Lab 8: Inference for Proportions and Chi-Square Tests using R
STAT 630, Fall 2021

Data set: The General Social Survey (GSS) is a major survey that tracks American demographics, characteristics, and views on social and cultural issues. It is conducted by the National Opinion Research Center (NORC) at the University of Chicago.

The 2002 GSS data can be accessed from the `resampledata` library. It contains the responses of 2765 participants.

```
library(resampledata)
data("GSS2002")
```

Variable descriptions:

Variable	Description
Region	Interview location
Gender	Gender of respondent
Race	Race of respondent
Education	Highest level of education
Marital	Marital status
Religion	Religious affiliation
Happy	General happiness
Income	Respondent's income
Politics	Political views
Marijuana	Legalize marijuana?
Death Penalty	Death penalty for murder?
OwnGun	Have gun at home?
GunLaw	Require permit to buy a gun?
Pres00	Whom did you vote for in the 2000 presidential election?
Postlife	Believe in life after death?

Example: two proportion z-test and confidence interval
Using this data, let's test the hypothesis that men and women differ in their beliefs in an afterlife. Specifically, we are testing
H_0: $p_f = p_m$
H_A: $p_f \neq p_m$
where p_f is the population proportion of females who believe in an afterlife, and p_m is the population proportion of men who believe in an afterlife.

First, remove the missing data.

```
index <- which(is.na(GSS2002$Gender) | is.na(GSS2002$Postlife))
gender <- GSS2002$Gender[-index]
postlife <- GSS2002$Postlife[-index]
```

Next, we can examine some contingency tables:

```
addmargins(table(gender, postlife))

##         postlife
## gender     No  Yes  Sum
##   Female   98  550  648
##   Male    138  425  563
##   Sum     236  975 1211

prop.table(table(gender, postlife), margin=1)

##         postlife
## gender          No       Yes
##   Female 0.1512346 0.8487654
##   Male   0.2451155 0.7548845
```

The sample sizes are large enough in each cell (greater than 5), so the conditions for the z-test are satisfied. Based on the table of row proportions, it appears that a higher proportion of females believe in an afterlife than males. We can proceed with the hypothesis test to find out if this is statistically significant.

```
n <- length(postlife)
n_f <- sum(gender == "Female")
n_m <-  sum(gender == "Male")
# sample proportion of females that believe in afterlife:
phat_f <- sum(gender == "Female" & postlife == "Yes") / n_f; phat_f

## [1] 0.8487654

# sample proportion of males that believe in afterlife:
phat_m <- sum(gender == "Male" & postlife == "Yes") / n_m; phat_m

## [1] 0.7548845

# pooled sample proportion
phat <- sum(postlife == "Yes") / n; phat

## [1] 0.8051197
```

```r
# compute test statistic
SE <- sqrt(phat*(1-phat)*(1/n_f + 1/n_m))
z <- (phat_f - phat_m) / SE; z

## [1] 4.1137

# p-value
pvalue <-  2 * (1 - pnorm(z)); pvalue

## [1] 3.893671e-05
```

The result is highly significant. We reject H_0 since the p-value ≈ 0. Conclusively, the data suggest that men and women differ in their views of an afterlife.

Next, let's calculate a 95% confidence interval for the difference between the proportion of males and females who believe in an afterlife:

```r
SE <- sqrt(phat_f*(1-phat_f)/n_f + phat_m*(1-phat_m)/n_m)
ci_l <- phat_f - phat_m - 1.96 * SE
ci_u <- phat_f - phat_m + 1.96 * SE
round(c(ci_l, ci_u), 3)

## [1] 0.049 0.139
```

We are 95% confident that the population proportion of females who believe in an afterlife is between 0.049 and 0.139 higher than the population proportion of males who believe in an afterlife.

Example: chi-square test for independence
Going back to the example from lecture:
H_0 : Education and position on death penalty are independent.
H_A : Education and position on death penalty are not independent.

First, let's examine some contingency tables:

```
education <- GSS2002$Education
death_penalty <- GSS2002$DeathPenalty
addmargins(table(education, death_penalty))
```

```
##            death_penalty
## education   Favor Oppose  Sum
##   Left HS     117     72  189
##   HS          511    200  711
##   Jr Col       71     16   87
##   Bachelors   135     71  206
##   Graduate     64     50  114
##   Sum         898    409 1307
```

```
prop.table(table(education, death_penalty), margin=1)
```

```
##            death_penalty
## education       Favor    Oppose
##   Left HS   0.6190476 0.3809524
##   HS        0.7187060 0.2812940
##   Jr Col    0.8160920 0.1839080
##   Bachelors 0.6553398 0.3446602
##   Graduate  0.5614035 0.4385965
```

We can use the function `chisq.test()` to implement the chi-square test for independence. Note that this function will automatically remove any missing data.

```
chisq1 <- chisq.test(education, death_penalty)
chisq1
```

```
##
##  Pearson's Chi-squared test
##
## data:  education and death_penalty
## X-squared = 23.451, df = 4, p-value = 0.0001029
```

The chi-square test statistic is 23.45 with a p-value = 0.0001029. Since the p-value < 0.05 we reject H_0, and conclude that the survey provides convincing evidence that education level and position on the death penalty are not independent. Note that the results from the R function are the same as the manual computations on the lecture slides.

The object `chisq1` also has attributes that we can extract by name.

```
attributes(chisq1)

## $names
## [1] "statistic" "parameter" "p.value"   "method"    "data.name" "observed"
## [7] "expected"  "residuals" "stdres"
##
## $class
## [1] "htest"

# test statistic
chisq1$statistic

## X-squared
##  23.45093

# p-value
chisq1$p.value

## [1] 0.0001028891

# table of expected counts
chisq1$expected

##            death_penalty
## education       Favor    Oppose
##   Left HS   129.85616  59.14384
##   HS        488.50650 222.49350
##   Jr Col     59.77506  27.22494
##   Bachelors 141.53634  64.46366
##   Graduate   78.32594  35.67406
```