Univariate EDA

Derek H. Ogle

## Background

[Karagas et al. (1996)](http://www.ncbi.nlm.nih.gov/pubmed/8896897) conducted a pilot study to assess the utility of [arsenic](https://en.wikipedia.org/wiki/Arsenic) concentrations in the [toenail](https://en.wikipedia.org/wiki/Nail_%28anatomy%29) as an indicator of ingestion of arsenic-containing water. They interviewed 21 individuals whose household drinking water supply was provided by a private (unregulated) well, including 10 individuals who lived in areas of New Hampshire where elevated water levels of arsenic had been reported previously. Each participant also provided a sample of water and toenail clippings.

The data are recorded in [Arsenic.csv](https://github.com/droglenc/NCData/raw/master/Arsenic.csv). Descriptions of the variables are below.

* age: Age (yrs) of person
* sex: Sex of person
* usedrink: How much (fraction of time) the well is used for drinking -- A="", B="", C="", D="", E=""
* usecook: How much (fraction of time) the well is used for cooking -- A="", B="", C="", D="", E=""
* arswater: Arsenic in water (ppm)
* arsnails: Arsenic in toenails (ppm)

## Getting the Data

> library(NCStats)  
> setwd("C:/aaaWork/Web/GitHub/NCMTH107/resources/class/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 ...

> headtail(ars)

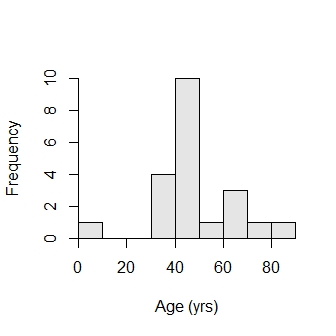
age sex usedrink usecook arswater arsnails  
1 44 F E E 0.00087 0.119  
2 45 F D E 0.00021 0.118  
3 44 M E E 0.00000 0.099  
19 42 M E E 0.01650 0.275  
20 62 M E E 0.00012 0.135  
21 36 M E E 0.00410 0.175

## 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)")

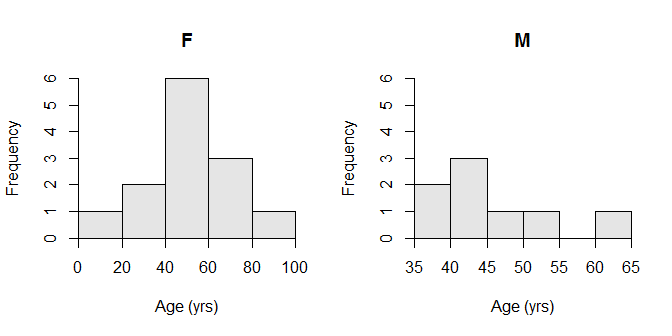


## Univariate EDA -- Quantitative (Separated by Groups)

> 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")

