Missing Data

In R, missing values are represented by the symbol **NA** (not available). Impossible values (e.g., dividing by zero) are represented by the symbol **NaN** (not a number). Unlike SAS, R uses the same symbol for character and numeric data.

For more practice on working with missing data, try this course on cleaning data in R.

Testing for Missing Values

is.na(x) # returns TRUE of x is missing  
y <- c(1,2,3,NA)  
is.na(y) # returns a vector (F F F T)

Recoding Values to Missing

# recode 99 to missing for variable v1  
# select rows where v1 is 99 and recode column v1   
mydata$v1[mydata$v1==99] <- NA

Excluding Missing Values from Analyses

Arithmetic functions on missing values yield missing values.

x <- c(1,2,NA,3)  
mean(x) # returns NA  
mean(x, na.rm=TRUE) # returns 2

max(x, na.rm=TRUE) # returns 2

The function **complete.cases()** returns a logical vector indicating which cases are complete.

# list rows of data that have missing values   
mydata[!complete.cases(mydata),]

The function **na.omit()** returns the object with listwise deletion of missing values.

# create new dataset without missing data   
newdata <- na.omit(mydata)