Epidemiology Using R Chapter 3 Problem Set

Chelsea McMullen

5 February 2015

3.1) Using RStudio and the data from Table 3.1 on page 107 Create the following data frame:

Status <- c("Dead", "Survived", "Dead", "Survived", "Dead", "Survived",   
 "Dead", "Survived")   
Treatment <- c("Tolbutamide", "Tolbutamide", "Placebo", "Placebo", "Tolbutamide", "Tolbutamide", "Placebo", "Placebo")   
Agegrp <- c("<55", "<55", "<55", "<55", "55+", "55+", "55+", "55+")  
Freq <- c(8, 98, 5, 115, 22, 76, 16, 69)  
dat <- data.frame(Status, Treatment, Agegrp, Freq)  
dat

## Status Treatment Agegrp Freq  
## 1 Dead Tolbutamide <55 8  
## 2 Survived Tolbutamide <55 98  
## 3 Dead Placebo <55 5  
## 4 Survived Placebo <55 115  
## 5 Dead Tolbutamide 55+ 22  
## 6 Survived Tolbutamide 55+ 76  
## 7 Dead Placebo 55+ 16  
## 8 Survived Placebo 55+ 69

3.2) Select 3 to 5 classmates and collect data on first name, last name, affiliation, two email addresses, and today’s date. Using a text editor, create a data frame with this data.

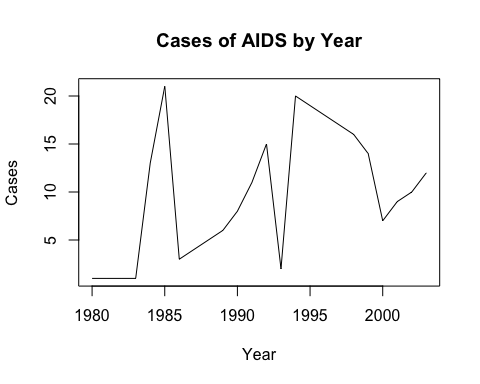
First.name <- c("John", "Lucy", "Sandra")  
Last.name <- c("Smith", "Bone", "Key")  
Affiliation <- c("Duke", "UNC", "Duke")  
Email <- c("NA", "lb123@unc.edu", "sk345@duke.edu")  
Date <- Sys.Date()  
dat2 <- data.frame(First.name, Last.name, Affiliation, Email, Date)  
dat2

## First.name Last.name Affiliation Email Date  
## 1 John Smith Duke NA 2015-02-05  
## 2 Lucy Bone UNC lb123@unc.edu 2015-02-05  
## 3 Sandra Key Duke sk345@duke.edu 2015-02-05

3.3) Review the United States data on AIDS cases by year available at http: //www.medepi.net/data/aids.txt. Read this data into a data frame. Graph a calendar time series of AIDS cases.

aids <- read.table("http://www.medepi.net/data/aids.txt", header = TRUE)

plot(aids$year, aids$cases, type = "l", xlab = "Year", ylab = "Cases",   
 main = "Cases of AIDS by Year")



The number of cases ploted is much lower than the numbers given in the raw data. Therefore, I checked the structure and summary of the data frame:

str(aids)

## 'data.frame': 24 obs. of 2 variables:  
## $ cases: Factor w/ 21 levels ".","103691","12932",..: 1 1 1 1 13 21 3 4 5 6 ...  
## $ year : int 1980 1981 1982 1983 1984 1985 1986 1987 1988 1989 ...

Results show that the variable aids$cases is classified as a factor. If we convert the local data frame to a different type of data frame using the tbl\_df() function from dplyr, the cases are displayed as characters, which are more easily plotted. We should also account for the missing variables in years 1980-1984.

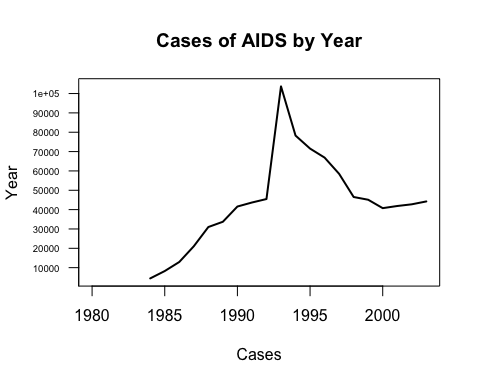
library(dplyr)

aids.cases <- as.character(aids$cases)  
aids.cases[aids.cases == "."] <- NA  
str(aids.cases)

## chr [1:24] NA NA NA NA "4445" "8249" "12932" "21070" ...

Now we have a more accurate plot.

plot(aids$year, aids.cases, type = "l", xlab = "Cases", lwd =2,   
 ylab = "Year", main = "Cases of AIDS by Year", yaxt = "n")  
axis(2, at = seq(0, 101000, 10000), labels = seq(0, 101000, 10000),  
 tick = TRUE, cex.axis = 0.6, las = 2)

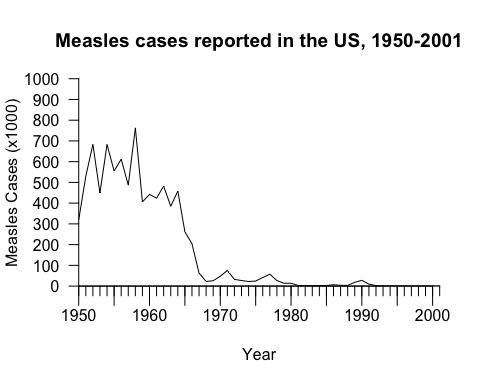


3.4) Review the United States data on measles cases by year available at <http://www.medepi.net/data/measles.txt>. Read this data into a data frame. Graph a calendar time series of measle cases using an arithmetic and semi-logarithmic scale.

measles <- read.table("http://www.medepi.net/data/measles.txt", header = TRUE)

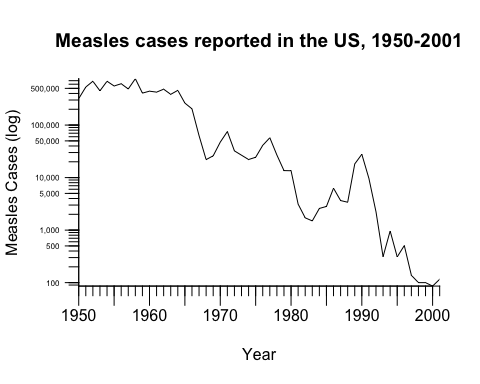
**Arithmetic Plot**

plot(measles$year, measles$cases/1000, type = "l",   
 xlab = "Year", ylab = "Measles Cases (x1000)",   
 main = "Measles cases reported in the US, 1950-2001",  
 axes = FALSE, xlim = c(1950, 2001), ylim = c(0,1000),  
 xaxs = "i", yaxs = "i")  
axis(1, at = measles$year, labels = FALSE, tick = TRUE)  
axis(1, at = seq(1950, 2000, 5), labels = seq(1950, 2000, 5),  
 tick = TRUE, tcl = -1)  
axis(2, at = seq(0, 1000, 100), labels = seq(0, 1000, 100),  
 las = 2, tick = TRUE)



**Semi-logarithmic Plot**

plot(measles$year, measles$cases, type = "l", log = "y",  
 xlab = "Year", ylab = "Measles Cases (log)",   
 main = "Measles cases reported in the US, 1950-2001",  
 axes = FALSE, xlim = c(1950, 2001),  
 xaxs = "i", yaxs = "i")  
axis(1, at = measles$year, labels = FALSE, tick = TRUE)  
axis(1, at = seq(1950, 2000, 5), labels = seq(1950, 2000, 5),  
 tick = TRUE, tcl = -1, cex.axis = 1)  
axis(2, at = c(seq(1, 10, 1), seq(10, 100, 10),  
 seq(100, 1000, 100), seq(1000, 10000, 1000),  
 seq(10000, 100000, 10000), seq(100000, 1000000, 100000)),  
 labels = FALSE, tick = TRUE, cex.axis = 1)  
axis(2, at = c(1, 5, 10, 50, 100, 500, 1000, 5000, 10000,   
 50000, 100000, 500000, 1000000),  
 labels = c("1", "5","10", "50", "100", "500", "1,000",   
 "5,000", "10,000", "50,000", "100,000", "500,000",  
 "1,000,000"),   
 las = 2, tick = TRUE, tcl = -0.7, cex.axis = 0.5)



3.5) Review the United States data on hepatitis B cases by year available at <http://www.medepi.net/data/hepb.txt>. Read this data into a data frame. Using the R code below, plot a times series of AIDS and hepatitis B cases.

library(dplyr)  
library(splitstackshape)

hepb <- read.csv("http://www.medepi.net/data/hepb.txt", header = TRUE)  
hepb <- tbl\_df(hepb)  
str(hepb)

## Classes 'tbl\_df', 'tbl' and 'data.frame': 24 obs. of 1 variable:  
## $ cases.year: Factor w/ 24 levels "10258 1998","10416 1997",..: 9 11 12 15 18 19 17 16 13 14 ...

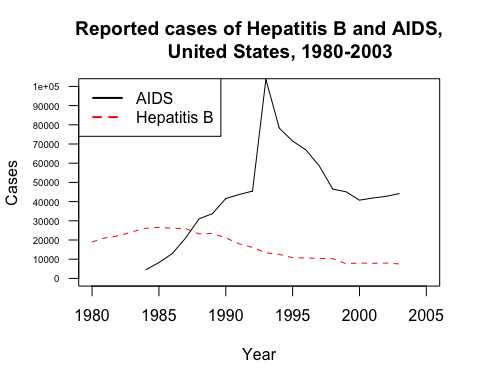
We can see from the structure of hepb that it onle contains one variable column, cases.year. This is not useful to us, as we want just one variable per column. We can use the splitstackshape package to split the two columns.

hepb.split <- cSplit\_f(hepb, "cases.year", sep = " ")  
names(hepb.split) <- c("cases", "year")  
str(hepb.split)

## Classes 'tbl\_df', 'tbl', 'data.table' and 'data.frame': 24 obs. of 2 variables:  
## $ cases: int 19015 21152 22177 24318 26115 26611 26107 25916 23177 23419 ...  
## $ year : int 1980 1981 1982 1983 1984 1985 1986 1987 1988 1989 ...  
## - attr(\*, ".internal.selfref")=<externalptr>

hepb.split now has one column for the cases variable, and one column for year.

matplot(c(1980, 2005), c(0,100000), type = "n", xlab = "Year",  
 ylab = "Cases",  
 main = "Reported cases of Hepatitis B and AIDS,  
 United States, 1980-2003", yaxt = "n")  
matlines(hepb.split, hepb.split$cases, type = "l")  
matlines(aids, aids$cases, type = "l", lty = 1, col = "black")  
legend(x = "topleft", legend = c("AIDS", "Hepatitis B"),   
 lwd = 2, lty = 1:2, col = 1:2)  
axis(2, at = seq(0, 100000, 10000), labels = seq(0, 101000, 10000),   
 tick = TRUE, las = 2, cex.axis = 0.6)



3.6) # Review data from the Evans cohort study in which 609 white males were followed for 7 years, with coronary heart disease as the outcome of interest (<http://www.medepi.net/data/evans.txt>). The data dictionary is provided in Table 3.10.

chd <- read.table("http://www.medepi.net/data/evans.txt", header = TRUE)

1. Recode the binary variables (0, 1) into factors with 2 levels.

chd$chd <- ifelse(chd$chd == 0, "no", "yes")  
chd$cat <- ifelse(chd$cat == 0, "normal", "high")  
chd$smk <- ifelse(chd$smk == 0, "never", "ever")  
chd$ecg <- ifelse(chd$ecg == 0, "no abnormality", "abnormality")  
chd$hpt <- ifelse(chd$hpt == 0, "no", "yes")  
chd <- tbl\_df(chd)  
chd

## Source: local data frame [609 x 12]  
##   
## id chd cat age chl smk ecg dbp sbp hpt ch cc  
## 1 21 no normal 56 270 never no abnormality 80 138 no 0 0  
## 2 31 no normal 43 159 ever no abnormality 74 128 no 0 0  
## 3 51 yes high 56 201 ever abnormality 112 164 yes 1 201  
## 4 71 no high 64 179 ever no abnormality 100 200 yes 1 179  
## 5 74 no normal 49 243 ever no abnormality 82 145 no 0 0  
## 6 91 no normal 46 252 ever no abnormality 88 142 no 0 0  
## 7 111 yes normal 52 179 ever abnormality 80 128 no 0 0  
## 8 131 no normal 63 217 never no abnormality 92 135 no 0 0  
## 9 141 no normal 42 176 ever no abnormality 76 114 no 0 0  
## 10 191 no normal 55 250 never abnormality 114 182 yes 0 0  
## .. ... ... ... ... ... ... ... ... ... ... .. ...

1. Discretize age into a factor with more than 2 levels.

agelab <- c("30-40", "40-50", "50-60", "60-70", ">70")  
agegrp <- cut(chd$age, breaks = c(30, 40, 50, 60, 70, 80),   
 right = FALSE, labels = agelab)  
table(agegrp)

## agegrp  
## 30-40 40-50 50-60 60-70 >70   
## 0 247 203 115 44

1. Create a new hyptertension categorical variable based on the current classification scheme17:

* Normal: SBP< 120 and DBP< 80;
* Prehypertension: SBP=[120, 140) or DBP=[80, 90);
* Hypertension-Stage 1: SBP=[140, 160) or DBP=[90, 100);
* Hypertension-Stage 2: SBP≥ 160 or DBP≥ 100.

hyplab <- c("Normal", "Pre-HT", "HT-Stage 1", "HT-Stage 2")  
hyp <- c(chd$dbp & chd$sbp)  
hyp[chd$dbp < 80 & chd$sbp < 120] <- "Normal"  
hyp[(chd$dbp >= 80 & chd$dbp <= 90) | (chd$sbp >= 120 & chd$sbp <= 140)] <- "Pre-HT"  
hyp[(chd$dbp >= 90 & chd$dbp <= 100) | (chd$sbp >= 140 & chd$sbp <= 160)] <- "HT-Stage 1"  
hyp[chd$dbp >= 100 | chd$sbp >= 160] <- "HT-Stage 2"  
table(hyp)

## hyp  
## HT-Stage 1 HT-Stage 2 Normal Pre-HT   
## 177 211 56 165

1. Using R, construct a contigency table comparing the old and new hyper- tension variables.

hyp <- factor(hyp, levels = hyplab, ordered = T)  
hpt <- ifelse(chd$hpt == "no", "Not Hypertensive", "Hypertensive")  
hpt.vs.hyp <- table(hpt, hyp)  
hpt.vs.hyp

## hyp  
## hpt Normal Pre-HT HT-Stage 1 HT-Stage 2  
## Hypertensive 0 0 44 211  
## Not Hypertensive 56 165 133 0

3.7) Review the California 2004 surveillance data on human West Nile virus cases available at <http://www.medepi.net/data/wnv/wnv2004raw>. txt. Read in the data, taking into account missing values. Convert the calendar dates into the international standard format. Using the write.table function export the data as an ASCII text file.

wnv <- read.csv("http://www.medepi.net/data/wnv/wnv2004raw.txt",   
 header = TRUE)  
wnv[wnv == "." | wnv == "Unknown"] <- NA  
wnv$date.onset <- as.Date(wnv$date.onset, "%m/%d/%Y")  
wnv$date.tested <- as.Date(wnv$date.tested, "%m/%d/%Y")  
str(wnv)

## 'data.frame': 779 obs. of 8 variables:  
## $ id : int 1 2 3 4 5 6 7 8 9 10 ...  
## $ county : Factor w/ 23 levels "Butte","Fresno",..: 14 14 14 14 14 14 14 14 8 12 ...  
## $ age : Factor w/ 92 levels ".","1","10","11",..: 35 61 12 5 5 10 58 72 69 20 ...  
## $ sex : Factor w/ 3 levels ".","F","M": 2 2 3 3 3 3 3 2 3 3 ...  
## $ syndrome : Factor w/ 3 levels "Unknown","WNF",..: 2 2 2 2 2 2 3 3 2 3 ...  
## $ date.onset : Date, format: "2004-05-19" "2004-05-22" ...  
## $ date.tested: Date, format: "2004-06-02" "2004-06-16" ...  
## $ death : Factor w/ 3 levels ".","No","Yes": 2 2 2 2 2 2 2 2 2 2 ...

Create a new csv file in the working directory.

write.csv(wnv, "wnv.csv")

3.8)

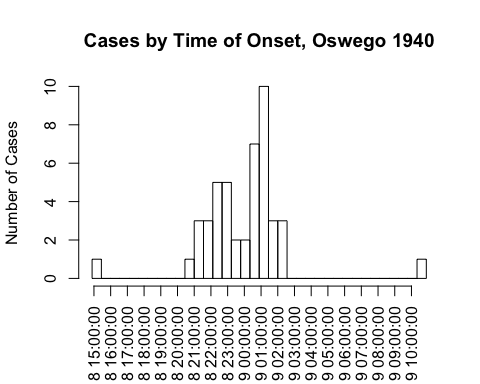
oswego <- read.table("http://www.medepi.net/data/oswego/oswego.txt", header = TRUE)

1. Using RStudio plot the cases by time of onset of illness (include appropriate labels and title). What does this graph tell you? (Hint: Process the text data and then use the hist function.)

onset.date <- paste(paste(oswego$onset.date, "/1940", sep = "",  
 oswego$onset.time), sep = " ")  
onset.dt <- strptime(onset.date, "%m/%d/%Y%I:%M %p")  
onset <- onset.dt[!is.na(onset.dt)]  
onset

## [1] "1940-04-19 00:30:00 WAT" "1940-04-19 00:30:00 WAT"  
## [3] "1940-04-19 00:30:00 WAT" "1940-04-18 22:30:00 WAT"  
## [5] "1940-04-18 22:30:00 WAT" "1940-04-19 02:00:00 WAT"  
## [7] "1940-04-19 01:00:00 WAT" "1940-04-18 23:00:00 WAT"  
## [9] "1940-04-19 02:00:00 WAT" "1940-04-19 10:30:00 WAT"  
## [11] "1940-04-19 00:30:00 WAT" "1940-04-18 22:15:00 WAT"  
## [13] "1940-04-18 22:00:00 WAT" "1940-04-19 01:00:00 WAT"  
## [15] "1940-04-18 23:00:00 WAT" "1940-04-18 21:45:00 WAT"  
## [17] "1940-04-18 21:45:00 WAT" "1940-04-19 01:00:00 WAT"  
## [19] "1940-04-18 23:00:00 WAT" "1940-04-18 21:00:00 WAT"  
## [21] "1940-04-19 01:00:00 WAT" "1940-04-19 01:00:00 WAT"  
## [23] "1940-04-18 21:15:00 WAT" "1940-04-18 23:30:00 WAT"  
## [25] "1940-04-19 01:00:00 WAT" "1940-04-18 21:30:00 WAT"  
## [27] "1940-04-19 02:30:00 WAT" "1940-04-19 02:00:00 WAT"  
## [29] "1940-04-18 21:30:00 WAT" "1940-04-19 00:30:00 WAT"  
## [31] "1940-04-19 01:00:00 WAT" "1940-04-18 22:30:00 WAT"  
## [33] "1940-04-18 15:00:00 WAT" "1940-04-19 00:00:00 WAT"  
## [35] "1940-04-18 23:00:00 WAT" "1940-04-18 22:30:00 WAT"  
## [37] "1940-04-19 01:00:00 WAT" "1940-04-19 02:30:00 WAT"  
## [39] "1940-04-18 23:30:00 WAT" "1940-04-19 01:00:00 WAT"  
## [41] "1940-04-19 00:30:00 WAT" "1940-04-19 00:30:00 WAT"  
## [43] "1940-04-19 01:00:00 WAT" "1940-04-19 00:00:00 WAT"  
## [45] "1940-04-19 02:15:00 WAT" "1940-04-18 23:00:00 WAT"

hist(onset, breaks = 35, xlab = " ",   
 ylab = "Number of Cases",   
 main = "Cases by Time of Onset, Oswego 1940",   
 freq = TRUE, xaxt = "n")  
axis.POSIXct(1, at = seq(min(onset), max(onset), by = "hour"),   
 labels = seq(min(onset), max(onset), by = "hour"),   
 cex.lab = 0.3, las = 2)



1. Are there any cases for which the times of onset are inconsistent with the general experience? How might they be explained?

There is one early onset case, and one late onset case. The early onset could have been someone who prepared the infected food source and consumed it before the rest, or an anomoly in incubation period. The late onset outlier could have been someone who ate the source of infection later than they rest, someone with a longer than normal incubation period, or someone who acquired a secondary infection from a sick family member or friend.

1. How could the data be sorted by illness status and illness onset times?

oswego.df <- tbl\_df(oswego)  
ill.onset <- select(oswego.df, onset.date, onset.time)  
ill.onset <- arrange(oswego.df, desc(ill), onset.date, onset.time)  
ill.onset

## Source: local data frame [75 x 21]  
##   
## id age sex meal.time ill onset.date onset.time baked.ham spinach  
## 1 20 33 F NA Y 4/18 10:00 PM Y Y  
## 2 18 36 M NA Y 4/18 10:15 PM Y Y  
## 3 6 63 F 7:30 PM Y 4/18 10:30 PM Y Y  
## 4 7 70 M 7:30 PM Y 4/18 10:30 PM Y Y  
## 5 49 52 F NA Y 4/18 10:30 PM Y Y  
## 6 57 74 M NA Y 4/18 10:30 PM Y Y  
## 7 10 33 F 7:00 PM Y 4/18 11:00 PM Y Y  
## 8 22 7 M NA Y 4/18 11:00 PM Y Y  
## 9 29 37 F NA Y 4/18 11:00 PM Y Y  
## 10 55 25 M NA Y 4/18 11:00 PM Y N  
## .. .. ... ... ... ... ... ... ... ...  
## Variables not shown: mashed.potato (fctr), cabbage.salad (fctr), jello  
## (fctr), rolls (fctr), brown.bread (fctr), milk (fctr), coffee (fctr),  
## water (fctr), cakes (fctr), vanilla.ice.cream (fctr),  
## chocolate.ice.cream (fctr), fruit.salad (fctr)

1. Where possible, calculate incubation periods and illustrate their distribution with an appropriate graph. Use the truehist function in the MASS package. Determine the mean, median, and range of the incubation period.

library(MASS)

meal.dt <- paste("4/18/1940", oswego$meal.time, sep = " ")  
meal.dt <- strptime(meal.dt, "%m/%d/%Y%I:%M %p")

inc.period <- onset.dt - meal.dt  
inc.period[!is.na(inc.period)]

## Time differences in hours  
## [1] 4.5 6.0 6.0 3.0 3.0 6.5 3.0 4.0 6.5 3.0 3.0 3.0 3.0 3.0 6.0 4.0 3.0  
## [18] 7.0 4.0 3.0 5.5 4.5

mean.inc.period <- mean(inc.period, na.rm = TRUE)  
median.inc.period <- median(as.numeric(inc.period[!is.na(inc.period)]))  
range.inc.period <- range(as.numeric(inc.period[!is.na(inc.period)]))

mean.inc.period  
median.inc.period  
range.inc.period

* Mean Incubation Period: 4.2954545
* Median Incubation Period: 4
* Range of Incubation Periods: 3, 7

truehist(as.numeric(inc.period), nbins = 10, xlab = "Incubation Period (hours)",  
 ylab = "Proportion of Cases", main = "Range of Incubation Periods, Oswego 1940",  
 col = "skyblue", xlim = c(3, 8), ylim = c(0, 1), labels = TRUE)

