Categorical Data with R

In this session, you will handle categorical data with R.

Load data

```
## Load vcd package
library(vcd)

## Load Arthritis dataset (data frame)
data(Arthritis)
```

Indexing (1st to 17th rows only)

```
Arthritis[1:17, ]
```

```
   ID Treatment    Sex Age Improved
1  57   Treated   Male  27     Some
2  46   Treated   Male  29     None
3  77   Treated   Male  30     None
4  17   Treated   Male  32   Marked
5  36   Treated   Male  46   Marked
6  23   Treated   Male  58   Marked
7  75   Treated   Male  59     None
8  39   Treated   Male  59   Marked
9  33   Treated   Male  63     None
10 55   Treated   Male  63     None
11 30   Treated   Male  64     None
12  5   Treated   Male  64     Some
13 63   Treated   Male  69     None
14 83   Treated   Male  70   Marked
15 66   Treated Female  23     None
16 40   Treated Female  32     None
17  6   Treated Female  37     Some
```

summary() on data frame (dataset)

```
summary(Arthritis)
```

```
       ID         Treatment      Sex          Age         Improved
 Min.   : 1.0   Placebo:43   Female:59   Min.   :23.0   None  :42
 1st Qu.:21.8   Treated:41   Male  :25   1st Qu.:46.0   Some  :14
 Median :42.5                           Median :57.0   Marked:28
 Mean   :42.5                           Mean   :53.4
 3rd Qu.:63.2                           3rd Qu.:63.0
 Max.   :84.0                           Max.   :74.0
```

Access to a vector (variable) within a data frame

```
Arthritis$Treatment
```

```
 [1] Treated Treated Treated Treated Treated Treated Treated Treated Treated Treated Treated Treated Treated Treated Treated
[15] Treated Treated Treated Treated Treated Treated Treated Treated Treated Treated Treated Treated Treated Treated Treated
[29] Treated Treated Treated Treated Treated Treated Treated Treated Treated Treated Treated Treated Treated Placebo Placebo
[43] Placebo Placebo Placebo Placebo Placebo Placebo Placebo Placebo Placebo Placebo Placebo Placebo Placebo Placebo Placebo
[57] Placebo Placebo Placebo Placebo Placebo Placebo Placebo Placebo Placebo Placebo Placebo Placebo Placebo Placebo Placebo
[71] Placebo Placebo Placebo Placebo Placebo Placebo Placebo Placebo Placebo Placebo Placebo Placebo Placebo Placebo Placebo
Levels: Placebo Treated
```

Check factor levels (categories)

```
levels(Arthritis$Improved)
```

```
[1] "None"   "Some"   "Marked"
```

Ordered categorical variable

```
Arthritis$Improved
```

```
 [1] Some   None   None   Marked Marked Marked None   Marked None   None   None   Some   None   Marked None   None
[17] Some   None   Marked None   Marked Marked Marked Marked Marked Marked Marked None   Marked Marked Marked Marked
[33] Some   Marked Marked Marked Some   Marked None   Some   Some   None   None   None   None   None   None   None
[49] None   None   None   Marked None   None   None   Some   None   Marked None   None   None   None   None   None
[65] None   None   None   None   Marked Marked None   Some   Some   Some   Marked None   Some   None   Marked None
[81] None   Some   Some   Marked
Levels: None < Some < Marked
```

Check length (number of patients) of a vector (variable)

```
length(Arthritis$Improved)
```

```
[1] 84
```

Table for a single variable

```
## table()
table(Arthritis$Improved)
```

```
  None   Some Marked
    42     14     28
```

```
## summary()
summary(Arthritis$Improved)
```

```
  None   Some Marked
    42     14     28
```

Proportions for a single variable table

```
tab1 <- table(Arthritis$Improved)
prop.table(tab1)
```

```
  None   Some Marked
0.5000 0.1667 0.3333
```

Cross table by two variables

```
xtab1 <- xtabs(~ Treatment +Improved, Arthritis)
xtab1
```

```
         Improved
Treatment None Some Marked
  Placebo   29    7      7
  Treated   13    7     21
```

Add margins (sums)

```
addmargins(xtab1)
```

```
         Improved
Treatment None Some Marked Sum
  Placebo   29    7      7  43
  Treated   13    7     21  41
  Sum       42   14     28  84
```

Proportions in cross table (margin 1: row proportion; 2: column proportion)

```
prop.table(xtab1)              # proportion to total
```

```
         Improved
Treatment    None    Some  Marked
  Placebo 0.34524 0.08333 0.08333
  Treated 0.15476 0.08333 0.25000
```

```
prop.table(xtab1, margin = 1)   # proportion to row sum
```

```
         Improved
Treatment   None   Some Marked
  Placebo 0.6744 0.1628 0.1628
  Treated 0.3171 0.1707 0.5122
```

```
prop.table(xtab1, margin = 2)   # proportion to column sum
```

```
         Improved
Treatment   None   Some Marked
  Placebo 0.6905 0.5000 0.2500
  Treated 0.3095 0.5000 0.7500
```

Stratified table and flat table

```
## 3rd variable as stratified variable
xtab2 <- xtabs(~ Treatment +Improved +Sex, Arthritis)
xtab2
```

```
, , Sex = Female

         Improved
Treatment None Some Marked
  Placebo   19    7      6
  Treated    6    5     16

, , Sex = Male

         Improved
Treatment None Some Marked
  Placebo   10    0      1
  Treated    7    2      5
```

```
## flat table
ftable(xtab2)
```

```
                  Sex Female Male
Treatment Improved
Placebo   None             19   10
          Some              7    0
          Marked            6    1
Treated   None              6    7
          Some              5    2
          Marked           16    5
```

SAS-like cross table

```
library(gmodels)
tab1 <- xtabs(~ Treatment +Improved, Arthritis)
CrossTable(tab1)
```

```
   Cell Contents
|-------------------------|
|                       N |
| Chi-square contribution |
|           N / Row Total |
|           N / Col Total |
|         N / Table Total |
|-------------------------|

Total Observations in Table:  84

             | Improved
   Treatment |      None |      Some |    Marked | Row Total |
-------------|-----------|-----------|-----------|-----------|
     Placebo |        29 |         7 |         7 |        43 |
             |     2.616 |     0.004 |     3.752 |           |
             |     0.674 |     0.163 |     0.163 |     0.512 |
             |     0.690 |     0.500 |     0.250 |           |
             |     0.345 |     0.083 |     0.083 |           |
-------------|-----------|-----------|-----------|-----------|
     Treated |        13 |         7 |        21 |        41 |
             |     2.744 |     0.004 |     3.935 |           |
             |     0.317 |     0.171 |     0.512 |     0.488 |
             |     0.310 |     0.500 |     0.750 |           |
             |     0.155 |     0.083 |     0.250 |           |
-------------|-----------|-----------|-----------|-----------|
Column Total |        42 |        14 |        28 |        84 |
             |     0.500 |     0.167 |     0.333 |           |
-------------|-----------|-----------|-----------|-----------|
```

Epidemiologists' favorite 2x2 table with RR, OR, and RD

```
library(epiR)
tab.2by2 <- xtabs(~ Sex +Treatment, Arthritis)
tab.2by2
```

```
        Treatment
Sex      Placebo Treated
  Female      32      27
  Male        11      14
```

```
epi.2by2(tab.2by2, units = 1)
```

```
             Disease +    Disease -      Total        Inc risk *        Odds
Exposed +           32           27         59             0.542       1.185
Exposed -           11           14         25             0.440       0.786
Total               43           41         84             0.512       1.049

Point estimates and 95 % CIs:
-------------------------------------------------------------
Inc risk ratio                          1.23 (0.75, 2.03)
Odds ratio                              1.51 (0.59, 3.87)
Attrib risk *                           0.1 (-0.13, 0.33)
Attrib risk in population *             0.07 (-0.15, 0.29)
Attrib fraction in exposed (%)          18.87 (-33.82, 50.82)
Attrib fraction in population (%)       14.05 (-24.81, 40.81)
-------------------------------------------------------------
 * Cases per population unit
```