# IEHC0046 BASIC STATISTICS FOR MEDICAL SCIENCES

Analysis of Categorical Data II: Practical

12 October, 2020

In this practical we will use R to assess single proportions and to compare two proportions using the ELSA dataset.

Remember to use a script to save your code and to change your working directory so you can load the ELSA dataset easily.

```
load("elsa.Rdata")
```

**1. Let's start with basic demographic characteristics. How many men and women are there in the data?**

We can use the `table()` and `prop.table()` function for this task.

```
table(elsa$sex)
```

```
##
##   male female
##   1368   1761
```

```
prop.table(table(elsa$sex))
```

```
##
##      male    female
## 0.4372004 0.5627996
```

Alternatively, we can use the function `freq()` from the package `summarytools`, which provides more information.

```
library(summarytools)
freq(elsa$sex)
```

```
## Frequencies
## elsa$sex
## Label: Sex
## Type: Factor
##
##                Freq   % Valid   % Valid Cum.   % Total   % Total Cum.
## ------------ ------ --------- -------------- --------- --------------
##         male   1368     43.72          43.72     43.72          43.72
##       female   1761     56.28         100.00     56.28         100.00
##         <NA>      0                               0.00         100.00
##        Total   3129    100.00         100.00    100.00         100.00
```

**2. Now, let's focus on one of the health variables – `past_cvd`. Study participants were asked whether they had any cardiovascular condition diagnosed by their doctor in the past. How many people reported any such condition? Answer in terms of frequencies and proportions.**

We can use the functions from the previous question.

```r
table(elsa$past_cvd)
```

```
##
##  Yes   No
## 1216 1911
```

```r
prop.table(table(elsa$past_cvd))
```

```
##
##       Yes        No
## 0.3888711 0.6111289
```

```r
freq(elsa$past_cvd)
```

```
## Frequencies
## elsa$past_cvd
## Label: had cardiovascular condition in the past
## Type: Factor
##
##               Freq   % Valid   % Valid Cum.   % Total   % Total Cum.
## ----------- ------ --------- -------------- --------- --------------
##         Yes   1216     38.89          38.89     38.86          38.86
##          No   1911     61.11         100.00     61.07          99.94
##        <NA>      2                               0.06         100.00
##       Total   3129    100.00         100.00    100.00         100.00
```

**3. Now, let's use this dataset to estimate the proportion and 95% CIs of individuals with cardiovascular conditions diagnosed by their doctor**

We can do this using the `prop.test()` function. This function can take a table produced by the `table()` function as an input, if the number of successes is in the first cell.

```r
cvd <- table(elsa$past_cvd)
cvd
```

```
##
##  Yes   No
## 1216 1911
```

```r
prop.test(cvd) # Total with CVD data
```

```
##
##  1-sample proportions test with continuity correction
##
## data:  cvd, null probability 0.5
## X-squared = 154.02, df = 1, p-value < 2.2e-16
## alternative hypothesis: true p is not equal to 0.5
## 95 percent confidence interval:
##  0.3717725 0.4062450
## sample estimates:
##         p
## 0.3888711
```

38.9% of individuals have cardiovascular conditions. The 95% CI for this proportion is 37.2-40.6%.

**4.Let's focus on CVD history a little bit more, and evaluate whether there is a difference between reported history of CVD in men and women.**

We want to estimate proportions by sex. We can do this using subsetting.

```r
cvd_male <- table(elsa$past_cvd[elsa$sex == "male"])
cvd_male
```

```
##
## Yes  No
## 537 830
```

```r
prop.test(cvd_male)
```

```
##
##  1-sample proportions test with continuity correction
##
## data:  cvd_male, null probability 0.5
## X-squared = 62.373, df = 1, p-value = 2.842e-15
## alternative hypothesis: true p is not equal to 0.5
## 95 percent confidence interval:
##  0.3669160 0.4193551
## sample estimates:
##        p
## 0.392831
```

```r
cvd_female <- table(elsa$past_cvd[elsa$sex == "female"])
cvd_female
```

```
##
##  Yes   No
##  679 1081
```

```r
prop.test(cvd_female)
```

```
##
##  1-sample proportions test with continuity correction
##
## data:  cvd_female, null probability 0.5
## X-squared = 91.364, df = 1, p-value < 2.2e-16
## alternative hypothesis: true p is not equal to 0.5
## 95 percent confidence interval:
##  0.3630453 0.4090492
## sample estimates:
##         p
## 0.3857955
```

The proportions are very similar.

**5. Now compare these proportions. What conclusions can you make?**

We can use the `table()` function to cross-tabulate two vectors. `prop.test()` can then be used comparing the proportion of successes in the first vector.

```r
cvd_sex <- table(elsa$sex, elsa$past_cvd)
cvd_sex
```

```
##
##           Yes   No
##   male    537  830
##   female  679 1081
```

```r
prop.test(cvd_sex)
```

```
##
##  2-sample test for equality of proportions with continuity correction
##
## data:  cvd_sex
## X-squared = 0.13202, df = 1, p-value = 0.7163
## alternative hypothesis: two.sided
## 95 percent confidence interval:
##  -0.02807374  0.04214487
## sample estimates:
##    prop 1    prop 2
## 0.3928310 0.3857955
```

The output provides the proportion of individuals with CVD conditions among males (39.3%) and females (38.6%), and the 95% CI for the difference in these proportions (-0.03 - 0.04%).

**6. Compare CVD history for the age groups 45-64 and 65+. Do you think that there is any difference age groups?**

We will need to create a new age group variable. We can do this using the `mutate()` function from the `tidyverse`.

```
library(tidyverse)
elsa <- elsa %>%
  mutate(age_group = case_when(age >= 45 & age <= 64 ~ "45 - 64",
                               age >= 65 ~ "65+"),
         age_group = factor(age_group))
summary(elsa$age[elsa$age_group == "45 - 64"]) # Check we made the variable correctly
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   45.00   50.00   54.00   54.56   59.00   64.00
```

```
summary(elsa$age[elsa$age_group == "65+"])
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   65.00   67.00   71.00   71.67   75.00   90.00
```

The `case_when()` function set the value of the `age_group` variable conditional on the values of age.

Now we can compare proportions in the CVD groups.

```
cvd_age <- table(elsa$age_group, elsa$past_cvd)
cvd_age
```

```
##
##            Yes   No
##   45 - 64  672 1438
##   65+      544  473
```

```
prop.test(cvd_age)
```

```
##
##  2-sample test for equality of proportions with continuity correction
##
## data:  cvd_age
## X-squared = 134.34, df = 1, p-value < 2.2e-16
## alternative hypothesis: two.sided
## 95 percent confidence interval:
##  -0.2536877 -0.1791587
## sample estimates:
```

```
##    prop 1    prop 2
## 0.3184834 0.5349066
```

**7. Look more closely at those aged 45-64, and see whether there are any difference between those who are 45-54 and those who are 55-64. What are your conclusions about the age difference in reported CVD history?**

We'll need to create another age_group variable.

```
elsa <- elsa %>%
  mutate(age_group2 = case_when(age >= 45 & age <= 54 ~ "45 - 54",
                                age >= 55 & age <= 64 ~ "55 - 64"),
         age_group2 = factor(age_group2))
summary(elsa$age[elsa$age_group2 == "45 - 54"]) # Check we made the variable correctly
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA's
##   45.00   48.00   50.00   50.15   52.00   54.00    1017
```

```
summary(elsa$age[elsa$age_group2 == "55 - 64"])
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA's
##   55.00   57.00   60.00   59.47   62.00   64.00    1017
```

Note, the `age_group2` is equal to `NA` (the missing value) if none of the conditions in the `case_when()` function are met (e.g. if age 65+).

```
cvd_age2 <- table(elsa$age_group2, elsa$past_cvd)
cvd_age2
```

```
##
##           Yes  No
##   45 - 54 286 827
##   55 - 64 386 611
```

```
prop.test(cvd_age2)
```

```
##
##  2-sample test for equality of proportions with continuity correction
##
## data:  cvd_age2
## X-squared = 40.475, df = 1, p-value = 1.991e-10
## alternative hypothesis: two.sided
## 95 percent confidence interval:
##  -0.17081253 -0.08958411
## sample estimates:
##    prop 1    prop 2
## 0.2569632 0.3871615
```

**Formative Exercise: use the variable `smok` to create a binary variable that can be used in prop.test() command, and then assess whether reported CVD history differs by smoking status.**