R Workshop 2

I. Workshop 1 Review

A. options (max.print = 200)

B. assignments: <-

C. Vectors ~~arithmetic~~ (homogeneous)
1. constant: c(...)
2. recycling: multiples
3. scan
4. 1-based indexing

D. Lists (heterogeneous)
[ ] [[ ]] $
m1 <- list(a=1:3, b='david', c=pi)

E. Data Sets (explore with data())
1. USArrests (?USArrests)
2. ChickWeight

F. data.frame
1. dim 3. length 5. head
2. names 4. str 6. tail

G. Random Variables
1. outcomes $\xrightarrow{X}$ $\mathbb{R}$
assign outcomes to numbers
2. cdf - cumulative distribution function
$$F_X(x) = P[X \le x]$$
3. pdf - probability density function (discrete)
$$f_X(x) = P[X = x]$$
4. binomial $f_X(i; n, p) = \binom{n}{i} p^i (1-p)^{n-i}$
5. discrete uniform
- $\Omega_1 = \{1, \dots, n\}$
- $f_X(i) = \begin{cases} \frac{1}{n}, & i \in \{1, \dots, n\} \\ 0, & \text{otherwise} \end{cases}$
- $F_X(i) = \frac{\text{Li}}{n}$
- sample ( )

II Random Variables

1. For discrete RV with pdf $f_X(i)$

a. Expectation (mean): $E[X] = \sum_{i=-\infty}^{\infty} i f_X(i)$ μ_X

b. Variance: $\text{Var}[X] = \sum_{i=-\infty}^{\infty} (i - E[X])^2 f_X(i)$ σ_X^2 (std dev)

c. n^{th} Moment $M_n = \sum_{i=-\infty}^{\infty} i^n f_X(i) = E[X^n]$

2. Examples

a. Uniform (discrete)
$$E[X] = \sum_{i=1}^{n} i \frac{1}{n} = \frac{1}{n} \sum_{i=1}^{n} i = \frac{1}{n} \cdot \frac{n(n+1)}{2} = \frac{n+1}{2}$$
$$\text{Var}[X] = \sum_{i=1}^{n} \left(i - \frac{n+1}{2}\right)^2 \frac{1}{n} = \frac{n^2-1}{12}$$
Verify next time

b. Binomial Distribution
$$E[X] = \sum_{i=0}^{n} i \binom{n}{i} p^i (1-p)^{n-i} = np$$
$$\text{Var}[X] = \sum_{i=0}^{n} (i - np)^2 \binom{n}{i} p^i (1-p)^{n-i} = npq$$
where $q = 1-p$

3. Tricks

a. $\text{Var}[X] = E[X^2] - E[X]^2$

Check: $\text{Var}[X] = \sum_{i=-\infty}^{\infty} (i-\mu)^2 f_X(i)$
$$= \sum_{i=-\infty}^{\infty} (i^2 - 2i\mu + \mu^2) f_X(i)$$
$$= \sum_{i=-\infty}^{\infty} i^2 f_X(i) - 2\mu \underbrace{\sum_{i=-\infty}^{\infty} i f_X(i)}_{E[X]} + \mu^2 \sum_{i=-\infty}^{\infty} f_X(i)$$
$$= E[X^2] - 2\mu\mu + \mu^2 \cdot 1$$
Verify next time

b. Moment Generating Function
$$m_X(t) = E[e^{it}] = \sum_{i=-\infty}^{\infty} e^{it} f_X(i)$$
Use analytic function theory to show:
$E[X] = m'(0)$ ← 1st derivative at $t=0$.
$E[X^2] = m''(0)$ ← 2nd derivative at $t=0$.
Verify for Binomial RV.

R Workshop 2 — 5/24/2017

III. R Factors

A. myF = as.factor(c('one','one','two','two','one'))

levels() - shows the levels for a factor

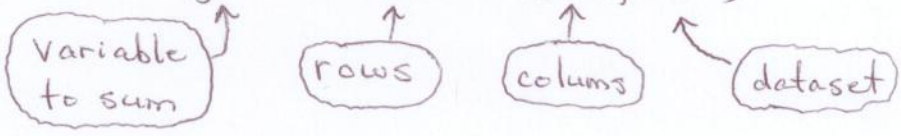
levels(myF); levels(myF) <- c('ONE','TWO')

Note: the assignment to a function call. This is unusual in most programming.

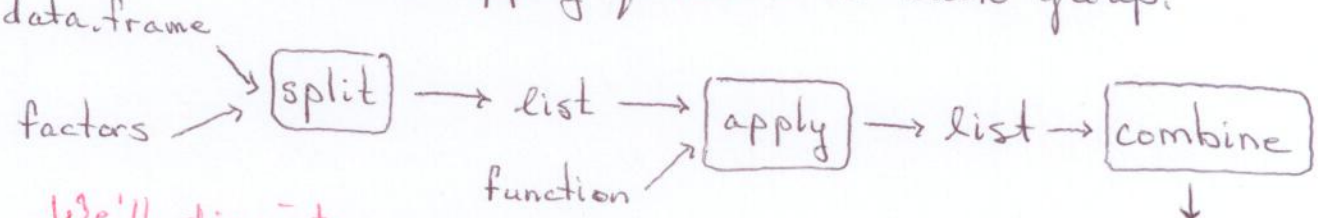
cw <- ChickWeight

B. table, xtabs

1. table - good to quickly tabulate counts of values
   a. table(cw$Diet) — one dimension count
   b. table(cw$Time, cw$Diet)
      i. makes 2d table
      ii. first argument is axis 0 (along rows), 2nd along columns
2. xtabs - a little more flexible than table.
   a. xtabs(~Diet, cw)
   b. xtabs(~Time + Diet, cw)
   c. xtabs(weight ~ Time + Diet, cw)



C. aggregate - aggregate based on like value of a factor; then apply function to each group.



We'll dig into this more formally later.

aggregate(weight ~ Diet, cw, mean)

ag <- aggregate(weight ~ Time + Diet, cw, mean)

This gives a single column for all diets (1→4). We can pivot these value so that each has its own column.

library(reshape2)

dcast(ag, Time ~ Diet, value.var='weight')

More later

this was skipped

R Workshop 2

IV File I/O

Using manual data entry, random number generators, and prepackaged data sets is fine for a while, but eventually you need to read data from a file or a database.

A. Directories

1. getwd() - prints the current working directory
2. setwd('~') - sets the current dir
3. list.files('.')

B. CSV Files - input read.csv

1. file name
2. header = TRUE
3. sep = ","
4. stringAsFactors = True ← usually want FALSE
5. col.names = c(~~~)

C. Common Conversions

df$col = as.factor(df$col)

df$ts = strptime(df$ts, format='%Y-%m-%d %H:%M:%S')

D. CSV Files - output write.csv

1. x
2. file
3. append

skipped

V RStudio

VI R Graphics

Graphing Data with R
John Hilfiger
Chapter 3+4

A. Three Plotting Systems
1. Base
2. Lattice
3. ggplot

We will work with base through the summer and switch to ggplot in Fall.

B. Two classes of plots
1. exploratory plots for your own insights
2. polished plots for presentation

C. Two basic types of single variable plots
1. stripchart
2. dotchart

skipped

D. Plot options - there are many plotting options available. We'll introduce a handful with each workshop.

1. pch = 19 [point character]
2. xlab = "~~" [Labels]
   ylab = "~~"
3. xlim = c(·,·) [Axis Limits]
   ylim = c(·,·)