Homework Week 1

Nikhil Guruji

October 11, 2017

Q1. Review the basics of summation notation and covariance formulas. Show that:

a. $\sum_{i=1}^{N}(Y_i - \overline{Y}) = 0$

b. $\sum_{i=1}^{N}(X_i - \overline{X})(Y_i - \overline{Y}) = \sum_{i=1}^{N}(X_i - \overline{X})Y_i$

Answer:

a. $\sum_{i=1}^{N}(Y_i - \overline{Y})$

$$= (\sum_{i=1}^{N}(Y_i)) - (N\overline{Y})$$

$$= (Y_0 + Y_1 + ... + Y_N) - N\overline{Y}$$

$$= N\overline{Y} - N\overline{Y}$$

$$= 0$$

b. $\sum_{i=1}^{N}(X_i - \overline{X})(Y_i - \overline{Y})$

$$= \sum_{i=1}^{N}(X_iY_i) - \sum_{i=1}^{N}(\overline{X}Y_i) - \sum_{i=1}^{N}(X_i\overline{Y}) + N\overline{XY}$$

$$= \sum_{i=1}^{N}(X_iY_i) - N\overline{XY} - N\overline{XY} + N\overline{XY}$$

$$= \sum_{i=1}^{N}(X_iY_i) - N\overline{XY}$$

$$= \sum_{i=1}^{N}(X_iY_i) - \overline{X}(N\overline{Y})$$

$= \sum_{i=1}^{N}(X_iY_i) - \overline{X}(\sum_{i=1}^{N} Y_i)$ (from part a.)

$$= \sum_{i=1}^{N}(X_i - \overline{X})Y_i$$

Q2. Define both (and explain the difference between) the expectation of a random variable and the sample average?

Answer:

The expectation of a random variable is the expected value (or an estimate). For example, if we want to estimate the height of a person in a classroom, our best estimate is called the expectation of the random variable (in this case, height). The sample average is the sum of all values divided by the number of values. So, essentially, sample average can be derived mathematically but the expectation is just an estimate.

The best possible estimate will, in most cases, be equal to the sample average value if no other information is provided.

Q3. Review the normal distribution and the mean and variance of a linear combination of two normally distributed random variables. Let $X \sim \N(1,2)$ and $Y \sim \N(2,3)$. Note that the second parameter is variance. X and Y are independent. Compute:

a. $E(3X)$

b. $Var(3X)$

c. $Var(2X - 2Y)$ and $Var(2X + 2Y)$

d. Explain why in part (c) you get the same answer no matter whether you add or subtract. (Your answer should discuss both the coefficient on Y and why independence between X and Y is important.)

Answer:

a. $E(3X) = 3E(X) = 3.1 = 3$

b. $Var(3X) = 3^2.Var(X) = 3^2.2$

= 18

c. $Var(2X - 2Y) = 2^2.Var(X) + 2^2.Var(Y) - 2.2.2.cov(X,Y)$

$= 2^2.2 + 2^2.3$

= 20

$Var(2X + 2Y) = 2^2.Var(X) + 2^2.Var(Y) + 2.2.2.cov(X,Y)$

= 20

d. The reason why we get the same answers for both $Var(2X - 2Y)$ and $Var(2X + 2Y)$ is because the coefficients of X and Y don't change for both cases, and because the covariance of X and Y is zero (as they are independent).

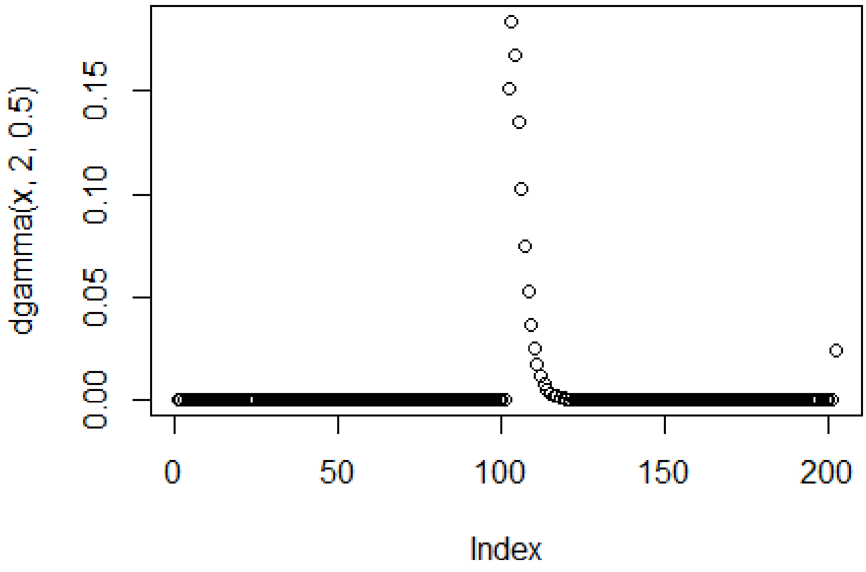
Q4. a. Describe the Central Limit Theorem as simply as you can. b. Let $X \sim \text{Gamma}(\alpha = 2,\ \beta = 2)$. For the Gamma distribution, α is often called the "shape" parameter, β is often called the "scale" parameter, and the $\mathbb{E}[X] = \alpha\beta$. Plot the density of X. You may find the functions `dgamma()` or `curve()` to be helpful. c. Let n be the number of draws from that distribution in one sample and r be the number of times we repeat the process of sampling from that distribution. Draw an iid sample of size $n = 10$ from the Gamma(2,2) distribution and calculate the sample average; call this $\overline{X}_n^{(1)}$. Repeat this process r times where $r = 1000$ so that you have $\overline{X}_n^{(1)}, \dots, \overline{X}_n^{(r)}$. Plot a histogram of these r values and describe what you see. This is the sampling distribution of $\overline{X}_{(n)}$. d. Repeat part (c) but with $n = 100$. Be sure to produce and describe the histogram. e. Let's say you were given a dataset for 2,000 people with 2 variables: each person's height and weight. What are the values for n and r in this "real world" example?

Answer:

a. The central limit theorem states that when the size of the sample is large, the sample average approximately equals the mean of the population from which the sample is chosen.

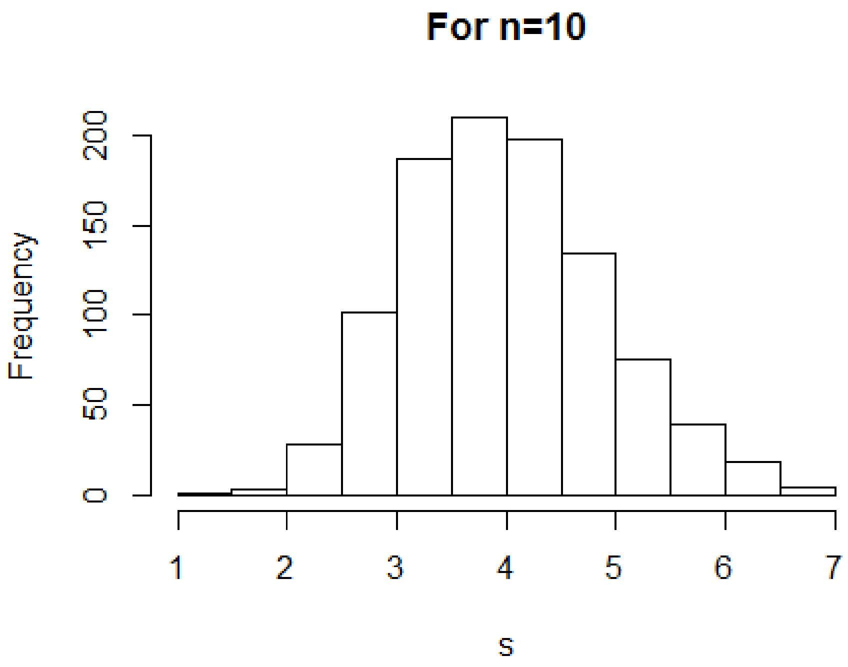
For distributions, as the size of the population increases, the distribution tends towards a normal distribution.

```
x<- c(-100:100, by=0.1) #choosing a vector for input to the dgamma function
plot(dgamma(x,2,0.5)) #where 2 is the shape of the gamma distribution and 0.5 is the rate (ie 1/scale)
```

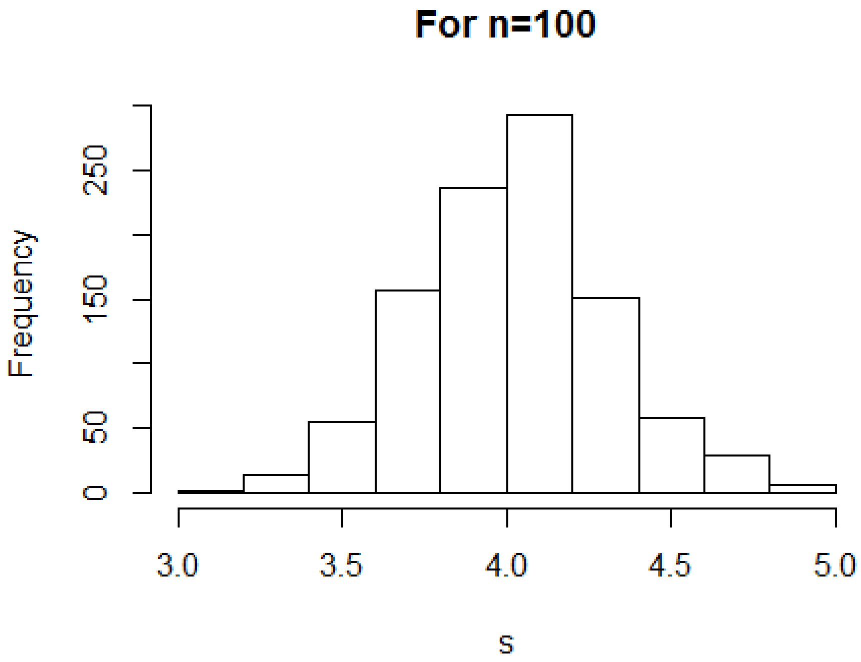


```
#code to output a histogram with n=10
p<-list(mode="vector",length=1000) #initializing a list to store the "r" vectors containing n=10 values
s<-0 #initializing a vector to store the means of all r vectors
```

```r
for(i in 1:1000){
  p[[i]]<-rgamma(10,2,0.5) #n=10, alpha=2, beta=2
  s[i]<-mean(p[[i]])
}
hist(s, main="For n=10")
```



```r
#code to output a histogram with n=100
p<-list(mode="vector",length=1000) #initializing a list to store the "r" vectors containing n=100 values
s<-0 #initializing a vector to store the means of all r vectors
for(i in 1:1000){
  p[[i]]<-rgamma(100,2,0.5) #n=100, alpha=2, beta=2
  s[i]<-mean(p[[i]])
}
hist(s, main="For n=100")
```



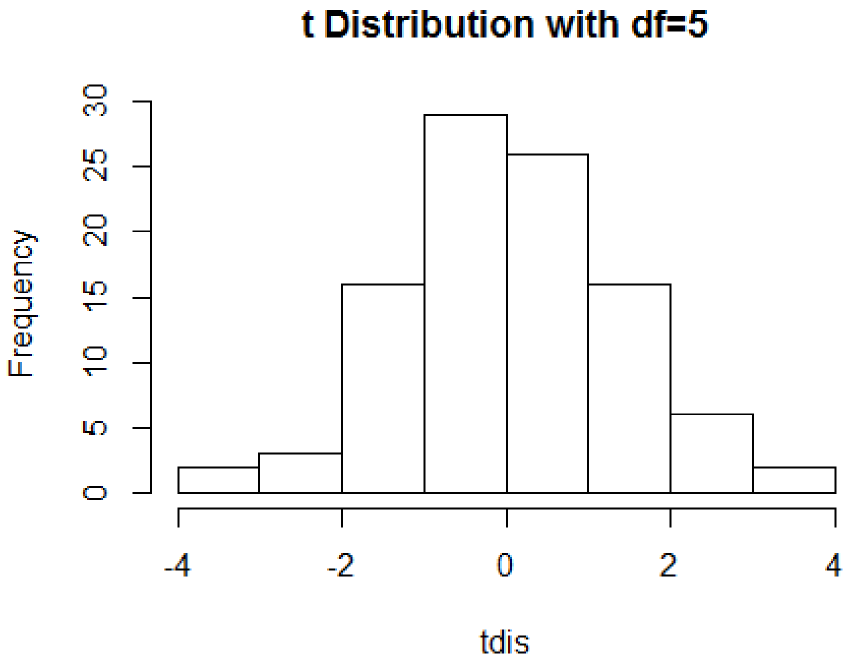
As seen from the two histograms for n=10 and n=100, we can notice that as n increases, the histogram starts showing a normal distribution instead of a gamma distribution.

e. For this case, n=1000, r=2 because we have 2 types of distributions and 1000 random values for each distribution.

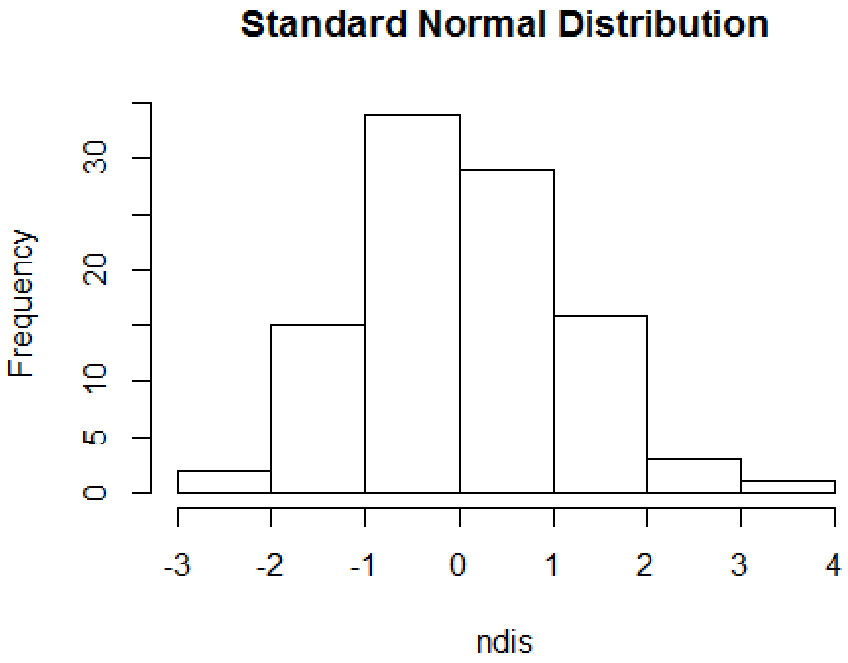
Q5. The normal distribution is often said to have "thin tails" relative to other distributions like the *t*-distribution. Use random number generation in R to illustrate that a $\N(0,1)$ distribution has much thinner tails than a *t*-distribution with 5 degrees of freedom. (Note that `rnorm()` and `rt()` are the functions in R to draw from a normal distribution and a *t*-distribution.)

Answer:

```r
tdis<-rt(100,5) #t distribution with degrees of freedom = 5
ndis<-rnorm(100,0,1) #normal distribution with mean=0, sd=1
hist(tdis, main="t Distribution with df=5")
```



```
hist(ndis, main="Standard Normal Distribution")
```



Q6. a. From the Vanguard dataset, compute the standard error of the mean for the VFIAX index fund return. b. For this fund, the mean and the standard error of the mean are almost exactly the same. Why is this a problem for a financial analyst who wants to assess the performance of this fund? c. Calculate the size of the sample which would be required to reduce the standard error of the mean to 1/10th of the size of the mean return.

Answer:

```
library(devtools)
library(ggplot2)


install_bitbucket("perossichi/DataAnalytics")

## Downloading bitbucket repo perossichi/DataAnalytics@master

## Installing DataAnalytics

## "C:/PROGRA~1/R/R-33~1.3/bin/i386/R" --no-site-file --no-environ  \
##   --no-save --no-restore --quiet CMD INSTALL  \
## 
"C:/Users/NIKHIL/AppData/Local/Temp/RtmpGugwG8/devtoolsf205bf16834/perossichi
-dataanalytics-a2ca2153b6ff"  \
##   --library="C:/Users/NIKHIL/Documents/R/win-library/3.3"  \
##   --install-tests

## 

library(DataAnalytics)

#
# reshape Vanguard data using reshape2
#
# or use the reshape package
#
library(reshape2)
data(Vanguard)
Van=Vanguard[,c(1,2,5)]   # grab relevant cols
V_reshaped=dcast(Van,date~ticker,value.var="mret")

#
# now let's plot mean std deviation.
#
mat=descStat(V_reshaped)

##         Mean Median    SD   IQR SE Mean 95% CI-L 95% CI-U NMissing
## VEIPX 0.009  0.012 0.037 0.043   0.002    0.005    0.013       46
## VFIAX 0.004  0.011 0.045 0.050   0.004   -0.003    0.011      198
## VGENX 0.012  0.012 0.060 0.073   0.003    0.005    0.018        0
## VGHCX 0.014  0.016 0.041 0.046   0.002    0.010    0.018        0
## VMGRX 0.010  0.013 0.072 0.073   0.005    0.000    0.021      163
## VQNPX 0.009  0.014 0.045 0.055   0.003    0.004    0.014      31
## VSMAX 0.009  0.017 0.059 0.072   0.005    0.000    0.018     198
## VTSAX 0.005  0.012 0.046 0.053   0.004   -0.003    0.012     198
## VWNFX 0.009  0.014 0.043 0.048   0.002    0.005    0.014      13
## Number of Observations =  349
```

As seen from the table above, standard error of the mean is 0.004 for VFIAX.

b. This is because the confidence interval is too large to make accurate predictions.

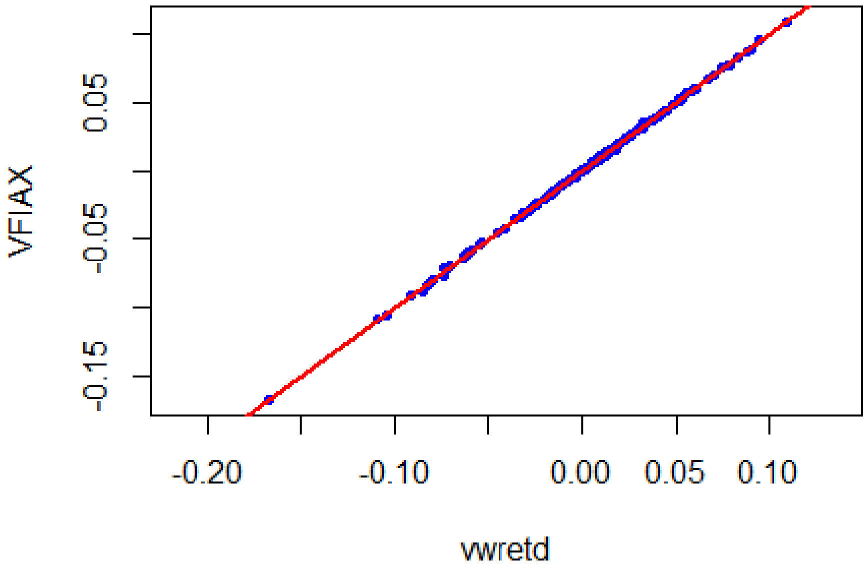
c. To reduce the standard error by n, we need to increase the sample size by n^2. So, here, we need to increase the sample size by 100.

Q7. a. Plot the `VFIAX` index fund return against the `ewretd` (equal-weighted market return) and add the fitted regression line to the plot. You might find the function `abline()` to be helpful. b. Provide the regression output using the `lmSumm()` function from the `DataAnalytics` package.

Answer:

```
data(marketRf)
Van_mkt=merge(V_reshaped,marketRf,by="date")
with(Van_mkt,
     plot(vwretd,VFIAX,pch=20,col="blue")
)

out=lm(VFIAX~vwretd,data=Van_mkt)
abline(out$coef,col="red",lwd=2)
with(Van_mkt,
     points(mean(vwretd),mean(VFIAX),pch=20,cex=2,col="green")
)
```



```
library(DataAnalytics)
lmSumm(out)

## Multiple Regression Analysis:
##     2 regressors(including intercept) and 151 observations
##
## lm(formula = VFIAX ~ vwretd, data = Van_mkt)
##
## Coefficients:
##                Estimate Std Error t value p value
## (Intercept) -0.0001314 6.475e-05   -2.03   0.044
## vwretd       1.0040000 1.440e-03  696.94   0.000
## ---
## Standard Error of the Regression:  0.0007924
## Multiple R-squared:  1  Adjusted R-squared:  1
## Overall F stat: 485730.9 on 1 and 149 DF, pvalue= 0
```