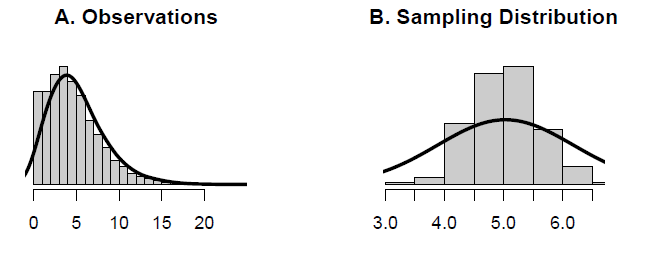
IS 606 Fall 2015 - Final Exam

**Part 1**

Figure A below represents the distribution of an observed variable. Figure B below represents the distribution of the mean from 500 random samples of size 30 from A. The mean of A is 5.05 and the mean of B is 5.04. The standard deviations of A and B are 3.22 and 0.58, respectively.



**a**. Describe the two distributions.

Distribution A is unimodal with a right skew

Distribution B is unimodal and symmetric, a normal distribution

**b**. Explain why the means of these two distributions are similar but the standard deviations are not.

The sampling distribution consists of at least 30 independent observations and the samples are large enough to allow the sample means to closely approximate the mean of the observed variable.

The standard deviations are different because the observed variable has a notable right skew with prominent outliers that will stretch the standard deviation, while the sampling distribution is close to normal and more closely hugs the mean on both sides, allowing for a smaller standard deviation.

**c**. What is the statistical principal that describes this phenomenon?

These principles are described by the central limit theorem, which says that if a sample consists of at least 30 independent observations and the data are not strongly skewed, then the distribution of the sample mean is well approximated by a normal model.

**Part II**

Consider the four datasets, each with two columns (x and y), provided below.

**options**(digits=2)

data1 <- **data.frame**(x=**c**(10,8,13,9,11,14,6,4,12,7,5), y=**c**(8.04,6.95,7.58,8.81,8.33,9.96,7.24,4.26,10.84,4.82,5.68))

data2 <- **data.frame**(x=**c**(10,8,13,9,11,14,6,4,12,7,5), y=**c**(9.14,8.14,8.74,8.77,9.26,8.1,6.13,3.1,9.13,7.26,4.74))

data3 <- **data.frame**(x=**c**(10,8,13,9,11,14,6,4,12,7,5), y=**c**(7.46,6.77,12.74,7.11,7.81,8.84,6.08,5.39,8.15,6.42,5.73))

data4 <- **data.frame**(x=**c**(8,8,8,8,8,8,8,19,8,8,8), y=**c**(6.58,5.76,7.71,8.84,8.47,7.04,5.25,12.5,5.56,7.91,6.89))

For each column, calculate (to two decimal places):

**a. The mean (for x and y separately).**

IE: mean(data1$x)

Data1: x = 9, y = 7.5

Data2: x = 9, y = 7.5

Data3: x = 9, y = 7.5

Data4: x = 9, y = 7.5

**b. The median (for x and y separately).**

IE: median(data1$x)

Data1: x = 9, y = 7.6

Data2: x = 9, y = 8.1

Data3: x = 9, y = 7.1

Data4: x = 8, y = 7

**c. The standard deviation (for x and y separately).**

IE: sd(data1$x)

Data1: x = 3.3, y = 2

Data2: x = 3.3, y = 2

Data3: x = 3.3, y = 2

Data4: x = 3.3, y = 2

**For each x and y pair, calculate (also to two decimal places):**

**d. The correlation.**

IE: cor(data1$x, data1$y)

Data1: .82

Data2: .82

Data3: .82

Data4: .82

**e. Linear regression equation.**

IE:lm(data1$y~data1$x)

Data1: y = 3 + .5x

Data2: y = 3 + .5x

Data3: y = 3 + .5x

Data4: y = 3 + .5x

**f. R-Squared**

IE: summary(lm(data1$y~data1$x))

Data1: .667

Data2: .666

Data3: .666

Data4: .667

**For each pair, is it appropriate to estimate a linear regression model? Why or why not? Be specific as to why for each pair!**

While they all neatly have similar basic descriptive statistics, only data pair 1 is appropriate for linear regression analysis as x and y appear to have a linear relationship as demonstrated in its scatter plot.

Data set 2 appears to be best estimated by a curve, not a line.

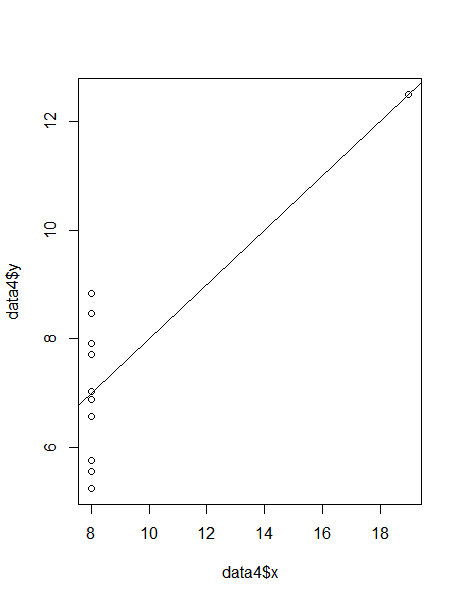
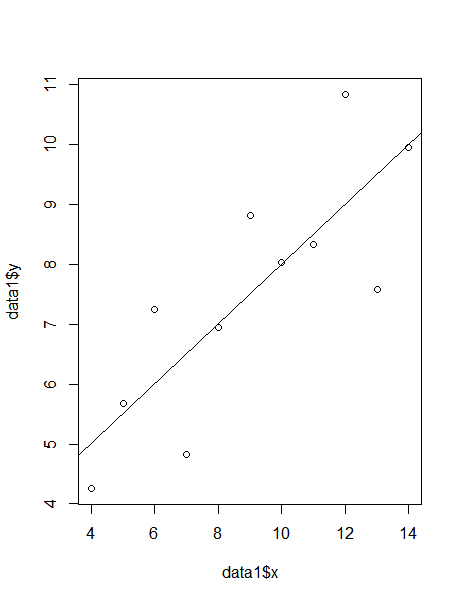
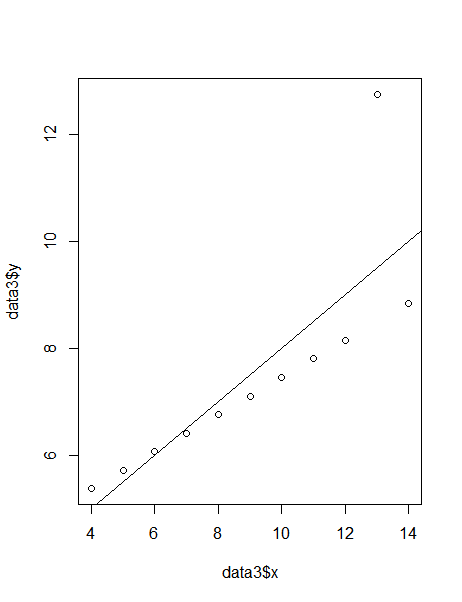
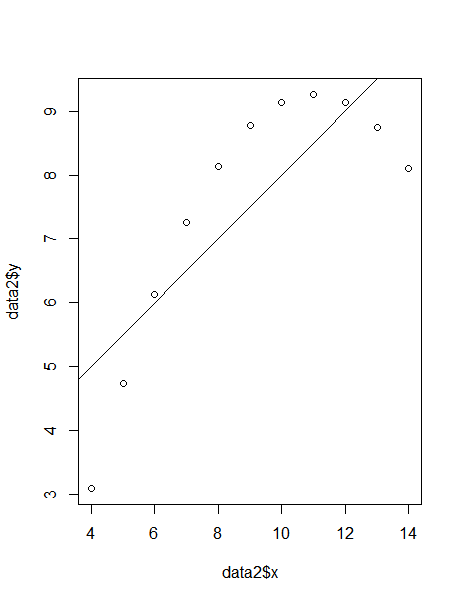
Dataset3 also appears to have a curved line, with a very notable outlier, although if the outlier is disregarded, linear regression can give some approximation of the curved line.

Dataset4 is a vertical line with a very distinguished outlier, which far too heavily skews the linear approximation.

**Explain why it is important to include appropriate visualizations when analyzing data. Include any visualization(s) you create.**

Appropriate visualizations allow for a more intuitive understanding of the data, allowing one to manually inspect the dataset in ways that basic descriptive statistics doesn’t allow for. In this case, the basic statistics implied that the datasets are very similar, but visualizations show a very different story.

IE: plot(data1$y~data1$x)

abline(d1)