# Assignment 1

***Callin Switzer***

***September 8, 2014***

```
## Loading required package: ggplot2
## Loading required package: data.table
```

normal

**bold**

*italic*

1. (Not Graded) Complete the following Conceptual Exercises in Chapter 1 (2nd or 3rd edition of R&S): 2, 7, 9. Feel free to include your answers in the write-up. However, this question will not be graded.
2. (20 points) For each of the following surveys, specify study units, the target population, and the sampling frame. Discuss in 2-3 sentences any possible sources of selection bias, specifically, undercoverage or overcoverage. Finally, specify what type of sampling was used.
   a. (10 points) To estimate how many books in the library need rebinding, a librarian uses a random number generator to select 100 locations on library shelves. He then walks to each location, looks at the book that resides at that spot, and records whether the book needs rebinding or not.

      * Study units = books in the library
      * Target population = all the books in the library
      * Sampling frame = all the books in the library
      * Selection Bias: Undercoverage could occur if books were checked out or misplaced , and the librarian was unable to find them on the shelves. Overcoverage could occur if the librarian found books that were not supposed to be in the libary – for example, someone could have returned a book to the library that didn't belong to the library.
      * Sampling = simple random sample

   b. (10 points) The Arizona Intrastate Travel Committee commissioned a study to identify in-state travel patterns of residents of major metropolitan cities and to evaluate different sources of vacation planning information. The plan was to conduct phone interviews with Phoenix and Tucson residents. Landline telephone numbers with Phoenix and Tucson area codes were generated randomly so that listed and unlisted telephone numbers could be reached. (Arizona Office of Tourism, 1991.)

      * Study units = residents of major Phoenix and Tucson with phone numbers in those area codes
      * Target population = all the road-users in the major metropolitan cities
      * Sampling frame = all the people with a landline in Phoenix and Tucson
      * Selection bias: Undercoverage could occur if people lived in Phoenix or Tucson and didn't have a landline phone, or didn't answer their phone. Overcoverage could occur if someone who wasn't a resident of a major metropolitan area answered the phone – for example, someone could have holiday home in a major city, and not live there for most of the year.
      * Sampling = simple random sample

3. (5 points) Consider two securities, the first having expected return $\mu_1 = 1$ and standard deviation $\sigma_1 = 0.1$, and the second having $\mu_2 = 0.8$ and $\sigma = 0.12$. Also, let the correlation between securities be $\rho = 0.1$. Suppose you invest $\pi = 0.8$, or 80%, of your money in the first security and $1-\pi$ in the second security. What is your expected return and what is the standard deviation of your return?

$E(X) = \pi(X1) + (1-\pi)(X2)$
$= 0.8(1) + 0.2(0.8)$
$= 0.96$ is the expected return

$Var(X) = sqrt(\pi) * Var(X1) + sqrt(1-\pi) * Var(X2) + 2 * \pi * (1-\pi)\ Cov(X1,X2)$
$= sqrt(0.8) * 0.1^2 + sqrt(0.2) * 0.12^2 + 2 * 0.8 * 0.2 * 0.1 * sqrt(0.1 + 0.12)$
$= 0.0304$ is the standard deviation of the return

4. (10 points) Solve the following equations for the matrix X. You can assume that the matrices A and B (when needed) are all invertible n x n matrices.

a. $AXB = C$

$X = A^{-1} * C * B^{-1}$

b. $(AX) + B = D$

$AX = D - B$

$X = A^{-1} * (D-B)$

c. Solve equation (b) for X by hand

See work by hand attached on other paper

5. (15 points) The csdata.txt data set on the course web-site contains information on 224 computer science students. Use R to perform the following tasks:
   a. Split the students into two groups with GPA $< 3$ and GPA $\geq 3$ and provide the following numerical summaries of the distributions of SAT Math (SATM) scores for each of the two groups: sample mean, sample SD, min, median, max, 1st and 3rd quartiles.

```r
# import data
stud <- read.table("data/CSDATA.txt", header = T)
stud$gpaGroup <- ifelse(stud$GPA > 3, yes = "High", no = "Low")

# split into low and high
high <- stud[stud$gpaGroup == "High", ]
low <- stud[ stud$gpaGroup == "Low",]

# custom summary function to include sd
mysum <- function(vector) {
     s <- numeric(7)
     names(s) <- c("sample mean", "sample SD", "min", "median",
                   "max", "1st quartile","3rd quartile")
     s[c(1,3,4,5,6,7)] <- summary(vector)[c(4, 1, 3, 6, 2, 5)]
     s[2] <- round(sd(vector), digits = 1)
     return(s)
}
```

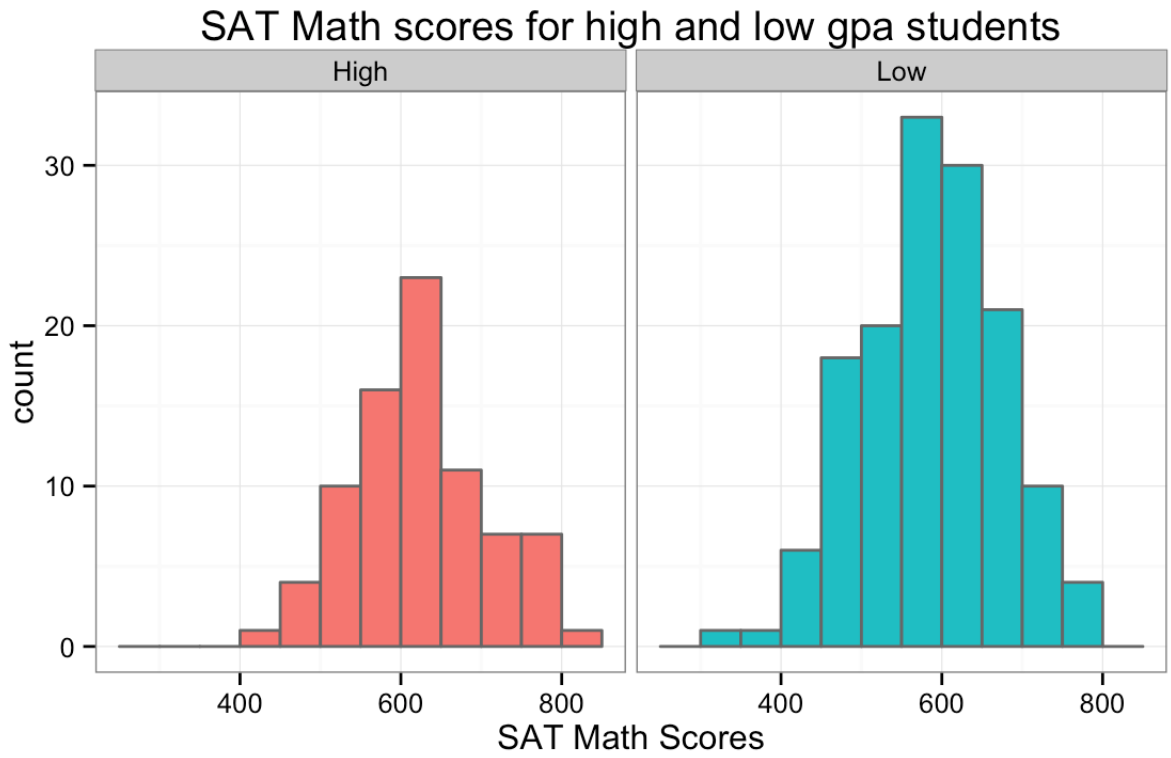
Summary for high-gpa students

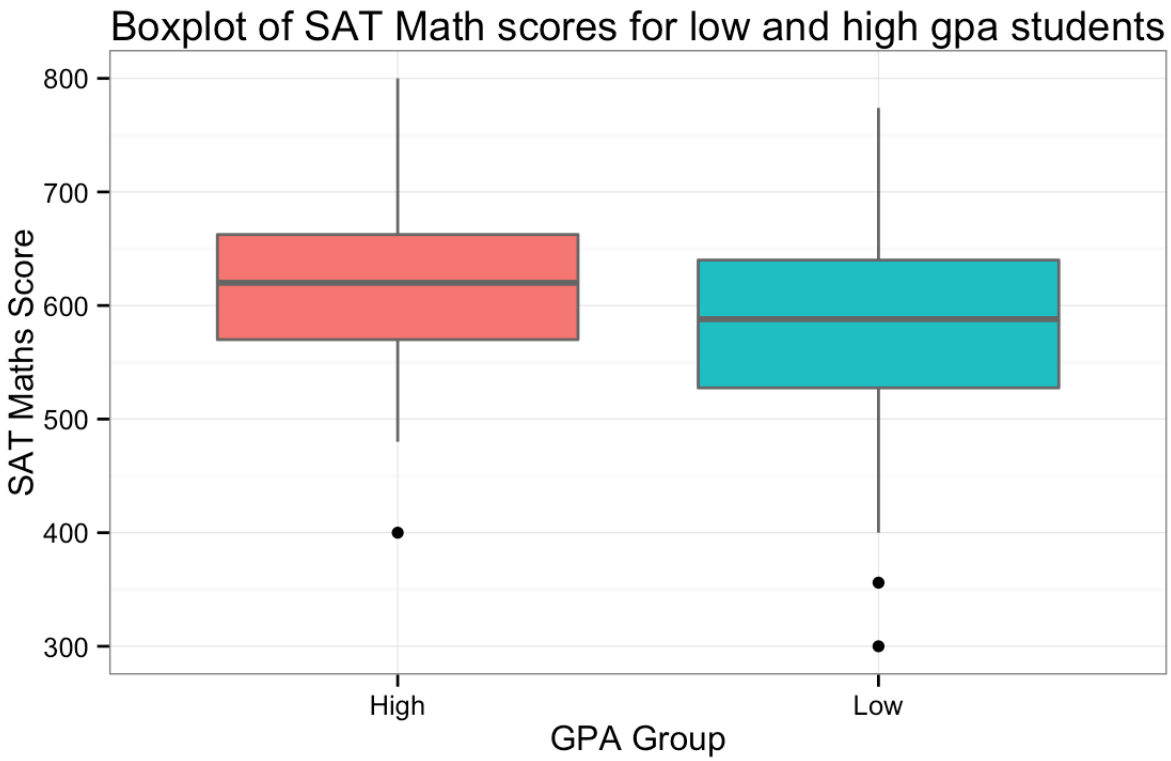
```
##  sample mean    sample SD          min       median          max
##        618.0         80.6        400.0        620.0        800.0
## 1st quartile 3rd quartile
##        570.0        662.0
```

Summary for low-gpa students

```
##  sample mean    sample SD          min       median          max
##        583.0         87.2        300.0        588.0        774.0
## 1st quartile 3rd quartile
##        528.0        640.0
```

b. Plot histograms and box-plots of SATM scores for both groups side-by-side and describe the shapes of their distributions. Are there any visible differences?




Both distributions seem to be fairly normally distributed. They both have similar variances and neither seems especially skewed. There are more poeple in the low-gpa group, as evidenced by the taller histogram. The distribution of SATM scores from the high-gpa group has a slightly higher mean than the low-gpa group.

c. Comment on whether you think the group means are very different or not (without con- ducting any formal tests). The sample group means differ by about 40 points. This is quite a large difference – I suspect that it's highly unlikely to find that large of a difference if the sample was random. Thus, I think the group means are different.

d. Calculate the overall median SATM score in the sample and interpret it.

The overall median SATM score is 600, which is basically right in the middle of two groups. Since the data are distributed fairly normally, the median is quite close to the mean. The median is the score that 50% of the other scores fall below, and 50% fall above.

e. SAT is calibrated such that scores in the entire student population are distributed approx- imately normally with mean 500 and standard deviation 110. Identifying what fraction of student population is expected to score above the sample median found in the previous part? (Hint: use the R command pnorm() to find normal probabilities.) Are your findings consistent with an assertion that CS students have stronger math skills?