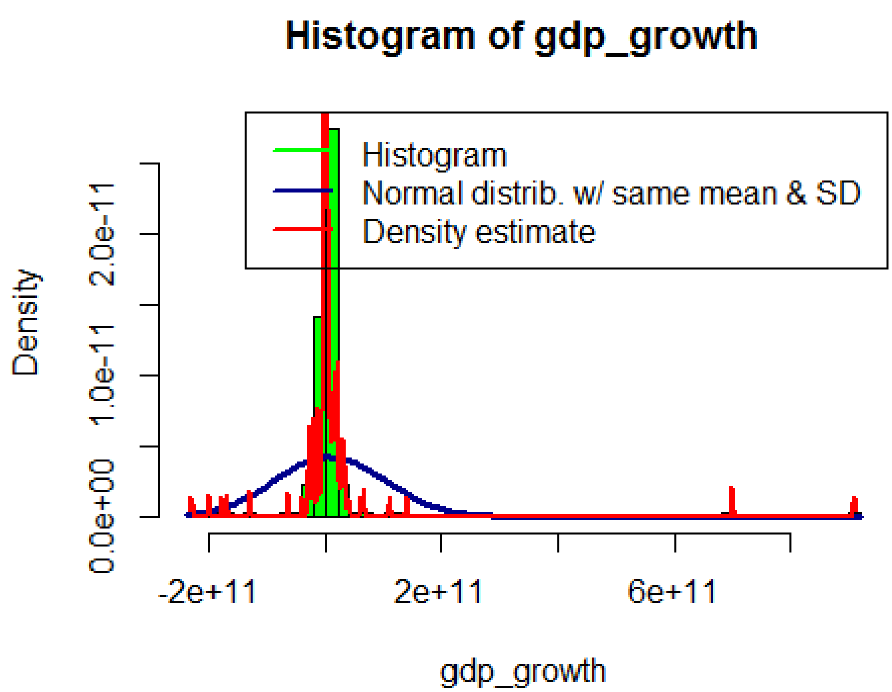
Lab1_JuanjoCarin.R

JuanJose

Thu Sep 25 18:42:13 2014

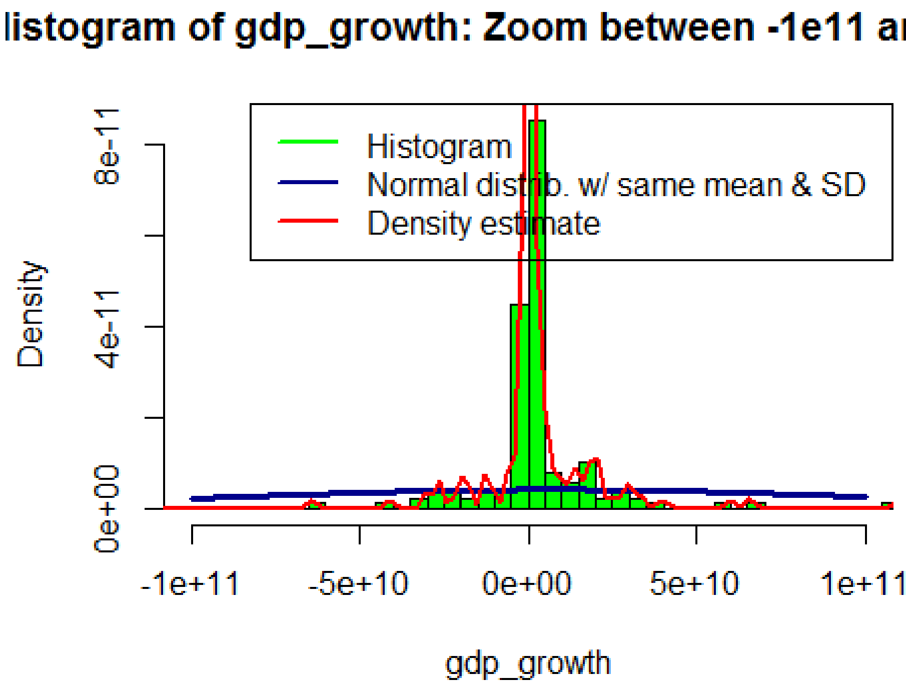
```
#### INIT ####

rm(list = ls())
route <- "C:/Users/JuanJose/Google Drive/ANL/MASTERS/Berkeley/"
route <- paste(route, "DATASCI W203 Exploring and Analyzing Data/Labs/Lab1",
               sep="")
# Change the route if needed
setwd(route)


#### LIBRARIES ####

library(car)

## Warning: package 'car' was built under R version 3.1.1

library(e1071)
library(XLConnect)

## XLConnect 0.2-7 by Mirai Solutions GmbH
## http://www.mirai-solutions.com ,
## http://miraisolutions.wordpress.com

#### 1. VARIABLE MANIPULATION ####

#### 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

nrow(GDP_World_Bank)

## [1] 212

sum(!complete.cases(GDP_World_Bank$gdp_growth))

## [1] 39

sum(is.na(GDP_World_Bank$gdp_growth)) # Another way of counting NA observations

## [1] 39

# There are missing valures 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
gdp_growth <- na.omit(GDP_World_Bank$gdp_growth)
names(gdp_growth) <- GDP_World_Bank$Country[!is.na(GDP_World_Bank$gdp_growth)]
```

```r
gdp_growth_mean <- mean(GDP_World_Bank$gdp_growth, na.rm=TRUE)
mean(gdp_growth) # Another possible way of obtaining the mean

## [1] 7.172e+09

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) 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
hist(gdp_growth, breaks=50, freq=F, col="green")
curve(dnorm(x, mean=mean(gdp_growth), sd=sd(gdp_growth)),
      add=TRUE, col="darkblue", lwd=3)
lines(density(gdp_growth), col="red", lwd=2)
legend("topright", legend=c("Histogram", "Normal distrib. w/ same mean & SD",
                            "Density estimate"),
       col=c("green","darkblue","red"),lwd=2)
```



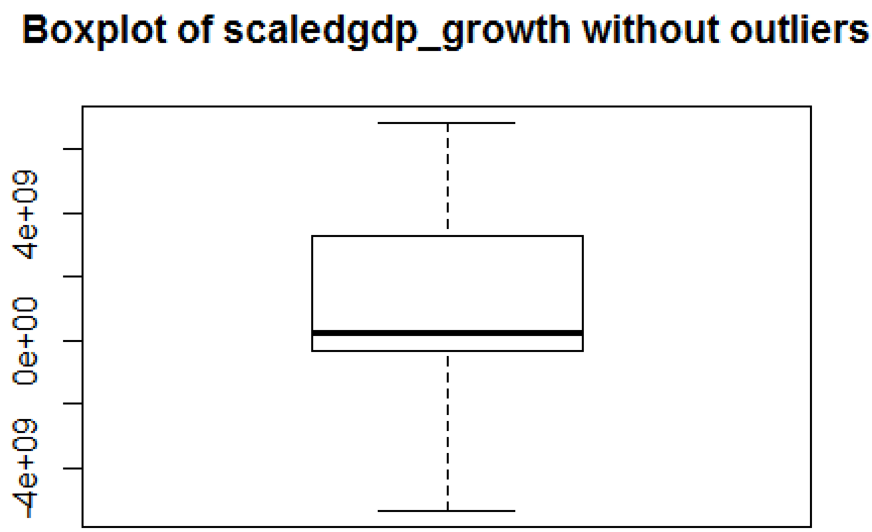

```r
# Since there are some observations far from the mean, we now zoom the
  # histogram
hist(gdp_growth, breaks=200, freq=F, col="green", xlim=c(-1e11,1e11),
     main="Histogram of gdp_growth: Zoom between -1e11 and +1e11")
curve(dnorm(x, mean=mean(gdp_growth), sd=sd(gdp_growth)),
      add=TRUE, col="darkblue", lwd=3)
lines(density(gdp_growth), col="red", lwd=2)
```

```r
legend("topright", legend=c("Histogram", "Normal distrib. w/ same mean & SD",
                            "Density estimate"),
       col=c("green","darkblue","red"),lwd=2)
```

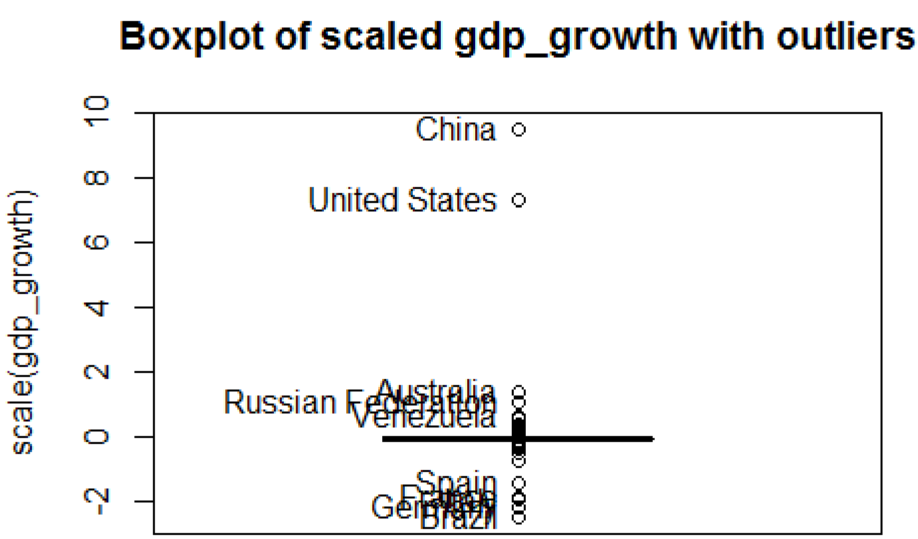



```r
# The first histogram shows a long right tail, i.e., the distribution is
  # positively skewed (the corresponding function confirms this)
skewness(gdp_growth, type=2)

## [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
Boxplot(gdp_growth,labels=names(gdp_growth),id.n=5,
        main="Boxplot of gdp_growth with outliers")
```

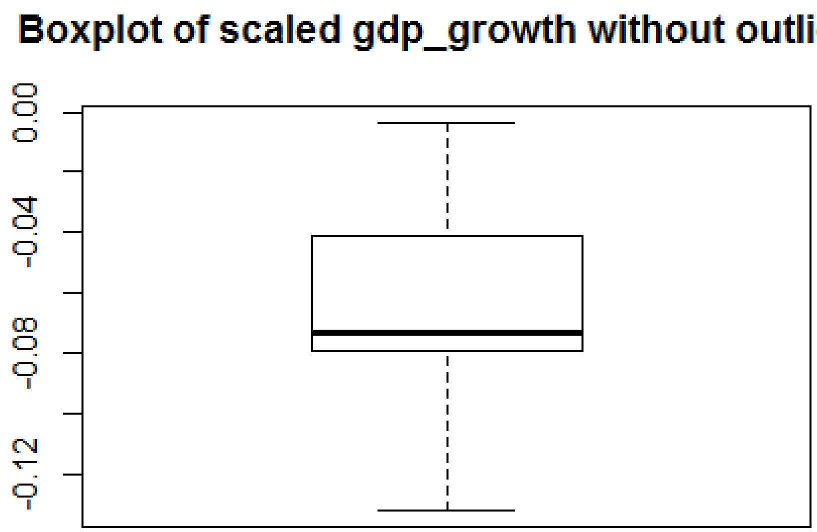

```
##  [1] "Brazil"             "Germany"            "Italy"
##  [4] "France"             "Spain"              "China"
##  [7] "United States"      "Australia"          "Russian Federation"
## [10] "Venezuela"

boxplot(gdp_growth, outline=F,
        main="Boxplot of scaledgdp_growth without outliers")
```




```r
  # With & without outliers, and scaling the variable
Boxplot(scale(gdp_growth),labels=names(gdp_growth),id.n=5,
        main="Boxplot of scaled gdp_growth with outliers")
```

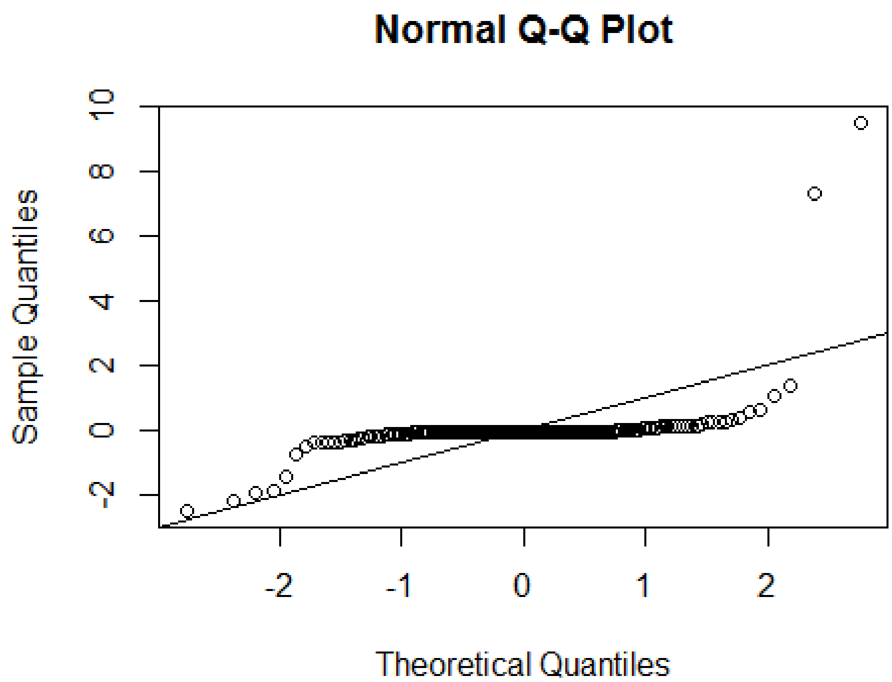


```
##  [1] "Brazil"             "Germany"            "Italy"
##  [4] "France"             "Spain"              "China"
##  [7] "United States"      "Australia"          "Russian Federation"
## [10] "Venezuela"

boxplot(scale(gdp_growth), outline=F,
        main="Boxplot of scaled gdp_growth 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)
```

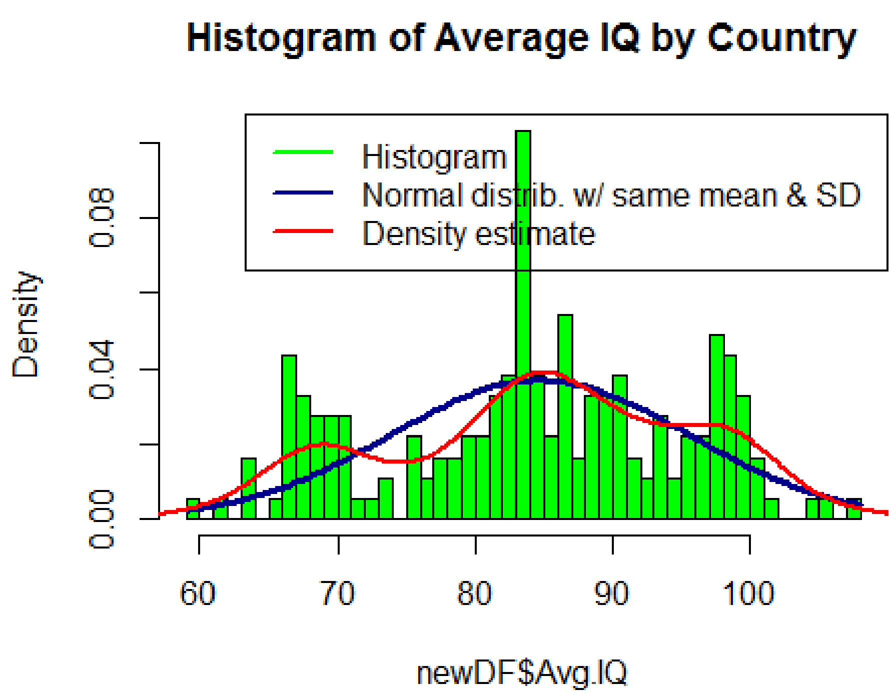
```
## [1] 0.03832

  # 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
qqnorm(scale(gdp_growth))
abline(0,1)
```




```
# 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) 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"
```

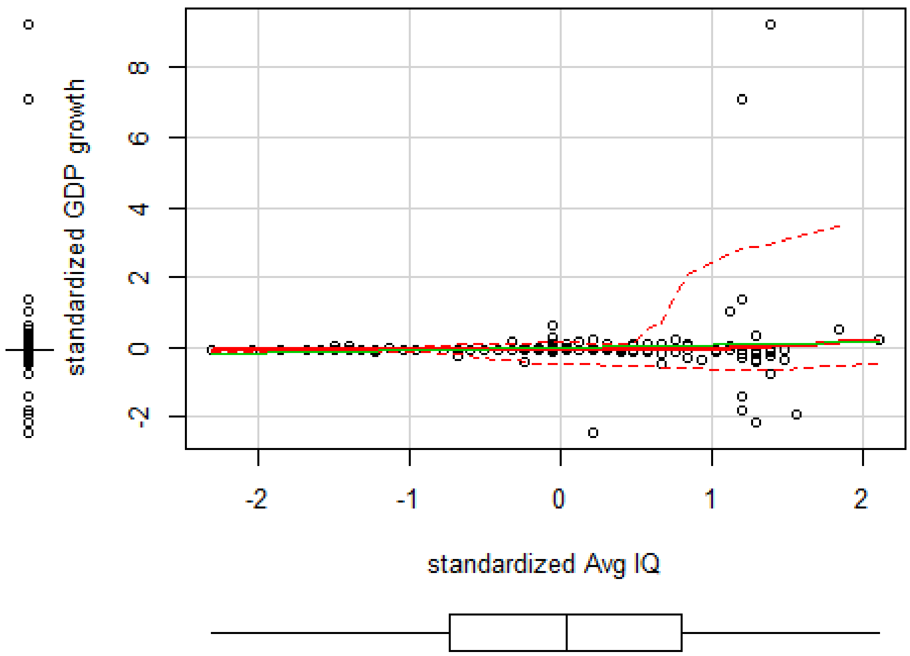
```r
print(paste(length(high_growth[high_growth==FALSE]),
            "countries have below average growth", sep=" "))

## [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

as.numeric(format(quantile(gdp_growth,probs=prob_below_avg),
                  digits=4, scientific=T))

## [1] 7.221e+09

as.numeric(format(mean(gdp_growth), digits=4, scientific=T))

## [1] 7.172e+09

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

#### 2. DATA IMPORT ####

#### 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)
nrow(newDF[!is.na(newDF$Avg.IQ), ])

## [1] 184

mean(newDF[!is.na(newDF$Avg.IQ), ]$Avg.IQ)

## [1] 84.7

sum(newDF[!is.na(newDF$Avg.IQ), ]$Avg.IQ < 70)

## [1] 25

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)
legend("topright", legend=c("Histogram", "Normal distrib. w/ same mean & SD",
                            "Density estimate"),
       col=c("green","darkblue","red"),lwd=2)
```




```
# Regardless of whether the IQ variable is normal, we are working with samples of
  # 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
```

```r
  # 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

scatterplot(as.vector(scale(newDF$Avg.IQ[complete.cases(newDF)])),
            as.vector(scale(newDF$gdp_growth[complete.cases(newDF)])),
            xlab="standardized Avg IQ", ylab="standardized GDP growth")
```



standardized GDP growth
-2
0
2
4
6
8
-2
-1
0
1
2
standardized Avg IQ