Problem Set 2

Felipe Riveroll Aguirre*

October 28, 2012

1 BMI and CHD prevalence

The following table uses data from the NHLBI teaching data set and displays categories of body mass index 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 with existing Coronary Heart Disease (CHD) at that exam (prevchd1=1)

Category	BMI	Number of subjects	Cases of CHD at exam 1
Under Weight	$BMI < 18.5$	57	0
Normal Weight	$18.5 \leq BMI \geq 25$	1936	66
Overweight	$25 < BMI \geq 30$	1848	90
Obese	$30 < BMI$	574	38
	Total	4415	194

1. What is the prevalence of obesity among the 4415 participants at the 1956 exam?

   $A = \frac{574}{4415} = 0.13001$

```
> library("foreign") #needed for read.dta function
> ### Import Data from STATA ###
> data <- read.dta("https://dl.dropbox.com/u/4828275/fhs.dta",
+                  convert.factors = TRUE,
+                  missing.type = TRUE)
> attach(data)
>
```

*friveroll@gmail.com

```
> bmi1_cat <- NA
> bmi1_cat[bmi1<18.5] <- 1
> bmi1_cat[bmi1>=18.5 & bmi1 <= 25] <- 2
> bmi1_cat[bmi1>25 & bmi1 <= 30] <- 3
> bmi1_cat[bmi1>30 & !is.na(bmi1)] <- 4
> library("gmodels")
> with(data, CrossTable(bmi1_cat, prevchd1, prop.chisq=F, digits= 4))
```

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


Total Observations in Table:  4415


             | prevchd1
    bmi1_cat |        No |       Yes | Row Total |
-------------|-----------|-----------|-----------|
           1 |        57 |         0 |        57 |
             |    1.0000 |    0.0000 |    0.0129 |
             |    0.0135 |    0.0000 |           |
             |    0.0129 |    0.0000 |           |
-------------|-----------|-----------|-----------|
           2 |      1870 |        66 |      1936 |
             |    0.9659 |    0.0341 |    0.4385 |
             |    0.4430 |    0.3402 |           |
             |    0.4236 |    0.0149 |           |
-------------|-----------|-----------|-----------|
           3 |      1758 |        90 |      1848 |
             |    0.9513 |    0.0487 |    0.4186 |
             |    0.4165 |    0.4639 |           |
             |    0.3982 |    0.0204 |           |
-------------|-----------|-----------|-----------|
           4 |       536 |        38 |       574 |
             |    0.9338 |    0.0662 |    0.1300 |
             |    0.1270 |    0.1959 |           |
             |    0.1214 |    0.0086 |           |
-------------|-----------|-----------|-----------|
Column Total |      4221 |       194 |      4415 |
             |    0.9561 |    0.0439 |           |
-------------|-----------|-----------|-----------|
```

2. What is the prevalence of CHD at the 1956 exam among the 4415 participants at the 1956 exam?
   R $= \frac{194}{4415} = 0.0439$

3. What is the prevalence of CHD at the 1956 exam for each of the body mass index classes?

   (a) Under Weight Participants
   $A = 0$

   (b) Normal Weight Participants
   $A = \frac{66}{1936} = 0.0341$

   (c) Overweight Participants
   $A = \frac{90}{1848} = 0.0487$

   (d) Obese Participants
   $A = \frac{38}{574} = 0.0662$

```
> bmi1_Vs_prevchd1 <- with(data, table(bmi1_cat, prevchd1))
```

```
> # Row proportions
> prop.table(bmi1_Vs_prevchd1, 1)
```

	CHD at exam 1	
Category	No	Yes
Under Weight	1.0000	0.0000
Normal Weight	0.9659	0.0341
Overweight	0.9513	0.0487
Obese	0.9338	0.0662

Diabetes prevalence. Use R and the NHLBI data set to calculate the prevalence of diabetes among participants who attended and had non-missing data on diabetes at all three examinations. **(Hint: There were 3,206 such participants.)**

```
> diabetestot <- na.exclude(data.frame(diabetes1, diabetes2, diabetes3))
> library("epicalc")
```

1. What is the prevalence of diabetes at the first exam **(diabetes1=1)**?

```
> tab1(diabetestot$diabetes1, graph=F, digits=4)

diabetestot$diabetes1 :
        Frequency Percent Cum. percent
No           3148    98.2         98.2
Yes            58     1.8        100.0
  Total      3206   100.0        100.0
```

2. What is the prevalence of diabetes at the second exam **(diabetes2=1)**?

```
> tab1(diabetestot$diabetes2, graph=F, digits=4)

diabetestot$diabetes2 :
        Frequency Percent Cum. percent
No           3101    96.7         96.7
Yes           105     3.3        100.0
  Total      3206   100.0        100.0
```

3. What is the prevalence of diabetes at the third exam **(diabetes3=1)**?

```
> tab1(diabetestot$diabetes3, graph=F, digits=4)

diabetestot$diabetes3 :
        Frequency Percent Cum. percent
No           2955    92.2         92.2
Yes           251     7.8        100.0
  Total      3206   100.0        100.0
```

BMI and hypertension prevalence. Use Stata and the BMI1 variable in the NHLBI data set to create the four categories of body mass index as defined in the first question.

1. What is the prevalence of hypertension **(prevhyp1=1)** at the 1956 exam for each of the body mass index classes?

```
> bmi1_Vs_prevhyp1 <- with(data, table(bmi1_cat, prevhyp1))
```

   (a) Under Weight Participants

   (b) Normal Weight Participants

   (c) Overweight Participants

   (d) Obese Participants

```
> prop.table(bmi1_Vs_prevhyp1, 1)
```

	Hypertension at exam 1	
Category	No	Yes
Under Weight	0.8947	0.1053
Normal Weight	0.7944	0.2056
Overweight	0.6304	0.3696
Obese	0.4146	0.5854

Hypertension and high blood pressure. Use R to create a binary variable (highbp1) to represent the presence/absence of high blood pressure at the 1956 examination

```
> highbp1 <- NA
> 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.)

```
> highbp1_Vs_prevchd1 <- with(data, table(highbp1, prevchd1))
```

1. What is the prevalence of CHD (prevchd1=1) at the 1956 exam for participants with high blood pressure at the 1956 exam (highbp1=1)?

2. What is the prevalence of CHD (prevchd1=1) at the 1956 exam for participants without high blood pressure at the 1956 exam (highbp1=0)?

   ```
   > prop.table(highbp1_Vs_prevchd1, 1)
   ```

	Hypertension at exam 1	
blood pressure	No	Yes
Normal	0.9687	0.0313
High	0.9345	0.0655

2 HYPOTHETICAL LIFE TABLE

The table below lists the number of individuals at age x for a hypothetical population in 1950-1952 and 1990 - 1992

Age	1950-1952	1990-1992
0	100000	100000
20	73412	96902
40	56884	92638
70	31744	79873

```
> life_table <- read.csv("https://dl.dropbox.com/u/4828275/life_table_hwk2.csv")
```

1. What is the probability of surviving from birth to age 20 in 1950-1952?

```
> X1950.X1952 <- matrix(nrow = 3, ncol = 3)
```

Number of people dying between ages x and $x + n$
${}_nd_x = l_x - l_{x+n}$

```
> X1950.X1952[,1] <- (embed(life_table[,2],2)
+                     - embed(life_table[,2],2)[,1])[,2]
```

Probability of dying between ages x and x + n
${}_nq_x = n_{dx}/l_x$

```
> X1950.X1952[,2] <- X1950.X1952[,1]/life_table[1:3,2]
```

Probability of surviving between ages x and x + n
${}_np_x = 1 - n_{qx}$

```
> X1950.X1952[,3] <- 1 - X1950.X1952[,2]
```

1950 − 1952

Age	${}_nd_x$	${}_nq_x$	${}_np_x$
0 − 20	26588	0.2659	0.7341
20 − 40	16528	0.2251	0.7749
40 − 70	25140	0.4420	0.5580

2. What is the probability of surviving from age 40 to age 70 in 1990-1992?

```
> X1990.X1992 <- matrix(nrow = 3, ncol = 3)
> #ndx Number of people dying between ages x and x + n
> X1990.X1992[,1] <- (embed(life_table[,3],2)
+                     - embed(life_table[,3],2)[,1])[,2]
> #nqx Probability of dying between ages x and x + n
> X1990.X1992[,2] <- X1990.X1992[,1]/life_table[1:3,3]
> #npx Probability of surviving between ages x and x + n
> X1990.X1992[,3] <- 1 - X1990.X1992[,2]
```

1990 − 1992			
Age	${}_nd_x$	${}_nq_x$	${}_np_x$
0 − 20	3098	0.0310	0.9690
20 − 40	4264	0.0440	0.9560
40 − 70	12765	0.1378	0.8622

3. Define the absolute survival increase over the 40 year span as $p_1 - p_2$, where p_1 is the chance of surviving from age x to age $x + n$ in 1990-1992 and p_2 is the chance of surviving from age x to age $x + n$ in 1950-1952. Which age group has the greatest absolute survival increase?

   (a) 0 − 20 , (b) 20 − 40, (c) 40 − 70

```
> abs_surv_inc <- X1990.X1992[,3] -X1950.X1952[,3]
```

4. Define the relative survival increase over the 40 year span as $\frac{p_1 - p_2}{p_2}$, where p_1 is the chance of surviving from age x to age x+n in 1990-1992 and p2 is the chance of surviving from age x to age $x + n$ in 1950-1952. Which age group has the greatest relative survival increase?

   (a) 0 − 20 , (b) 20 − 40, (c) 40 − 70

```
> rel_surv_inc <- abs_surv_inc/X1950.X1952[,3]

> data.frame(abs_surv_inc, rel_surv_inc)
```

Age	$p_1 - p_2$	$\frac{p_1 - p_2}{p_2}$
0 − 20	0.2349	0.3200
20 − 40	0.1811	0.2338
40 − 70	0.3042	0.5450