Univariate EDA

Derek H. Ogle

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 https://github.com/droglenc/NCData/blob/master/Arsenic.csv file located on the R Resources web page.

- `age`: Age (yrs) of person
- `sex`: Sex of person
- `usedrink`: Household well used for drinking – A="$< \frac{1}{4}$", B="$\approx \frac{1}{4}$", C="$\approx \frac{1}{2}$", D="$\approx \frac{3}{4}$", E="$> \frac{3}{4}$"
- `usecook`: Household well used for cooking – A="$< \frac{1}{4}$", B="$\approx \frac{1}{4}$", C="$\approx \frac{1}{2}$", D="$\approx \frac{3}{4}$", E="$> \frac{3}{4}$"
- `arswater`: Arsenic in water (ppm)
- `arsnails`: Arsenic in toenails (ppm)

Getting the Data

```
> library(NCStats)
> setwd("C:/aaaWork/Web/GitHub/NCMTH107/lecture/HOs")
> Ars <- read.csv("Arsenic.csv")
> 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)
```

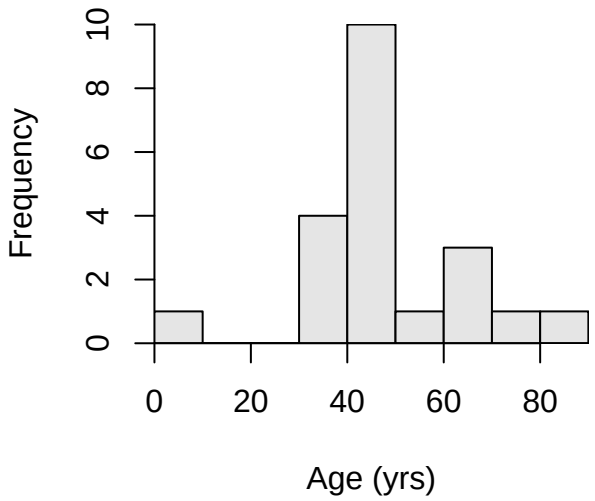
```
   age sex usedrink usecook arswater arsnails
1   44   F        E       E  0.00087    0.119
2   45   F        D       E  0.00021    0.118
4   66   F        C       E  0.00115    0.118
5   37   M        B       E  0.00000    0.277
13  53   M        E       E  0.01940    0.517
14  86   F        E       E  0.13700    2.252
```

Univariate EDA – Quantitative

```
> Summarize(~age,data=Ars,digits=2)
```

```
      n   nvalid     mean       sd      min       Q1   median       Q3      max percZero
  21.00    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)")
```

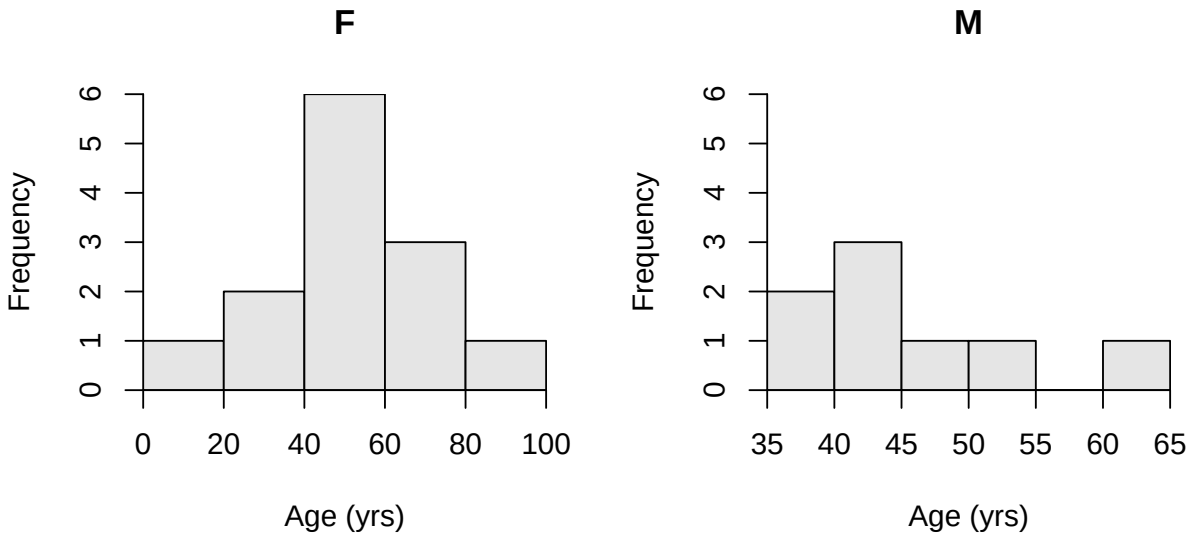


Univariate EDA – Quantitative (Separated by Groups)

```
> Summarize(age~sex,data=Ars,digits=2)
```

```
  sex  n nvalid  mean    sd min    Q1 median   Q3 max percZero
1   F 13     13 48.77 19.60   8 41.00     45 63.0  86        0
2   M  8      8 45.62  8.53  36 40.75     44 48.5  62        0
```

```
> hist(age~sex,data=Ars,xlab="Age (yrs)",col="gray90")
```



Univariate EDA – Categorical

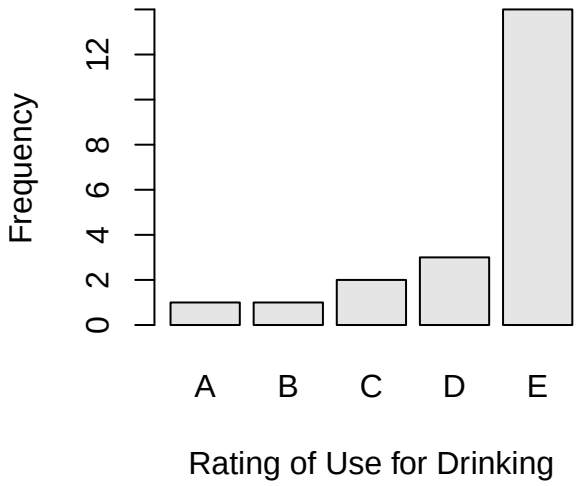
```
> ( tbl.drink <- xtabs(~usedrink,data=Ars) )
```

```
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")
```


Frequency
0
2
4
6
8
12
A
B
C
D
E
Rating of Use for Drinking