# Descriptive Statistics II: Association
# Assignment

1. Exam scores for $n_J$=10 students in a Jenny's chemistry course were 61, 97, 84, 91, 78, 74, 76, 86, 82, 94. Exam scores for $n_S$=8 students in Smith's chemistry course were 48, 51, 52, 81, 76, 65, 80, 82. Exam scores for Montgomery's $n_M$=9 chemistry students were 78, 88, 84, 82, 82, 85, 87, 81, 85.

   (a) Calculate the mean exam score for each class.

   (b) Calculate the sum of squared deviations from the sample mean for each class:

$$\sum_{i=1}^{n}(y_i - \bar{y})^2$$

   (c) Calculate the sample variance and sample standard deviation for each of the three classes.

   (d) Calculate the ranges (maximum – minimum) for each set of exam scores.

   (e) Calculate the coefficient of variation (CV) for each of the three classes.

   (f) Make dotplots for the exam scores by class, and put them on the same scale (0 to 100).

   (g) Which professor had the highest average exam score?

   (h) Which professor's exam scores are "most predictable" in the sense that there is little predictive uncertainty?

   (i) The range and the standard deviation are both measures of spread. How can it be that Smith's standard deviation exceeds Jenny's, but Jenny's range exceeds Smiths? Explain.

   (j) Which professor had the highest CV? Which had the lowest?

   (k) Which professor would you choose to take and why?

   (l) The ages of Jenny's students (rounded to the neared year) were 19, 20, 20, 18, 21, 26, 20, 23, 19, 22. What is the standard deviation? Why is it smaller than the standard deviation of Jenny's students' exam scores?

2. Forty waiting times (rounded to the nearest minute) at a doctor's office were

| | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| 8 | 12 | 21 | 16 | 15 | 17 | 5 | 26 | 29 | 18 |
| 27 | 9 | 15 | 14 | 20 | 14 | 41 | 39 | 11 | 25 |
| 13 | 22 | 25 | 37 | 16 | 19 | 11 | 31 | 18 | 19 |
| 10 | 12 | 15 | 14 | 24 | 10 | 17 | 12 | 17 | 18 |

(a) Make a dotplot of the data. Comment on the shape of the distribution.
(b) Find the 25th, 50th, and 75th percentiles (Q1, Q2, and Q3).
(c) Make a boxplot of the data.
(d) The doctor's office advertises that 90% of its patients wait 25 minutes or less. Do the data support this claim?

3. (Use R)Open the dataset called lumber.csv. A land owner wanted to estimate how much lumber his land would yield. He randomly selected seventy equal-sized tracts (50' x 50') from his land, and the number of trees with diameters in excess of 12" were counted for each.

(a) Make a relative frequency histogram.
*Hint:*

```
> prob30 <- read.csv("lumber.txt")
> hist(prob30$X.Number., main="Relative Frequency Distribution of Tree Counts Per 50' X 50' Plot", freq=FALSE, ylab="Relative Frequencies", xlab="Number of Trees")
```

(b) Calculate the sample mean. What does this sample mean represent in the context of the problem? That is, what can the land owner use this to estimate?
(c) Calculate the sample standard deviation.
(d) Construct the intervals $\overline{y} \pm s$, $\overline{y} \pm 2s$, and $\overline{y} \pm 3s$. Calculate the percentages of the seventy land tracts that fall within each interval. Compare these percentages with those of the empirical rule. Are they close?

4. (Use R)The numbers of blood donors for twenty consecutive Fridays are recorded in the dataset blood.csv. Load this dataset into R.

   (a) Make a stem and leaf plot. ***Hint:***

   ```
   > stem(blood$X.Donors., scale=2)
   ```

   Try changing the scale parameter and see what happens.

   (b) Make a boxplot and describe the apparent distribution shape.

   *Hint:* try

   ```
   > boxplot(blood$X.Donors., main="Numbers of Blood Donors on Fridays")
   ```

5. (Use R)The tribal populations of Manipur have recently been transitioning from a subsistence economy to a market-oriented one (Demographic implications of socioeconomic transition among the tribal populations of Manipur, India. *Human Biology,* (1998) 70(3): 597-619). The table below gives information on the literacy levels among subsistence groups for a random sample of 614 married men and women from Manipur.

| **Subsistence \ Literacy Group Level** | **Illiterate** | **Primary Schooling** | **Middle School or More** |
|---|---|---|---|
| **Shifting Cultivators** | 114 | 10 | 45 |
| **Settled Agriculturalists** | 76 | 2 | 53 |
| **Town Dwellers** | 93 | 13 | 208 |

(a) Make a stacked bar chart.

(b) Compare the percentages based on row and column totals. What conclusions can you reach about the association between subsistence group and literacy level?

***Hint:*** I found this site useful:
http://www.cs.grinnell.edu/~rebelsky/Courses/MAT115/2008S/R/stacked-bar-graphs.html
and so I did this:

```
> litlev = data.frame(
+ ShftngCultivators = c(114, 10, 45),
+ SettledAgs = c(76, 2, 53),
+ TownDwellers = c(93, 13, 208)
+ )
```

```
> barplot(as.matrix(litlev), main="Literacy Levels for Three Subsistence
Groups", legend=c("Illiterate", "Primary Schooling", "At Least Middle
School"), xlim=c(0, 10), width=1.8)
```

6. (Use  R)The dataset papertowels.csv contains price data on twenty four brands of paper towels from the February 1998 issue of *Consumer Reports*. Because some brands had more sheets per roll, the prices are given in cost per roll as well as cost per sheet.

(a) Compute the standard deviations for price per roll and price per sheet.

(b) Which variable (price per roll, price per sheet) is more variable?

(c) Would it be better to use the coefficient of variation (CV) to compare? Explain.

(d) Make a scatterplot of price per roll vs. number of sheets per roll.

(i) Is there an obvious linear trend?

(ii) If not, is there any kind of apparent relation between the two variables?

(iii) What factors might explain why the ratio of price per roll to number of sheets per roll is not constant? Explain.

***Hints:*** The R command for getting the linear correlation coefficient is

```
> cor(variable1, variable2)
```

Maybe there is a correlation between price per roll and sheets per roll (or maybe not). Also, you can get standard deviation with the sd() command. You can also investigate the variation in the variables with dotplots and boxplots. You can check out the ratio of one variable to another by making a scatterplot and checking to see if there seems to be a constant slope: plot(x.variable, y.variable) or plot(y.variable ~ x.variable)

7. (Use R)Open the dataset called AIDSSyphTuber.csv. This dataset contains the numbers of reported cases of AIDS, syphilis, and tuberculosis by state in 2001.

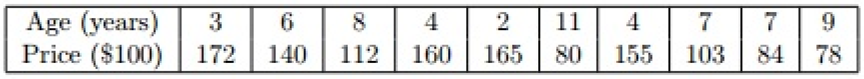
(a) Make a scatterplot of the number AIDS cases vs. the number of syphilis cases. Also make a scatterplot for the number of AIDS cases vs. the number of tuberculosis cases, and make a scatterplot for the number of syphilis cases vs. tuberculosis cases. Comment on what you observe.

(b) Compute the correlation coefficients for each of the three variable pairs.

(c) Do the correlation coefficients indicate any kind of associations? If so, why do you think this is the case?

(d) Now look up the state populations for 2001 on the internet. Add these numbers to your data set in R. Make three new scatterplots: one for AIDS cases vs. state population, one for syphilis cases vs. state population, and one for tuberculosis cases vs. state population. Now what do you think the reason is that some states have more AIDS, tuberculosis, and syphilis cases?

8. ***Used Cars. (Use R)***To estimate a fair used car price, you collect the following data:



| | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| Age (years) | 3 | 6 | 8 | 4 | 2 | 11 | 4 | 7 | 7 | 9 |
| Price ($100) | 172 | 140 | 112 | 160 | 165 | 80 | 155 | 103 | 84 | 78 |

(a) Using R, get the equation of regression line and predict the price of a 4.5 year-old car.

(b) Use the regression line we obtained with R to predict the price of a 60 year-old car. Does this prediction make any sense? Explain. What went wrong?

(c) Suppose your friend just sold a 70 year-old car for $50,000, and so we add that to our data (70, 500). You can use the rbind() command.

(d) What is the new regression line equation?

(e) What is the new *r* value?

(f) Make a scatterplot of the new data, and include both regression lines (original and new one). Is the new regression line equation better or worse than the original one for predicting used car prices? Explain.

(g) How is this new observation affecting our model?

(h) Remark on the fact that the new coefficient of determination is reasonably high, yet the new model is not very good.

9. ***What in the World?*** The famous humorist Will Rogers once said, "When the Oakies left Oklahoma and moved to California, they raised the average intelligence level in both states." Explain how this could be. Feel free to use pictures or made-up data in your explanation.

***10.NFL Penalty Yards vs. Perceived Uniform Malevolence: A First look at Permutation-Based Testing.*** In the paper, "The Dark Side of Self- and Social Perception: Black Uniforms and Aggression in Professional Sports," Frank and Gilovich studied whether or not perceived malevolence of NFL team jerseys was related to penalty yardage. People with no knowledge of the teams assigned scores to the jerseys that indicated how mean or nice they appeared, and then z-scores of the number of average penalty yards over seasons 1970 – 1986 for each team were calculated. Let's see if there was indeed a correlation between penalty yardage and the malevolence scores. But beyond that, even if the correlation is weak, we want to know if it is statistically significantly different from zero- that is, we want to know if the r-value we got is stranger than what we would see than if the penalty yards were assigned randomly regardless of the jersey. This will be our first hypothesis test, and we will take a permutation-based approach.

We'll use StatKey to do this, and then we'll do it ourselves with R. So go to http://www.lock5stat.com/ and click on StatKey. Then click on Test for Slope, Correlation under the Randomization Hypothesis Tests area. Click the light-blue bubble box in the upper left-hand corner to load the dataset called, "Malevolent Uniforms…". Look to the right where it says, "Original Sample". You'll see an r value of .43, and a scatterplot of rescaled penalty yards vs. some violence rating. Even though r is not high, we want to test to see if it is indeed statistically significantly larger than 0. Here's how.

Let's hypothesize that there really is no correlation, and so there is no relationship between penalty yards and what's on a uniform. Then given the uniforms along with their malevolence scores, and given the penalty yard scores, shouldn't each team be equally likely to be assigned the penalty yard scores? For example, shouldn't the Dolphins be just as likely to receive the penalty yard score of 1.19 as the score they actually received of -1.6?

So remember that the actual r value we observed it .43. Now, click Generate 1 Sample a few times. Each time you do so, the StatKey is randomly reassigning all the penalty yard scores to the malevolence scores, and recomputing the r value. You can mouse over the dots in your dotplot and see the scatterplot that pertains to each r-value to the right underRandomization Sample. Now click Generate 1000 Samples four or five times.

(a) What is happening to the big dotplot and why?
(b) Click the little box next to "Two-Tail" in the upper left-hand corner of the plot. Then click the light-blue bubble box at the bottom right, below the horizontal axis, and enter .43 (this is the actual correlation value you observed in the original sample). The area in the red should total to about

2% (plus or minus). This area in red is called the ***p-value***. It is the probability that you would observe a statistic (say, r) as strange or more as the one you actually observed (r = .43) if indeed your initial hypothesis (that the actual r is 0) is true. So based on the p-value, what is your conclusion? Is the r-value in this case significantly different from 0? Explain.

(c) Include a copy of your StatKey page with your homework.

Now figure out how to do this in R. That is, use the MalevolentUniforms data in R to carry out a permutation-based hypothesis test for the correlation coefficient r. You should be able to highlight the data in StatKey, copy it, and paste it into NotePad, and open it in R from there. Be sure to include your R code in your homework. What is your p-value now? Is it much different from the one you got with StatKey? Explain.

***Hints:*** first obtain the actual correlation with

```
> the.correlation <- cor(uniforms)
> the.correlation
```

Note this results in a matrix whose elements are the correlations between penalty yards and itself (1), penalty yards and malevolence (.429796), malevolence and penalty yards (.429796), and malevolence with itself (1). So this is a symmetric matrix because the correlation function is a symmetric function (examine the formula for r).

Then write a loop to approximate the permutation distribution of the linear correlation coefficient:

```
> uniforms <- read.csv(“MalNFLUniforms.txt”)
> i <- 1
> rpermdist <- matrix(NA, 10000, 1)
> while(i <= 10000){
+ sam1 <- sample(uniforms[,1], 28, replace = FALSE,
prob = NULL)
+ sam2 <- sample(uniforms[,2], 28, replace = FALSE,
prob = NULL)
+ r <- cor(sam1, sam2)
+ rpermdist[i,1] <- r
+ i <- i+1
+ }
```

Make a histogram to get a visual:

```
> hist(rpermdist)
```

You can approximate a p-value (two-tailed) with

```
> sum(rpermdist >= the.correlation[1,2])/10000 +
sum(rpermdist <=
-1*the.correlation[1,2])/10000
```