# R Demo on Prevalence

Elizabeth Mostofsky[*] Felipe Riveroll Aguirre[†]

October 28, 2012

## 1 Part I

1. Objectives
   (a) Calculate the prevalence of smoking in the Framingham Data Set and interpret the results
   (b) Restrict an analysis to non-missing data
   (c) Create a 2 way table to examine changes in self-reported smoking status between visit 1 and visit 2
2. Calculate the proportion of people at each visit that report current smoking (NA+) and the proportion of people at each visit that report current smoking among those with data on smoking status at that visit (NA-). In this data set, current smoking status us coded as "0 = not current smoker, 1= current smoker"
   (a) Install the required package Foreign to read the dataset

```
> install.packages("foreign", dependencies = TRUE)
```

   (b) Load the library Foreign

```
> library("foreign")
```

   (c) Load and attach the dataset in a dataframe named data.

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

   (d) Install and load the package epicalc

```
> install.packages("epicalc", dependencies = TRUE)
> library("epicalc")
```

   (e) Use tab1 from epicalc to get one-way tabulation to get the frequency table for cursmoke 1,2 and 3.

[*] STATA tutorial

[†] R Version friveroll@gmail.com

```
> tab1(cursmoke1, graph=F, cum.percent = any(is.na(cursmoke1)))

cursmoke1 :
        Frequency Percent
No           2253    50.8
Yes          2181    49.2
  Total      4434   100.0

> tab1(cursmoke2, graph=F, cum.percent = any(is.na(cursmoke2)))

cursmoke2 :
        Frequency   %(NA+) cum.%(NA+)   %(NA-) cum.%(NA-)
No           2203     49.7       49.7     56.1       56.1
Yes          1727     38.9       88.6     43.9      100.0
NAs          504     11.4      100.0      0.0      100.0
  Total      4434    100.0      100.0    100.0      100.0

> tab1(cursmoke3, graph=F, cum.percent = any(is.na(cursmoke3)))

cursmoke3 :
        Frequency   %(NA+) cum.%(NA+)   %(NA-) cum.%(NA-)
No           2142     48.3       48.3     65.6       65.6
Yes          1121     25.3       73.6     34.4      100.0
NAs         1171     26.4      100.0      0.0      100.0
  Total      4434    100.0      100.0    100.0      100.0
```

NA+ proportion of people with missing data
NA- proportion of people among those with data

3. Calculate the proportion of people at each visit that report current smoking among those with data on smoking status at all 3 visits.

   (a) We can create a dataframe excluding those with missing data (NA's)

   ```
   > cursmokenotmiss <- na.exclude(data.frame(cursmoke1, cursmoke2, cursmoke3))
   ```

   (b) Use `tab1` to get the proportions from the new dataframe cursmokenotmiss

   ```
   > tab1(cursmokenotmiss$cursmoke1, graph=F)
   ```

   ```
   cursmokenotmiss$cursmoke1 :
           Frequency Percent Cum. percent
   No           1681    52.4         52.4
   Yes          1525    47.6        100.0
     Total      3206   100.0        100.0
   ```

   ```
   > tab1(cursmokenotmiss$cursmoke2, graph=F)
   ```

   ```
   cursmokenotmiss$cursmoke2 :
           Frequency Percent Cum. percent
   No           1812    56.5         56.5
   Yes          1394    43.5        100.0
     Total      3206   100.0        100.0
   ```

   ```
   > tab1(cursmokenotmiss$cursmoke3, graph=F)
   ```

   ```
   cursmokenotmiss$cursmoke3 :
           Frequency Percent Cum. percent
   No           2109    65.8         65.8
   Yes          1097    34.2        100.0
     Total      3206   100.0        100.0
   ```

4. What could explain the declining prevalence of smoking?

   (a) Over time, the prevalence of smoking is declining in the population

   (b) Current smokers have a shorter life

   (c) Several smokers choose not to participate in the 2nd and 3rd visits

5. Calculate the change in smoking prevalence between the 1st and 2nd visit.

   (a) Install and load the package `gmodels`

   ```
   > install.packages("gmodels", dependencies = TRUE)
   > library("gmodels")
   ```

(b) Use the command with to generate a 2 way frequency table with CrossTable from package gmodels, including missing values.

```
> with(data, CrossTable(cursmoke1,
+                       cursmoke2,
+                       missing.include=TRUE,
+                       format="SPSS",
+                       prop.chisq=FALSE))

   Cell Contents
|-------------------------|
|                   Count |
|             Row Percent |
|          Column Percent |
|           Total Percent |
|-------------------------|

Total Observations in Table:  4434

             | cursmoke2
   cursmoke1 |        No  |       Yes  |        NA  | Row Total |
-------------|-----------|-----------|-----------|-----------|
          No |      1898  |       131  |       224  |      2253  |
             |   84.243% |    5.814% |    9.942% |   50.812% |
             |   86.155% |    7.585% |   44.444% |           |
             |   42.806% |    2.954% |    5.052% |           |
-------------|-----------|-----------|-----------|-----------|
         Yes |       305  |      1596  |       280  |      2181  |
             |   13.984% |   73.177% |   12.838% |   49.188% |
             |   13.845% |   92.415% |   55.556% |           |
             |    6.879% |   35.995% |    6.315% |           |
-------------|-----------|-----------|-----------|-----------|
Column Total |      2203  |      1727  |       504  |      4434  |
             |   49.684% |   38.949% |   11.367% |           |
-------------|-----------|-----------|-----------|-----------|

>
```

6. Calculate the change in smoking prevalence between the 1st and 2 nd visit among those with data on smoking status at both visits.

```
> with(data, CrossTable(cursmoke1,
+                       cursmoke2,
+                       format="SPSS"))

   Cell Contents
|-------------------------|
|                   Count |
| Chi-square contribution |
|             Row Percent |
|          Column Percent |
|           Total Percent |
|-------------------------|

Total Observations in Table:  3930

             | cursmoke2
   cursmoke1 |       No  |      Yes  | Row Total |
-------------|-----------|-----------|-----------|
          No |     1898  |      131  |     2029  |
             |  508.670  |  648.871  |           |
             |   93.544% |    6.456% |   51.628% |
             |   86.155% |    7.585% |           |
             |   48.295% |    3.333% |           |
-------------|-----------|-----------|-----------|
         Yes |      305  |     1596  |     1901  |
             |  542.920  |  692.561  |           |
             |   16.044% |   83.956% |   48.372% |
             |   13.845% |   92.415% |           |
             |    7.761% |   40.611% |           |
-------------|-----------|-----------|-----------|
Column Total |     2203  |     1727  |     3930  |
             |   56.056% |   43.944% |           |
-------------|-----------|-----------|-----------|
```

7. Conclusions

   (a) Smoking prevalence declined over time

      i. Smokers are quitting

      ii. Smokers have a shorter life

      iii. Smokers are less likely to participate

   (b) R can be used to

      i. Restrict an analysis to non-missing data

      ii. Create a 2 way table to cross-classify two nominal variables

# 2 Part II

1. Objectives

    (a) Create an ordinal variable from continuous data

    (b) Calculate the prevalence of CHD for different levels of smoking at visit 1

2. Calculate the prevalence of coronary heart disease (CHD) at visit 1 by categories of cigarettes per day

    "PREVCHD is defined as pre-existing angina pectoris, myocardial infarction (hospitalized, silent or unrecognized), or coronary insufficiency (unstable angina) 0 = Free of disease, 1 = Prevalent disease"

    (a) Create 4 categories of cigarette packs per day ( 0 , 1-20 , 21-40, $\geq$ 41). $Sincethevaluesreflect, aparticularordering, itisanordinalvariable.$

    ```
    > data$packs1 <- NA # initialize packs1
    > data$packs1 [data$cigpday1==0] <- 0
    > data$packs1 [data$cigpday1>=1 & data$cigpday1 <= 20] <- 1
    > data$packs1 [data$cigpday1>=21 & data$cigpday1 <= 40] <- 2
    > data$packs1 [data$cigpday1>=41 & !is.na(data$cigpday1)] <- 3
    ```

(b) Use CrossTable to get a 2 way table from packs1 and prevchd1

```
> with(data, CrossTable(packs1, prevchd1, format="SPSS"))

   Cell Contents
|-------------------------|
|                   Count |
| Chi-square contribution |
|             Row Percent |
|          Column Percent |
|           Total Percent |
|-------------------------|

Total Observations in Table:  4402

             | prevchd1
      packs1 |       No  |      Yes  | Row Total |
-------------|-----------|-----------|-----------|
           0 |     2145  |      108  |     2253  |
             |    0.049  |    1.073  |           |
             |   95.206% |    4.794% |   51.181% |
             |   50.938% |   56.545% |           |
             |   48.728% |    2.453% |           |
-------------|-----------|-----------|-----------|
           1 |     1606  |       65  |     1671  |
             |    0.035  |    0.777  |           |
             |   96.110% |    3.890% |   37.960% |
             |   38.138% |   34.031% |           |
             |   36.483% |    1.477% |           |
-------------|-----------|-----------|-----------|
           2 |      383  |       15  |      398  |
             |    0.014  |    0.298  |           |
             |   96.231% |    3.769% |    9.041% |
             |    9.095% |    7.853% |           |
             |    8.701% |    0.341% |           |
-------------|-----------|-----------|-----------|
           3 |       77  |        3  |       80  |
             |    0.003  |    0.064  |           |
             |   96.250% |    3.750% |    1.817% |
             |    1.829% |    1.571% |           |
             |    1.749% |    0.068% |           |
-------------|-----------|-----------|-----------|
Column Total |     4211  |      191  |     4402  |
             |   95.661% |    4.339% |           |
-------------|-----------|-----------|-----------|
```

3. What could explain the higher prevalence of CHD among non-smokers compared to those who smoke 1 or more cigarettes per day?
    (a) High incidence, Long duration
    (b) Cross-sectional data is susceptible to reverse causation
    (c) Other common suspects
        i. Bias
        ii. Confounding
        iii. Chance
4. Conclusions
    (a) R can be used to create an ordinal variable based on continuous data.
    (b) CHD prevalence was lower among people with higher levels of smoking.
    (c) Prevalence is a function of incidence and duration.
    (d) In addition to a causal effect of exposure on disease risk, there are several alternative explanations for observing an association between two factors of interest.