W203: Exploring and Analyzing Data - 1

Lab 1

Juan Jose Carin

Part 1: Multiple Choice

1. **e)** Ordinal
2. **d)** the ecological fallacy
3. **d)** Standard deviations can be directly compared to the individual deviation of one data point away from the mean.
4. **b)** stratified random sampling
5. **b)** Choosing 100 San Franciscans at random and finding that they drink an average of over 52oz of beer a week.
6. **b)** For larger samples, it suggests that the normal distribution is a good model for the distribution of the mean and other statistics.
7. **f)** None of the above.
8. **d)** The distribution of your age variable is platykurtic.
9. **b)** $.01 \leq P(H|A_1) < .02$

Some comments about Part 1:

2. **d)** The toughest question, because I wasn't familiar with the Ecological Fallacy… but all other options are discarded: a variable that considered all the possible genders wouldn't be neither interval nor ordinal but nominal, the male/female variable is not ordinal, by considering only 2 genders you give more priority to social aspects than to psychological ones, etc.

   My reasoning is: in a study in which gender is measured using a strict male/female dichotomy, you wouldn't be able to deduce nothing about an individual from inference for the group to which that individual had been assigned, as that person may feel he or she doesn't belong to that group.

9. **b)**

$$P(H|A_1) = \frac{P(A_1|H) \times P(H)}{P(A_1)} = \frac{1 \times 0.01}{P(A_1)} = \frac{0.01}{P(A_1)}$$

$$P(\bar{H}|A_1) = \frac{P(A_1|\bar{H}) \times P(\bar{H})}{P(A_1)} = \frac{0.5 \times 0.99}{P(A_1)} = \frac{0.495}{P(A_1)}$$

$$P(H|A_1) + P(\bar{H}|A_1) = 1 \Rightarrow P(A_1) = 0.01 + 0.495$$

$$P(\bar{H}|A_1) = \frac{0.01}{0.01 + 0.495} = \frac{0.01}{0.505} = \frac{2}{101} \approx 0.0198$$

Part 2: Data Analysis and Short Answer

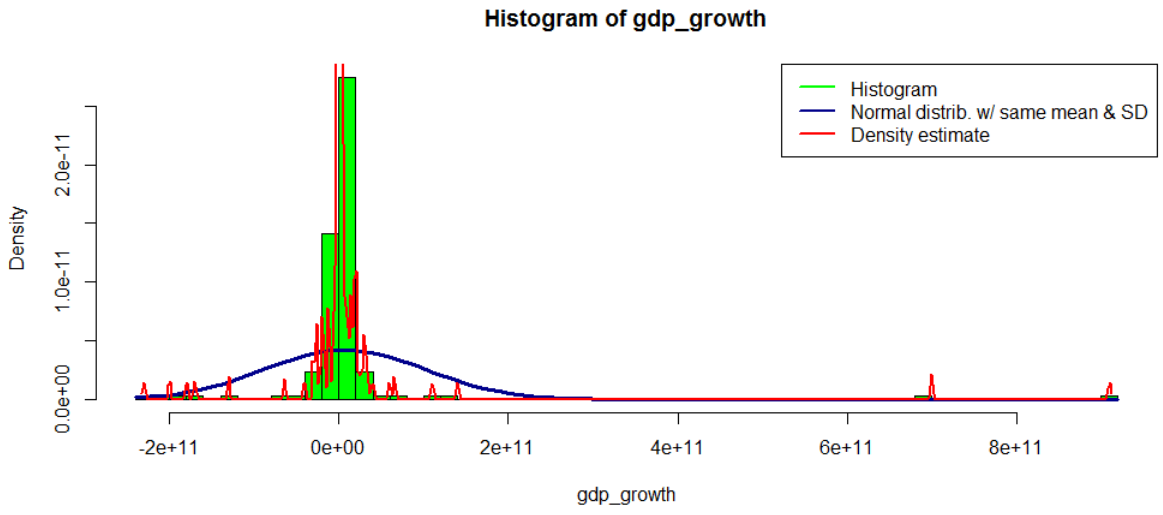
1. Variable Manipulation

a.

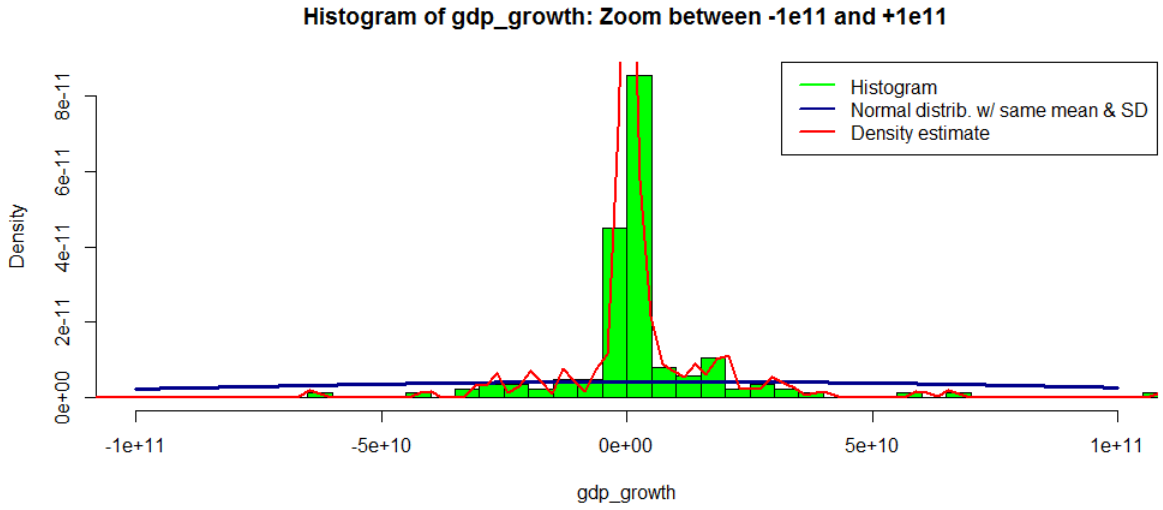
```
#### a) gdp_growth MEAN

GDP_World_Bank<-read.csv("GDP_World_Bank.csv")
GDP_World_Bank$gdp_growth <- GDP_World_Bank$gdp2012 - GDP_World_Bank$gdp2011

# There are missing values of GDP in 2011 or 2012 or both for 39 out of the
  # 212 countries. So the new variable has a numerical value for 173 countries
# We can keep these 39 observations in the original dataframe (and use
  # "na.rm=TRUE" when required by the functions) or create a new variable
  #  without the missing values

print(paste("The mean of the gdp_growth variable is",
            format(gdp_growth_mean, digits=4, scientific=T)), sep=" ")

## [1] "The mean of the gdp_growth variable is 7.172e+09"
```

b.

```
#### b) gdp_growth HISTOGRAM & FIT TO NORMAL

# Now we plot the histogram, comparing it with an estimate of its Probability
  # Density Function and the normal PDF with the same mean and standard deviation
```




```
# Since there are some observations far from the mean, we now zoom the
  # histogram
```



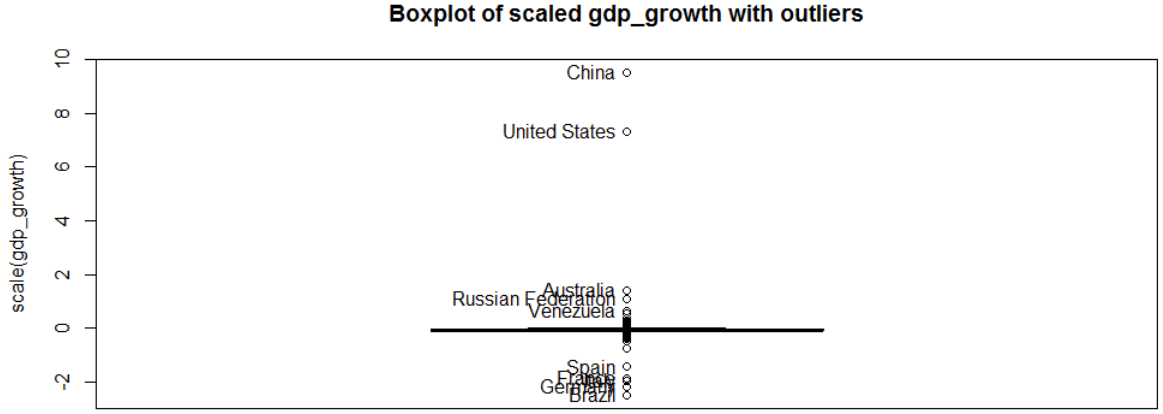

```
# The first histogram shows a long right tail, i.e., the distribution is
  # positively skewed (the corresponding function confirms this)
skewness(gdp_growth, type=2)
```

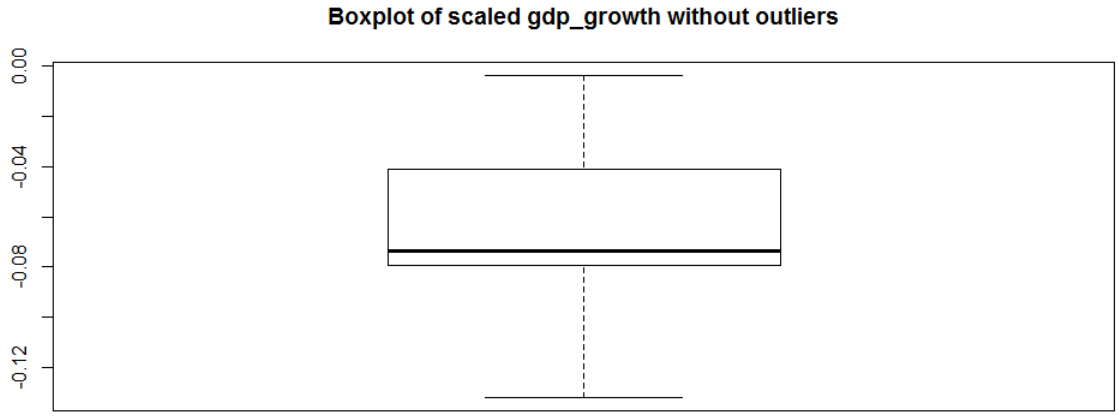
```
## [1] 7.151
```

```
# The comparison against the Probability Density Function of a normal
  # distribution with the same mean and standard deviation shows that the
  # distribution under analysis is clearly leptokurtic (i.e., the kurtosis is
  # positive and high), and hence far from normal
kurtosis(gdp_growth, type=2)
```

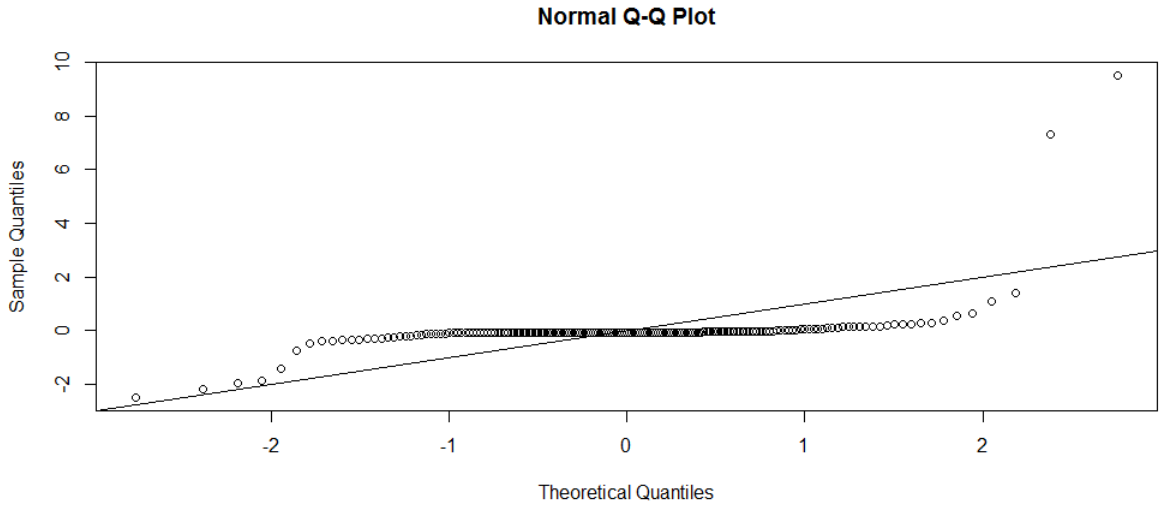
```
## [1] 64.35
```

```
  # The numerous outliers causes the standard deviation to be higher
  # than it would be in the absence of those outliers, so a normal distribution
  # with the same standard deviation (and mean) has a more rounded peak and
  # thinner tails

# Now we display the boxplot (which says the same about normality)
  # With & without outliers
```






```
# Conclusions:
# There are 2 outliers (China and U.S.) almost 8 and 10 standard deviations
  # far from the mean. The probability of 2 observations like these in a sample
  # of size 173 from a normal distribution is infinitesimal (the probability of
  # a single observation more than 3 or 4 standard deviations far from the mean
  # is already close to null.
# The boxplot shows that the median is much lower than the mean...
median(GDP_World_Bank$gdp_growth, na.rm=T)

## [1] 201700000

mean(GDP_World_Bank$gdp_growth, na.rm=T)

## [1] 7.172e+09

  # ... and that's another indication that the data doesn't fit to the normal
  # distribution very well... but nothing can be said now about skewness
  # (though the older notion of skew implied a relationship between the mean
  # the median, the modern definition does not, and for some distributions the
  # relationship we observe doesn't necessarily mean that the skewness is
  # positive)
# The IQR range is very small (in terms of standard deviations): about 0.4 SD
IQR(gdp_growth)/sd(gdp_growth)

## [1] 0.03832

  # while the IQR range of a normal distribution is about 1.5 SD. That's
  # another proof of the normal distribution being a poor fit to our data, and
  # also means that the distribution is leptokurtic

# The Q-Q plot is another indication of non-normality
```




```
# We finally use the Shapiro-Wilkes test
shapiro.test(gdp_growth)$p.value

## [1] 1.341e-25
```

```
  # The p-value is so low that we would have to reject the null hypothesis that
  # the samples came from a normal distribution
```

C.

```
#### c) high_growth & COUNTRIES ABOVE MEAN

high_growth <- gdp_growth > mean(gdp_growth)
# or
# GDP_World_Bank$high_growth <-
#    GDP_World_Bank$gdp_growth > mean(GDP_World_Bank$gdp_growth,na.rm=T)

print(paste(length(high_growth[high_growth==TRUE]),
            "countries have above average growth", sep=" "))

## [1] "31 countries have above average growth"

print(paste(length(high_growth[high_growth==FALSE]),
            "countries have below average growth", sep=" "))

## [1] "142 countries have below average growth"

# As previously explained, the mean was much larger than the median, so a lot
  # more than 50% of the countries have below average growth
  # Actually, less than 18% of the countries have above average growth
(prob_below_avg <- 1-table(high_growth)[2]/(length(high_growth)))

##    TRUE
## 0.8208
```

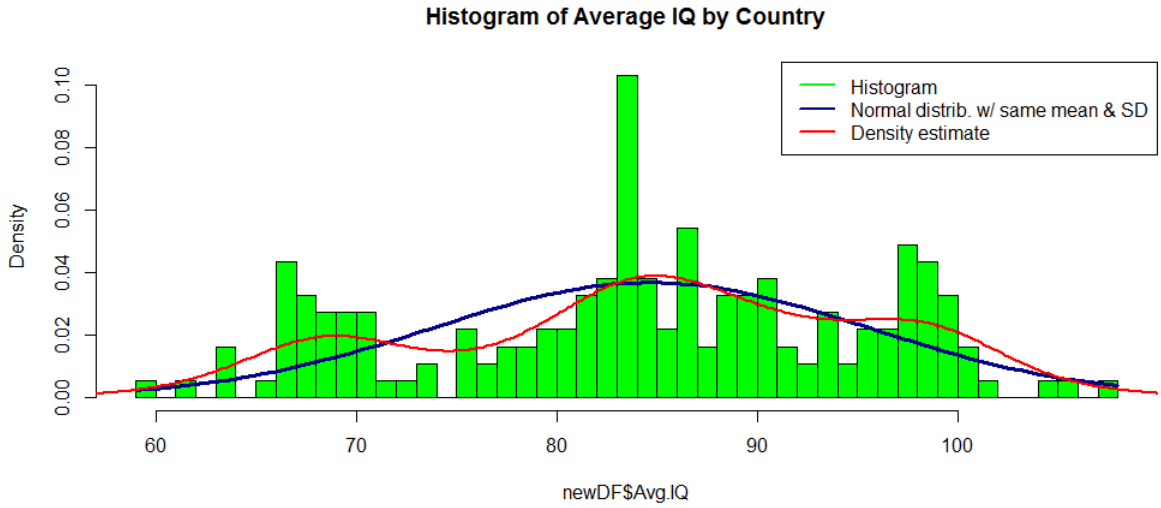
```
  # That is coherent with what the histogram showed: a long right tail,
  # causing the distribution to be positively skewed
```

---

2. Variable Manipulation

a.

```
#### a) New metric country-level variable

# I found a very powerful source (needless to say, GOOGLE:
  # http://www.google.com/publicdata), which not only gathers data from
  # multiple sources but also allows to plot them in multiple ways.
# My first thought was considering one of the sources mentioned there: the
  # International Monetary Fund (IMF):
  # http://www.google.com/publicdata/explore?ds=k3s92bru78li6_
  # http://www.imf.org/external/pubs/ft/weo/2014/01/weodata/download.aspx
  # http://www.imf.org/external/pubs/ft/weo/2014/01/weodata/WEOApr2014all.xls
  # Where you can download information by countries about several measures of
  # GDP, Purchasing-Power-Parity (PPP), Inflation, Volume of Imports & Exports
  # of Goods & Services, Unemployment Rate, General Government Revenue, etc.
  # (from 1980 to 2011 or 2013 depending on the country, and estimates until
  # 2019)
# But most of those variables are  related to GDP, so it would have been a
  # similar exercise

# As we've previously tried to fit a sample distribution to normality, I thought
  # of possible variables (on a country level) that may fit to that distribution
# Intelligence Quotient (IQ) is actually designed to fit to a normal
  # distribution (http://en.wikipedia.org/wiki/Intelligence_quotient), where the
  # mean is 100 points, and the standard deviation is 15 points
# And found what I was looking for in this other source (IBM):
  # http://www-958.ibm.com/software/analytics/manyeyes/datasets
  # http://www-958.ibm.com/software/analytics/manyeyes/datasets/national-iq-scores-country-ranking/versions/1.txt
# Since the number of Countries slightly varied from the ones mentioned in
  # "GDP_World_Bank.csv", I modified the tab-delimited text file (and used that
  # modified version instead)

IQ <- read.delim("new1.txt",header = T, sep = "\t")
newDF <- merge(GDP_World_Bank, IQ, by="Country", all=T)

hist(newDF$Avg.IQ, breaks=50, freq=F, col="green",
     main="Histogram of Average IQ by Country")
curve(dnorm(x, mean=mean(newDF$Avg.IQ, na.rm=T), sd=sd(newDF$Avg.IQ, na.rm=T)),
      add=TRUE, col="darkblue", lwd=3)
lines(density(newDF$Avg.IQ, na.rm=T), col="red", lwd=2)
```




```
# Regardless of whether the IQ variable is normal, we are working with sample
  # means. Therefore, by the Central Limit Theorem, the distribution of the
  # sample mean should approach the normal distribution... and that's not the
  # case. We just have to look at the histogram to see that the distribution is
  # multimodal, and far from normal
# For example, we can observe the following facts:
  # The mean of the sampling distribution is 84.7, 1 standard deviation -of the
  # the population, according to IQ test design, not of the sampling
  # distribution- below the expected mean of the population (100 points). That
  # contradicts the Law of Large Numbers.
  # And 25 out of the 184 countries have an average IQ 2 standard deviations (30
  # points) below the expected mean. That's the 13.6% of the samples, when the
  # expected probability in a normal distribution would be less than 2.5%
# Was sir Francis Galton wrong? Probably. But there are 3 main PROBLEMS here:
# 1. IQ tests are not 100% infallible. That leads to effects like
  # http://en.wikipedia.org/wiki/Flynn_effect, or the fact that there are
  # huge differences between countries (some people argue that IQ tests are not
  # perfectly adapted to each country)
# 2. We don't have evidence that the samples were purely random or well selected
  # in each country
# 3. Each sample was drawn from completely different groups: we should not make
  # conclusions about individuals between groups (Ecological Fallacy). Moreover,
  # the tool (the test) used to measure the variable was not always the same
# CONCLUSIONS:
# The IQ tests do not seem to be a valid tool to measure intelligence (based on
  # which countries have a lowest average IQ, all of them in Africa, it seems
  # it measures culture & education rather than intelligence)
# OR the assumption that IQ is normally distributed is wrong (maybe it is
  # within a country -we can't know it with just the aveage- ...but probably not
  #  with a mean of 100 points in many cases)
# AND, unless the same tool is used to measure a variable, we're working with
  # samples of different variables
# AND FINALLY we should know if the samples have been drawn appropriately
  # within each country
```