Problem Set 1
MFE 402: Econometrics
Professor Rossi
Student: Sean Xiahao Wang (305229864)

This is designed to review material on summation, covariance, and the normal distribution.

Question 1

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

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

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

Answer:

a.
$$\sum_{i=1}^{N} Y_i - \sum_{i=1}^{N} \bar{Y} = \frac{N}{N}\sum_{i=1}^{N} Y_i - \sum_{i=1}^{N} \bar{Y} = N\bar{Y} - N\bar{Y} = 0$$

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

Using the proof from a: $\sum_{i=1}^{N}(X_i - \bar{X}) = 0$

Hence:
$$\sum_{i=1}^{N}(X_i - \bar{X})(Y_i - \bar{Y}) = \sum_{i=1}^{N} X_iY_i - \bar{X}Y_i = \sum_{i=1}^{N}(X_i - \bar{X})Y_i$$

Question 2

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

a). Expectation of a random variable is the theoretical mean of the distribution or is the long-run average value of repetitions of the experiments it represents.

b). Sample Average Sample average is the mean from a group of observations in a large population.

The key difference is that sample mean is based on observations and represents the estimate for the mean of the popluation while expectation is the theoretical mean for that distribution. Sample mean will deviate from the expectation of the random variable by a standard error. The standard error will decrease as the sample size increases.

Question 3

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

a. $\mathbb{E}[3X]$
b. $\text{Var}(3X)$
c. $\text{Var}(2X - 2Y)$ and $\text{Var}(2X + 2Y)$
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. $E[3X] = 3E[X] = 3 \times 1 = 3$

b. $Var(3X) = 3^2 Var(X) = 9Var(X) = 9 \times 2 = 18$

c. Since X and Y are independent, $Cov(X, Y) = 0$

$$Var(2X - 2Y) = 2^2 Var(X - Y) = 4Var(X - Y) = 4(Var(X) + Var(Y)) = 4 \times (2 + 3) = 20$$

$$Var(2X + 2Y) = 2^2 Var(X + Y) = 4Var(X + Y) = 4 \times (2 + 3) = 20$$

d. Since:

$$Var(2X - 2Y) = 2^2 Var(X) + (-2)^2 Var(Y) - 8Cov(X, Y)$$

$$Cov(X, Y) = 0$$

Since X,Y are independent.

$Var(2X + 2Y) = 2^2 Var(X) + 2^2 Var(Y) = Var(2X - 2Y)$

Question 4

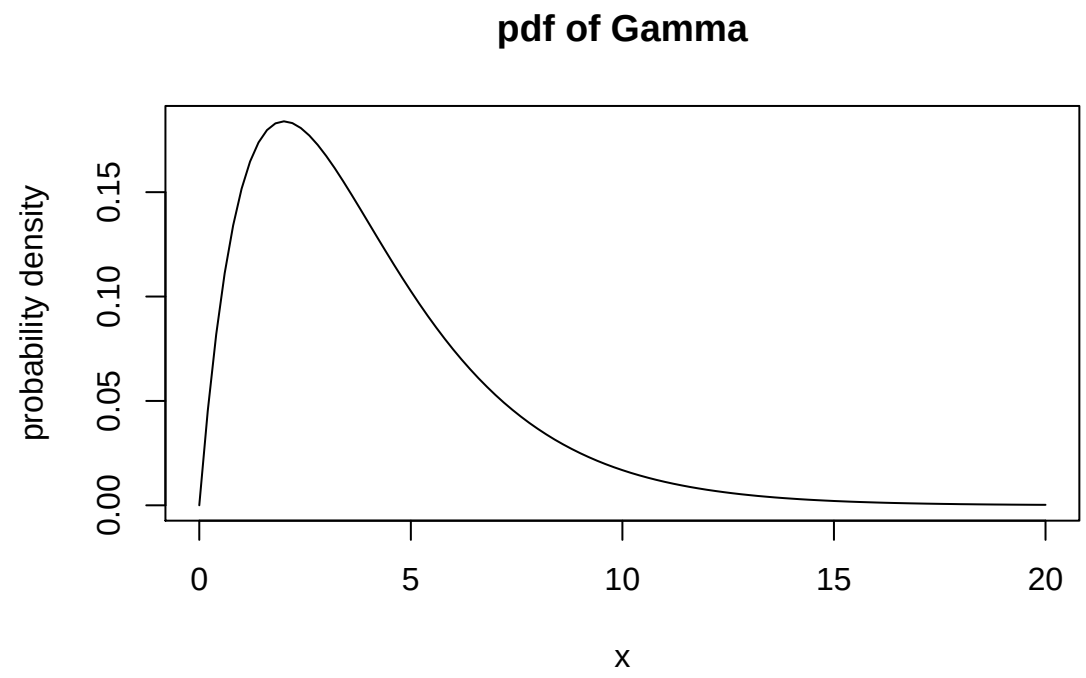
a. Describe the Central Limit Theorem as simply as you can.
b. Let $X \sim \text{Gamma}(\alpha = 2,\ \beta = 2)$. For the Gamma distribution, α is often called the "shape" parameter, β is often called the "scale" parameter, and the $\mathbb{E}[X] = \alpha\beta$. Plot the density of X and describe what you see. 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 $\bar{X}_n^{(1)}$. Repeat this process r times where $r = 1000$ so that you have $\bar{X}_n^{(1)}, \ldots, \bar{X}_n^{(r)}$. Plot a histogram of these r values and describe what you see. This is the sampling distribution of $\bar{X}_{(n)}$.
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. Central limit theorem states that when independent random variables are added, their properly normalized sum tends toward a normal distribution even if the random variables are not normally distributed

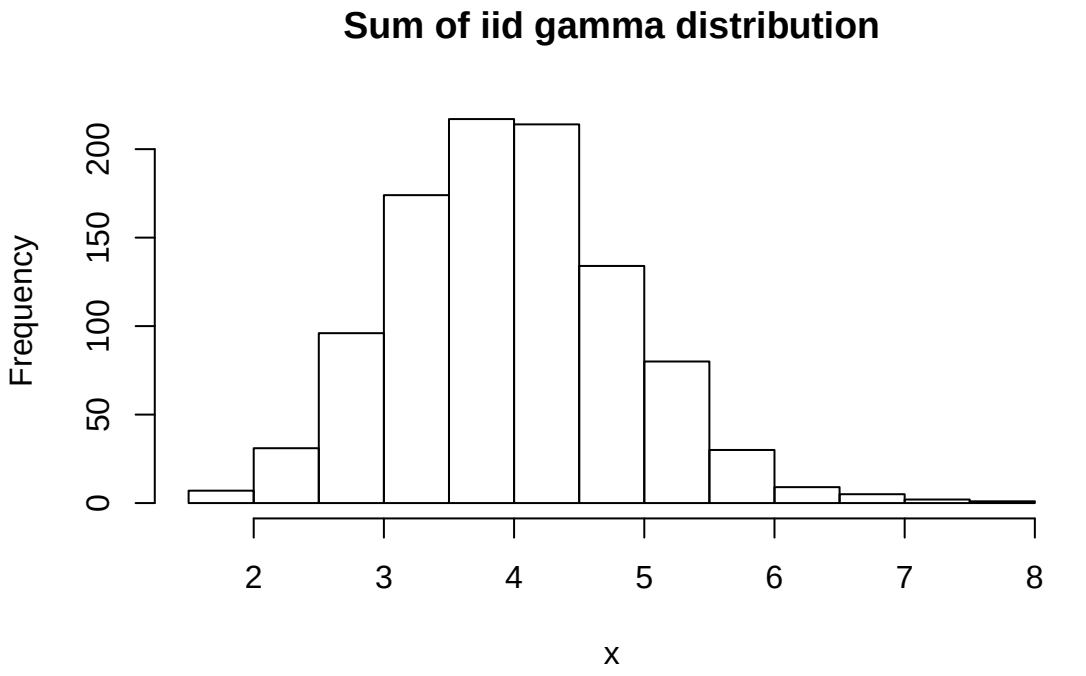
b.

```r
curve(dgamma(x, shape = 2, scale=2), from=0, to =20, main="pdf of Gamma", ylab="probability density")
```



c.

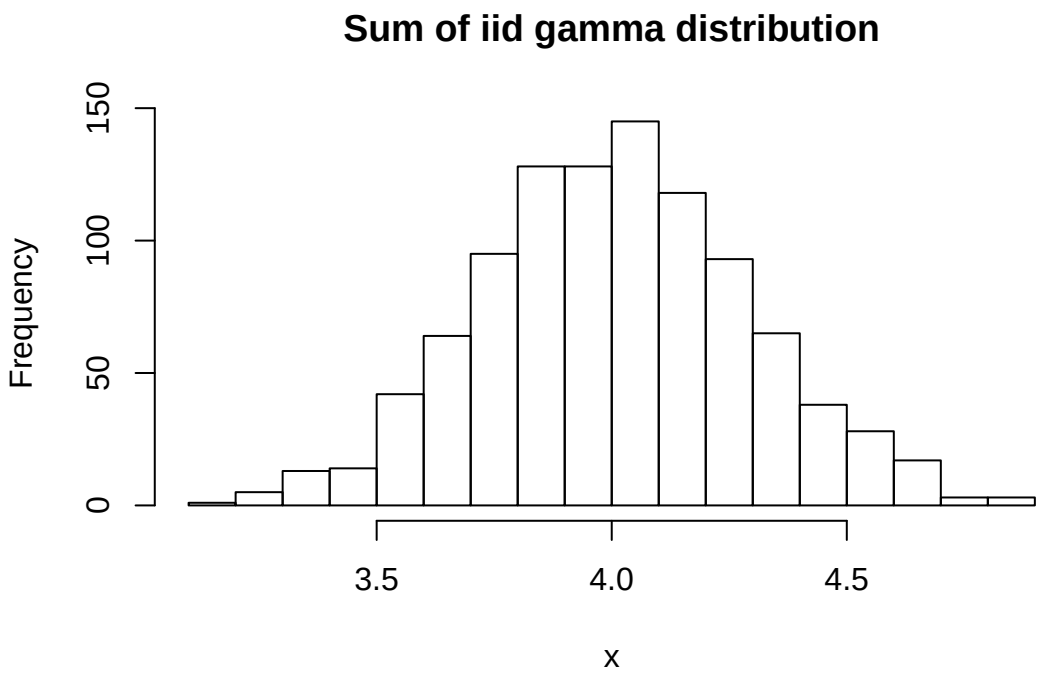
```r
n <- 10
r <- 1000
gammalist <- replicate(r, mean(rgamma(n=n, shape =2, scale = 2)))
hist(gammalist,breaks = 20, main="Sum of iid gamma distribution", xlab="x")
```



The sum of the iid gamma distribution approximates to a normal distribution, but slightly skewed to the right as the tail on the curve's right-hand side is longer than the tail on the left hand side.

d.

```r
n <- 100
r <- 1000
gammalist <- replicate(r, mean(rgamma(n=n, shape =2, scale = 2)))
hist(gammalist, breaks=20, main="Sum of iid gamma distribution", xlab="x")
```



The histogram approximates to a even more pronounced normal distribution as it is less skewed.

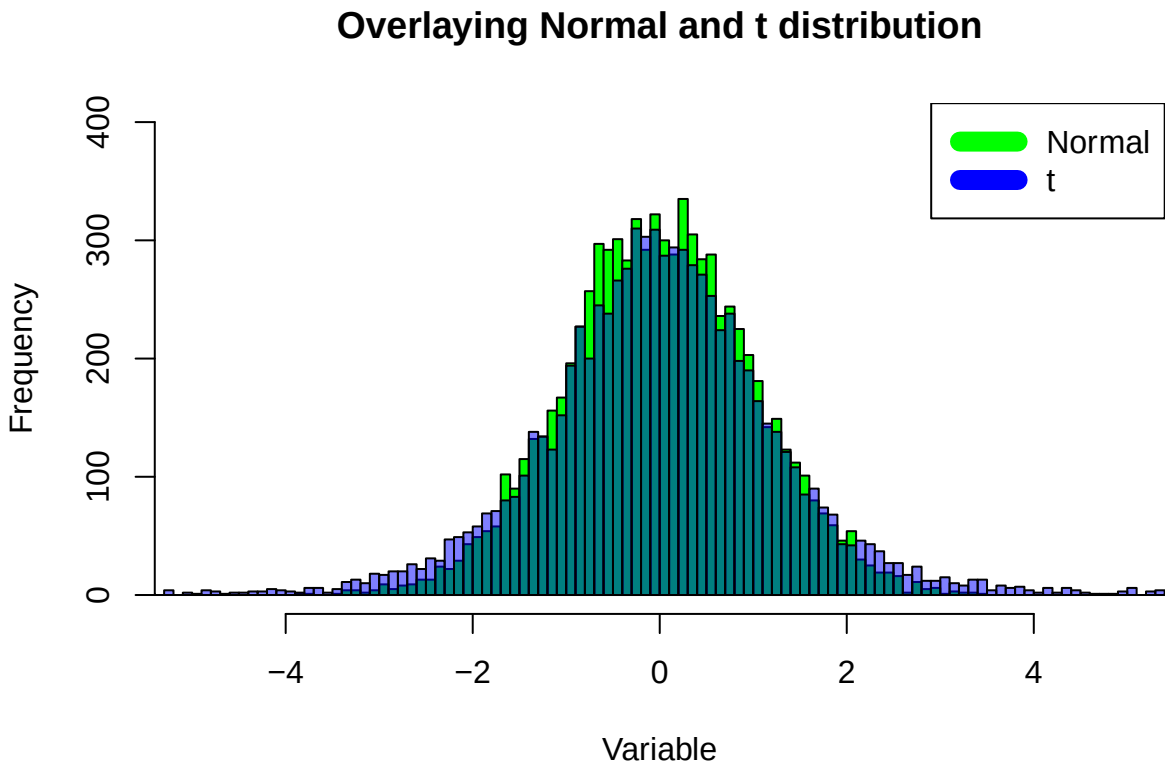
e. n = 2000 as we have a sample size of 2000. We need to make 1 draws from the sample to get the distribution of height and weight respectively. Hence, r = 1.

Question 5

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 $\mathcal{N}(0,1)$ distribution has much thinner tails than a t-distribution with 5 degrees of freedom.

A few coding hints: `rnorm()` and `rt()` are the functions in R to draw from a normal distribution and a t-distribution. The option `add=TRUE` for the `hist()` command can be used to overlay a second histogram on top of another histogram, and after installing the `scales` package, you can make a blue histogram 50% transparent with the option `col=scales::alpha("blue",0.5)`. Alternatively, you can put two plots side-by-side by first setting the plotting parameter with the code `par(mfrow=c(1,2))`. You can set the range of the x-axis to go from -5 to 5 with the plotting option `xlim=c(-5,5)`.

```
mynorm <- rnorm(8000, mean=0, sd=1)
hist(mynorm,breaks = seq(-5,5,by=0.1),col="green",xlab="Variable", main="Overlaying Normal and t distril
     ylim=c(0,400),xlim=c(-5,5))
myt <- rt(8000,5)
hist(myt,breaks = seq(-15,15,by=0.1),col=scales::alpha("blue",0.5), add= TRUE)
legend("topright", c("Normal", "t"), col=c("green", "blue"), lwd=10)
```



Question 6

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.

a).

```
data(Vanguard)
Van=Vanguard[,c(1,2,5)]
V_reshaped=dcast(Van,date~ticker,value.var="mret")
mat=descStat(V_reshaped["VFIAX"])
```

```
##        Mean Median    SD  IQR SE Mean 95% CI-L 95% CI-U NMissing
## VFIAX 0.004  0.011 0.045 0.05   0.004   -0.003    0.011      198
## Number of Observations =  349
```

The Standard error of the mean for fund VFIAX is 0.004

b). Standard Error measures the accuracy with which a sample represents a popluation. In this case the mean is equal to the standard error of the mean, which is a very significant number. This means the estimate for mean is not accurate, hence it is diffcult for financial analyst to get any statistically significant result.

c).

```
(sd(V_reshaped$VFIAX, na.rm=T)/(mean(V_reshaped$VFIAX, na.rm = T)/10))^2
```

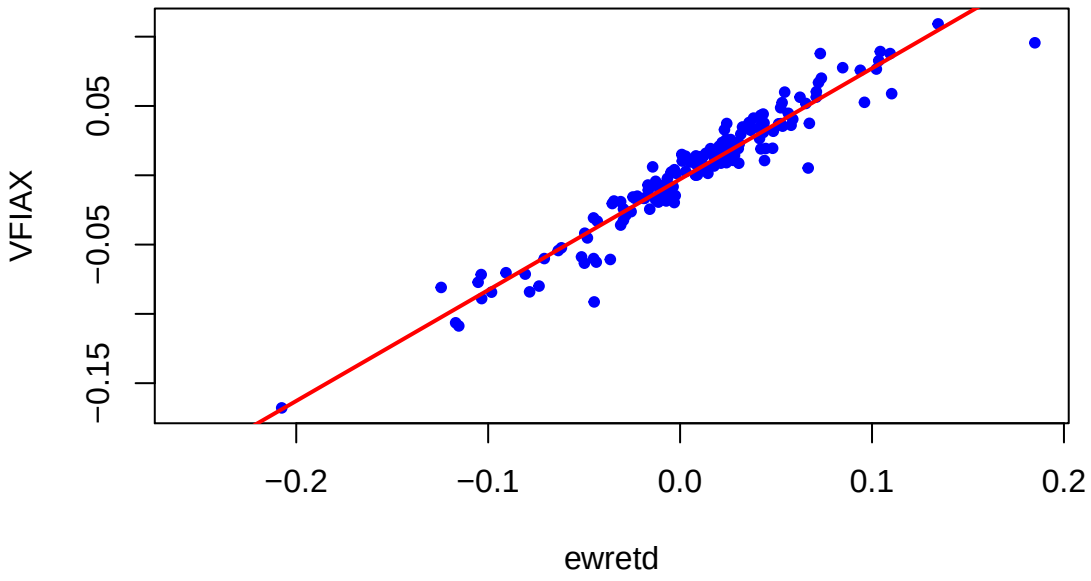
```
## [1] 12970.32
```

Question 7

a. Plot the `VFIAX` index fund return (as the Y variable) against the `ewretd` (equal-weighted market return, as the X variable) 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.

a).

```
data(marketRf)
Van_mkt=merge(V_reshaped,marketRf,by="date")
with(Van_mkt,
     plot(ewretd,VFIAX,pch=20,col="blue")
)
out=lm(VFIAX~ewretd,data=Van_mkt)
abline(out$coef,col="red",lwd=2)
```



b).

```
lmSumm(out)
```

```
## Multiple Regression Analysis:
##     2 regressors(including intercept) and 151 observations
##
## lm(formula = VFIAX ~ ewretd, data = Van_mkt)
##
## Coefficients:
##              Estimate Std Error t value p value
## (Intercept) -0.002855  0.001014   -2.82   0.006
## ewretd       0.799900  0.018520   43.19   0.000
## ---
## Standard Error of the Regression:  0.01231
## Multiple R-squared:  0.926  Adjusted R-squared:  0.926
## Overall F stat: 1865.32 on 1 and 149 DF, pvalue= 0
```