 Apex Legends                      2
##  7 Call Of Duty                      2
##  8 Grand Theft Auto 5                2
##  9 Nba 2k19                          2
## 10 Red Dead Redemption 2             2
## # … with 57 more rows
```

From the table, it is clear that the most common video game is Don't Have One. This is the most popular games among our student samples.

6: It is hypothesised that 90% of the population are right handed. Does your sample of data support this hypothesis?

```
# set a dataframe for handedness
handedness<-recode(mydata$handedness, right ="right",left ="left",.default=NA_charact
er_)
handedness<-table(handedness)
handedness
```

```
## handedness
##  left right
##    26   159
```

```
handed<-as.data.frame(handedness)
n=sum(handed$Freq)
n
```

```
## [1] 185
```

To test whether the 90% population are right-handed.

1: $H_0 : p_r = 0.9$ vs H_1 : The percentage of population using right hand is not equal to 90%.

2: Assumption: $X_1,…,X_n N(0.9,variance)$.

3: Test statistic: $T = (\bar{X} - 0.9)/(S/\sqrt{N})$ Under H_0, $T t with df=n-1 $ approx.

4: Observed test statistic: $t_0 = 3.37$.

5: p-value: $P(t\ with\ df = n - 1 >= t_0) = 0.06$.

6: Decision: Since the p-value is approximately 0.06, we do not reject the null hypothesis at the 5% significance level. We conclude that the proportion of population using right hand is equal to 90%.

```
p=c(0.1,0.9)
e=p*n
e
```

```
## [1]  18.5 166.5
```

```
t0=sum(((handed$Freq-e)^2)/e)
t0
```

```
## [1] 3.378378
```

```
(pval=1-pchisq(t0,df=1))
```

```
## [1] 0.06605702
```

7: Is handedness independent of gender?

1: $H_0 : p_{ij} = p_{i\bullet}p_{\bullet j}$ for $i = 1, 2$ and $j = 1, 2$ vs H_1 : At least one of the equalities does not hold.

2: Assumption: $e_{ij} = y_{i\bullet}y_{\bullet j}/n \geq 5$

3: Test statistic: $T = \sum_{i=1}^{2}\sum_{j=1}^{2}\frac{(y_{ij}-e_{ij})^2}{e_{ij}}$ Under H_0, $T \sim \chi_1^2$ approx.

4: Observed test statistic: $t_0 = 0.53$.

5: p-value: $P(\chi_1^2 \geq 0.53) = 0.46$.

6: Decision: Since the p-value is approximately 0.46, we do not reject the null hypothesis at the 5% significance level. We conclude that gender and handedness are independent.

```
mydata$handedness=recode(mydata$handedness,right="right",left="left",.default = NA_ch
aracter_)
mydata$gender=recode(mydata$gender,female="female",male="male",.default = NA_characte
r_)
table_gender_handedness<-table(mydata$handedness,mydata$gender)
table_gender_handedness
```

```
##
##         female male
##   left      13   12
##   right     95   64
```

```
r=c=2
(yr=apply(table_gender_handedness,1,sum))
```

```
##  left right
##    25   159
```

```
(yc=apply(table_gender_handedness,2,sum))
```

```
## female   male
##    108     76
```

```
(yr.mat = matrix(yr, r, c, byrow = FALSE))
```

```
##      [,1] [,2]
## [1,]   25   25
## [2,]  159  159
```

```
(yc.mat = matrix(yc, r, c, byrow = TRUE))
```

```
##      [,1] [,2]
## [1,]  108   76
## [2,]  108   76
```

```
# calculate expected value
(ey.mat = yc.mat * yr.mat / sum(table_gender_handedness))
```

```
##          [,1]     [,2]
## [1,] 14.67391 10.32609
## [2,] 93.32609 65.67391
```

```
# check the assumption
all(ey.mat >=5)
```

```
## [1] TRUE
```

```
# calculate test statistic
(t0 = sum((table_gender_handedness - ey.mat)^2 / ey.mat))
```

```
## [1] 0.5349889
```

```
#calculate p-value
(pval = pchisq(t0, 1, lower.tail = FALSE))
```

```
## [1] 0.4645169
```

```
# double check
chisq.test(table_gender_handedness,correct=FALSE)
```

```
##
##  Pearson's Chi-squared test
##
## data:  table_gender_handedness
## X-squared = 0.53499, df = 1, p-value = 0.4645
```

8: What proportion of students own a cell phone? Is this proportion constant across the different year groups? Perform a test to see if there is a statistically significant difference in cell phone ownership across year groups.

1: $H_0 : p_{1j} = p_{2j} = p_{3j} = \ldots = p_{10j}$ for $j = 1, 2$ vs H_1 : Not all equalities holds.

2: Assumption: $e_{ij} = y_{i\bullet}y_{\bullet j}/n \geq 5$

3: Test statistic: $T = \sum_{i=1}^{10}\sum_{j=1}^{2}\frac{(y_{ij}-e_{ij})^2}{e_{ij}}$ Under H_0, $T \sim \chi^2_9$ approx.

4: Observed test statistic: $t_0 = 83.84$.

5: p-value: $P(\chi^2_1 \geq 83.84) =$ 2.775558e-14.

6: Decision: Since the p-value is approximately 0, we do reject the null hypothesis at the 1% significance level. Proportion of students own a cell phone is constant across the different year groups.

```
table(mydata$own_cell_phone)
```

```
##
##  no yes
##  62 134
```

```
table_cell_phone_age<-table(mydata$year,mydata$own_cell_phone)
table_cell_phone_age
```

```
##
##      no yes
##   4  16   4
##   5  12   8
##   6  15   5
##   7   9  11
##   8   6  13
##   9   0  19
##   10  1  19
##   11  3  15
##   12  0  20
##   13  0  20
```

```
n=sum(table_cell_phone_age)
(yr=apply(table_cell_phone_age,MARGIN = 1,FUN=sum))
```

```
##  4  5  6  7  8  9 10 11 12 13
## 20 20 20 20 19 19 20 18 20 20
```

```
(yc = apply(table_cell_phone_age, MARGIN = 2,FUN = sum))
```

```
##  no yes
##  62 134
```

```
(yr.mat = matrix(yr, nrow = 10, ncol = 2,
                 byrow = F))
```

```
##       [,1] [,2]
##  [1,]   20   20
##  [2,]   20   20
##  [3,]   20   20
##  [4,]   20   20
##  [5,]   19   19
##  [6,]   19   19
##  [7,]   20   20
##  [8,]   18   18
##  [9,]   20   20
## [10,]   20   20
```

```
(yc.mat = matrix(yc, nrow = 10, ncol = 2,
                 byrow = T))
```

```
##       [,1] [,2]
##  [1,]   62  134
##  [2,]   62  134
##  [3,]   62  134
##  [4,]   62  134
##  [5,]   62  134
##  [6,]   62  134
##  [7,]   62  134
##  [8,]   62  134
##  [9,]   62  134
## [10,]   62  134
```

```
(etab = yr.mat * yc.mat / n)
```

```
##           [,1]     [,2]
##  [1,] 6.326531 13.67347
##  [2,] 6.326531 13.67347
##  [3,] 6.326531 13.67347
##  [4,] 6.326531 13.67347
##  [5,] 6.010204 12.98980
##  [6,] 6.010204 12.98980
##  [7,] 6.326531 13.67347
##  [8,] 5.693878 12.30612
##  [9,] 6.326531 13.67347
## [10,] 6.326531 13.67347
```

```
# check the assumption
etab >= 5
```

```
##       [,1] [,2]
##  [1,] TRUE TRUE
##  [2,] TRUE TRUE
##  [3,] TRUE TRUE
##  [4,] TRUE TRUE
##  [5,] TRUE TRUE
##  [6,] TRUE TRUE
##  [7,] TRUE TRUE
##  [8,] TRUE TRUE
##  [9,] TRUE TRUE
## [10,] TRUE TRUE
```

```
# calculate test statistics
(t0 = sum((table_cell_phone_age - etab)^2/etab))
```

```
## [1] 83.84422
```

```
# p-value
(p.value = 1 - pchisq(t0, 9))
```

```
## [1] 2.775558e-14
```

```
# double check
chisq.test(table_cell_phone_age,correct=F)
```

```
##
##  Pearson's Chi-squared test
##
## data:  table_cell_phone_age
## X-squared = 83.844, df = 9, p-value = 2.775e-14
```

9: Restricting attention to students in years 7 to 12 who own a cell phone, is there an association between the tendency to check messages as soon as they wake up and feeling anxious when they're without their phone. Do you get the same answer if you use a Monte Carlo p-value calculation?

1: $H_0 : p_{1j} = p_{2j}$ for $j = 1, 2$ vs H_1 : Not all equalities holds.

2: Assumption: $e_{ij} = y_{i\bullet} y_{\bullet j}/n \geq 5$

3: Test statistic: $T = \sum_{i=1}^{2} \sum_{j=1}^{2} \frac{(y_{ij} - e_{ij})^2}{e_{ij}}$ Under H_0, $T \sim \chi_1^2$ approx.

4: Observed test statistic: $t_0 = 0.74$.

5: p-value: $P(\chi_1^2 \geq 0.74) = 0.39$.

6: Decision: Since the p-value is approximately 0.39, we do not reject the null hypothesis at the 5% significance level. There is no association between these two variables.

```
mydata7_12<- mydata %>% filter(year>6 & year<13)
mydata7_12$check_messages_as_soon_as_you_wake_up=fct_collapse(mydata7_12$check_messag
es_as_soon_as_you_wake_up,
                                                              yes=c("always","often",
"sometimes"),
                                                              no=c("never","rarely"))
(checkphone_feelinganxious<-table(mydata7_12$check_messages_as_soon_as_you_wake_up, m
ydata7_12$feeling_without_phone_anxious))
```

```
##
##       no yes
##   yes 54  14
##   no  25   3
```

```
(yr=apply(checkphone_feelinganxious,MARGIN = 1,FUN=sum))
```

```
## yes  no
##  68  28
```

```
(yc = apply(checkphone_feelinganxious, MARGIN = 2,FUN = sum))
```

```
##  no yes
##  79  17
```

```
(yr.mat = matrix(yr, nrow = 2, ncol = 2,
                 byrow = FALSE))
```

```
##      [,1] [,2]
## [1,]   68   68
## [2,]   28   28
```

```
(yc.mat = matrix(yc, nrow = 2, ncol = 2,
                 byrow = TRUE))
```

```
##      [,1] [,2]
## [1,]   79   17
## [2,]   79   17
```

```
(etab = yr.mat * yc.mat / n)
```

```
##          [,1]     [,2]
## [1,] 27.40816 5.897959
## [2,] 11.28571 2.428571
```

```
# check the assumption
etab >= 5
```

```
##       [,1]  [,2]
## [1,] TRUE  TRUE
## [2,] TRUE FALSE
```

```
# violate the assumption
chisq.test(checkphone_feelinganxious)
```

```
##
##  Pearson's Chi-squared test with Yates' continuity correction
##
## data:  checkphone_feelinganxious
## X-squared = 0.73584, df = 1, p-value = 0.391
```

```
fisher.test(checkphone_feelinganxious) # only assumes that the row and column are fix
ed
```

```
##
##  Fisher's Exact Test for Count Data
##
## data:  checkphone_feelinganxious
## p-value = 0.3789
## alternative hypothesis: true odds ratio is not equal to 1
## 95 percent confidence interval:
##  0.07886483 1.89001639
## sample estimates:
## odds ratio
##  0.4661785
```

```
rcounts=rowSums(checkphone_feelinganxious)
ccounts=colSums(checkphone_feelinganxious)
B=10000
set.seed(123)
x_list=r2dtable(B,rcounts,ccounts)
rnd.chisq = numeric(B)
for (i in 1:B){
  rnd.chisq[i]=chisq.test(x_list[[i]])$statistic
}
sum(rnd.chisq>0.73584)/B
```

```
## [1] 0.3742
```

```
fisher.test(checkphone_feelinganxious,simulate.p.value=TRUE,B)
```

```
##
##  Fisher's Exact Test for Count Data
##
## data:  checkphone_feelinganxious and B
## p-value = 0.3789
## alternative hypothesis: true odds ratio is not equal to 1
## 95 percent confidence interval:
##  0.07886483 1.89001639
## sample estimates:
## odds ratio
##  0.4661785
```

```
# We can see that the assumption is being violated. Therefore, chi-squared test would
give a inccorect result. Therefore, the chi-squared test is different from the Monte
 Carlo p-value. However, fisher test has no assumption about expected values, hence,
 the fisher test result is consistent with Monte Carlo p-value calculation.

# Decision: By Monte Carlo p-value and fisher test, the alternative hypothesis is not
significant at 5% significance level, we do not reject null hypothesis. Hence, there
 is no association between these two variables.
```

10: What other questions could you ask of this data? Pick one and perform an appropriate test.

Question 10: Is gender independent of own cell phone?

1: $H_0 : p_{ij} = p_{i\bullet}p_{\bullet j}$ for $i = 1, 2$ and $j = 1, 2$ vs H_1 : At least one of the equalities does not hold.

2: Assumption: $e_{ij} = y_{i\bullet}y_{\bullet j}/n \geq 5$

3: Test statistic: $T = \sum_{i=1}^{2}\sum_{j=1}^{2}\frac{(y_{ij}-e_{ij})^2}{e_{ij}}$ Under H_0, $T \sim \chi_1^2$ approx.

4: Observed test statistic: $t_0 = 0.38987$.

5: p-value: $P(\chi_1^2 \geq 0.38987) = 0.5324$.

6: Decision: Since the p-value is approximately 0.5324, we do not reject the null hypothesis at the 5% significance level. We conclude that gender and own_cell_phone are independent.

```
mydata$gender=recode(mydata$gender,female="female",male="male",.default = NA_characte
r_)
mydata$own_cell_phone=recode(mydata$own_cell_phone,yes="yes",no="no",.default = NA_ch
aracter_)
table_gender_phone<-table(mydata$own_cell_phone,mydata$gender)
table_gender_phone
```

```
##
##       female male
##   no      32   30
##   yes     75   58
```

```
r=c=2
(yr=apply(table_gender_phone,1,sum))
```

```
##  no yes
##  62 133
```

```
(yc=apply(table_gender_phone,2,sum))
```

```
## female   male
##    107     88
```

```
(yr.mat = matrix(yr, r, c, byrow = FALSE))
```

```
##      [,1] [,2]
## [1,]   62   62
## [2,]  133  133
```

```
(yc.mat = matrix(yc, r, c, byrow = TRUE))
```

```
##      [,1] [,2]
## [1,]  107   88
## [2,]  107   88
```

```
# calculate expected value
(ey.mat = yc.mat * yr.mat / sum(table_gender_phone))
```

```
##          [,1]     [,2]
## [1,] 34.02051 27.97949
## [2,] 72.97949 60.02051
```

```
# check the assumption
all(ey.mat >=5)
```

```
## [1] TRUE
```

```
# calculate test statistic
(t0 = sum((table_gender_phone - ey.mat)^2 / ey.mat))
```

```
## [1] 0.3898677
```

```
#calculate p-value
(pval = pchisq(t0, 1, lower.tail = FALSE))
```

```
## [1] 0.5323689
```

```
# double check
chisq.test(table_gender_phone,correct=FALSE)
```

```
##
##  Pearson's Chi-squared test
##
## data:  table_gender_phone
## X-squared = 0.38987, df = 1, p-value = 0.5324
```

11: Are there any limitations of this data set?

This dataset contains a lot of aspects about the 200 samples, eg. year, region, country. It gives very detailed description about each sample. However, there are a few missing values in this dataset, which consumes an amount of time to do data cleaning. Several variables, such as how wrong, are totally missing. It doesn't provide any information about the samples and use extra space for these missing value. It is suggested that if these variables are useful for analysis, then make a multiple choice (degree of wrong) to students, then they will easily fill out the survey. If it is not useful for analysis, then just dropping these variables is good for data cleaning.

Reference:

CensusAtSchool New Zealand https://new.censusatschool.org.nz/ (https://new.censusatschool.org.nz/)

Reorder bars in geom_bar ggplot2. StackOverFlow(2015)
https://stackoverflow.com/questions/25664007/reorder-bars-in-geom-bar-ggplot2/25664367
(https://stackoverflow.com/questions/25664007/reorder-bars-in-geom-bar-ggplot2/25664367)

The R Stats Package https://stat.ethz.ch/R-manual/R-devel/library/stats/html/00Index.html
(https://stat.ethz.ch/R-manual/R-devel/library/stats/html/00Index.html)

The R Project for Statistical Computing https://www.r-project.org/ (https://www.r-project.org/)