Problem Set 3

Felipe Riveroll Aguirre[*]

November 6, 2012

Probability and BMI. The following table uses data from the NHLBI teaching data set and displays the body mass index for 3,909 participants in the Framingham Heart Study with BMI measurements at the first two exams, in 1956 and in 1962.

	$BMI \leq 25$, exam 2	$BMI > 25$, exam 2	Total
$BMI \leq 25$, exam 1	1,492	278	1,770
$BMI > 25$, exam 1	249	1,890	2,139
Total	1,741	2,168	3,909

Assume a study participant has been randomly selected from this subset of 3,909 participants.

- Define A as the event that this participant has a high BMI at exam 1.
- Define B as the event that this participant has a high BMI at exam 2.
- Define C as the event that this participant has a low BMI at exam 2.

1. What is the probability of A?
   A: $P(A) = \frac{2139}{3909} = 0.5472$
2. What is the probability of B?
   A: $P(B) = \frac{2168}{3909} = 0.5546$
3. What is the probability of A and B?
   A: $P(A \cap B) = \frac{1890}{3909} = 0.4835$
4. Are A and B independent?
   A: No

[*] `friveroll@gmail.com`

5. In a randomly selected participant, what is the probability that A and/or B occurs (namely, that the participant?s BMI is high during at least one of the first two exams)?
   A: $P(A \cup B) = P(A) + P(B) - P(A \cap B)$
   $P(A \cup B) = 0.5472 + 0.5546 - 0.4835 = 0.6183$

6. What is the probability that B occurs, given that A occurs?
   A: $P(B \mid A) = \frac{P(A \cap B)}{P(A)} = \frac{0.4835}{0.5472} = 0.8836$

7. What is the probability that C occurs, given that A occurs?
   A: $P(C \mid A) = \frac{249}{2139} = 0.1164$

We can get those values with a 2way table using R

```
> library("foreign")
> data <- read.dta("https://dl.dropbox.com/u/4828275/fhs.dta",
+                  convert.factors = TRUE,
+                  missing.type = TRUE)

> attach(data)
> bmi1high <- NA
> bmi1high[bmi1 > 25 & !is.na(bmi1)] <- 1
> bmi1high[bmi1 <= 25] <- 0
> bmi2high <- NA
> bmi2high[bmi2 > 25 & !is.na(bmi2)] <- 1
> bmi2high[bmi2 <= 25] <- 0
> library("gmodels")
> with(data, CrossTable(bmi1high, bmi2high, prop.chisq=F, digits= 4))
```

```
   Cell Contents
|-------------------------|
|                       N |
|           N / Row Total |
|           N / Col Total |
|         N / Table Total |
|-------------------------|


Total Observations in Table:  3909


             | bmi2high
    bmi1high |         0 |         1 | Row Total |
-------------|-----------|-----------|-----------|
           0 |      1492 |       278 |      1770 |
             |    0.8429 |    0.1571 |    0.4528 |
             |    0.8570 |    0.1282 |           |
             |    0.3817 |    0.0711 |           |
-------------|-----------|-----------|-----------|
           1 |       249 |      1890 |      2139 |
             |    0.1164 |    0.8836 |    0.5472 |
             |    0.1430 |    0.8718 |           |
             |    0.0637 |    0.4835 |           |
-------------|-----------|-----------|-----------|
Column Total |      1741 |      2168 |      3909 |
             |    0.4454 |    0.5546 |           |
-------------|-----------|-----------|-----------|
```

Probability of age and smoking events. The following table uses data from the NHLBI teaching data set using the 4,434 participants in the Framingham Heart Study who attended the exam in 1956. Below, we show a table that contains the probabilities that a study participant falls into one of the eight categories defined by all possible combinations of the age categories and smoking status at the first exam in 1956. Recall that all study participants are between 30 and 70 years old.

Define A as the event that a randomly chosen study participant is a smoker at exam 1 in 1956.

Define B as the event that the person chosen is between 60 and 70 years old at exam 1.

Age,	Smoker, exam 1	
exam 1	No	Yes
30-39	0.0519	0.0742
40-49	0.1554	0.2262
50-59	0.1809	0.1346
60-70	0.1200	0.0568

```
> age_smoking <- read.csv("https://dl.dropbox.com/u/4828275/age_smoking.csv")
```

1. Eight categories representing age and smoking status groups are shown in the table above. Are these groups:
   (a) mutually exclusive, (b) exhaustive, **(c) both**

   ```
   > sum(age_smoking[,2]) + sum(age_smoking[,3])

   [1] 1
   ```

2. What is the probability of A?
   A: $P(A) = 0.0742 + 0.2262 + 0.1346 + 0.0568 = 0.4918$

   ```
   > sum(age_smoking[,3])

   [1] 0.4918
   ```

3. What is the probability of B^C, the complement of B?
   A: $P(B) = 0.1200 + 0.0568 = 0.1768$
   $P(B^C) = 1 - P(B) = 1 - 0.1768 = 0.8232$

   ```
   > 1 - sum(age_smoking[4,c(2,3)])

   [1] 0.8232
   ```

4. What is the probability that a randomly selected individual is a non-smoker who is younger than 60 years old at exam 1?
   A: $P(C) = 0.0519 + 0.1554 + 0.1809 = 0.3882$

   ```
   > sum(age_smoking[1:3,2])
   ```

   ```
   [1] 0.3882
   ```

5. Are the events A and B independent?
   A: No

Diagnostic Testing. Screening for prostate cancer in men is a controversial topic. One of the most common screening mechanisms is the PSA test (prostate antigen test). In a meta-analysis, Mistry and Cable (2003) report that the sensitivity of the PSA test is 72.1% and the specificity is 93.2%. In the United States, it is estimated that 16.1% of men will have prostate cancer at some point in their life (America Cancer Society 2012). Assume that the prevalence of prostate cancer among men ages 75 and older is 16.1%. We examine the properties of the PSA screening test in men ages 75 and older, using the sensitivity and specificity values above.

Sensitivity: $P(T^+ \mid D^+) = 0.721$
Specificity: $P(T^- \mid D-) = 0.932$
Prevalence: $P(D^+) = 0.161$

1. What is the probability of a false negative test result?
   A: $P(T^- \mid D^+) = 1 - P(T^+ \mid D^+) = 1 - 0.721 = 0.279$

2. What is the probability of a false positive result?
   A: $P(T^+ \mid D^-) = 1 - P(T^- \mid D-) = 1 - 0.932 = 0.068$

3. What is the probability that a randomly selected man who is 75 years or older DOES NOT have prostate cancer, given that his PSA screening was positive?
   A: $NPV = 1 - PPV = 1 - P(D^+ \mid T^+) = 1 - \frac{P(D^+)P(T^+|D^+)}{P(D^+)P(T^+|D^+)+P(D^-)P(T|D^-)}$
   $= 1 - \frac{(0.161)(0.721)}{(0.161)(0.721)+(0.839)(0.068)} = 1 - 0.6705 = 0.3295$

4. What is the probability that a randomly selected man who is 75 years or older has prostate cancer, given that his PSA screening was negative?
   $P(D^+ \mid T^-) = \frac{P(D^+)P(T^-|D^+)}{P(D^+)P(T^-|D^+)+P(D^-)P(T^-|D^-)} = \frac{(0.279)(0.161)}{(0.279)(0.161)+(0.932)(0.839)}$
   $= 0.0543$

Titanic Survival. The following table describes the survival status of passengers on the Titanic, stratified by Passenger Class (First, Second, or Third), Sex/Age (Child, Women, or Man), and Survival Status. The Frequency column indicates the number of passengers in each stratum. (For example there were 4 1st class women passengers who did not survive and 140 1st class women passengers who did survive). These data were obtained from the website anesi.com and refers to British Parliamentary Papers, Shipping. Casualties (Loss of the Steamship ”Titanic”), 1912. cmd 6352 Report of a Formal Investigation into the circumstances attending the foundering on the 15 th April 1912 of the British Steamship ”Titanic” of Liverpool after striking ice in or near Latitude 41 46 N., Longitude 50 14 W., North Atlantic Ocean, whereby loss of life ensued (London; His Majesty?s Stationary Office, 1912) page 42.

Passenger Class	Age/Sex	Survival Status	Frequency
First	Child	Survived	6
First	Child	Did not survive	0
First	Women	Survived	140
First	Women	Did not survive	4
First	Man	Survived	57
First	Man	Did not survive	118
Second	Child	Survived	24
Second	Child	Did not survive	0
Second	Women	Survived	80
Second	Women	Did not survive	13
Second	Man	Survived	14
Second	Man	Did not survive	154
Third	Child	Survived	27
Third	Child	Did not survive	52
Third	Women	Survived	76
Third	Women	Did not survive	89
Third	Man	Survived	75
Third	Man	Did not survive	387

Use these data to calculate the cumulative incidence of surviving for each of the following groups of individuals:

```
> titanic <- read.csv("https://dl.dropbox.com/u/4828275/titanic.csv")
```

1. All Women

```
> Women <- subset(titanic, Age.Sex == "Women")
> sum(subset(Women, Survival.Status == "Survived")[,4])/sum(Women[,4])

[1] 0.7363184
```

2. All Children

```
> Child <- subset(titanic, Age.Sex == "Child")
> sum(subset(Child, Survival.Status == "Survived")[,4])/sum(Child[,4])
```

```
[1] 0.5229358
```

3. All Women or Children

```
> Child_Women <- rbind(Child, Women)
> sum(subset(Child_Women, Survival.Status == "Survived")[,4])/sum(Child_Women[,4])
```

```
[1] 0.6908023
```

4. All First Class Passengers

```
> First_class <- subset(titanic, Passenger.Class == "First")
> sum(subset(First_class, Survival.Status == "Survived")[,4])/sum(First_class[,4])
```

```
[1] 0.6246154
```

BMI and Cumulative Incidence. The following table uses data from the NHLBI teaching data set and displays categories of body mass index (used in the previous homework assignment) for 4,415 participants in the Framingham Heart Study attending an examination in 1956 with non-missing values for body mass index. For each body mass index category, the table displays the number of subjects who died (death=1) during follow-up and the total person-years of follow-up (timedeath) until death or the end of the follow-up period (24 years). Assume that all deaths and time of death were recorded among the 4415 participants.

BMI	Number of Subjects	Number of Deaths	Total Person-years
$BMI < 18.5$	57	18	1181.44
$18.5 \le BMI < 25$	1936	571	40708.74
$25 \le BMI < 30$	1848	691	37728.41
$30 \le BMI$	574	257	11254.52
Total	4415	1537	90873.11

```
> bmi_inc <- read.csv("https://dl.dropbox.com/u/4828275/bmi_inc.csv")
```

1. What is the cumulative incidence of death among the 4415 participants at the 1956 exam?

```
> bmi_inc[5,3]/bmi_inc[5,2]

[1] 0.3481314
```

2. What is the cumulative incidence of death during the 24 years of follow-up for each of the body mass index class?

```
> bmi_inc[5,3]/bmi_inc[5,2]

[1] 0.3481314
```

   (a) Under Weight Participants

   ```
   > bmi_inc[1,3]/bmi_inc[1,2]
   [1] 0.3157895
   ```

   (b) Normal Weight Participants

   ```
   > bmi_inc[2,3]/bmi_inc[2,2]
   [1] 0.294938
   ```

   (c) Overweight Participants

   ```
   > bmi_inc[3,3]/bmi_inc[3,2]
   ```

```
[1] 0.3739177
```

(d) Obese Participants

```
> bmi_inc[4,3]/bmi_inc[4,2]
[1] 0.4477352
```

3. What is the incidence rate of death among the 4415 participants during the 24 years of followup? (Express your answer as #deaths/(1000 person-years))

```
> bmi_inc[5,3]/bmi_inc[5,4]
[1] 0.01691369
```

4. What is the incidence rate of death during the 24 years of follow-up for each of the body mass index classes? (Express your answers as #deaths/(1000 person-years))

   (a) Under Weight Participants

```
> bmi_inc[1,3]/bmi_inc[1,4]*1000
[1] 15.23564
```

   (b) Normal Weight Participants

```
> bmi_inc[2,3]/bmi_inc[2,4]*1000
[1] 14.02647
```

   (c) Overweight Participants

```
> bmi_inc[3,3]/bmi_inc[3,4]*1000
[1] 18.31511
```

   (d) Obese Participants

```
> bmi_inc[4,3]/bmi_inc[4,4]*1000
[1] 22.83527
```

Also We can get all the calculations at a time with:

```
> C_I <- bmi_inc[,3]/bmi_inc[,2]
> I_R <- bmi_inc[,3]/bmi_inc[,4] * 1000

> data.frame("BMI"=bmi_inc[,1], "Cummulative Inc"=C_I, "Inc Rate"=I_R)
```

BMI	Cummulative Incidence	Incidence Rate
Underweight	0.3158	15.2356
Normal	0.2949	14.0265
Overweight	0.3739	18.3151
Obese	0.4477	22.8353
Total	0.3481	16.9137

BMI and CHD Incidence. Use Stata and the NHLBI data set to create a separate variable for ünderwtẗhat equals 1 if a person's BMI was less than 18.5 and 0 if a person's BMI was ≥ 18.5). To create the separate variables for each of the four categories of body mass index, use the BMI1 variable in the NHLBI dataset. What is the incidence rate for developing CHD (anychd=1) during the 24-years of follow-up for participants in each of the body mass index categories? (Express your answers as deaths/(1000 person-years)) Hint: Number of years a person was followed for CHD is recorded in the ẗimechdv̈ariable in the NHLBI dataset.

```
> underwt <- ifelse(bmi1<18.5, c(1), c(0))
> normalwt <- ifelse(bmi1>=18.5 & bmi1 <= 25, c(1), c(0))
> overwt <- ifelse(bmi1>25 & bmi1 <= 30, c(1), c(0))
> obese <- ifelse(bmi1>30 & !is.na(bmi1), c(1), c(0))
```

1. Under Weight Participants
2. Normal Weight Participants
3. Overweight Participants
4. Obese Participants

```
> if (!"epiR" %in% installed.packages())
+ {
+   install.packages("epiR", dependencies = TRUE)
+ }
> library(epiR)
```

We can get all the required tables for Incidence rate using a loop (see next question to see a single table and some comments for the code)

```
> bmi_cat <- c("underwt", "normalwt", "overwt", "obese")
> for (i in 1:4)
+ {
+   eval(parse(text = paste(bmi_cat[i],
+                           _t <- matrix(nrow =2, ncol=2),
+                           sep="")))
+         eval(parse(text = paste(bmi_cat[i],
+                           _t[,1] <- tapply(death=="Yes",,
+                           bmi_cat[i],, sum), sep="")))
+         eval(parse(text = paste(bmi_cat[i],
+                           _t[,2] <- tapply(timedth,,
+                           bmi_cat[i],, sum), sep="")))
+         eval(parse(text = paste(bmi_cat[i],
+                           _t <- rbind(,bmi_cat[i],
+                           _t[2,],,,bmi_cat[i],_t[1,]), sep="")))
+   cat("\n\n")
+   print(bmi_cat[i])
+   eval(parse(text = paste(epi.2by2(,bmi_cat[i],
+                           _t, method = "cohort.time",
+                           conf.level = 0.95, units = 1000,
+                           homogeneity = "breslow.day",
+                           verbose = F), sep="")))
+ }

[1] "underwt"
             Disease +    Time at risk        Inc rate *
Exposed +           18            1181              15.2
Exposed -         1519           89692              16.9
Total             1537           90873              16.9

Point estimates and 95 % CIs:
---------------------------------------------------------
Inc rate ratio                           0.9 (0.53, 1.43)
Attrib rate *                            -1.7 (-8.79, 5.39)
Attrib rate in population *              -0.02 (-1.22, 1.18)
Attrib fraction in exposed (%)           -11.16 (-88.07, 29.91)
Attrib fraction in population (%)        -0.13 (-0.16, -0.1)
---------------------------------------------------------
 * Cases per 1000 units of population time at risk


[1] "normalwt"
             Disease +    Time at risk        Inc rate *
Exposed +          571           40709              14.0
Exposed -          966           50164              19.3
Total             1537           90873              16.9

Point estimates and 95 % CIs:
---------------------------------------------------------
Inc rate ratio                           0.73 (0.66, 0.81)
Attrib rate *                            -5.23 (-6.9, -3.56)
Attrib rate in population *              -2.34 (-3.82, -0.86)
Attrib fraction in exposed (%)           -37.29 (-52.52, -23.67)
Attrib fraction in population (%)        -13.85 (-15.35, -12.38)
---------------------------------------------------------
 * Cases per 1000 units of population time at risk


[1] "overwt"
             Disease +    Time at risk        Inc rate *
Exposed +          691           37728              18.3
Exposed -          846           53145              15.9
```

```
Total             1537          90873             16.9

Point estimates and 95 % CIs:
---------------------------------------------------------
Inc rate ratio                          1.15 (1.04, 1.27)
Attrib rate *                           2.4 (0.66, 4.13)
Attrib rate in population *             0.99 (-0.37, 2.36)
Attrib fraction in exposed (%)          13.08 (3.76, 21.49)
Attrib fraction in population (%)       5.88 (4.23, 7.5)
---------------------------------------------------------
 * Cases per 1000 units of population time at risk


[1] "obese"
             Disease +    Time at risk        Inc rate *
Exposed +          257           11255              22.8
Exposed -         1293           79861              16.2
Total             1550           91116              17.0

Point estimates and 95 % CIs:
---------------------------------------------------------
Inc rate ratio                          1.41 (1.23, 1.61)
Attrib rate *                           6.64 (3.72, 9.57)
Attrib rate in population *             0.82 (-0.4, 2.04)
Attrib fraction in exposed (%)          29.1 (18.62, 38.03)
Attrib fraction in population (%)       4.82 (4.37, 5.27)
---------------------------------------------------------
 * Cases per 1000 units of population time at risk
```

High Blood Pressure and CHD. Use Stata and the NHLBI data set to create the two categories of high blood pressure (highbp1).

```
> highbp1 <- NULL
> highbp1[sysbp1>=140 | diabp1 >= 90] <- 1
> highbp1[sysbp1<140 & diabp1 < 90] <- 0
```

(Note: There are no missing data on sysbp1 and diabp1. If data were missing on both sysbp1 and diabp1 then they should also be missing for highbp1. If data were missing on diabp1 only and sysbp1 $>$ 140 then highbp1 =1, otherwise highbp1 should be missing. Similarly, if data were missing on sysbp1 only and diabp1 $>$ 90 then highbp1 =1, otherwise highbp1 should be missing.)

1. What is the incidence rate for developing CHD (anychd=1) during the 24-years of follow-up for participants in each of the blood pressure categories? (Express your answers as deaths/(1000 person-years))

   (a) Participants with high blood pressure at the 1956 exam (highbp1=1)

   (b) Participants without high blood pressure at the 1956 exam (highbp1=0)

```
> # Generate the 2 by 2 Table
> anychd_highbp1 <- matrix(nrow =2, ncol=2)
> # Add column 1
> anychd_highbp1[,1] <- tapply(anychd=="Yes",highbp1,sum)
> # Add column 2
> anychd_highbp1[,2] <- tapply(timechd,highbp1,sum)
> # Now the table look like this
> #    Disease +   Time at risk
> # Expose -   c   d
> # Expose +   a   b
>
> # Get the correct table
> anychd_highbp1 <- rbind(anychd_highbp1[2,], anychd_highbp1[1,])
>
> #    Disease +   Time at risk
> # Expose +   a   b
> # Expose -   c   d
> #Use anychd_highbp1 to get Incidence Rate per 1000 persons (units = 1000)
> epi.2by2(anychd_highbp1, method = "cohort.time", conf.level = 0.95,
+          units = 1000, homogeneity = "breslow.day", verbose = F)

             Disease +    Time at risk        Inc rate *
Exposed +          605           25541              23.7
Exposed -          635           55384              11.5
Total             1240           80925              15.3

Point estimates and 95 % CIs:
---------------------------------------------------------
Inc rate ratio                           2.07 (1.85, 2.31)
Attrib rate *                            12.22 (10.13, 14.31)
Attrib rate in population *              3.86 (2.62, 5.09)
Attrib fraction in exposed (%)           51.6 (45.81, 56.77)
Attrib fraction in population (%)        25.17 (23.5, 26.81)
---------------------------------------------------------
 * Cases per 1000 units of population time at risk
```