R Handout - Data Frames

Dr. Derek Ogle
Northland College

Mar 2014, Vermont CFWRU Workshop

Load Necessary Packages

```
> library(FSA)      # for Subset(), view(), lencat()
```

Reading External Data

When beginning with R it is easiest to load external data with the following steps:

1. Enter data in an external software (e.g., spreadsheet or database) and save as a "comma separated values" (CSV) file.
2. Start an initial script in RStudio. Save this script to the **exact same folder** as the CSV file.
3. Use the *"Session"*, *"Set Working Directory ..."*, *"To Source File Location"* menu items to print the appropriate `setwd()` function to the *Console* pane.
4. Copy the appropriate `setwd()` function from the *Console* pane to your script.
5. Use `read.csv()` to load the external file into the R environment (described below).
6. Use `str()`, `head()`, `view()`[1], or view the file from the *Environment* tab (upper-right pane of RStudio) to make sure the data appears proper.

```
> setwd("C:/aaaWork/Web/fishR/courses/Vermont2014/CourseMaterial/") # Derek's computer only
> d <- read.csv("Data/MNBCData.csv",header=TRUE)
> str(d)
'data.frame': 2422 obs. of  20 variables:
 $ species: Factor w/ 9 levels "BLC","BLG","LMB",..: 1 1 1 1 1 1 1 1 1 1 ...
 $ gear   : Factor w/ 5 levels "All","GN","GN, TN",..: 5 5 5 5 5 5 5 5 3 3 ...
 $ lake   : Factor w/ 24 levels "Bean Lake","Benton",..: 2 2 2 2 2 2 3 3 16 16 ...
 $ yearcap: int  2006 2006 2006 2006 2006 2006 2006 2006 2006 2006 ...
 $ fish   : int  67 43 45 44 42 41 63 78 56 55 ...
 $ agecap : int  1 2 4 4 4 4 4 4 1 1 ...
 $ lencap : int  108 198 258 247 249 235 278 284 135 127 ...
 $ anu1   : num  1.15 1.42 1.52 1.28 1.43 ...
 $ anu2   : num  2.28 3.06 3.28 3.35 2.53 ...
 $ anu3   : num  NA 3.88 4.27 3.86 3.36 ...
 $ anu4   : num  NA NA 4.88 4.11 3.68 ...
 $ anu5   : num  NA NA 5.66 4.34 3.99 ...
 $ anu6   : num  NA NA NA NA NA NA NA NA NA NA ...
 $ anu7   : num  NA NA NA NA NA NA NA NA NA NA ...
 $ anu8   : num  NA NA NA NA NA NA NA NA NA NA ...
 $ anu9   : num  NA NA NA NA NA NA NA NA NA NA ...
 $ anu10  : num  NA NA NA NA NA NA NA NA NA NA ...
 $ anu11  : num  NA NA NA NA NA NA NA NA NA NA ...
 $ anu12  : num  NA NA NA NA NA NA NA NA NA NA ...
 $ radcap : num  2.28 3.88 5.66 4.34 3.99 ...
> view(d)
```

[1] This requires the `FSA` package.

```
     species gear        lake yearcap fish agecap lencap   anu1  anu2  anu3  anu4
127      BLG   TN        Long    2006   29      3    162 0.9370 1.966 3.034 3.536
407      YEP  All   Bean Lake    1998   72      2    237 1.4616 3.344 4.228    NA
1058     BLC  All   Iowa Lake    1998  120      4    295 0.9971 2.453 3.697 4.831
1277     BLC  All  Lake Sarah    1998   46      2    193 1.4169 3.555 3.753    NA
1809     BLC  All   Long Lake    1998  128      1    107 0.5355 1.660    NA    NA
2334     BLC  All Timber Lake    1998   38      2    192 0.9008 3.043 3.789    NA
      anu5 anu6 anu7 anu8 anu9 anu10 anu11 anu12 radcap
127     NA   NA   NA   NA   NA    NA    NA    NA  3.536
407     NA   NA   NA   NA   NA    NA    NA    NA  4.228
1058 5.303   NA   NA   NA   NA    NA    NA    NA  5.303
1277    NA   NA   NA   NA   NA    NA    NA    NA  3.753
1809    NA   NA   NA   NA   NA    NA    NA    NA  1.660
2334    NA   NA   NA   NA   NA    NA    NA    NA  3.789
> nrow(d)
[1] 2422
```

Subsets of the Data

```
> d[5,]
  species gear   lake yearcap fish agecap lencap  anu1  anu2 anu3  anu4  anu5 anu6
5     BLC   TN Benton    2006   42      4    249 1.431 2.534 3.36 3.676 3.992   NA
  anu7 anu8 anu9 anu10 anu11 anu12 radcap
5   NA   NA   NA    NA    NA    NA  3.992
> d[c(5,11,17),]
   species   gear   lake yearcap fish agecap lencap  anu1  anu2 anu3  anu4  anu5
5      BLC     TN Benton    2006   42      4    249 1.431 2.534 3.36 3.676 3.992
11     BLC GN, TN   Long    2006   54      1    123 1.452 2.185   NA    NA    NA
17     BLC GN, TN   Long    2006   41      1    118 1.221 1.928   NA    NA    NA
   anu6 anu7 anu8 anu9 anu10 anu11 anu12 radcap
5    NA   NA   NA   NA    NA    NA    NA  3.992
11   NA   NA   NA   NA    NA    NA    NA  2.185
17   NA   NA   NA   NA    NA    NA    NA  1.928
> d$age[1:25]
 [1] 1 2 4 4 4 4 4 4 1 1 1 1 1 1 1 1 1 1 1 1 1 2 2 3
```

```
> levels(d$species)
[1] "BLC" "BLG" "LMB" "NOP" "PMK" "SMB" "WAE" "WHC" "YEP"
> levels(d$lake)
 [1] "Bean Lake"         "Benton"            "Bingham"           "Bingham Lake"
 [5] "Buff Lake"         "Cottonwood Lake"   "Fish Lake"         "Fox Lake"
 [9] "Hills Reservoir"   "Iowa Lake"         "Kinbrae"           "Lake Okamanpeedan"
[13] "Lake Sarah"        "Lake Shetek"       "Lake Yankton"      "Long"
[17] "Long Lake"         "Okabena Lake"      "Rock Lake"         "South Silver"
[21] "South Silver Lake" "Summit Lake"       "Talcot"            "Timber Lake"
```

```
> dBLC <- Subset(d,species=="BLC")
> xtabs(~species,data=dBLC)

species
BLC
563

> dBLCTL <- Subset(d,species=="BLC" & lake=="Talcot")
> xtabs(~species+lake,data=dBLCTL)

       lake
species Talcot
    BLC     68

> dBLCBLG <- Subset(d,species=="BLC" | species=="BLG")
> xtabs(~species,data=dBLCBLG)

species
BLC BLG
563 174

> d2 <- Subset(d,species!="BLC")
> xtabs(~species,data=d2)

species
BLG LMB NOP PMK SMB WAE WHC YEP
174  27  72  75  42 550 175 744

> dPred <- Subset(d,species %in% c("LMB","NOP","SMB","WAE"))
> xtabs(~species,data=dPred)

species
LMB NOP SMB WAE
 27  72  42 550

> dgt500 <- Subset(d,lencap>=500)
> nrow(dgt500)

[1] 173

> min(dgt500$lencap)

[1] 500
```

Adding Variables I

```
> d$lenin <- d$lencap/25.4
> d$loglen <- log(d$lencap)
> view(d)

     species  gear              lake yearcap fish agecap lencap   anu1  anu2  anu3
151      NOP    TN   Hills Reservoir    2006   26      3    511 1.3490 2.039 2.401
244      WAE GN,TN            Talcot    2006   27      2    313 1.4650 3.207 4.255
1233     YEP   All Lake Okamanpeedan    1998   44      2    139 1.2602 2.458 2.745
1488     BLC   All       Lake Shetek    1998  131      4    268 0.9118 2.651 3.961
1799     YEP   All      Lake Yankton    1998  170      3    254 1.1317 2.928 4.344
2344     BLC   All       Timber Lake    1998   94      2    224 1.1785 3.656 4.432
      anu4  anu5 anu6 anu7 anu8 anu9 anu10 anu11 anu12 radcap  lenin loglen
151  2.875    NA   NA   NA   NA   NA    NA    NA    NA  2.875 20.118  6.236
244     NA    NA   NA   NA   NA   NA    NA    NA    NA  4.255 12.323  5.746
1233    NA    NA   NA   NA   NA   NA    NA    NA    NA  2.745  5.472  4.934
1488 4.645 4.964   NA   NA   NA   NA    NA    NA    NA  4.964 10.551  5.591
1799 4.693    NA   NA   NA   NA   NA    NA    NA    NA  4.693 10.000  5.537
2344    NA    NA   NA   NA   NA   NA    NA    NA    NA  4.432  8.819  5.412
```

```
> # Create a year factor (categorical) variable
> d$fyearcap <- factor(d$yearcap)
> str(d)
'data.frame': 2422 obs. of  23 variables:
 $ species : Factor w/ 9 levels "BLC","BLG","LMB",..: 1 1 1 1 1 1 1 1 1 1 ...
 $ gear    : Factor w/ 5 levels "All","GN","GN, TN",..: 5 5 5 5 5 5 5 5 3 3 ...
 $ lake    : Factor w/ 24 levels "Bean Lake","Benton",..: 2 2 2 2 2 2 3 3 16 16 ...
 $ yearcap : int  2006 2006 2006 2006 2006 2006 2006 2006 2006 2006 ...
 $ fish    : int  67 43 45 44 42 41 63 78 56 55 ...
 $ agecap  : int  1 2 4 4 4 4 4 4 1 1 ...
 $ lencap  : int  108 198 258 247 249 235 278 284 135 127 ...
 $ anu1    : num  1.15 1.42 1.52 1.28 1.43 ...
 $ anu2    : num  2.28 3.06 3.28 3.35 2.53 ...
 $ anu3    : num  NA 3.88 4.27 3.86 3.36 ...
 $ anu4    : num  NA NA 4.88 4.11 3.68 ...
 $ anu5    : num  NA NA 5.66 4.34 3.99 ...
 $ anu6    : num  NA NA NA NA NA NA NA NA NA NA ...
 $ anu7    : num  NA NA NA NA NA NA NA NA NA NA ...
 $ anu8    : num  NA NA NA NA NA NA NA NA NA NA ...
 $ anu9    : num  NA NA NA NA NA NA NA NA NA NA ...
 $ anu10   : num  NA NA NA NA NA NA NA NA NA NA ...
 $ anu11   : num  NA NA NA NA NA NA NA NA NA NA ...
 $ anu12   : num  NA NA NA NA NA NA NA NA NA NA ...
 $ radcap  : num  2.28 3.88 5.66 4.34 3.99 ...
 $ lenin   : num  4.25 7.8 10.16 9.72 9.8 ...
 $ loglen  : num  4.68 5.29 5.55 5.51 5.52 ...
 $ fyearcap: Factor w/ 2 levels "1998","2006": 2 2 2 2 2 2 2 2 2 2 ...
> levels(d$fyearcap)
[1] "1998" "2006"
```

```
> # Create a length categorization variable
> d <- lencat(~lencap,data=d,startcat=75,w=25)
> view(d)
     species  gear        lake yearcap fish agecap lencap   anu1  anu2  anu3  anu4
161      NOP GN,TN     Kinbrae    2006   28      2    628 1.1290 2.552 3.035    NA
951      WAE   All    Fox Lake    1998  108      3    415 2.6896 4.362 5.059 5.456
1096     WHC   All   Iowa Lake    1998  106      2    254 0.6068 3.138 3.956    NA
1102     WHC   All   Iowa Lake    1998   91      3    292 0.8688 3.026 3.983 4.300
1373     WAE   All  Lake Sarah    1998  151      6    585 2.1288 4.105 5.638 6.319
1605     YEP   All Lake Shetek    1998   40      3    203 0.9115 2.230 3.072 3.410
      anu5  anu6  anu7 anu8 anu9 anu10 anu11 anu12 radcap  lenin loglen fyearcap
161     NA    NA    NA   NA   NA    NA    NA    NA  3.035 24.724  6.443     2006
951     NA    NA    NA   NA   NA    NA    NA    NA  5.456 16.339  6.028     1998
1096    NA    NA    NA   NA   NA    NA    NA    NA  3.956 10.000  5.537     1998
1102    NA    NA    NA   NA   NA    NA    NA    NA  4.300 11.496  5.677     1998
1373 6.717 7.015 7.168   NA   NA    NA    NA    NA  7.168 23.031  6.372     1998
1605    NA    NA    NA   NA   NA    NA    NA    NA  3.410  7.992  5.313     1998
     LCat
161   625
951   400
1096  250
1102  275
1373  575
1605  200
```

```
> xtabs(~species+LCat,data=d)
       LCat
species  75 100 125 150 175 200 225 250 275 300 325 350 375 400 425 450 475 500 525
    BLC   0  26  39  78  94  91  84  67  56  17   7   4   0   0   0   0   0   0   0
    BLG   6  16  25  65  48  10   4   0   0   0   0   0   0   0   0   0   0   0   0
    LMB   0   0   0   0   0   0   0   0   2   3   3   3   4   3   3   5   1   0   0
    NOP   0   0   0   0   0   0   0   0   0   0   0   2   2   6   3   2   5   3   6
    PMK   1  20  28  18   7   1   0   0   0   0   0   0   0   0   0   0   0   0   0
    SMB   1   3   0   0   9   7   0   5   6   6   2   0   0   0   2   1   0   0   0
    WAE   0   0   0   0  11  18  18  17  28  56  54  48  54  30  35  27  33  37  26
    WHC   0   5   2  15  37  50  23  18  24   1   0   0   0   0   0   0   0   0   0
    YEP   0   4  66 141 152 158 119  59  35  10   0   0   0   0   0   0   0   0   0
       LCat
species 550 575 600 625 650 675 700 725 750 775 800 825
    BLC   0   0   0   0   0   0   0   0   0   0   0   0
    BLG   0   0   0   0   0   0   0   0   0   0   0   0
    LMB   0   0   0   0   0   0   0   0   0   0   0   0
    NOP   2   6  10   9   4   4   1   2   2   0   2   1
    PMK   0   0   0   0   0   0   0   0   0   0   0   0
    SMB   0   0   0   0   0   0   0   0   0   0   0   0
    WAE  22  13   5   9   2   4   3   0   0   0   0   0
    WHC   0   0   0   0   0   0   0   0   0   0   0   0
    YEP   0   0   0   0   0   0   0   0   0   0   0   0
```

Back-Calculation Example

```
> # Focus hereafter on Talcot Lake Black Crappie in 2006 (only year sampled)
> #   and eliminate several variables not used (for illustration & simplicity)
> dBLC <- Subset(d,species=="BLC" & lake=="Talcot" & yearcap==2006,
                 select=-c(gear,yearcap,lenin,loglen,fyearcap,LCat))
> levels(dBLC$species)
[1] "BLC"
> levels(dBLC$lake)
[1] "Talcot"
```

Reshaping

Some definitions:

- *"Wide" Data* – Rows contains repeated measurements on same individuals. This is currently the cases ... each row has multiple scale radii measurements for the same fish in the `anuX` variables.
- *"Long" Data* – Each row contains only one measurement from an individual. Thus, multiple measurements on the same fish will be in multiple rows.

The `reshape()` function can be used to change the shape of a data frame from wide to long, or vice versa. Within this function several items have to be defined when moving from "wide" to "long" format.

- `idvar` – The single variable name that identifies an individual (a fish in this case).
- `varying` – A vector of names for the variables that contain the repeated measurements (i.e., the variables names containing the scale radius measurements).
- `v.names` – A single name for repeated measurements variable in the long format. This will usually be very closely related to the common portion of the names in `varying`.
- `timevar` – A single name for the labels for the repeated measurements values in the long format. This is likely a descriptive name for specific portion of the names in `varying`.
- `times` – A vector of values for the repeated measurements in the long format (i.e., the ages corresponding to the radial measurements in this case).

```
> # list the variables that contain the repeated measurements
> varying1 <- c("anu1","anu2","anu3","anu4","anu5","anu6",
                "anu7","anu8","anu9","anu10","anu11","anu12")
> # this is an alternative to the above that eliminates repetitive typing
> ( varying2 <- which(grepl("anu",names(dBLC))) )
 [1]  6  7  8  9 10 11 12 13 14 15 16 17
>
> # do the reshaping
> ldBLC <- reshape(dBLC,direction="long",
                   idvar="fish",          # what identifies unique fish
                   varying=varying1,      # declare the repeated measurements
                   v.names="anu",         # name for repeat meas in long format
                   timevar="age",         # name of var that identifies the repeat
                   times=1:12)            # values in timevar for repeat
```

```
>
> view(ldBLC)
      species   lake fish agecap lencap radcap age   anu
149.2     BLC Talcot  149      1    128  2.369   2 2.369
38.4      BLC Talcot   38      1    168  3.363   4    NA
81.6      BLC Talcot   81      2    193  3.668   6    NA
84.8      BLC Talcot   84      1    127  2.469   8    NA
48.10     BLC Talcot   48      1    144  2.855  10    NA
49.12     BLC Talcot   49      1    137  2.756  12    NA
>
> ldBLC[ldBLC$fish==165,]                # example for one fish
       species   lake fish agecap lencap radcap age   anu
165.1      BLC Talcot  165      5    276  5.159   1 1.757
165.2      BLC Talcot  165      5    276  5.159   2 3.265
165.3      BLC Talcot  165      5    276  5.159   3 4.308
165.4      BLC Talcot  165      5    276  5.159   4 4.728
165.5      BLC Talcot  165      5    276  5.159   5 5.045
165.6      BLC Talcot  165      5    276  5.159   6 5.159
165.7      BLC Talcot  165      5    276  5.159   7    NA
165.8      BLC Talcot  165      5    276  5.159   8    NA
165.9      BLC Talcot  165      5    276  5.159   9    NA
165.10     BLC Talcot  165      5    276  5.159  10    NA
165.11     BLC Talcot  165      5    276  5.159  11    NA
165.12     BLC Talcot  165      5    276  5.159  12    NA
>
> # remove all of the NAs
> ldBLC <- Subset(ldBLC,!is.na(anu))
> ldBLC[ldBLC$fish==165,]                # same example for one fish
    species   lake fish agecap lencap radcap age   anu
67      BLC Talcot  165      5    276  5.159   1 1.757
135     BLC Talcot  165      5    276  5.159   2 3.265
172     BLC Talcot  165      5    276  5.159   3 4.308
193     BLC Talcot  165      5    276  5.159   4 4.728
208     BLC Talcot  165      5    276  5.159   5 5.045
216     BLC Talcot  165      5    276  5.159   6 5.159
>
> # remove the "plus" growth
> ldBLC <- Subset(ldBLC,agecap-age>=0)
> ldBLC[ldBLC$fish==165,]                # same example for one fish
    species   lake fish agecap lencap radcap age   anu
67      BLC Talcot  165      5    276  5.159   1 1.757
104     BLC Talcot  165      5    276  5.159   2 3.265
125     BLC Talcot  165      5    276  5.159   3 4.308
140     BLC Talcot  165      5    276  5.159   4 4.728
148     BLC Talcot  165      5    276  5.159   5 5.045
```

Adding Variables II

```
> k <- 35  # use Carlander intercept of k=35 mm
> ldBLC <- within(ldBLC, {
   bcFL <- (anu/radcap)*(lencap-k)+k
 })
> view(ldBLC)
```

```
    species   lake fish agecap lencap radcap age   anu   bcFL
42      BLC Talcot   61      2    213  4.259   1 0.842  70.19
45      BLC Talcot   77      2    218  4.384   1 0.980  75.91
88      BLC Talcot  143      3    241  4.736   2 3.186 173.58
101     BLC Talcot   80      5    290  5.023   2 2.710 172.58
121     BLC Talcot   30      5    279  5.023   3 3.787 218.96
140     BLC Talcot  165      5    276  5.159   4 4.728 255.87
```

```
> Summarize(bcFL~age,data=ldBLC)
```

```
Warning:  To continue, variable(s) on RHS of formula were converted to a factor.
  age  n   mean     sd   min    Q1 median    Q3 max percZero
1   1 68  89.32 14.167  63.8  78.4   88.7  98.9 117        0
2   2 37 173.03 16.442 134.0 162.0  171.0 184.0 217        0
3   3 21 226.02 11.139 205.0 221.0  227.0 231.0 249        0
4   4 15 255.28  8.737 247.0 250.0  252.0 261.0 273        0
5   5  8 277.12  8.414 266.0 271.0  276.0 285.0 288        0
```