BSTAT 5301 - Homework H01

By Craig W. Slinkman

September 4, 2015

# Homework H01: Descriptive statistics

This homework does contain a relatively large number of questions. However, if you can answer every question correctly then I will guarantee that you will do well on the first exam.

## 1. 2010 United States Census

In the 2010 US Census, we see that 37,342,870 people, or 12.4% of all residents, are foreign-born. Give the relevant mathematical symbol and value for the proportion using correct notation.

## 2. Student survey data

Of the 361 students who answered the question about the number of piercings they had in their body, 188 had no piercings, 82 had one or two piercings, and the rest had more than two. Construct a frequency distribution of the number of piercings using this data.

## 3. Airborne Antibiotics

A recent study shows that antibiotics added to animal feed to accelerate growth can become airborne. Some of these drugs can be toxic if inhaled and may increase the evolution of antibiotic-resistant bacteria. Scientists analyzed 20 samples of dust particles from animal farms. Tylosin, an antibiotic used in animal feed that is chemically related to erythromycin, showed up in 16 of the samples.

1. What is the variable in this study? What are the individual cases?
2. Display the results in a frequency table.
3. Make a bar chart of the data.
4. Give a relative frequency table of the data.

## 4. Is There a Genetic Marker for Dyslexia?

A disruption of a gene called DYXC1 on chromosome 15 for humans may be related to an increased risk of developing dyslexia. Researchers studied the gene in 109 people diagnosed with dyslexia and in a control group of 195 others who had no learning disorder. The DYXC1 break occurred in 10 of those with dyslexia and in 5 of those in the control group.

1. Is this an experiment or an observational study? What are the variables?
2. How many rows and how many columns will the data table have? Assume rows are the cases and columns are the variables. (There might be an extra column for identification purposes; do not count this column in your total.)
3. Display the results of the study in a two-way table.
4. To see if there appears to be a substantial difference between the group with dyslexia and the control group, compare the proportion of each group who have the break on the DYXC1 gene.
5. Does there appear to be an association between this genetic marker and dyslexia for the people in this sample? (We will see in Chapter 4 whether we can generalize this result to the entire population.)
6. If the association appears to be strong, can we assume that the gene assume that the gene disruption causes dyslexia? Why or why not?

## 5. Smoking and Pregnancy Rate

Studies have concluded that smoking while pregnant can have negative consequences, but could smoking also negatively affect one’s ability to become pregnant? A study collected data on 678 women who had gone off birth control with the intention of becoming pregnant. Smokers were defined as those who smoked at least one cigarette a day prior to pregnancy. We are interested in the pregnancy rate during the first cycle off birth control. The results are summarized in below.

## Smoker on-smoker Total  
## Pregnant 38 206 244  
## Not Pregnant 97 337 434  
## Total 135 543 678

1. Is this an experiment or an observational study?  
   Can we use these data to determine whether smoking influences one’s ability to get pregnant? Why or why not?
2. What is the population of interest?
3. What is the proportion of women successfully pregnant after their first cycle disregarding their smoking habits? Denote this quantity by.
4. Now consider the non-smokers. What proportion of non-smokers where successful pregnant after their first cycle? Denote this proportion by. We read this as the proportion of successful pregnancy given the women is a non-smoker.
5. Now consider the smokers. What proportion of smokers were successfully pregnant after their first cycle? Denote this proportion by. We read this as the estimated proportion of smokers who became pregnant.

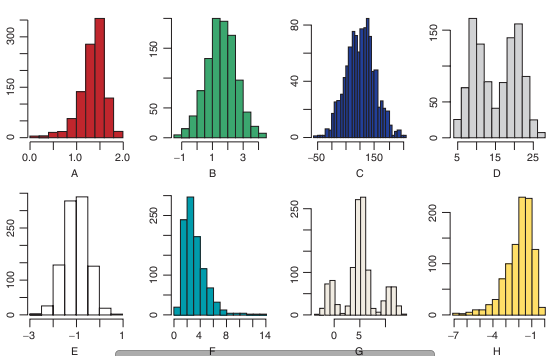
## 6. Vaccine for Malaria

In order for a vaccine to be effective, it should reduce a person’s chance of acquiring a disease. Consider a hypothetical vaccine for malaria tropical disease that kills between 1.5 and 2.7 million people every year. Suppose the vaccine is tested with 500 volunteers in a village who are malaria free at the beginning of the trial. Two hundred of the volunteers will get the experimental vaccine and the rest will not be vaccinated. Suppose that the chance of contracting malaria is 10% for those who are not vaccinated. Construct a two-way table to show the results of the experiment if:

1. The vaccine has no effect.
2. The vaccine cuts the risk of contracting malaria in half.

## 7. Distributional shapes

Which histograms below are approximately symmetric and bell-shaped?



## 8. Distributional shape

Draw any dotplot to show a dataset that is:

1. Clearly skewed to the left
2. Approximately symmetric and bell-shaped
3. Clearly positively skewed
4. Approximately symmetric but not bell-shaped

## 9. Measures of centrality and outliers

For each set of data below:

1. Find the mean x.
2. Find the median m.
3. Indicate whether there appear to be any outliers. If so, what are they?

* 2.43 8, 12, 3, 18, 15
* 2.44 41, 53, 38, 32, 115, 47, 50
* 2.45 15, 22, 12, 28, 58, 18, 25, 18
* 2.46 110, 112, 118, 119, 122, 125, 129, 135, 138, 140

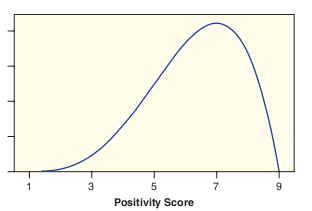
## 10. Statistical notation

Give the appropriate statistical notation for the following fact:

The average number of television sets owned per household for all households in the US is 2.6.

## 11. Is Language Biased toward Happiness?

"Are natural languages neutrally, positively, or negatively biased?  That is the question a recent study set out to answer. They found the top 5000 words used in English in each of four different places: Twitter, books on the Google Book Project, The New York Times, and music lyrics. The resulting complete list was 10,222 unique words in the English language. Each word was then evaluated independently by 50 different people, each giving a rating on how the word made them feel on a 1 to 9 scale where 1 = least happy, 5 = neutral, and 9 = most happy. (The highest rated word was "laughter while the lowest was "terrorist.) The distributions of the ratings for all 10,222 words for each of the four media sources were surprisingly similar, and all had approximately the shape shown below.



1. Describe the shape of the distribution.
2. Which of the following values is closest to the median of the distribution:
3. Will the mean be smaller or larger than the value you gave for the median in part (b)?

## 12. 2.60 Time Spent Exercising, between Males and Females

Females Often we are interested not just in a single mean but in a difference in means between two groups. In the data, there are 36 seniors: 26 males and 10 females. The data for males and females spent exercising is:

## [1] 4 2 5 6 12 15 10 5 0 5

## [1] 10 10 6 5 7 8 4 12 12 4 15 10 5 5 2 2 7 3 5 15 6 6 5  
## [24] 0 8 5

The time spent exercising is measured in hours.

1. Compute, the mean number of hours spent exercising by the females.
2. Compute, the mean number of hours spent exercising by the males.
3. Compute the difference- , and interpret it in context.

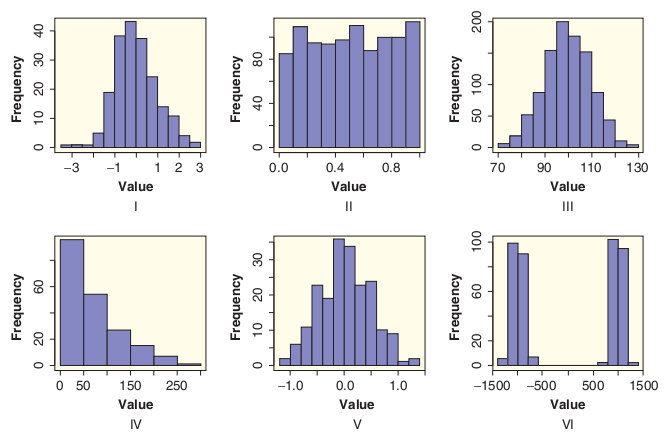
## 13. Using StatKey to compute means and standard deviations

Use **StatKey** or other technology to compute the mean and standard deviations of the following data sets.

1. 64, 57, 78, 73, 62
2. 9, 28, 24, 24, 4, 1, 1, 2, 1, 7
3. 10, 9, 7, 8, 8, 8, 7, 7, 6, 10, 9, 8

## 14. Five number displays and histograms

Match each five number summary with one of the histograms below.

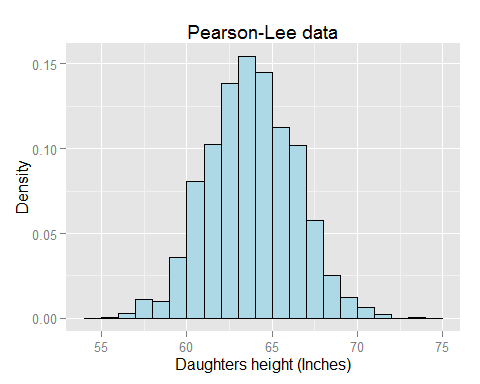


1. (0, 0.25, 0.5, 0.75, 1)
2. (-1.08,-0.30, 0.01, 0.35, 1.27)
3. (0.64, 27.25, 53.16, 200, 275.7)
4. (-3.5,-0.63,-0.11, 0.59, 2.66)
5. (71.45, 92.77, 99.41, 106.60, 129.70)

## 15. Heights of daughters

Karl Pearson organized the collection of data on over 1200 families in England in the period 1893 to 1898. This particular data set gives the Heights in inches of mothers and their daughters, with up to two daughters per mother. All daughters are at least age 18, and all mothers are younger than 65. Data were given in the source as a frequency table to the nearest inch. Rounding error has been added to remove discreteness from graph. Use the histogram below to estimate the mean and she standard deviation of the daughters.

## Loading required package: alr4  
## Loading required package: car  
## Loading required package: effects  
##   
## Attaching package: 'effects'  
##   
## The following object is masked from 'package:car':  
##   
## Prestige  
##   
## Loading required package: ggplot2



Estimate the mean and the standard deviation of the daughters in the sample.

## Q16. Computing z scores

For the following data sets with given mean and standard deviation compute the z value for the observed value of x.

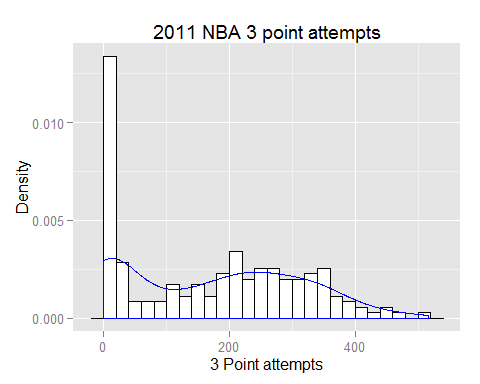
## Q17. Laptop computer studies and male fertility

Laptop Compute Studies have shown that heating the scrotum by just 1 degree C can reduce sperm count and sperm quality, so men concerned about fertility are cautioned to avoid too much time in the hot tub or sauna. A new study suggests that men also keep their laptop computers off their laps. The study measured scrotal temperature in 29 healthy male volunteer as they sat with legs together and a laptop computer on the lap. Temperature increase in the left scrotum over a 60-minute session is given as 2.31 ± 0.96 and a note tells us that Temperatures are given as degrees C; values are shown as mean ± SD. The abbreviation SD stands for standard deviation. (Men who sit with their legs together without a laptop computer do not show an increase in temperature.)

If we assume that the distribution of the temperature increases for the 29 men is symmetric and bell-shaped, find an interval that we expect to contain about 95% of the temperature increases.

## Q18. Distribution of Three-Point Attempts in the NBA

In basketball, a basket is awarded three points (rather than the usual two) if it is shot from farther away. Some players attempt lots of three-point shots and quite a few attempt none, as we see in the distribution of number of three-point attempts by players in the NBA is shown below. The data are available in data set. The 0 variable name . This data was plotted using the R-package



Is it appropriate to use the 95% rule with this dataset? Why or why not?

## Q19. Rough Rule of Thumb for the Standard Deviation

According to the 95% rule, the largest value in a sample from a distribution which is approximately symmetric and bell-shaped should be between 2 and 3 standard deviations above the mean, while the smallest value should be between 2 and 3 standard deviations below the mean. Thus the range should be roughly 4 to 6 times the standard deviation.

As a rough rule of thumb, we can get a quick estimate of the standard deviation for a bell-shaped distribution by dividing the range by 5.

Check how well this quick estimate works in the following situations.

(a) Pulse rates from the dataset discussed in Example 2.17 on page 77. The five number display summary of pulse rates is

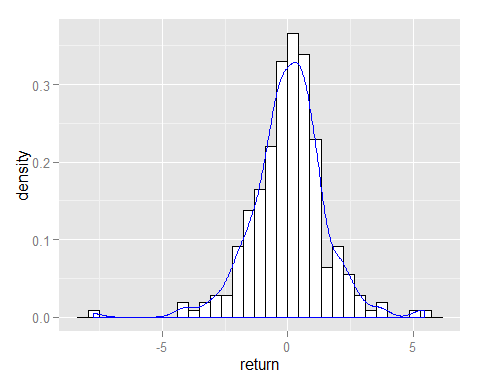
The standard deviation is s = 12.2 bpm. Find the rough estimate using all the data, and then excluding the two outliers at 120 and 130, which leaves the maximum at 96. Find the rough estimate of standard deviation and compare to the computed deviation for the two cases (a) and (b) below.

1. Number of hours a week spent exercising from the dataset discussed in Example 2.21 on page 81. The five number summary of this dataset is
2. The standard deviation is s = 5.741 hours.

## Q20. Returns on McDonald's stock

The weekly market return on McDonald's stock for 10 years is given in the file. A histogram of the weekly returns in percent is shown below.

## stat\_bin: binwidth defaulted to range/30. Use 'binwidth = x' to adjust this.



The five number display for the weekly returns (plus the mean) is given below:

with( MCD,summary(return))

## Min. 1st Qu. Median Mean 3rd Qu. Max.   
## -7.76000 -0.74150 0.12750 0.03388 0.81260 5.46100

The standard deviation is

1. Use the approximation for quick estimation of the standard deviation based on the range suggested prior problem.
2. Compare that value to the computed standard deviation given above.
3. Does it appear that there are outliers in the weekly returns?

## Q25.

Four data sets are given blow along with the six number displays summarizing the data set. A six number display include the dataset mean in the data. Determine if there are any potential outliers in the data using the

Data set 1

## [1] 81 78 87 97 56 72 75 77 101 86

## Min. 1st Qu. Median Mean 3rd Qu. Max.   
## 56.00 75.50 79.50 81.00 86.75 101.00

### Data set 2

## Min. 1st Qu. Median Mean 3rd Qu. Max.   
## 0.0 1.0 1.5 5.2 6.5 25.0

### Data set 3

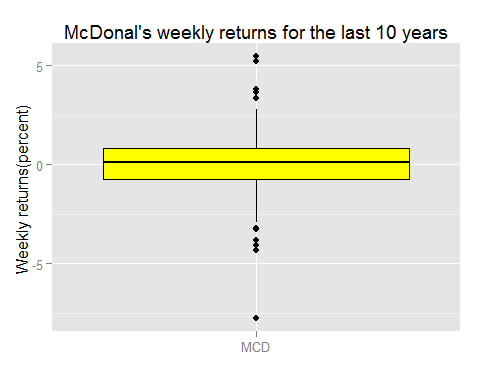
## Min. 1st Qu. Median Mean 3rd Qu. Max.   
## 6.00 10.00 10.50 10.75 12.25 14.00

## Dataset 4

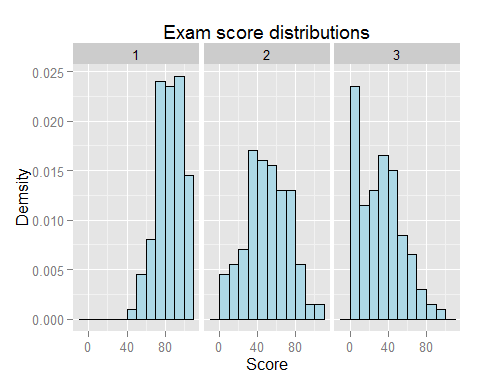
## Min. 1st Qu. Median Mean 3rd Qu. Max.   
## 0.03713 0.35400 0.53430 0.47540 0.66250 0.72490

## Q27 Using the boxplot to assess a distribution

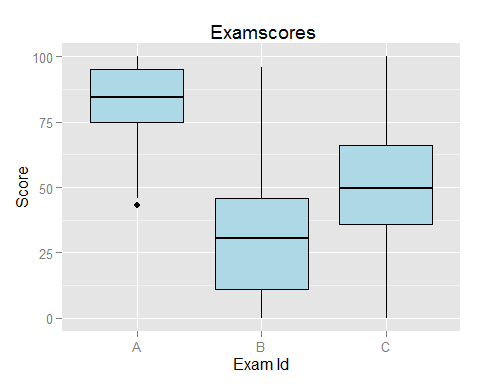
Let's reconsider the distribution of weekly McDonald’s weekly stock returns for the last 10 years, this time we will use a boxplot. This is shown below.

 Use the boxplot to give an approximate description of the weekly McDonalds' returns on stock for the last 10 years.

## Q28. Boxplot and distributional shapes

Consider the three histograms plotted below. 

Match the above histograms using the histogram number with the boxplot letter found on the boxplots below.

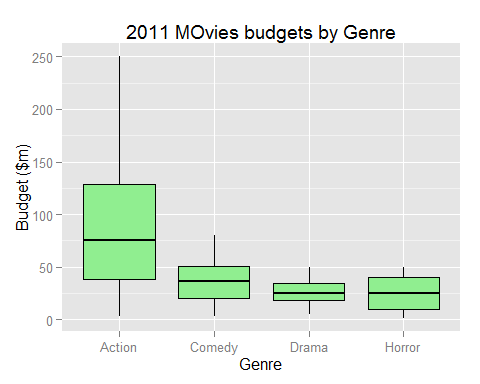


## 29. Do Movie Budgets Differ Based on the Genre of the Movie?

The dataset includes a quantitative variable on the Budget of the movie, in millions of dollars, as well as a categorical variable classifying each movie by its Genre. (We have limited this analysis to the four most popular genres).

The plot below shows side-by-side boxplots investigating a relationship between these two variables.

## Loading required package: data.table



1. Which genre appears to have the largest budgets? Which appears to have the smallest?
2. Which genre has the biggest spread in its budgets? Which has the smallest spread?
3. Does there appear to be an association between genre of a movie and size of the budget? Explain.

## Q30. Infection in Dialysis Patients

The table below gives the data showing the time to infection (days?), at the point of insertion of the catheter, for kidney patients using portable dialysis equipment. There are 38 patients, and the data give the first observation for each patient.

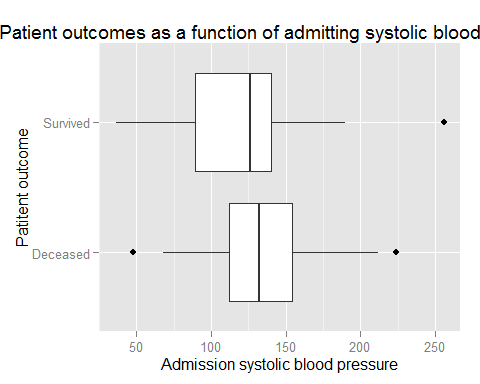
The five number summary for these data is (2, 15, 46, 149, 536).

## [1] 2 5 6 7 7 8 12 13 15 15 17 22 22 23 24 27 30  
## [18] 34 39 53 54 63 96 113 119 130 132 141 149 152 152 185 190 292  
## [35] 402 447 511 536

1. Identify any outliers in the data. Justify your answer.
2. Draw the boxplot.

## 31. Systolic Blood Pressure and Survival

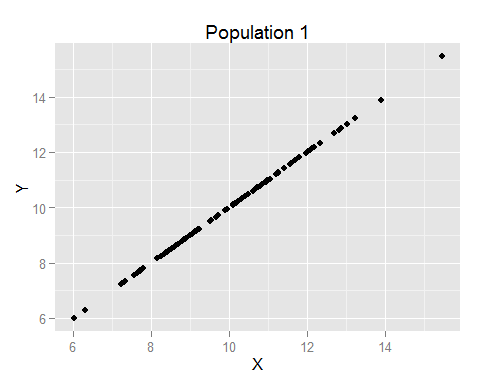
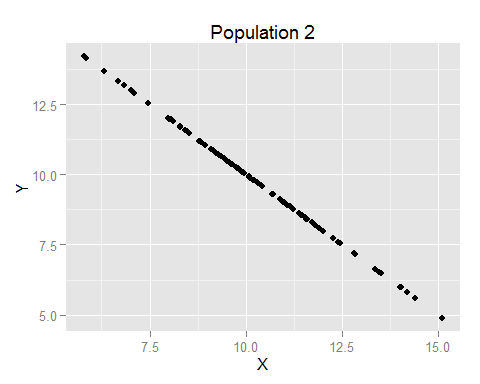
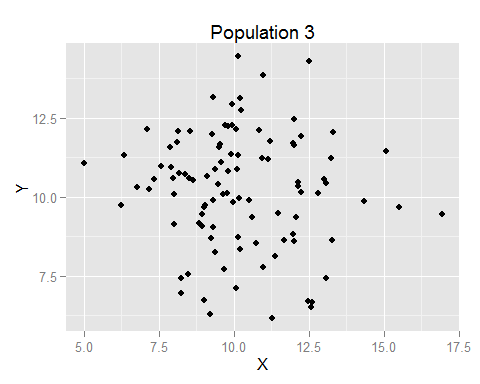
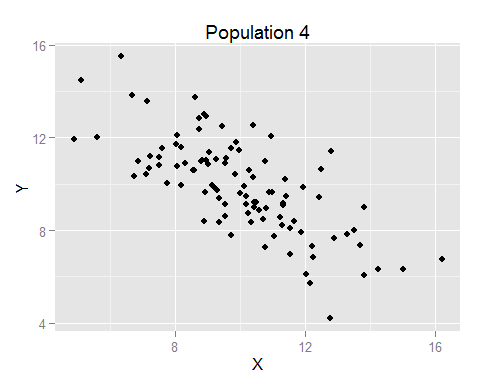
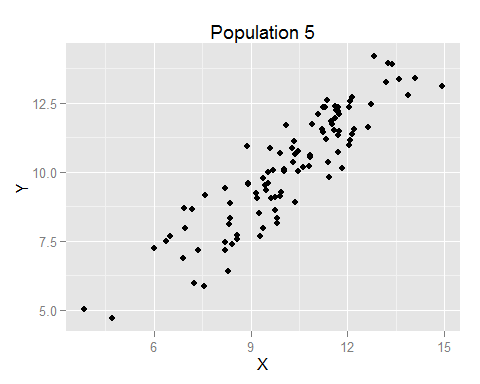
The data in data set contains a categorical variable Status indicating whether each patient lived (0) or died (1). Is there a relationship between the status (lived/died) and the systolic blood pressures? Use the side-by-side boxplots showing the systolic blood pressures for these two groups of patients in the figure below to discuss how the distributions compare.



## 32. Scatterplots ad correlations

Match the correlations and scatterplots below:

The scatter plots for the five populations are given below:

Match the correlations with the above scatterplots.

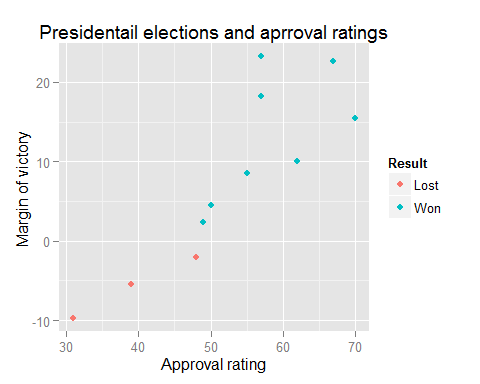
1. -0.083
2. -1
3. -0.726
4. 1
5. 0.899

## 33. Presidential approval rating and margin of victory or defeat

When a US president runs for re-election, how strong is the relationship between the president’s approval rating and the outcome of the election? The data we have in includes all the presidential elections since 1940 in which an incumbent was running and shows the presidential approval rating at the time of the election and the margin of victory or defeat for the president in the election. We show this data set below.

## Year Candidate Approval Margin Result  
## 1: 1940 Roosevelt 62 10.0 Won  
## 2: 1948 Truman 50 4.5 Won  
## 3: 1956 Eisenhower 70 15.4 Won  
## 4: 1964 Johnson 67 22.6 Won  
## 5: 1972 Nixon 57 23.2 Won  
## 6: 1976 Ford 48 -2.1 Lost  
## 7: 1980 Carter 31 -9.7 Lost  
## 8: 1984 Reagan 57 18.2 Won  
## 9: 1992 G.H.W.Bush 39 -5.5 Lost  
## 10: 1996 Clinton 55 8.5 Won  
## 11: 2004 G.W.Bush 49 2.4 Won

The scatterplot of margin of victory versus presidential approval rating is shown below.



1. In how many of the 11 elections listed did the incumbent president lose? Since 1940, what percent of the time has the sitting president lost his bid for re-election?
2. Which president had the highest approval rating? Which president had the highest margin of victory? Identify these two points on the scatterplot.

## 34. Mothers Love, Hippocampus, and Resiliency

Multiple studies in both animals and humans show the importance of a mother’s love (or the unconditional love of any close person to a child) in a child’s brain development. A recent study shows that children with nurturing mothers had a substantially larger area of the brain called the hippocampus than children with less nurturing mothers. This is important because other studies have shown that the size of the hippocampus matters: People with large hippocampus area are more resilient and are more likely to be able to weather the stresses and strains of daily life. These observations come from experiments in animals and observational studies in humans.

1. Is the amount of maternal nurturing one receives as a child positively or negatively associated with hippocampus size?
2. Is hippocampus size positively or negatively associated with resiliency and the ability to weather the stresses of life?
3. How might a randomized experiment be designed to test the effect described in part (a) in humans? Would such an experiment be ethical?
4. Can we conclude that maternal nurturing in humans causes the hippocampus to grow larger? Can we conclude that maternal nurturing in animals (such as mice, who were used in many of the experiments) causes the hippocampus to grow larger? Explain.

## 34. Is Your Body Language Closed or Open?

A closed body posture includes sitting hunched over or standing with arms crossed rather than sitting or standing up straight and having the arms more open. According to a recent study, people who were rated as having a more closed body posture "had higher levels of stress hormones and said they felt less powerful than those who had a more open pose."

1. What are the variables in this study? Is each variable categorical or quantitative? Assume participants had body language rated on a numerical scale from low values representing more closed to larger values representing more open. Assume also that participants were rated on a numerical scale indicating whether each felt less powerful (low values) or more powerful (higher values).
2. Do the results of the study indicate a positive or negative relationship between the body language scores and levels of stress hormones? Would your answer be different if the scale had been reversed for the body language scores?
3. Do the results of the study indicate a positive or negative relationship between the body language scores and the scores on the feelings of power? Would your answer be different if both scales were reversed? Would your answer be different if only one of the scales had been reversed?

## 35. Comparing Global Internet Connections

The Nielsen Company measured connection speeds on home computers in nine different countries in order to determine whether connection speed affects the amount of time consumers spend online. This data is shown below:

## Country PercentFastConnection HoursOnline  
## 1 Swizerland 88 20.18  
## 2 United States 70 26.26  
## 3 Germany 72 28.04  
## 4 Australia 64 23.02  
## 5 United Kingdom 75 28.48  
## 6 France 70 27.49  
## 7 Spain 69 26.97  
## 8 Italy 64 23.59  
## 9 Brazil 21 31.58

The data shows the percent of Internet users with a "fast connection (defined as 2Mb or faster) and the average amount of time spent online, defined as total hours connected to the web from a home computer during the month of February 2011. The data are also available in the dataset.

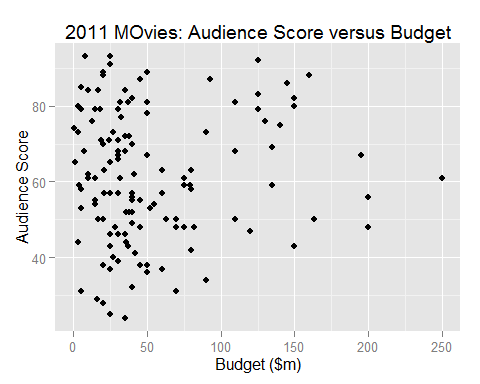
1. What would a positive association mean between these two variables? Explain why a positive relationship might make sense in this context.
2. What would a negative association mean between these two variables? Explain why a negative relationship might make sense in this context.

## 36. Do Movies with Larger Budgets Get Higher Audience Ratings?

We have previously dataset. This data set includes many variables for movies that were produced in Hollywood in 2011, including & .

1. Use technology to create a scatterplot to show the relationship between the budget of a movie, in millions of dollars, and the audience score. We want to see if the budget has an effect on the audience score. Your instructor has created this scatterplot for you,

## Warning: Removed 3 rows containing missing values (geom\_point).



1. Is there a linear relationship? How strong is it? Give your answer in the context of movies.
2. There is an outlier with a very large budget. What is the audience rating for this movie and what movie is it? There is another data value with a budget of about 125 million dollars and an audience score over 90. To what movie does that dot correspond? Hit you will need to download the data. I suggest you download the Excel version of the file.
3. Is there a linear relationship? How strong is it? Give your answer in the context of movies.
4. There is an outlier with a very large budget. What is the audience rating for this movie and what movie is it? There is another data value with a budget of about 125 million dollars and an audience score over 90. To what movie does that dot correspond?
5. The sample correlation coefficient for this data is 0.084. Is it a good strategy to only make high budget movies?

## 37. Predicting an exam score

Suppose we want to predict Grade the grade on an exam, as a function of the hours spent studying for the exam. The estimated regression equation is

Note the notation is read the estimated expected grade on the exam. This is the same

Suppose a student spends 10 hours studying for the exam. Predict what grade the student will earn.

## 38. Using StatKey for simple linear regression computations

Given the data in the table below use **StatKey** to compute the regression equation. Be sure to write your answer in standard form, that is.

### Data

## x y  
## 1 6 66  
## 2 3 55  
## 3 9 78  
## 4 7 72  
## 5 6 65  
## 6 9 76  
## 7 11 84  
## 8 6 66  
## 9 10 82  
## 10 5 62

## 39. Concentration of in the Atmosphere

Levels of carbon dioxide () in the atmosphere are rising rapidly, far above any levels ever before recorded. Levels were around 278 parts per million in 1800, before the Industrial Age, and had never, in the hundreds of thousands of years before that, gone above 300 ppm. Levels are now nearing 400 ppm. Table below shows the rapid rise of $CO2\_$ concentrations over the last 50 years. The dataset also available in Lock datasets . This dataset is also available in **StatKey** in the **Two Quantitative Variables** section of the main page. Our research question of interest can we use this information to predict levels in future years.

1. What is the explanatory variable? What is the response variable?
2. Draw a scatterplot of the data. Does there appear to be a linear relationship in the data?
3. Use technology to find the correlation between year and CO2 levels. Does the value of the correlation support your answer to part (b)?
4. Use technology to estimate the simple linear regression equation and write this equation in standard form.
5. Predict the amount of for the year 2015.
6. Use **Google** or another search engine to find the current concentration in parts per million (ppm). Assess the accuracy of your prediction.

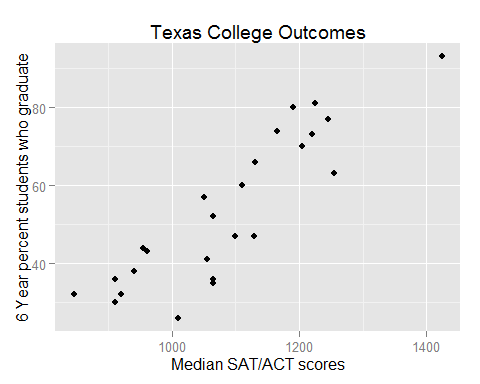
## 40. Texas College Results

The following data is from a study comparing the graduation rates versus demographic and student SAT/ACT scores.

## Warning: NAs introduced by coercion

The scatter plot below is a plot of 2009 6 year graduation rates versus the institutions estimated median SAT/ACT score.

## Warning: Removed 1 rows containing missing values (geom\_point).



1. Does there appear to be a linear relationship between the median SAT/ACT scores and the percentage of students who graduate within 6 years?
2. The estimated regression intercept coefficient for this data is

.

1. The estimated regression slope coefficient for this data is

Write the prediction equation for predicting the percent graduation as a linear function of the median SAT/ACT score.

1. For this data set the median UTA SAT score is the 1050. Predict the percent of students that you would expect to graduate from UTA within 6 years.