R Handout - Univariate EDA

Dr. Derek Ogle
Northland College

Fall 2014, MTH107 Statistics

Background

Measurements of the levels of arsenic in the drinking water, cooking water, and toenail samples, as well as related covariates, were measured on 21 individuals with private wells in a New Hampshire community. The variables below were recorded in the Arsenic.txt file located on the R Resources web page.

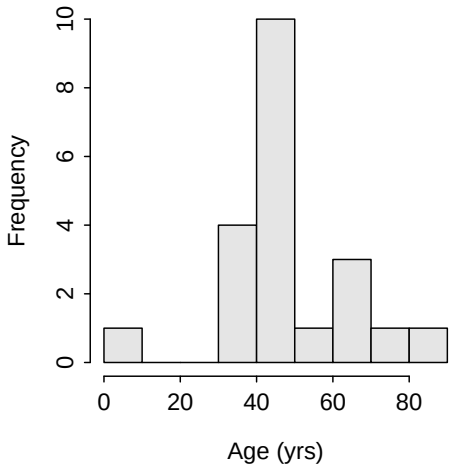
- *age*: Age (yrs) of person
- *sex*: Sex of person
- *usedrink*: Household well used for drinking (A="<1/4", B="≈1/4", C="≈1/2", D="≈3/4", E=">3/4")
- *usecook*: Household well used for cooking (A="<1/4", B="≈1/4", C="≈1/2", D="≈3/4", E=">3/4")
- *arswater*: Arsenic in water (ppm)
- *arsnails*: Arsenic in toenails (ppm)

Initialization

```
> library(NCStats)
> setwd("C:/aaaWork/Class Materials/MTH107/Lecture/HOs/")
> Ars <- read.table("Arsenic.txt",header=TRUE)
> str(Ars)
'data.frame': 21 obs. of  6 variables:
 $ age     : int  44 45 44 66 37 45 47 38 41 49 ...
 $ sex     : Factor w/ 2 levels "F","M": 1 1 2 1 2 1 2 1 1 1 ...
 $ usedrink: Factor w/ 5 levels "A","B","C","D",..: 5 4 5 3 2 5 5 4 3 4 ...
 $ usecook : Factor w/ 2 levels "B","E": 2 2 2 2 2 2 2 2 1 2 ...
 $ arswater: num  0.00087 0.00021 0 0.00115 0 0 0.00013 0.00069 0.00039 0 ...
 $ arsnails: num  0.119 0.118 0.099 0.118 0.277 0.358 0.08 0.158 0.31 0.105 ...
> view(Ars)
   age sex usedrink usecook arswater arsnails
2   45   F        D       E  0.00021    0.118
4   66   F        C       E  0.00115    0.118
11  72   F        E       E  0.00000    0.073
14  86   F        E       E  0.13700    2.252
18  63   F        E       E  0.00000    0.141
21  36   M        E       E  0.00410    0.175
```

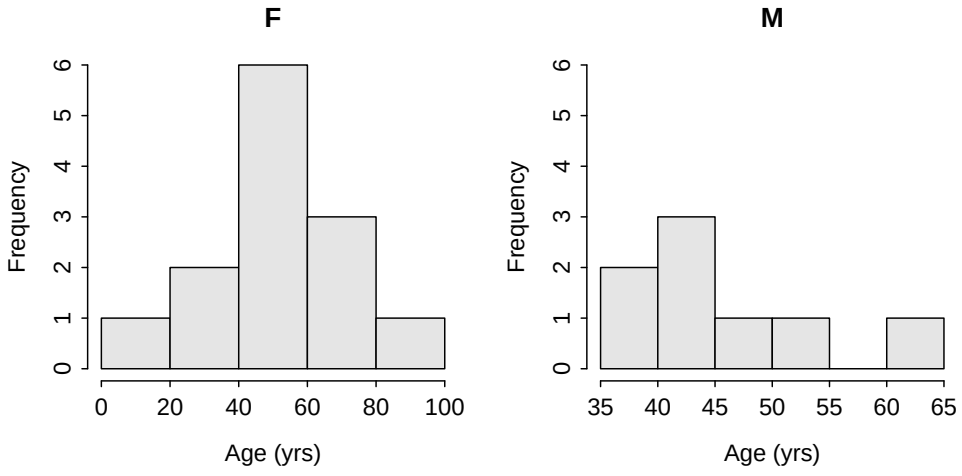
1 Univariate EDA – Quantitative

```
> Summarize(~age,data=Ars,digits=2)
      n     mean       sd      min       Q1   median       Q3      max percZero
  21.00    47.57    16.08     8.00    41.00    45.00    53.00    86.00     0.00
> hist(~age,data=Ars,main="",xlab="Age (yrs)")
```



2 Univariate EDA – Quantitative (Separated by Groups)

```
> Summarize(age~sex,data=Ars,digits=2)
  sex  n  mean    sd min   Q1 median   Q3 max percZero
1   F 13 48.77 19.60   8 41.0     45 63.0  86        0
2   M  8 45.62  8.53  36 40.8     44 48.5  62        0
> hist(age~sex,data=Ars,xlab="Age (yrs)",col="gray90")
```



3 Univariate EDA – Categorical

```
> ( tbl.drink <- xtabs(~usedrink,data=Ars) )
usedrink
 A  B  C  D  E
 1  1  2  3 14
> percTable(tbl.drink,digits=1)
usedrink
    A     B     C     D     E   Sum
  4.8   4.8   9.5  14.3  66.7 100.1
> barplot(tbl.drink,xlab="Rating of Use for Drinking",ylab="Frequency",col="gray90")
```