R Notebook

#load necessary libraries  
library(dplyr)

##   
## Attaching package: 'dplyr'

## The following objects are masked from 'package:stats':  
##   
## filter, lag

## The following objects are masked from 'package:base':  
##   
## intersect, setdiff, setequal, union

library(readxl)  
library(ggplot2)  
library(ggpubr)  
library(qqplotr)

##   
## Attaching package: 'qqplotr'

## The following objects are masked from 'package:ggplot2':  
##   
## stat\_qq\_line, StatQqLine

#library(car)  
library(e1071)  
library(nortest)  
library(BSDA)

## Loading required package: lattice

##   
## Attaching package: 'BSDA'

## The following object is masked from 'package:datasets':  
##   
## Orange

# 8.58

## Teeth and military service

## A) Solution:

# For recruits who were under 20  
x1 = 68  
n1 = 58952  
p1 = round(x1/n1,6)  
p1

## [1] 0.001153

# For recruits who were at the age of 40 or over  
x2 = 3801  
n2 = 43786  
p2 = round(x2/n2,6)  
p2

## [1] 0.086809

## B) Solution:

# Estimate for difference  
p1-p2

## [1] -0.085656

dat = data.frame(sample = c(1,2), x = c(x1,x2), n = c(n1,n2), P = c(p1,p2))  
dat

## sample x n P  
## 1 1 68 58952 0.001153  
## 2 2 3801 43786 0.086809

p\_test = prop.test(x = c(x1,x2), n = c(n1,n2), alternative = 'two.sided', conf.level = .99, correct = F)  
p\_test

##   
## 2-sample test for equality of proportions without continuity  
## correction  
##   
## data: c(x1, x2) out of c(n1, n2)  
## X-squared = 5086.4, df = 1, p-value < 2.2e-16  
## alternative hypothesis: two.sided  
## 99 percent confidence interval:  
## -0.08913961 -0.08217057  
## sample estimates:  
## prop 1 prop 2   
## 0.001153481 0.086808569

# As difference in population proportions is negative  
# So Z-value  
-unname(sqrt(p\_test$statistic))

## [1] -71.31916

# 99% confidence interval for the difference in the proportions  
p\_test$conf.int

## [1] -0.08913961 -0.08217057  
## attr(,"conf.level")  
## [1] 0.99

# mean / difference in population proportions  
m = p1-p2  
m

## [1] -0.085656

# Standard error of mean / difference  
sd = sqrt((p1\*(1-p1)/n1) + (p2\*(1-p2)/n2))  
sd

## [1] 0.001352777

# pooled estimate  
p = (p1\*n1+p2\*n2)/(n1+n2)  
p

## [1] 0.03765881

# Z-value  
z = (p1-p2)/sqrt(p\*(1-p)\*(1/n1+1/n2))  
z

## [1] -71.32

## C) Solution:

*The null Hypothesis:* H0: p1 = p2 *The alternate Hypothesis:* H1: p1 ≠ p2 *Here p1 = Proportion of rejects for recruits who were under 20 and p2 = Proportion of rejects for recruits who were 40 or above*

# P-value  
pnorm(z,m,sd, lower.tail = T)

## [1] 0

# Or  
p\_test$p.value

## [1] 0

*As the p-value in this context is way less than 0.01, the null hypothesis is rejected at 1% level of significance.There is sufficient evidence to indicate that there is a difference between proportions of rejection among recruits who were under age of 20 and who were age of 40 or above. The result is statistically significant.*

## D) Solution:

#### The guidelines for using large sample method for a 99% confidence interval is that to check the below conditions for recruits who were under the age of 20

#### n1*p1>=10 and n1*(1-p1)>=10

n1\*p1

## [1] 67.97166

n1\*(1-p1)

## [1] 58884.03

*As both of them are greater than 10, we can say the guideline is satisfied.*

#### The guidelines for using large sample method for a 99% confidence interval is that to check the below conditions for recruits who were at the age of 40 or above

#### n2*p2>=10 and n2*(1-p2)>=10

n2\*p2

## [1] 3801.019

n2\*(1-p2)

## [1] 39984.98

*As both of them are greater than 10, we can say the guideline is satisfied.*

# 8.71

## Gender bias in textbooks

## A) Solution:

# For female  
n1 = 60   
x1 = 48  
p1 = x1/n1  
p1

## [1] 0.8

se1 = sqrt(p1\*(1-p1)/n1)  
se1

## [1] 0.05163978

# For male  
n2 = 132   
x2 = 52  
p2 = x2/n2  
p2

## [1] 0.3939394

se2 = sqrt(p2\*(1-p2)/n2)  
se2

## [1] 0.04252906

## B) Solution:

# Estimate for difference  
p1-p2

## [1] 0.4060606

# pooled estimate  
p = (p1\*n1+p2\*n2)/(n1+n2)  
p

## [1] 0.5208333

options(scipen = n)  
p\_test = prop.test(x = c(x1,x2), n = c(n1,n2), alternative = 'two.sided', conf.level = .95, correct = F)  
p\_test

##   
## 2-sample test for equality of proportions without continuity  
## correction  
##   
## data: c(x1, x2) out of c(n1, n2)  
## X-squared = 27.253, df = 1, p-value = 1.785e-07  
## alternative hypothesis: two.sided  
## 95 percent confidence interval:  
## 0.2749423 0.5371789  
## sample estimates:  
## prop 1 prop 2   
## 0.8000000 0.3939394

*So the P-value is 0.0000001785 and the difference in the two proportions is somewhere between 0.2749423 and 0.5371789 using 95% confidence intervals.*

## C) Solution:

*The null Hypothesis:* H0: p1 = p2 *The alternate Hypothesis:* H1: p1 ≠ p2 *Here p1 = Proportion of juvenile among female and p2 = Proportion of juvenile among male* *Using level of significance a = .05, we can reject null hypothesis (H0) if z value from the difference in proportions is either less than -1.96 or greater than 1.96*

# Z-value  
unname(sqrt(p\_test$statistic))

## [1] 5.220477

*Our Z-value is 5.220477. So we can reject H0. Also P-value was really low.*

*The test statistic value falls in the rejection region, so we reject the null hypothesis. There is a difference in the proportion of female juveniles and male juveniles.*

# 9.37

## Health care fraud

## A) Solution:

#Here  
n1 = 57  
n2 = 17  
n3 = 5  
x1 = 6  
x2 = 5  
x3 = 1  
dat = data.frame(Stratum = c("Small", "Medium", "Large"), "Numbers\_Allowed" = c(n1-x1,n2-x2,n3-x3), "Numbers\_not\_allowed" = c(x1,x2,x3), Total = c(n1,n2,n3))  
#, Total = c(n1,n2,n3)  
dat

## Stratum Numbers\_Allowed Numbers\_not\_allowed Total  
## 1 Small 51 6 57  
## 2 Medium 12 5 17  
## 3 Large 4 1 5

dat <- rbind(dat, c("Total", colSums(dat[,2:4])))  
dat

## Stratum Numbers\_Allowed Numbers\_not\_allowed Total  
## 1 Small 51 6 57  
## 2 Medium 12 5 17  
## 3 Large 4 1 5  
## 4 Total 67 12 79

#3 × 2 table of counts  
dat = transform(dat, Numbers\_Allowed = as.numeric(Numbers\_Allowed),Numbers\_not\_allowed = as.numeric(Numbers\_not\_allowed),Total = as.numeric(Total))  
dat

## Stratum Numbers\_Allowed Numbers\_not\_allowed Total  
## 1 Small 51 6 57  
## 2 Medium 12 5 17  
## 3 Large 4 1 5  
## 4 Total 67 12 79

## B) Solution:

#the percent of claims that were not allowed in each of the three strata  
dat$pct\_not\_allowed = round(dat$Numbers\_not\_allowed/dat$Total\*100,2)  
dat[1:3,c(1,5)]

## Stratum pct\_not\_allowed  
## 1 Small 10.53  
## 2 Medium 29.41  
## 3 Large 20.00

## C) Solution:

*In a 3x2 table, the expected count corresponding to the cell (Large, Numbers\_not\_allowed) is less then 5. SO need to combine Medium and Large rows.*

# Combining medium and large row  
dat = subset(dat, select = -c(pct\_not\_allowed))  
dat <- rbind(dat, c("Medium & Large", colSums(dat[2:3,2:4])))  
dat = dat[-c(2,3),]  
dat = transform(dat, Numbers\_Allowed = as.numeric(Numbers\_Allowed),Numbers\_not\_allowed = as.numeric(Numbers\_not\_allowed),Total = as.numeric(Total))  
dat

## Stratum Numbers\_Allowed Numbers\_not\_allowed Total  
## 1 Small 51 6 57  
## 4 Total 67 12 79  
## 5 Medium & Large 16 6 22

# reordering the rows  
dat = dat[c(1,3,2),]  
row.names(dat) <- NULL  
  
dat

## Stratum Numbers\_Allowed Numbers\_not\_allowed Total  
## 1 Small 51 6 57  
## 2 Medium & Large 16 6 22  
## 3 Total 67 12 79

## D) Solution:

#### The null hypothesis:

*H0: There is no relationship between claims sizes and whether a claim is allowed.*

#### The alternate hypothesis:

*H1: There is a relationship between claims sizes and whether a claim is allowed.*

## E) Solution:

row.names(dat) = dat[,1]  
dat = dat[,c(2,3,4)]  
dat

## Numbers\_Allowed Numbers\_not\_allowed Total  
## Small 51 6 57  
## Medium & Large 16 6 22  
## Total 67 12 79

# Expected frequencies  
xsq = chisq.test(dat[1:2,1:2], correct = F)

## Warning in chisq.test(dat[1:2, 1:2], correct = F): Chi-squared approximation may  
## be incorrect

freq\_df = data.frame(xsq$expected)  
freq\_df <- cbind(freq\_df, Total = rowSums(freq\_df))  
freq\_df <- rbind(freq\_df, c(colSums(freq\_df[,1:3])))  
rownames(freq\_df)[rownames(freq\_df)==3] = c("Total")  
freq\_df

## Numbers\_Allowed Numbers\_not\_allowed Total  
## Small 48.34177 8.658228 57  
## Medium & Large 18.65823 3.341772 22  
## Total 67.00000 12.000000 79

# expected frequency without totals  
xsq$expected

## Numbers\_Allowed Numbers\_not\_allowed  
## Small 48.34177 8.658228  
## Medium & Large 18.65823 3.341772

# Chi -square test statistics  
xsq$statistic

## X-squared   
## 3.45551

# Degree of freedom  
xsq$parameter

## df   
## 1

# P-value  
xsq$p.value

## [1] 0.0630413

xsq

##   
## Pearson's Chi-squared test  
##   
## data: dat[1:2, 1:2]  
## X-squared = 3.4555, df = 1, p-value = 0.06304

*The smallest P-value suggests that there is not enough evidence to reject the null hypothesis. So we failed to reject the null hypothesis as p-value (.0634) is higher than 5 times in 100.*

# 9.41

## When do Canadian students enter private career colleges?

## A) Solution:

file\_name = "ex09-41canf.xls"  
dat = read\_xls(file\_name)  
dat

## # A tibble: 12 x 4  
## Time Field n Percent  
## <chr> <chr> <dbl> <dbl>  
## 1 AfterHS Trades 942 34  
## 2 Later Trades 942 66  
## 3 AfterHS Design 584 47  
## 4 Later Design 584 53  
## 5 AfterHS Health 5085 40  
## 6 Later Health 5085 60  
## 7 AfterHS MediaIT 3148 31  
## 8 Later MediaIT 3148 69  
## 9 AfterHS Service 1350 36  
## 10 Later Service 1350 64  
## 11 AfterHS Other 2255 52  
## 12 Later Other 2255 48

after\_df = filter(dat, Time == 'AfterHS')  
after\_df$after = after\_df$n\*after\_df$Percent/100  
after\_df = after\_df[c(2,5)]  
after\_df

## # A tibble: 6 x 2  
## Field after  
## <chr> <dbl>  
## 1 Trades 320.  
## 2 Design 274.  
## 3 Health 2034   
## 4 MediaIT 976.  
## 5 Service 486   
## 6 Other 1173.

later\_df = filter(dat, Time == 'Later')  
later\_df$later = later\_df$n\*later\_df$Percent/100  
later\_df = later\_df[c(2,5)]  
later\_df

## # A tibble: 6 x 2  
## Field later  
## <chr> <dbl>  
## 1 Trades 622.  
## 2 Design 310.  
## 3 Health 3051   
## 4 MediaIT 2172.  
## 5 Service 864   
## 6 Other 1082.

# 6x2 table with observed counts  
main\_df = merge(x = after\_df, y = later\_df, by = "Field", all = TRUE)  
main\_df <- cbind(main\_df, Total = rowSums(main\_df[2:3]))  
main\_df <- rbind(main\_df, c("Total", colSums(main\_df[2:4])))  
main\_df = transform(main\_df, after = as.numeric(after),later = as.numeric(later),Total = as.numeric(Total))  
row.names(main\_df) = main\_df[,1]  
main\_df = main\_df[,c(2,3,4)]  
main\_df

## after later Total  
## Design 274.48 309.52 584  
## Health 2034.00 3051.00 5085  
## MediaIT 975.88 2172.12 3148  
## Other 1172.60 1082.40 2255  
## Service 486.00 864.00 1350  
## Trades 320.28 621.72 942  
## Total 5263.24 8100.76 13364

## B) Solution:

# Total number of Canadian students surveyed  
main\_df[max(rownames(main\_df)),max(colnames(main\_df))]

## [1] 942

# The percentage of student enrolled right after high school is  
paste0(round(main\_df[max(rownames(main\_df)),2]/main\_df[max(rownames(main\_df)),max(colnames(main\_df))]\*100,2),"%")

## [1] "66%"

# The percentage of student enrolled later is  
paste0(round(main\_df[max(rownames(main\_df)),3]/main\_df[max(rownames(main\_df)),max(colnames(main\_df))]\*100,2),"%")

## [1] "100%"

## C) Solution:

#### The null hypothesis:

*H0: There is no relationship between field of study and when a Canadian enrolls for a college.*

#### The alternate hypothesis:

*H1: There is a relationship between field of study and when a Canadian enrolls for a college.*

# Expected frequencies  
xsq = chisq.test(main\_df[1:6,1:2], correct = F)  
freq\_df = data.frame(xsq$expected)  
freq\_df <- cbind(freq\_df, Total = rowSums(freq\_df))  
freq\_df <- rbind(freq\_df, c(colSums(freq\_df[,1:3])))  
rownames(freq\_df)[rownames(freq\_df)==7] = c("Total")  
freq\_df

## after later Total  
## Design 230.0009 353.9991 584  
## Health 2002.6620 3082.3380 5085  
## MediaIT 1239.7994 1908.2006 3148  
## Other 888.1028 1366.8972 2255  
## Service 531.6802 818.3198 1350  
## Trades 370.9946 571.0054 942  
## Total 5263.2400 8100.7600 13364

# Chi-square summary  
xsq

##   
## Pearson's Chi-squared test  
##   
## data: main\_df[1:6, 1:2]  
## X-squared = 275.94, df = 5, p-value < 2.2e-16

*The smallest P-value suggests that we have enough evidence to reject the null hypothesis.*

# 9.50

## Goodness of fit to a standard Normal distribution

## A) Solution:

#With the given data  
dat = data.frame(category = c(1,2,3,4,5), observed = c(139,102,41,78,140))  
dat

## category observed  
## 1 1 139  
## 2 2 102  
## 3 3 41  
## 4 4 78  
## 5 5 140

summary(dat)

## category observed   
## Min. :1 Min. : 41   
## 1st Qu.:2 1st Qu.: 78   
## Median :3 Median :102   
## Mean :3 Mean :100   
## 3rd Qu.:4 3rd Qu.:139   
## Max. :5 Max. :140

s = sd(dat$observed)  
m = mean(dat$observed)  
n = sum(dat[,2])  
n

## [1] 500

# For first category, the probability of less than or equal to -0.6  
p1 = pnorm(-.6,lower.tail = T)  
# For second category, the probability of greater than -0.6 and less than or equal to -0.1  
p2 = pnorm(-.1,lower.tail = T) - pnorm(-.6,lower.tail = T)  
# For third category, the probability of greater than -0.1 and less than or equal to 0.1  
p3 = pnorm(.1,lower.tail = T) - pnorm(-.1,lower.tail = T)  
# For fourth category, the probability of greater than 0.1 and less than or equal to 0.6  
p4 = pnorm(.6,lower.tail = T) - pnorm(.1,lower.tail = T)  
# For fifth category, the probability of greater than 0.6  
p5 = pnorm(.6,lower.tail = F)

dat$expected = c(n\*p1, n\*p2, n\*p3, n\*p4, n\*p5)  
dat$"(o-E)^2" = (dat$observed-dat$expected)^2  
dat$"(o-E)^2/E" = dat$`(o-E)^2`/dat$expected  
dat

## category observed expected (o-E)^2 (o-E)^2/E  
## 1 1 139 137.12656 3.509782 0.02559520  
## 2 2 102 92.95952 81.730234 0.87920238  
## 3 3 41 39.82784 1.373965 0.03449762  
## 4 4 78 92.95952 223.787313 2.40736298  
## 5 5 140 137.12656 8.256664 0.06021200

# Chi-square value  
chisq = sum(dat$`(o-E)^2/E`)  
chisq

## [1] 3.40687

# Degree of freedom (n-1)  
d\_f = length(dat)-1  
d\_f

## [1] 4

# The Chi-square value with α = 0.05 and 4 degrees of freedom is  
qchisq(p = .05, df = d\_f, lower.tail=F)

## [1] 9.487729

# P-value = P(X^2 >= 3.407)  
pchisq(chisq, df=d\_f, lower.tail=FALSE)

## [1] 0.4921795

*The P-value > a = 0.05 and the test statistic is lower than the Chi-square value. We fail to reject the hypothesis of equal proportions.*

*Therefore, the test is a good fit.*