**z-scores and Empirical Rule &**

**Regression, Correlation and Scatterplots**

We will begin this activity with some practice using the concepts of z-scores and the Empirical Rule. This will be short. The bulk of the activity will be dedicated to examining relationships between 2 quantitative variables.

**z-scores and Empirical Rule**

**#1** Car and truck speeds at a particular location have approximately a bell-shaped distribution with mean = 65 mph and standard deviation = 5 mph.

*For parts a-c, uses the empirical rule to fill in the blanks in each part:*

**a.** About 68% of cars and trucks travel between \_\_\_\_\_\_\_ and \_\_\_\_\_\_\_ at this location.

**b.** About 95% of cars and trucks travel between \_\_\_\_\_\_ and \_\_\_\_\_\_\_\_ at this location.

**c.** About 99.7% of cars and trucks travel between \_\_\_\_\_\_ and \_\_\_\_\_\_\_\_ at this location.

**d.** A z-score is a measure of how many standard deviations a value is from the mean. Later in the course, we will learn that a z-score is an important measure of the size of a value.

The formula for a z-score is z = .



Determine a z-score for a vehicle speed of 72 mph.

Determine a z-score for a vehicle speed of 65 mph.

Determine a z-score for a vehicle speed of 60 mph. (Note: A value below the mean has a negative z-score.)

**e.** Whatvehicle speed is one standard deviation below the mean vehicle speed?

**f.** Complete the following two sentences.

The z-scores for about 68% of the vehicle speeds will be between \_\_\_\_\_\_\_ and \_\_\_\_\_\_\_ .

The z-scores for about 95% of the vehicle speeds will be between \_\_\_\_\_\_\_ and \_\_\_\_\_\_\_ .

**#2** At the course website, access the folder for today’s lab, where I have placed a dataset called introReg\_ex2.R. On a Mac, you can double-click on this .R file directly. If R is not open, it will launch it and produce a line of code, and if it is already open, look in the console and see that it produced a line of code for you with the load() command. On a Windows machine, you must use the load command. First, download the data (say to your Desktop), and then change the working directory of R to the location where you downloaded the data (File<Change dir…<). Then simply use

load("introReg\_ex2.R")

The data now exists in R, and to see what it is called, use the ls() command to get a list of all objects currently in R. The data are from students in a statistics class at UC Davis. Variables are ***Sleep*** = hours of sleep the previous night, ***momheight*** = student’s guess at their mother’s height, and ***exercise*** = student’s self-reported hours of exercise in a typical week.

**a.** Draw a histogram of the ***Sleep*** variable (**with(ex2,hist(Sleep))**). Characterize the shape of the histogram (bell-shaped, symmetric but not bell-shaped, skewed, etc.)

**b.** Now suppose we want some basic summary information about the Sleep variable. Get the mean and standard deviation, and report them below (Reminder: functions mean() and sd() will produce these values).

Mean hours of sleep = standard deviation =

**c.** Assuming that the empirical rule applies, calculate an interval that should include about 95% of the data values for the hours of sleep variable.

**d.** Produce a histogram of the ***momheight*** variable (**with(ex2,hist(momheight))**). Characterize the shape of the histogram, and discuss any other noteworthy features of the data.

**e.** Draw a box plot of ***momheight*** using **with(ex2,boxplot(momheight))**. What noteworthy feature(s) of the data is indicated by the plot?

**f.** Using this boxplot, what are the values of the quartiles and IQR?

Q1 = Q3 = IQR =

**g.** A data value is marked as an outlier in a boxplot either if it is larger than Q3+(1.5×IQR) or smaller than Q1 −(1.5×IQR). For ***momheight***, calculate the two boundaries for marking outliers

Q3 + (1.5×IQR) = Q1 − (1.5×IQR) =

**h.** Refer to part g. Would a **momheight** = 57 inches be marked as an outlier? Why or why not?

**i.** Suppose that we were to delete the two outliers from the ***momheight*** data. For each of the following statistics, briefly explain whether you think that the value of the statistic would change or not.

Standard deviation

Range

Median

**j.** Produce a histogram of the ***exercise*** variable. Characterize the shape of the histogram and explain whether the empirical rule would apply to this variable.

**Regression, Correlation and Scatterplots**

**#3** At the course website, access the folder for today’s lab, where I have placed a dataset called introReg\_ex3.R, which, on a Mac, you can double-click directly to load it into R (with the load command produced in the console). On a Windows machine, use the load command again (if the data are saved in the same location). For a statistics class project at a large northeastern university (Penn State), a student examined the relationship between *x =*  body weight (in pounds) and *y =*  time to chug a 12-ounce beverage (in seconds). The student collected data from 13 individuals.

**a.** Produce a scatterplot of the measurements. The y-variable is “chug time” and the x-variable is weight. In R, we can use:

plot(ex3) ##This works here as a special case because there are only 2 variables

##that happen to be in order: x first, y second.

with(ex3,plot(x=Weight,y=ChugTime)) ## This is more explicit!

with(ex3,plot(x=Weight,y=ChugTime, main='Relationship between Weight and Time to Chug a 12oz. Beverage')) ## How about a nice descriptive title…

with(ex3,plot(x=Weight,y=ChugTime, main='Relationship between Weight and \nTime to Chug a 12oz. Beverage')) ## Wanna split that title into 2 lines??

- Describe the main features of the graph. Specifically, is there a negative or a positive association? Does the pattern look to be linear or curved? Are there any outliers? If there is an outlier, describe where it’s located on the graph.

**b.** In general, outliers should not be thrown out unless there’s a good reason, but there are several reasons why it may be legitimate to conduct an analysis without them. In this case, let’s ignore the data point for the heaviest person and then determine a regression line for the remainder of the data.

ex3\_tmp=subset(ex3,ex3$Weight<max(ex3$Weight))

with(ex3\_tmp,summary(lm(ChugTime~Weight)))

Write the estimated regression equation (look in the Estimates column of the Coefficients table for the intercept and slope, in that order).

**c.** Write a sentence that interprets what this slope says about the relationship between chug time and body weight.

**d.** Use the equation found in part b to estimate the chug time for somebody who weighs 160 pounds.

You might try to write an R function here. For example:

myreg=function(x){

return(m\*x+b) ##you fill in the slope (m) and the intercept (b)

}

myreg(160) ##this calculates the prediction

**e.** A prediction error, also called a residual, is calculated as “actual y-value – predicted y-value.” Suppose that a person (not in the dataset) who weighs 160 pounds can do a chug time of 6.5 seconds. What is the value of the prediction error for this person? Note: You got the predicted chug time in part d.

6.5-myreg(160) ##boom.

**f.** What is the value of the prediction error (residual) for a person who weights 200 pounds and can do a chug time of 5.2 seconds?

#See what I did in (e).

**g.** What is the predicted chug time for a person who weighs 300 pounds? What is (obviously) invalid about this prediction? Note: This part is about the problem caused by “extrapolation,” which is predicting too far beyond the observed range of the data.

**#4** At the course website, access the folder for today’s lab, where I have placed a dataset called introReg\_ex4.R, which, on a Mac, you can double-click directly to load it into R (with the load command produced in the console). On a Windows machine, use the load command again (if the data are saved in the same location). The dataset includes teenage mother birth rates and poverty rates for the 50 states of the U.S. and the District of Columbia. The variable ***PovPct***is the percent of a state’s population in 2000 living in households with incomes below the federally defined poverty level. The variable ***Brth15to17***is the birth rate for females 15 to 17 years old in 2002, calculated as births per 1000 persons in this age group.

**a.** Plot ***Brth15to17***(as y-variable) versus ***PovPct*** (as x-variable).

with(ex4,plot(PovPct,Brth15to17))

Describe the direction of the relationship, comment on whether the pattern appears to be linear or curved, and comment on whether there are any outliers.

**b.** Determine a regression line for these data with ***Brth15to17***as the y-variable and ***PovPct*** as the x-variable (see #3b). Write the equation.

**c.** Write a sentence that interprets what this slope says about the relationship between ***PovPct*** and ***Brth15to17*** .

**d.** The variable ***Brth18to19*** is the rate of giving birth for females in the 18 and 19 year old age group. Plot ***Brth18to19*** versus ***PovPct***.

Describe the direction of the relationship, comment on whether the pattern appears to be linear or curved, and comment on whether the relationship in this plot appears to be weaker, stronger, or about the same strength as the relationship between ***Brth15to17*** and ***PovPct***. Explain.

**e**. The variable ***ViolCrime*** is a measure of the rate of violent crimes. Plot ***ViolCrime*** versus ***PovPct***. Discuss the most obvious feature of this graph. If there’s an outlier, identify the location and the numerical values of the variables involved.

**#5** At the course website, access the folder for today’s lab, where I have placed a dataset called introReg\_ex5.R, which, on a Mac, you can double-click directly to load it into R (with the load command produced in the console). On a Windows machine, use the load command again (if the data are saved in the same location). The data are latitude and temperature data for 20 U.S. cities. ***Latitude*** is the geographic latitude of the city, ***JanTemp*** is the mean January temperature, ***AprTemp*** is the mean April temperature, and ***AugTemp*** is the mean August temperature.

**a.** Create a scatterplot as before to examine the connection between ***AugTemp*** (*y*-variable**)** and ***latitude*** (*x*-variable). Once you have made the plot, add the regression line.

with(ex5,plot(latitude,AugTemp))

abline(with(ex5,lm(AugTemp~latitude)))

Based on the resulting plot, answer the following questions:

Does it look like a straight line is a suitable description of the data, or do the data look to be curved?

Is the correlation between the two variables positive or is it negative? Briefly explain your answer.

Are there any outliers? If so, which city (or cities) is an outlier? Explain why.

**b.** What is the equation of the regression line? Write the equation that relates August temperature to latitude.

summary(with(ex5,lm(AugTemp~latitude)))

**c.** Write a complete sentence that describes how much mean August temperature changes per each one degree increase in latitude. Explain how you know this.

**d.** In the output that you generated for part (b), you obtained a value for “Multiple R-squared:”. What is that number?

**e.** The R2 value is the squared correlation value, written as a percent. It is interpreted as the fraction of the observed variation in y-values that can be explained by the x-variable. Apply that interpretation to this situation by writing a sentence that interprets the R2 value for the relationship between August temperature and latitude. The sentence structure might be something like “ \_\_\_\_ percent of the observed variation in \_\_\_\_ is explained by \_\_\_\_\_\_.”

**f.** Use R to find the correlation between ***AugTemp*** and ***latitude***.

with(ex5,cor(AugTemp,latitude))

with(ex5,cor(latitude,AugTemp)) ##Are these different??

Give the numerical value of the correlation.

Briefly discuss why this value shows that there is a moderately strong negative association between ***AugTemp*** and ***latitude***.

**g.** Square the value of the correlation that you found in part f.

with(ex5,cor(AugTemp,latitude))^2

Then compare the squared value to the R2 value that you reported in part d. How do they compare?

**h.** The equator has latitude = 0. Based on the regression equation found in part b, determine the predicted mean August temperature at the equator.

**i.** Refer to your answer to part h. Explain why this is probably a bad estimate of the mean August temperature at the equator. Hint: What is the interval of observed latitudes in the dataset?

**j.** Create a scatterplot to examine the connection between ***AugTemp*** (*y*-variable**)** and ***AprTemp*** (*x*-variable), and add the regression line as before.

with(ex5,plot(AprTemp,AugTemp))

abline(with(ex5,lm(AugTemp~AprTemp)))

Based on the resulting plot, answer these questions:

Is the correlation between the two variables positive or is it negative? Briefly explain your answer.

Are there any outliers? If so, which city (or cities) is an outlier? Explain why.

**k.** Write a complete sentence that describes how much mean August temperature changes per each one degree increase in April Temperature.

**l.** Find the correlation between ***AugTemp*** and ***AprTemp*** (see part f)***.*** What is the correlation value?

Explain why the correlation value indicates that there is fairly strong positive association between August temperature and April temperature.

**m.** Which variable, latitude or April temperature, is a stronger predictor of August temperatures? Explain why you think this.

**#6** At the course website, access the folder for today’s lab, where I have placed a dataset called introReg\_ex6.R, which, on a Mac, you can double-click directly to load it into R (with the load command produced in the console). On a Windows machine, use the load command again (if the data are saved in the same location). The data were collected in Stat 100 at Penn State and were used in some examples in the text. This activity shows how a third variable (confounding variable) can affect a correlation between two other variables.

**a.** Find the correlation between ***Fastest*** and ***RtSpan***. ***Fastest =*** self-reported fastest speed student has ever driven a car and ***RtSpan*** = student’s stretched right hand span (centimeters). What is the correlation value?

with(ex6,cor(Fastest,RtSpan)) ##What’s going on?? NA??

##We have some missing data, so we have to tell R to only use **com**plete cases

with(ex6,cor(Fastest,RtSpan,use="com")) ##Better.

Explain why the correlation value indicates that there is moderately strong positive association between fastest speed ever driven and stretched right hand span.

**b.** Now let’s visualize this relationship. Make the scatterplot using:

with(ex6,plot(RtSpan,Fastest)) ##Does being taller make you drive faster??

But we know that hand span is probably just a proxy for general body size, and that men tend to have larger bodies than women. So let’s explore the role of gender here. Create this plot and comment on the main feature you notice:

with(ex6,plot(RtSpan,Fastest,col=Sex)) ##Men will be colored red, women will be black

**c.** In this part, let’s separately estimate this correlation for males and females.

ex6M=subset(ex6,ex6$Sex=='M')

ex6F=subset(ex6,ex6$Sex=='F')

with(ex6M,cor(RtSpan,Fastest,use="com"))

with(ex6F,cor(RtSpan,Fastest,use="com"))

Correlation for males only = Correlation for females only =

Explain why these two correlation values indicate that, for each sex, there is a very weak relationship between fastest ever driven and right hand span.

**d.** When we used the Sex variable to color the points on the scatterplot, we could also have created two plots separately (or even put them side-by-side):

par(mfrow=c(1,2))

with(ex6M,plot(RtSpan,Fastest,main='Relationship Between Hand Span and Fastest \nSpeed Ever Driven for Males'))

with(ex6F,plot(RtSpan,Fastest,main='Relationship Between Hand Span and Fastest \nSpeed Ever Driven for Females'))

On the basis of these two graphs, explain how we can see that sex is a confounding variable in this problem.