NDH802 R Application - Live session 2

Huong Nguyen

Live session

Load your data

```
salaries = read.csv("https://bit.ly/3r918BW")
```

Probability of an event

From the data you have, what is the probability of RM/BE student? More formally, compute $P(RM)$ and $P(BE)$.

```
table(salaries$program)/nrow(salaries)
```

```
##
##        BE        RM
## 0.7142857 0.2857143
```

```
prop.table(table(salaries$program))
```

```
##
##        BE        RM
## 0.7142857 0.2857143
```

```
#these two commands give the same results
```

Similarly, can you compute $P(male)$ and $P(female)$?

```
table(salaries$whatisyourgender1isfemale)/nrow(salaries) #including NA
```

```
##
##         1         2
## 0.3308271 0.6090226
```

```
prop.table(table(salaries$whatisyourgender1isfemale)) #without NA observations
```

```
##
##     1     2
## 0.352 0.648
```

Joint and conditional probability

Now let's spice things up a little bit. Assume we want to compute $P(RM \cap female)$, $P(female \mid RM)$ and $P(RM \mid female)$

```
#If we don't put anything, the sample space is all of the observations included in our table
a = prop.table(table(salaries$program, salaries$whatisyourgender1isfemale))
sum(a) #all of them add up to 1
```

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

As some of you asked, you can write like this in the assignment. $P(RM \cap female)$ is

```
## [1] 0.184
```

```
#If you put margin =1, we change the sample space to the variable in the rows (in this case, the progra
b = prop.table(table(salaries$program, salaries$whatisyourgender1isfemale),
           margin = 1)
sum(b[1,]) #the probabilities in each row add up to 1
```

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

```
sum(b[2,]) #the probabilities in each row add up to 1
```

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

```
#If you put margin =2, we change the sample space to the variable in the columns (in this case, the gen
c = prop.table(table(salaries$program, salaries$whatisyourgender1isfemale),
           margin = 2)
sum(c[,1]) #the probabilities in each column add up to 1
```

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

```
sum(c[,2]) #the probabilities in each column add up to 1
```

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

Your turn:

Exercises and solutions

```
survey = read.csv("https://raw.githubusercontent.com/lanhuongnguyen276/NDH802/master/Live%20sessions/NDH
```

Probability of an event

$P(Q2 = \text{R})$

```
#We learn this in the morning session
prop.table(table(survey$Q2))
```

```
##
##     Excel         R
## 0.5909091 0.4090909
```

```
#If you want to "extract" a number from your prop.table
R = prop.table(table(survey$Q2))[2] #because R is the second element
```

$P(Q1 = \text{Love it})$

```
#We learn this in the morning session
prop.table(table(survey$Q1))
```

```
##
##    Crying   Love it       Meh
## 0.3181818 0.1363636 0.5454545
```

```
#If you want to "extract" a number from your prop.table
loveit = prop.table(table(survey$Q1))[2] #because loveit is the second element
```

$P(Q1 = \text{Crying} \cup \text{Meh})$

Formally, we learn that $P(Crying \cup Meh) = P(Crying) + P(Meh) - P(Crying \cap Meh)$. We know that Crying and Meh are mutually exclusive i.e. $P(Crying \cap Meh) = 0$. Therefore, $P(Crying \cup Meh) = P(Crying) + P(Meh)$

```
crying = prop.table(table(survey$Q1))[1]
meh = prop.table(table(survey$Q1))[3]
crying_union_meh = as.numeric(crying) + as.numeric(meh) #as crying and meh are not numbers, we "force"
```

We also know that Q1 has only 3 outcomes (Love it, Meh and Crying) therefore they are collectively exhaustive (and mutually exclusive). Therefore, $P(Crying) + P(Loveit) + P(Meh) = 1$

```
#Another way to calculate crying_union_meh
crying_union_meh_2 = 1 - as.numeric(loveit) #notice that it gives you the same results as before
```

Conditional probability

$P(Q2 = \text{R} \mid Q1 = \text{Love it})$

```
#This code gives you the prop.table
prop.table(table(survey$Q1, survey$Q2), margin = 1)
```

```
##
##                Excel         R
##    Crying   0.7857143 0.2142857
##    Love it  0.6666667 0.3333333
##    Meh      0.4583333 0.5416667
```

```
#Then you can extract speciic values from your prop.table
R_given_loveit = prop.table(table(survey$Q1, survey$Q2), margin = 1)[2,2] #because R|love it is second
```

$P(Q2 = \text{Excel} \mid Q1 = \text{Love it})$

```
Excel_given_loveit = prop.table(table(survey$Q1, survey$Q2), margin = 1)[2,1] #because R given love it
```

Notice that $P(Q2 = \text{R} \mid Q1 = \text{Love it}) + P(Q2 = \text{Excel} \mid Q1 = \text{Love it}) = 1$

```
R_given_loveit + Excel_given_loveit
```

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

Joint probability

$P(Q2 = \text{R} \cap Q1 = \text{Love it})$

```
#This code gives you the prop.table
prop.table(table(survey$Q1, survey$Q2))
```

```
##
##                Excel          R
##   Crying   0.25000000 0.06818182
##   Love it  0.09090909 0.04545455
##   Meh      0.25000000 0.29545455
```

```
#Then you can extract speciic values from your prop.table
R_and_loveit = prop.table(table(survey$Q1, survey$Q2))[2,2] #because R and love it is second row/second
```

$P(Q2 = \text{Excel} \cap Q1 = \text{Love it})$

```
Excel_and_loveit = prop.table(table(survey$Q1, survey$Q2))[2,1] #because Excel and love it is second ro
```

Notice that $P(Q2 = \text{Excel} \cap Q1 = \text{Love it}) + P(Q2 = \text{R} \cap Q1 = \text{Love it}) = P(Q1 = \text{Love it})$

```
print(paste(loveit, R_and_loveit + Excel_and_loveit))
```

```
## [1] "0.136363636363636 0.136363636363636"
```

```
#you don't need to learn to code print() and paste(). It is for illustration purpose only
```

$P(Q2 = \text{R} \cap Q1 = \text{Crying} \cup \text{Meh})$

```
as.numeric(R) - R_and_loveit
```

```
## [1] 0.3636364
```

```
#I'll let you think about why :) But you cannot do it to me in the assignments. Please explain why, it'
```