Math 227 First Exam Due: October 4 at 2:00 pm

This is a take-home exam. ***Please insert your answers into this document.*** *You may use a calculator, your notes, my notes, and R. You may not discuss any aspect of this exam with anybody but me until after 2:00 pm on October 4. Neatness and clarity will be appreciated. Do show the major steps in your work. Correct answers may not receive full credit without supporting work. Please present your answers in the form of complete sentences. Insert as much R code and output as is necessary to support your answers. Please write and sign the Honor Pledge [I have neither given nor received help on this exam] on your exam.*

Honor Pledge

I have neither given nor received help on this exam.

Name

Ellen Stanton

Please do not write on this page.

1. The data set *StudentSurvey.csv* contains data for most all of the students in a large statistics class at Purdue University in Indiana. All the values are self-reported.

Here are the variables contained in the data set:

Year	Year in school: FirstYear, Sophomore, Junior, or Senior
Gender	Student's gender: F or M
Smoke	Smoker? No or Yes
Award	Award the student would most prefer: Academy, Nobel, Olympic
Highest SAT	Which SAT is higher? Math or Verbal
Exercise	Hours of exercise per week
TV	Hours of TV viewing per week
Height	Height (in inches)
Weight	Weight (in pounds)
Siblings	Number of siblings
BirthOrder	Birth order, 1 = oldest, 2 = second oldest, etc.
VerbalSAT	Verbal SAT score
MathSAT	Math SAT score
SAT	Combined Verbal + Math SAT
Pulse	Pulse rate (beats per minute)
Piercings	Number of body piercing

Before attempting the questions below, you may want to use the code

s <- STUDENTSURVEY

to simplify the name of the data frame you analyze.

```
STUDENTSURVEY <- read.csv("~/Desktop/STUDENTSURVEY.csv")
View(STUDENTSURVEY)
s <- STUDENTSURVEY
View(s)
```

(a) How many students are included in this data set?

```
dim(s)
[1] 362  17
```

There are 362 students included in the data set, since each row represents a student and there are 362 rows.

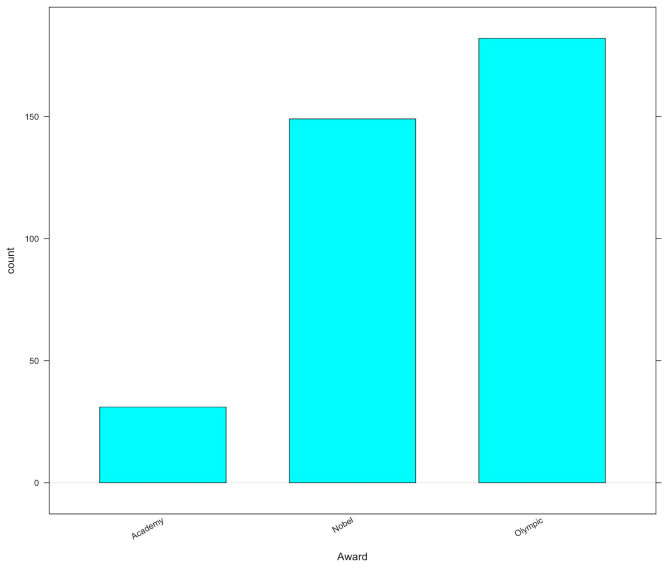
(b) Which of the three awards is the most popular? The least popular?

```
tally(~Award,data=s)
Award
Academy   Nobel Olympic
     31     149     182
```

The Olympic award is the most popular, as it was chosen by 182 students. The Academy award is the least popular, since only 31 students chose it.

(c) Produce a display that illustrates your answers in part (b).

```
bargraph(~Award,data=s)
```



(d) Does it matter in this context that R orders the three awards in alphabetical order? Explain.
No, in this context it doesn't matter that the awards are ordered alphabetically by R. This is because the categories in "Award" are nominal, so they have no inherent order in which they should be displayed.

(e) Does the choice of preferred award vary by gender? Obtain a table which allows you compare the distribution of preferred award by gender. Write a couple of sentences in answer to this question.

```
round(tally(Award~Gender,format="percent",margin=T,data=s),2)
         Gender
Award          F      M
  Academy  11.83   5.70
  Nobel    44.97  37.82
  Olympic  43.20  56.48
  Total   100.00 100.00
```

The choice of preferred award does vary by gender. Over half of male respondents (56.48%) would most prefer an Olympic award, which is thirteen percentage points higher than percent of female respondents (43.20%) who made the same choice. The most popular choice for female respondents was a Nobel award, which 44.97% of them chose, as compared to 37.82% of men. Also, the percentage of women who chose an Academy award (11.83%) is over two times the percentage of men who made the same choice (5.70%).

(f) Use the *favstats* function to obtain summary statistics for the variable MathSAT. What is the mean MathSAT? What is the standard deviation of this variable? Interpret the standard deviation of MathSAT in this case.

```
favstats(~MathSAT,data=s)
 min  Q1 median  Q3 max     mean       sd   n missing
 400 560    610 650 800 609.4365 68.49007 362       0
```

The mean MathSAT score is 609.44 points. The standard deviation of this variable is 68.49 points. This means that, roughly, the individual MathSAT scores differ from their mean (609.44 points) by an average of 68.49 points.

(g) The MathSAT score for the fifth student in the data set is how many standard deviations from the mean?

```
s[5,13]
[1] 450
609.44-450
[1] 159.44
159.44/68.49
[1] 2.327931
```

The Math SAT score for the fifth student in the data set, which is 450 points, differs from the mean of 609.44 points by approximately 2.33 standard deviations.

(h) Use a logical vector to find out the number and proportion of students with MathSAT scores of 600 or more.

```
good <- s$MathSAT>=600
sum(good)
[1] 229
mean(good)
[1] 0.6325967
```

The number of students with MathSAT scores of 600 points or more is 229. The proportion of students with MathSAT scores of 600 points or more is .633.

(i) If we select one of the students at random, what are the odds that the student has a MathSAT score of 600 or more?

```
229/(362-229)
[1] 1.721805
```

The odds that a student has a MathSAT score of 600 points or more are about 1.72 to 1.

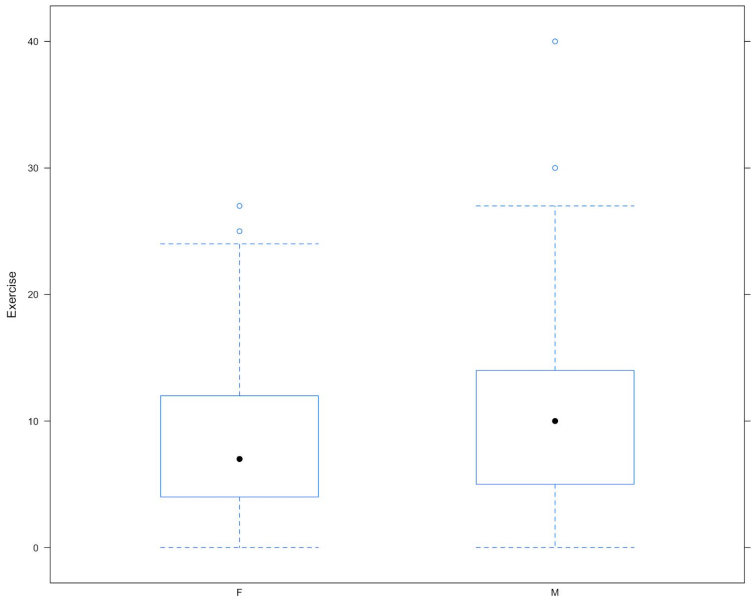
(j) You are to investigate how the time spent exercising varies by gender? Obtain appropriate numeric outcome and write a brief account of your results.

```
favstats(Exercise~Gender,data=s)
  Gender min Q1 median Q3 max     mean       sd   n missing
1      F   0  4      7 12  27 8.109467 5.183091 169       0
2      M   0  5     10 14  40 9.875648 6.068625 193       0
```

The mean amount of time exercising is higher for male respondents (9.88 hours) than for female respondents (8.11 hours). Similarly, the median amount of time exercising for male respondents (10 hours) is higher than it is for female respondents (7 hours). In addition to having a higher center by both conventional measures, the distribution of exercise time for male respondents has a greater spread. The standard deviation of exercise times for male respondents (6.07 hours) is greater than that of the distribution for female respondents (5.18 hours), and the range of exercise time for male respondents (40 hours) is greater than that of female respondents (27 hours) as well.

(k) Obtain a suitable display to support your account in part (j). What can you say about the shape of the two distributions of exercise times?

```
bwplot(Exercise~Gender,data=s)
```



As you can see, the box plot of male exercise times has a higher center and greater spread than the box plot of female exercise times. As far as the shape of the two distributions goes, for both male and female exercise times the distribution is skewed right.

2. According to the American Statistical Association the distribution of salaries for statisticians employed by pharmaceutical companies in the United States is approximately normal with a mean of $125,500 and a standard deviation of $12,600.

(a) What fraction of such employees earn more than $140,000?

```
1-pnorm(140000,125500,12600)
[1] 0.1249086
```

The fraction of statisticians employed by pharmaceutical companies in the US who earn more than $140,000 is approximately .125, or ⅛.

(b) What fraction earn between $120,000 and $130,000?

```
pnorm(130000,125500,12600)-pnorm(120000,125500,12600)
[1] 0.3082734
```

The fraction who earn between $120,000 and $130,000 is approximately .308.

(c) Only 5% of such employees earn more than $X. What is X?

```
qnorm(.95,125500,12600)
[1] 146225.2
```

Only 5% of such employees earn more than $146,225.20. X is $146,225.20.

(d) What is the interquartile range of salaries in this population?

```
qnorm(.75,125500,12600)-qnorm(.25,125500,12600)
[1] 16997.14
```

The interquartile range of salaries in this population is $16,997.14.

(e) Suppose we randomly select 20 such employees. What is the probability that exactly 2 of the 20 earn more than $140,000?

```
dbinom(2,20,.1249086)
[1] 0.2684725
```

The probability that exactly 2 of the 20 randomly selected employees earn more than $140,000 is .268.