MIDS W203 Lab 1

Matthew Potts

February 9, 2016

Part 1. Multiple Choice

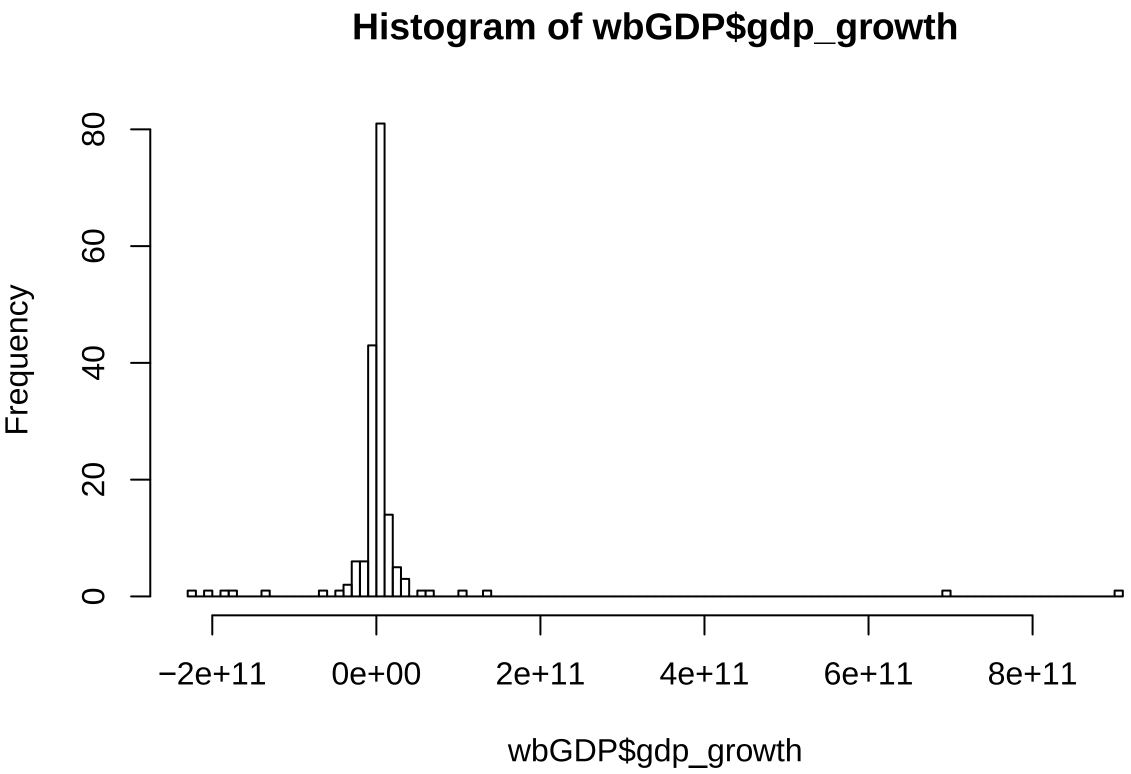
1. E
2. A
3. D
4. B
5. D
6. B
7. C
8. B
9. B

Part 2a. Variable Manipulations

```
file = "/Users/matthewpotts/Dropbox/UC Berkeley/Exploring and Analyzing Data/Lab1/Lab1Spring2016/GDP_Wo

read.csv(file, header = TRUE) -> wbGDP

# Create the gdp_growth variable that is the nominal increase in GDP from 2011 to 2012
wbGDP$gdp_growth <- wbGDP$gdp2012 - wbGDP$gdp2011

# Print the mean of gdp_growth by removing NA values
mean(wbGDP$gdp_growth, na.rm = TRUE)
```

```
## [1] 7172376796
```

```
# Create a histogram of the vaiable gdp_growth
hist(wbGDP$gdp_growth, breaks = 100)
```



This data is not normally distributed but rather positively skewed. There are two outliers in the positive direction that are skewing the data.

```
# Create a new variable that is a logical vector with TRUE equal to countries where
# gdp_growth is greater that the mean of gdp_growth
wbGDP$high_growth <- wbGDP$gdp_growth > mean(wbGDP$gdp_growth, na.rm = TRUE)

# Look at the count of countries above and below the mean
table(wbGDP$high_growth)
```

```
##
## FALSE  TRUE
##   142    31
```

There are 142 countries with GDP growth lower than the mean and 31 countries with GDP growth higher than the mean. This makes sense if the shape of the gdp_growth distribution is positively skewed. With a few positive outliers, the mean is higher, leaving more countries below the mean than in a normally distributed data set (where the distribution above and below the mean would be more even).

Part 2b. Data Import

Cousin Marriage Data

There was a recent story on FiveThirtyEight.com about the prevalence of marriage to cousins in the United States. This is called Consanguinity and is defined as marriages between individuals who are second cousins or closer. The article included data put together in 2001 for a number of countries. The data source and the article are listed below.

http://fivethirtyeight.com/features/how-many-americans-are-married-to-their-cousins/

The raw data behind the story [Dear Mona: How Many Americans Are Married To Their Cousins?] on FiveThirtyEight.com.

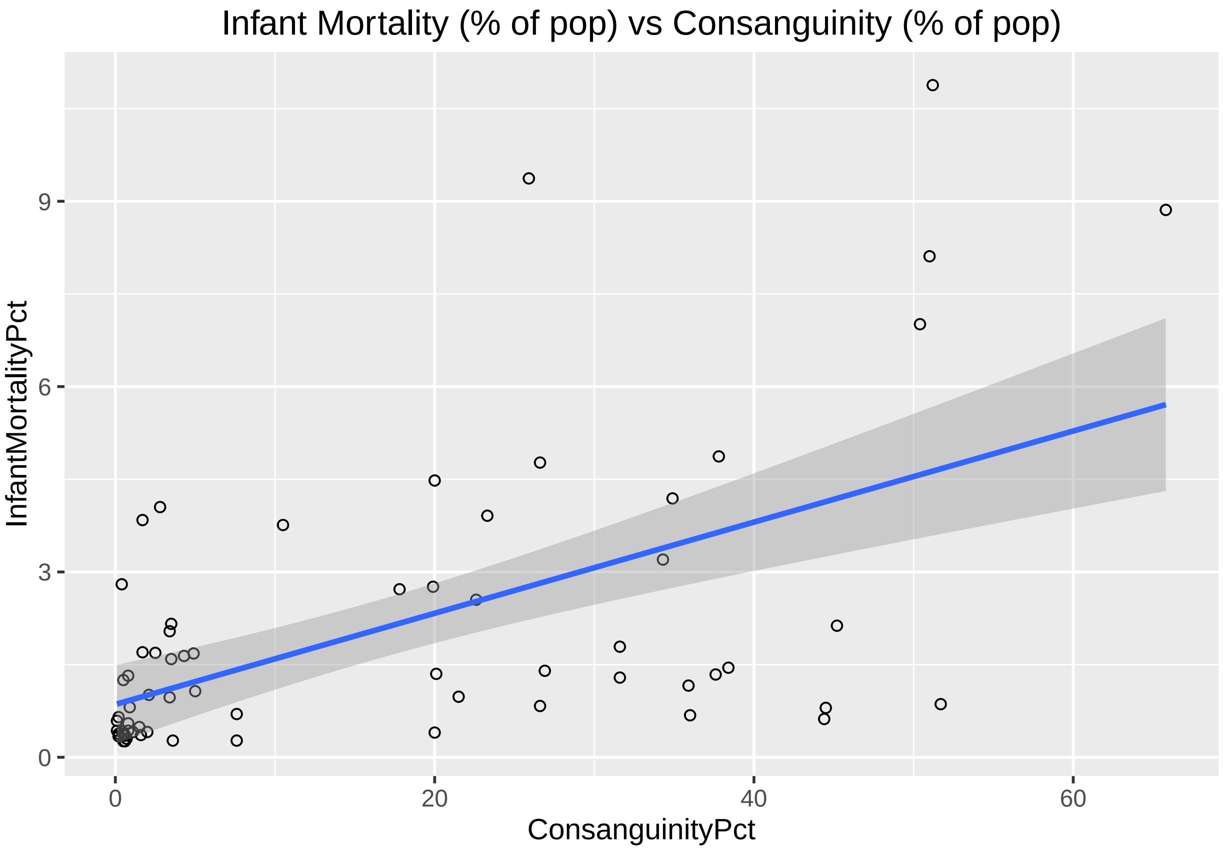
Header | Definition

`country` | Country names

`percent` | Percent of marriages that are consanguineous

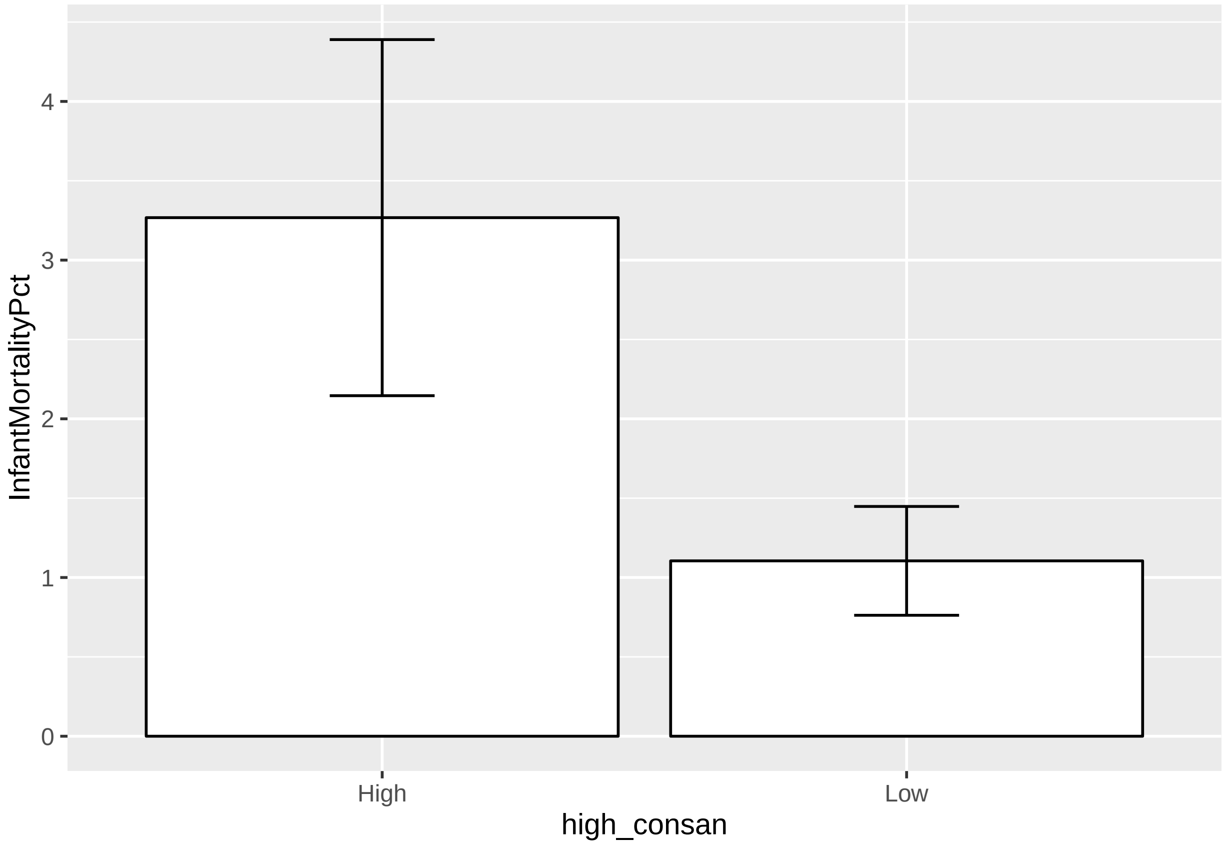
Source: cosang.net

```
# Load cousin marriage data and read to data frame
x <- "https://raw.githubusercontent.com/fivethirtyeight/data/master/cousin-marriage/cousin-marriage-data
consan <- read.csv(x, header = TRUE)

# Do a bit of cleaning
  # Find country name differences
consan$match <- match(consan$Country, wbGDP$Country)

  # Do some manual replacements
consan$Country <- as.character(consan$Country)
consan$Country[consan$Country == "Great Britain"] <- "United Kingdom"
consan$Country[consan$Country == "Kyrgyzstan"] <- "Kyrgyz Republic"
consan$Country[consan$Country == "Syria"] <- "Syrian Arab Republic"
consan$Country[consan$Country == "The Netherlands"] <- "Netherlands"
consan$Country[consan$Country == "Yemen"] <- "Yemen, Rep."

  # Drop the match column in consan
consan$match <- NULL

# Load infant mortality data from the World Bank (Children under 5 and per 1000 individuals)
file <- "/Users/matthewpotts/Downloads/Data_Extract_From_World_Development_Indicators/Data_Extract_From_
mortality <- read.csv(file = file, header = TRUE)

# Merge the World Bank data sets with the Consanguineous data.
merge <- merge(x = wbGDP, y = consan, all.y = TRUE, by = "Country")
merge <- merge(x = merge, y = mortality, all.x = TRUE, by.x = "Country", by.y = "Country.Name")

# Clean data set and create an infant mortality percentage

finaldata <- merge[c(1:7, 22)]
names(finaldata)[8] <- "Mortality"
names(finaldata)[7] <- "ConsanguinityPct"
finaldata$Mortality <- suppressWarnings(as.numeric(levels(finaldata$Mortality))[finaldata$Mortality])
finaldata$InfantMortalityPct <- (finaldata$Mortality / 10)
finaldata <- finaldata[-c(69, 19, 27), ]

# Create a plot of InfantMortality vs. Consanguinity
suppressWarnings(library(ggplot2))
ggplot(data = finaldata, aes(x = ConsanguinityPct, y = InfantMortalityPct)) + geom_point(shape = 1) + s
```

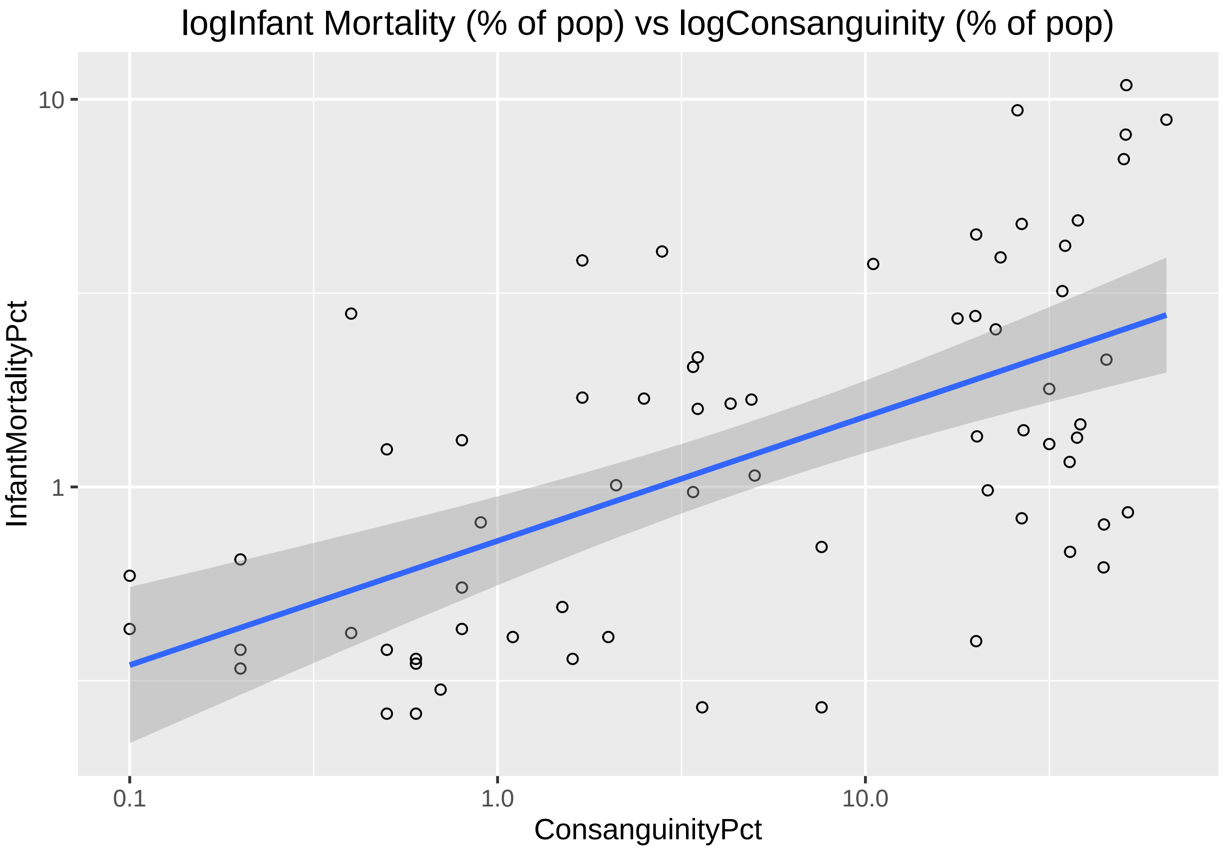


This graph shows consanguinity plotted along with infant mortality. The graph shows a positive correlation between the two variables. However, if you read the article, they come to the conclusion that consanguinity may not be as much of a genetic factor and popular belief. While the graph shows correlation, this may not be the result of causation.

```r
# Investigate effect of high and low consanguinity on infant mortality

finaldata$high_consan <- ifelse(finaldata$ConsanguinityPct > mean(finaldata$ConsanguinityPct,
                                na.rm = TRUE), "High", "Low")
ggplot(data = finaldata, aes(high_consan, InfantMortalityPct)) + stat_summary(fun.y = mean, geom = "bar
```



```
# Plot the log10 of both varables
suppressWarnings(library(ggplot2))
ggplot(data = finaldata, aes(x = ConsanguinityPct, y = InfantMortalityPct)) + geom_point(shape = 1) + s
```



InfantMortality and consanguinity percentages are clustered around 0 - 1. Taking the log10 of both variable spreads out the data but still preserves a clear correlation.