

# Univariate EDA in R Handout

## 1 Background

Measurements of drinking water and toenail levels of arsenic, 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 (1=<1/4 2=1/4 3=1/2 4=3/4 5=>3/4)
- *usecook* : Household well used for cooking (1=<1/4 2=1/4 3=1/2 4=3/4 5=>3/4)
- *arswater* : Arsenic in water (ppm)
- *arsnails* : Arsenic in toenails (ppm)

## 2 Initialization

```
> library(NCStats)
```

You must change the directory with the File...Change Dir menu or `setwd()` (as below, but for your directory) before the following command

```
> setwd("C://aaaWork//Class Materials//MTH107//F08//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: int   5 4 5 3 2 5 5 4 3 4 ...
 $ usecook  : int   5 5 5 5 5 5 5 5 2 5 ...
 $ 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 ...
```

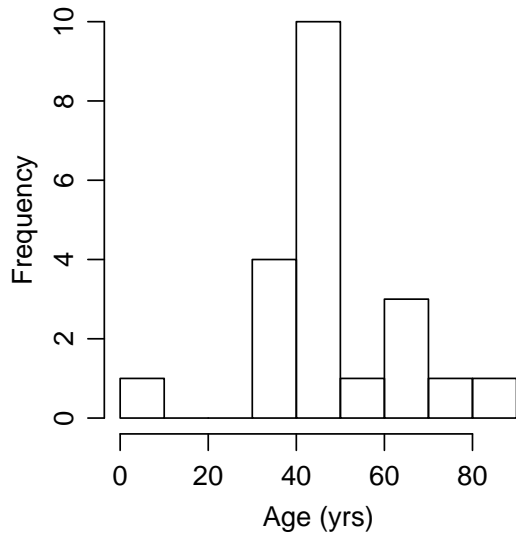
```
> Ars$f.usedrink <- factor(Ars$usedrink)
> Ars$f.usecook <- factor(Ars$usecook)
> str(Ars)
```

```
'data.frame':      21 obs. of  8 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 : int   5 4 5 3 2 5 5 4 3 4 ...
 $ usecook  : int   5 5 5 5 5 5 5 5 2 5 ...
 $ 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 ...
 $ f.usedrink: Factor w/ 5 levels "1","2","3","4",...: 5 4 5 3 2 5 5 4 3 4 ...
 $ f.usecook : Factor w/ 2 levels "2","5": 2 2 2 2 2 2 2 2 1 2 ...
```

```
> attach(Ars)
```

### 3 Univariate EDA – Quantitative

```
> hist(age,main="",xlab="Age (yrs)")
```

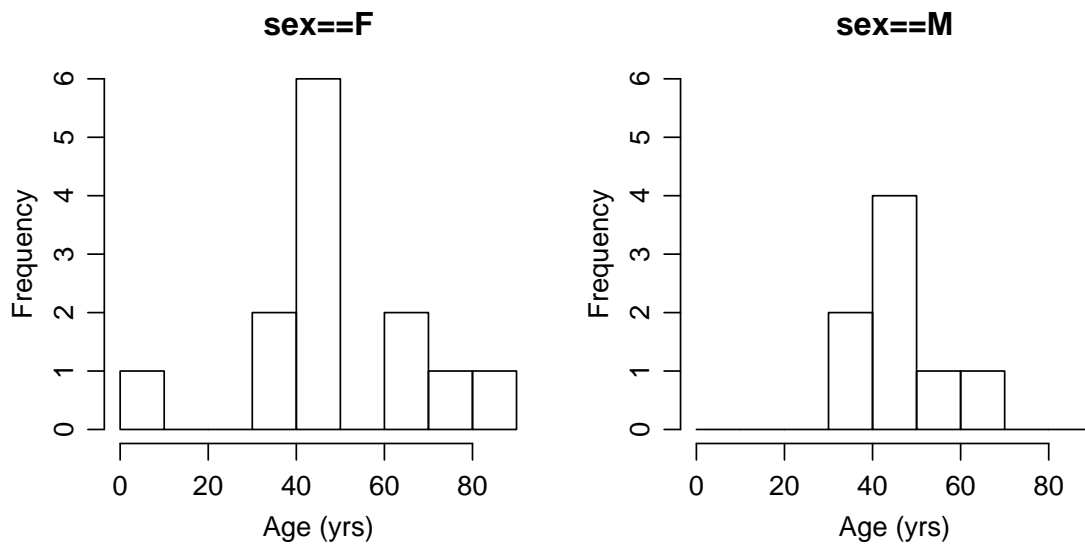


```
> Summarize(age,numdigs=2)
```

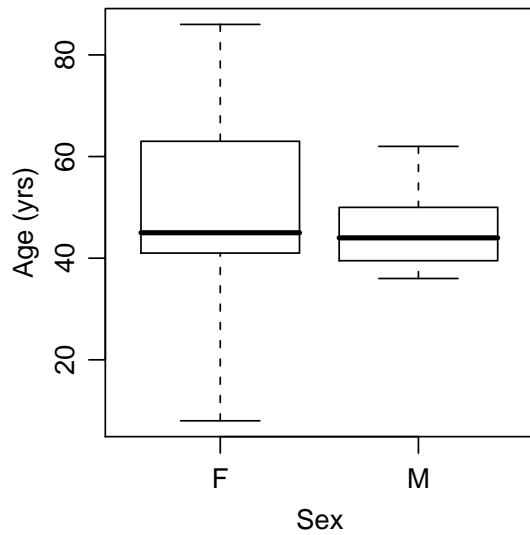
n	Mean	St. Dev.	Min.	1st Qu.	Median	3rd Qu.	Max.
21.00	47.57	16.08	8.00	41.00	45.00	53.00	86.00

### 4 Univariate EDA – Quantitative (Separated by Groups)

```
> hist(age~sex,xlab="Age (yrs)")
```



```
> boxplot(age~sex,ylab="Age (yrs)",xlab="Sex")
```



```
> Summarize(age~sex,numdigs=2)
```

```
   n Mean St. Dev. Min. 1st Qu. Median 3rd Qu. Max.
F 13 48.77  19.60   8  41.00   45  63.0  86
M  8 45.62   8.53  36  40.75   44  48.5  62
```

## 5 Univariate EDA – Categorical

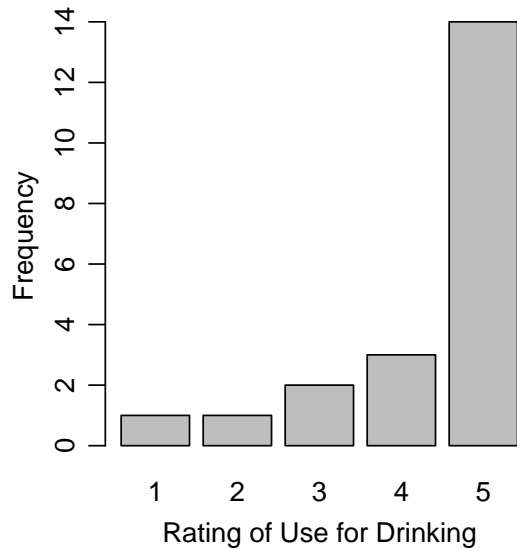
```
> tbl.drink <- table(f.usedrink)
> tbl.drink
```

```
f.usedrink
 1  2  3  4  5
 1  1  2  3 14
```

```
> prop.table(tbl.drink)*100
```

```
f.usedrink
 1      2      3      4      5
4.761905 4.761905 9.523810 14.285714 66.666667
```

```
> barplot(tbl.drink,xlab="Rating of Use for Drinking",ylab="Frequency")
```



## 6 Good “Ending” Commands

```
> detach(Ars)
```

## 7 Class Exercise

Coarse woody debris (CWD) in lakes is important for aquatic systems as it provides refuge for young fish and invertebrates as well as provide areas for periphyton to grow. CWD was studied in the north basin of Allequash Lake in northern Wisconsin. Among other things the researchers recorded the diameter of CWD found in the lake littoral zone and a qualitative measure of the degree to which the location where the CWD was found was exposed to winds (low or medium). The data (sampled from information at [North Temperate Lakes Long Term Ecological Research website](#)) they observed are shown below.

```
diameter 21 15 18 23 18 17 19 17 15 22 16 20 16 17 18 15 16 24
exposure med med med low med low med med med med med med low med med med med low
```

```
diameter 18 17 19 17 17 15 17 18 19 31 25 15 17 34 16 18 19 15
exposure med med med med med med med med low med med med med low low med med med
```

```
diameter 20 23 34 20 17 20 15 34 18 24 16 15 24 23
exposure med med low med med med med med med med low med med med
```

Use this information to answer the following questions.

1. Enter the data into Excel, save as a tab-delimited text file, and read the data into R. [HINT: see Section 2.3.1 in textbook or [this FAQ](#).]
2. Describe the distribution of the diameter of CWD. Use a histogram and summary statistics to support your description.
3. Describe the distribution of the diameter of CWD separately for the low- and medium-exposed sites. Use a histogram and summary statistics, constructed from one R command each, to support your description.

4. Describe the distribution of the exposure variable. Use a frequency table, percentage table, and bar chart to support your answer.
5. Show the `Subset()` command that you would use to isolate the following subsets of data.
  - (a) Only CWD at low-exposure sites.
  - (b) Only CWD where the diameter was greater than 20 cm.
  - (c) Only CWD where the diameter was greater than 20 cm and was observed in low-exposure sites.
  - (d) Only the third individual CWD.