Dragos Ailoae

Applied Statistics and Econometrics I

2021-10-03

Homework 1 - Suggested Solutions

Problem 1:

Show that $E(Y)=\left(\theta_1+\theta_2\right)/2$ when Y is a uniformly distributed random variable on the interval $\left(\theta_1, \theta_2\right)$.

Per Greene Appendix:

DEFINITION B.1 Mean of a Random Variable
The ***mean****, or* ***expected value****, of a random variable is*

$$E[x]=\begin{cases}\sum_x x f(x) & \text{if } x \text{ is discrete,}\\ \int_x x f(x)\,dx & \text{if } x \text{ is continuous.}\end{cases} \qquad \textbf{(B-11)}$$

For a uniformly distributed random variable the probability density function (PDF) is:

$$f(x)=\begin{cases}\frac{1}{\theta_2-\theta_1}, & \text{if} \quad \theta_1 \leq x \leq \theta_2\\ 0, & \text{otherwise}\end{cases}$$

Therefore:

$$\begin{aligned}E[Y] &= \int_{\theta_1}^{\theta_2} Y f(Y) dY=\int_{\theta_1}^{\theta_2} Y \frac{1}{\theta_2-\theta_1} dY=\frac{1}{\theta_2-\theta_1} \int_{\theta_1}^{\theta_2} Y dY \\ &= \frac{1}{\theta_2-\theta_1} \frac{Y^2}{2}\bigg|_{\theta_1}^{\theta_2}=\frac{1}{\theta_2-\theta_1} \frac{\theta_2^2-\theta_1^2}{2}=\frac{\left(\theta_2-\theta_1\right)\left(\theta_2+\theta_1\right)}{2\left(\theta_2-\theta_1\right)}=\frac{\theta_2+\theta_1}{2} \\ & QED\end{aligned}$$

Problem 2:

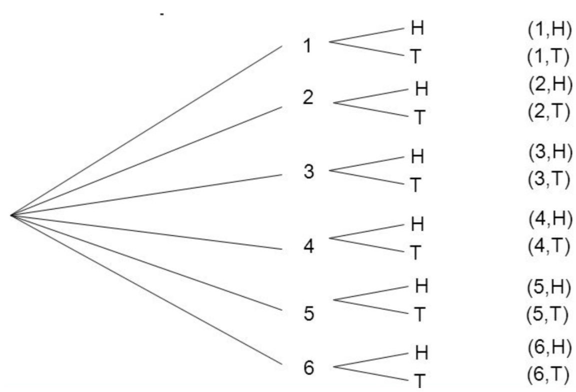
Consider an experiment that consists of tossing a die and flipping a coin at the same time. Consider the following random variables:

X_1: The number of dots appearing on the die.

X_2: The sum of the number of dots on the die and the indicator for the coin, where the value of a head is 1 and the value of a tail is zero.

a) How many possible events are there, and what are they? What are the possible outcomes and associated probabilities for X_2 ?

We can create a tree diagram for this experiment



Then, letting Head = 1 and Tail = 0, we can rewrite the possible event space E as:

$$\begin{pmatrix} (1,0),(2,0),(3,0),(4,0),(5,0),(6,0) \\ (1,1),(2,1),(3,1),(4,1),(5,1),(6,1) \end{pmatrix}$$

for a total of 12 possible events.

The corresponding values for X_2 are then:

$$\begin{pmatrix} (1),(2),(3),(4),(5),(6) \\ (2),(3),(4),(5),(6),(7) \end{pmatrix}$$

Tabulating, the possible outcomes and associated probabilities for X_2 are:

Value	Probability
1	1/12
2	1/6
3	1/6
4	1/6
5	1/6
6	1/6
7	1/12

Alternately, define random variable $C \equiv$ coin toss outcome. Then,

$$P(X_2=1)=P(X_1=1)\times P(C=0)=\frac{1}{6}\cdot\frac{1}{2}=\frac{1}{12}$$
$$P(X_2=2)=P(X_1=1)\times P(C=1)+P(X_1=2)\times P(C=0)=\frac{1}{12}+\frac{1}{12}=\frac{1}{6}$$
$$P(X_2=3)=P(X_1=2)\times P(C=1)+P(X_1=3)\times P(C=0)=\frac{1}{12}+\frac{1}{12}=\frac{1}{6}$$
$$\vdots$$
$$P(X_2=7)=P(X_1=6)\times P(C=1)=\frac{1}{12}$$

b) What is the $P(X_1=3, X_2=3)$?

First note that $P(X_1=3, X_2=3)\neq P(X_1=3)\times P(X_2=3)$. This is because X_2 is a function of X_1 which means the two random variables **ARE NOT** independent. Per Greene Appendix:

Two random variables are statistically independent if and only if their joint density is the product of the marginal densities:

$$f(x,y)=f_x(x)f_y(y)\Leftrightarrow x \text{ and } y \text{ are independent.} \qquad \textbf{(B-46)}$$

Also from Greene:

It follows from (B-46) that.

If x and y are independent, then $f(y|x)=f_y(y)$ and $f(x|y)=f_x(x)$. **(B-60)**

In our case $P(X_2)$ is dependent on X_1 i.e. $P(X_2|X_1)\neq P(X_2)$.

However, we know that the bivariate conditional density is defined as (see Greene Appendix or here):

In a bivariate distribution, there is a **conditional distribution** over y for each value of x. The conditional densities are

$$f(y|x)=\frac{f(x,y)}{f_x(x)}, \qquad \textbf{(B-59)}$$

In our case then,

$$P(X_2=3|X_1=3)=\frac{P(X_1=3,X_2=3)}{P(X_1=3)}$$

Or, rearranging,

$$P(X_1=3,X_2=3)=P(X_1=3)\times P(X_2=3|X_1=3)$$

But $P(X_2=3|X_1=3)=\frac{1}{2}$, i.e. the probability of the coin flip being 0. Thus we have:

$$\begin{aligned}P(X_1=3,X_2=3)&=P(X_1=3)\times P(X_2=3|X_1=3)\\&=\frac{1}{6}\times\frac{1}{2}\\&=\frac{1}{12}\end{aligned}$$

The complete bivariate distribution $f(X_1, X_2)$ is:

		X_2						
		1	**2**	**3**	**4**	**5**	**6**	**7**
X_1	**1**	1/12	1/12	0	0	0	0	0
	2	0	1/12	1/12	0	0	0	0
	3	0	0	1/12	1/12	0	0	0
	4	0	0	0	1/12	1/12	0	0
	5	0	0	0	0	1/12	1/12	0
	6	0	0	0	0	0	0	1/12

Problem 3:

A local supermarket has three checkout counters. Two customers arrive at the counters at different times when the counters are serving no other customers. Each customer chooses a counter at random, independently of the other. Let Y_1 denote the number of customers who select counter 1 and Y_2, the number who select counter 2. Find the joint distribution of Y_1 and Y_2.

The key is to note that the two customers choose "independently of the other". This means we can have 2 customers select the same counter and so we can have 0, 1, or 2 customers at a given counter.

Let's first look at the sample space in terms of the counter selected by each of the two customers. Let the pair {i, j} denote the simple event that the first customer chooses counter i and the second customer chooses counter j, with i = {1,2,3} and j = {1,2,3}. The sample space then consists of 9 possible checkout counter pairs:

$$S(i,j) = \{(1,1),(1,2),(1,3),(2,1),(2,2),(2,3),(3,1),(3,2),(3,3)\}$$

Now, in terms of the RVs defined in the problem

Y_1 = number of customers who select counter 1

Y_2 = number of customers who select counter 2

Therefore $S(i,j) = (1,1)$ is equivalent to $(Y_1 = 2, Y_2 = 0)$ and so $P(Y_1 = 2, Y_2 = 0) = \frac{1}{9}$

Similarly, the event $(Y_1 = 1, Y_2 = 1)$ is equivalent to $S(i,j) = \{(2,1),(1,2)\}$ therefore $P(Y_1 = 1, Y_2 = 1) = \frac{1}{9} + \frac{1}{9} = \frac{2}{9}$

The complete bivariate distribution $f(Y_1, Y_2)$ is:

		Y_2		
		0	**1**	**2**
	0	1/9	2/9	1/9
Y_1	**1**	2/9	2/9	0
	2	1/9	0	0

Problem 4:

Is the estimator Z shown below a biased and/or consistent estimator of the population mean, μ?

$$Z = \frac{1}{n+1}\sum_{i=1}^{n} X_i$$

Per Greene Appendix:

> **DEFINITION C.2 Unbiased Estimator**
> *An estimator of a parameter θ is unbiased if the mean of its sampling distribution is θ. Formally,*
>
> $$E[\hat{\theta}] = \theta$$
>
> *or*
>
> $$E[\hat{\theta} - \theta] = \text{Bias}[\hat{\theta} \mid \theta] = 0$$
>
> *implies that $\hat{\theta}$ is unbiased. Note that this implies that the expected sampling error is zero. If $\boldsymbol{\theta}$ is a vector of parameters, then the estimator is unbiased if the expected value of every element of $\hat{\boldsymbol{\theta}}$ equals the corresponding element of $\boldsymbol{\theta}$.*

We are given the true population mean $E[X] = \mu$. To prove that the estimator Z is unbiased we need to prove that $E[Z] = E[X] = \mu$.

The expected value of Z is then

$$E[Z] = E\left[\frac{1}{n+1}\sum_{i=1}^{n} X_i\right]$$

Since $\frac{1}{n+1}$ is a constant (i.e. not a random variable) we can pull it out of the expectation operator, thus:

$$E[Z] = E\left[\frac{1}{n+1}\sum_{i=1}^{n} X_i\right] = \frac{1}{n+1}\sum_{i=1}^{n} E[X_i] = \frac{1}{n+1} nE[X] = \frac{n}{n+1}\mu$$

Since $E[Z] \neq \mu$, estimator Z is **biased.** (Note $E[Z] < \mu$ so negatively biased)

As far as consistency, per Greene Appendix:

> **DEFINITION D.2 Consistent Estimator**
> *An estimator $\hat{\theta}_n$ of a parameter θ is a consistent estimator of θ if and only if*
>
> $$\text{plim}\ \hat{\theta}_n = \theta. \qquad \textbf{(D-4)}$$

We can see that, asymptotically

$$\lim_{n\to\infty} E[Z] = \lim_{n\to\infty} \frac{n}{n+1}\mu = \mu$$

Also,

$$
\begin{aligned}
var[Z] &= E\left[(Z - E[Z])^2\right] \\
&= E\left[\left(\left(\frac{1}{n+1}\sum_{i=1}^{n} X_i\right) - \left(\frac{n}{n+1}\mu\right)\right)^2\right] \\
&= E\left[\frac{1}{(n+1)^2}\left(\left(\sum_{i=1}^{n} X_i\right) - n\mu\right)^2\right] \\
&= \frac{1}{(n+1)^2} E\left[\left(\left(\sum_{i=1}^{n} X_i\right) - n\mu\right)^2\right] \\
&= \frac{1}{(n+1)^2} E\left[\left(\sum_{i=1}^{n} X_i - \mu\right)^2\right] \\
&= \frac{1}{(n+1)^2} nE\left[(X-\mu)^2\right]
\end{aligned}
$$

Note that $E\left[(X-\mu)^2\right] = var[X] \equiv \sigma_X^2$ so that

$$
\begin{aligned}
var[Z] &= \frac{1}{(n+1)^2} nE\left[(X-\mu)^2\right] \\
&= \frac{n}{(n+1)^2}\sigma_X^2
\end{aligned}
$$

We can see then that,

$$\lim_{n\to\infty} var\left[Z\right] = \lim_{n\to\infty} \frac{n}{(n+1)^2}\sigma_X^2 = 0$$

Therefore

$$\text{plim Z} = \mu$$

This means that as n grows the distributions of the Z estimates become more and more concentrated near the true value, so that the probability of the estimator being arbitrarily close to μ converges to 1. Estimator Z is therefore **consistent**.

Problem 5:

Download the R package AER. This package contains a number of data sets. Upload the data set "USInvest" into a data frame as a time series. You will find the data set listed under Greene (2003).

According to the documentation for the AER package available here the 'USInvest' dataset is sourced from the online complement to Greene(2003) Table F3.1. http://pages.stern.nyu.edu/~wgreene/Text/tables/tablelist5.htm . Note that Greene specifies "CPI 1967 is 79.06"

We start by loading the data into a time series object and checking that data was loaded properly.

In [1]:

```
rm(list=ls())     # Removes all items in Environment!
options(warn=-1) # Suppress warnings
qui <- suppressPackageStartupMessages # quiet! - suppress library load messages

# Install missing packages if necessary
# install.packages("AER")

# Load packages
# qui(library(AER))  # for 'USInvest' dataset

# install (if not already installed) and load package
qui(if(!require(AER)){install.packages('AER')}) # for 'USInvest' dataset

data(USInvest)                                  # Load dataset
mydata <- USInvest                              # Store locally

mydata
```

A Time Series: 15 × 4

	gnp	invest	price	interest
1968	873.4	133.3	82.54	5.16
1969	944.0	149.3	86.79	5.87
1970	992.7	144.2	91.45	5.95
1971	1077.6	166.4	96.01	4.88
1972	1185.9	195.0	100.00	4.50
1973	1326.4	229.8	105.75	6.44
1974	1434.2	228.7	115.08	7.83
1975	1549.2	206.1	125.79	6.25
1976	1718.0	257.9	132.34	5.50
1977	1918.3	324.1	140.05	5.46
1978	2163.9	386.6	150.42	7.46
1979	2417.8	423.0	163.42	10.28
1980	2633.1	402.3	178.64	11.77
1981	2937.7	471.5	195.51	13.42
1982	3057.5	421.9	207.23	11.02

a) Add real GNP, real invest, and inflation to the data frame.

In [2]:

```
#install (if not already installed) and load package
qui(if(!require(tfplot)){install.packages('tfplot')})  # for percentChange()

CPI_Base <- 79.06    # Per Greene note "CPI 1967 is 79.06"
rgnp <- mydata[,'gnp'] / mydata[,'price'] * CPI_Base
rinvest <- mydata[,'invest'] / mydata[,'price'] * CPI_Base
inflation <- percentChange(mydata[,'price'])
```

In [3]:

```
mydata<-ts.union(gnp=mydata[,'gnp'],rgnp,
                 invest=mydata[,'invest'],rinvest,
                 price=mydata[,'price'],interest=mydata[,'interest'],
                 inflation
                )
mydata[1,'inflation'] <- (mydata[1,'price']/CPI_Base - 1) * 100  #include 1968 inflation leve
mydata
```

A Time Series: 15 × 7

	gnp	rgnp	invest	rinvest	price	interest	inflation
1968	873.4	836.5763	133.3	127.6799	82.54	5.16	4.401720
1969	944.0	859.9221	149.3	136.0025	86.79	5.87	5.149019
1970	992.7	858.2052	144.2	124.6632	91.45	5.95	5.369282
1971	1077.6	887.3561	166.4	137.0231	96.01	4.88	4.986331
1972	1185.9	937.5725	195.0	154.1670	100.00	4.50	4.155817
1973	1326.4	991.6329	229.8	171.8013	105.75	6.44	5.750000
1974	1434.2	985.2959	228.7	157.1170	115.08	7.83	8.822695
1975	1549.2	973.6843	206.1	129.5355	125.79	6.25	9.306569
1976	1718.0	1026.3343	257.9	154.0696	132.34	5.50	5.207091
1977	1918.3	1082.9047	324.1	182.9586	140.05	5.46	5.825903
1978	2163.9	1137.3350	386.6	203.1950	150.42	7.46	7.404498
1979	2417.8	1169.6932	423.0	204.6407	163.42	10.28	8.642468
1980	2633.1	1165.3207	402.3	178.0443	178.64	11.77	9.313426
1981	2937.7	1187.9421	471.5	190.6644	195.51	13.42	9.443574
1982	3057.5	1166.4621	421.9	160.9584	207.23	11.02	5.994578

b) Calculate the mean and standard deviation for each of the series.

This can be calculated using the 'mean' and 'sd' functions while using 'sapply' to loop through our data.

In [4]:
```
#install (if not already installed) and load package
qui(if(!require(knitr)){install.packages('knitr')})         # for kable() - table formatting
qui(if(!require(tidyverse)){install.packages('tidyverse')}) # for pipes

means <- sapply(mydata,mean, na.rm=TRUE)

# output nicely formatted table
kable(means,col.names="Mean", digits=1, align="c", caption="Summary Statistics","html")
```

Summary Statistics

	Mean
gnp	1748.6
rgnp	1017.7
invest	276.0
rinvest	160.8
price	131.4
interest	7.5
inflation	6.7

Similarly for standard deviation:

In [5]:
```
std.devs <- sapply(mydata,sd, na.rm=TRUE)

# output nicely formatted table
kable(std.devs, col.names="Std. Dev.",digits=1, align="c",caption="Summary Statistics","html'
```

Summary Statistics

	Std. Dev.
gnp	738.1
rgnp	126.7
invest	117.6
rinvest	27.0
price	40.3
interest	2.8
inflation	1.9

We can confirm answers by comparing with the output of a pre-packaged summary stats function

In [6]:
```
qui(if(!require(pastecs)){install.packages('pastecs')}) # for stat.desc() function - comprehe

kable(stat.desc(mydata, basic=F), digits=1, align="c", caption="Summary Statistics","html")
```

Summary Statistics

	gnp	rgnp	invest	rinvest	price	interest	inflation
median	1549.2	991.6	229.8	157.1	125.8	6.2	5.8
mean	1748.6	1017.7	276.0	160.8	131.4	7.5	6.7
SE.mean	190.6	32.7	30.4	7.0	10.4	0.7	0.5
CI.mean.0.95	408.8	70.2	65.1	14.9	22.3	1.6	1.1
var	544859.2	16061.4	13825.7	727.4	1623.0	7.9	3.8
std.dev	738.1	126.7	117.6	27.0	40.3	2.8	1.9
coef.var	0.4	0.1	0.4	0.2	0.3	0.4	0.3

c) Include growth rates for real GNP and real investment in the data frame.

In [7]:

```
rgnp.growth <- percentChange(mydata[,'rgnp'])
rinvest.growth <- percentChange(mydata[,'rinvest'])
mydata<-ts.union(gnp=mydata[,'gnp'],rgnp,rgnp.growth,
                  invest=mydata[,'invest'],rinvest,rinvest.growth,
                  price=mydata[,'price'],interest=mydata[,'interest'],
                  inflation=mydata[,'inflation']
                 )
mydata
```

A Time Series: 15 × 9

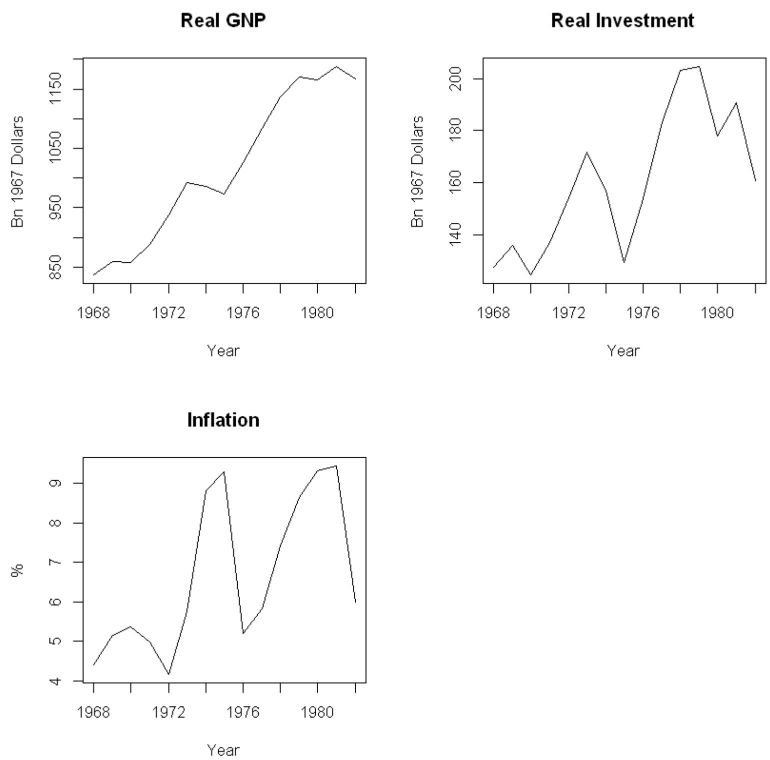
	gnp	rgnp	rgnp.growth	invest	rinvest	rinvest.growth	price	interest	inflation
1968	873.4	836.5763	NA	133.3	127.6799	NA	82.54	5.16	4.401720
1969	944.0	859.9221	2.7906430	149.3	136.0025	6.5183510	86.79	5.87	5.149019
1970	992.7	858.2052	-0.1996634	144.2	124.6632	-8.3375563	91.45	5.95	5.369282
1971	1077.6	887.3561	3.3967293	166.4	137.0231	9.9145792	96.01	4.88	4.986331
1972	1185.9	937.5725	5.6591119	195.0	154.1670	12.5117188	100.00	4.50	4.155817
1973	1326.4	991.6329	5.7659971	229.8	171.8013	11.4384434	105.75	6.44	5.750000
1974	1434.2	985.2959	-0.6390517	228.7	157.1170	-8.5472724	115.08	7.83	8.822695
1975	1549.2	973.6843	-1.1784853	206.1	129.5355	-17.5547644	125.79	6.25	9.306569
1976	1718.0	1026.3343	5.4072925	257.9	154.0696	18.9401104	132.34	5.50	5.207091
1977	1918.3	1082.9047	5.5118856	324.1	182.9586	18.7505709	140.05	5.46	5.825903
1978	2163.9	1137.3350	5.0263298	386.6	203.1950	11.0606849	150.42	7.46	7.404498
1979	2417.8	1169.6932	2.8450905	423.0	204.6407	0.7114609	163.42	10.28	8.642468
1980	2633.1	1165.3207	-0.3738206	402.3	178.0443	-12.9966127	178.64	11.77	9.313426
1981	2937.7	1187.9421	1.9412195	471.5	190.6644	7.0881458	195.51	13.42	9.443574
1982	3057.5	1166.4621	-1.8081661	421.9	160.9584	-15.5802276	207.23	11.02	5.994578

d) Plot real GNP, real invest, and inflation as lines on separate charts on a single page. Be sure to put time on the horizontal axis. Include a legends and a title.

In [8]:

```
par(mfrow=c(2,2))
plot(mydata[,'rgnp'],xlab="Year",ylab="Bn 1967 Dollars",
main="Real GNP",cex=0.7)
```

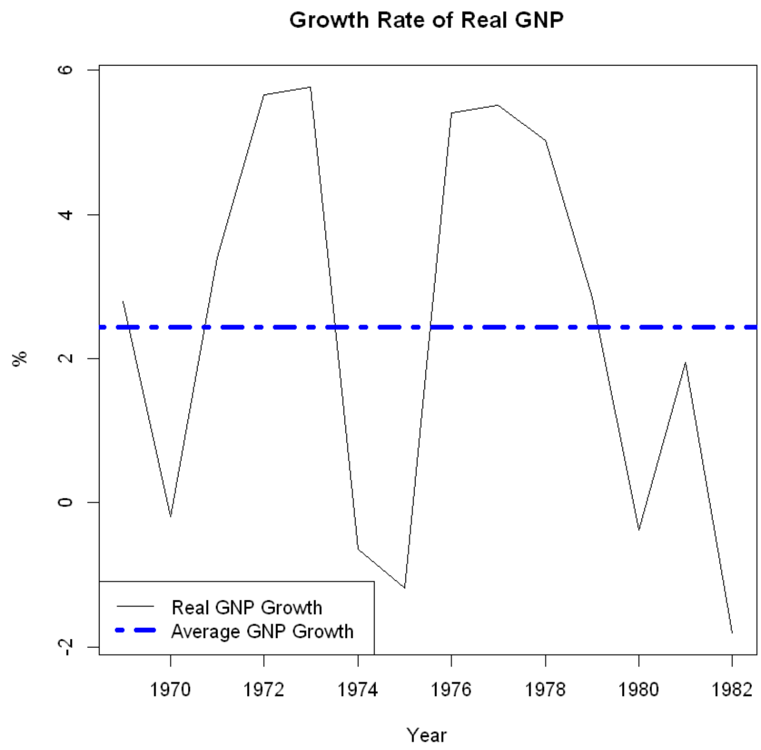
```
plot(mydata[,'rinvest'],xlab="Year",ylab="Bn 1967 Dollars",
main="Real Investment",cex=0.7)
plot(mydata[,'inflation'],xlab="Year",ylab="%",
main="Inflation",cex=0.7)
```



e) Plot the growth rates of real GNP and real investment on separate charts and include a horizontal line on each chart representing the mean rate of growth for the series.

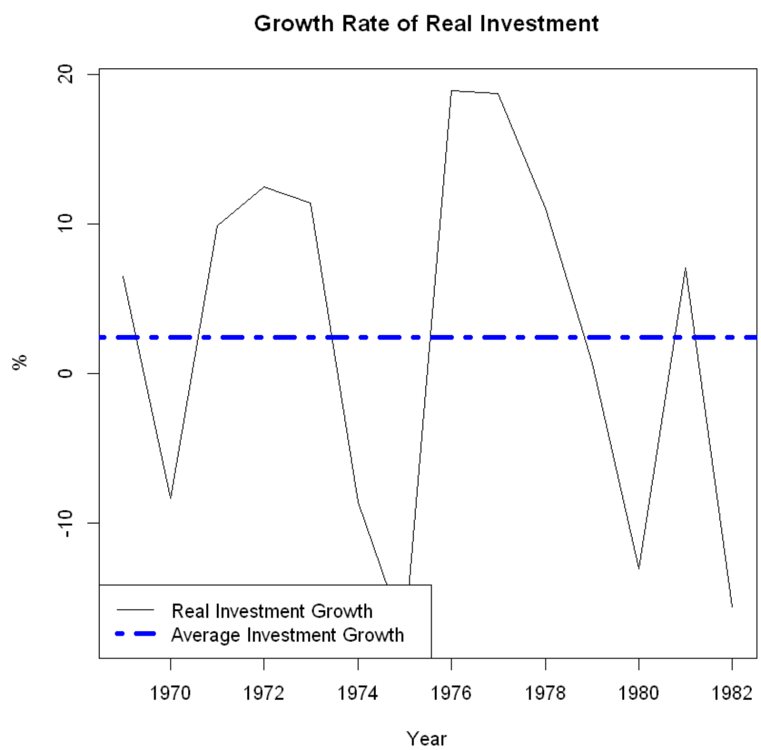
In [9]:

```
plot(rgnp.growth,
     Rate of Real GNP", xlab="Year",ylab="%")
 abline(h = mean(rgnp.growth), col = "blue", lwd = 4, lty = 4)
 legend("bottomleft", legend=c("Real GNP Growth","Average GNP Growth"),lty=c(1,4),lwd=c(1,4)
 par(cex = 1)
```



In [10]:

```
plot(rinvest.growth, main="Growth Rate of Real Investment", xlab="Year",ylab="%")
abline(h = mean(rinvest.growth), col = "blue", lwd = 4, lty = 4)
legend("bottomleft", legend=c("Real Investment Growth","Average Investment Growth"),
        lty=c(1,4),lwd=c(1,4),col =c("black","blue"))
par(cex = 1)
```

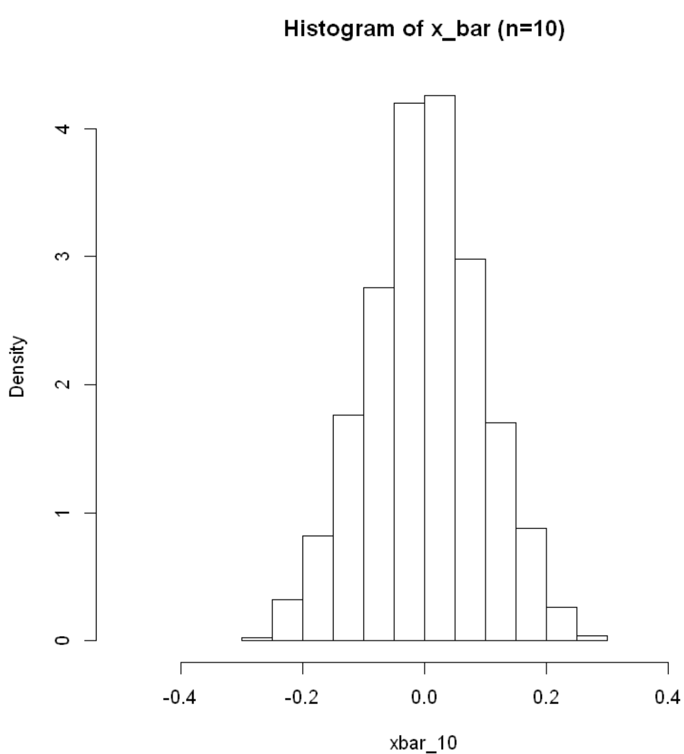


Problem 6

Write a program in R to illustrate the Central Limit Theorem by duplicating the results in the notes. That is, draw 1000 samples of size 10, 20, and 100 from a uniform distribution that ranges from -0.5 to +0.5. Calculate the mean for each sample and plot the 3 densities using the kernel density estimator.

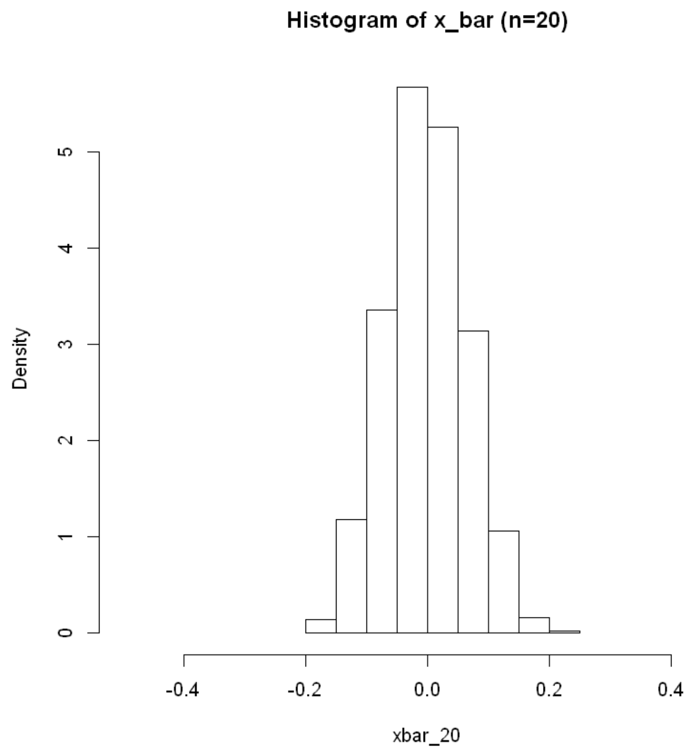
In [11]:

```
set.seed(12345)

n <- 10  # sample size
num_repetitions <- 1000  # experiment repetitions

u_min <- -0.5 # uniform distribution limits
u_max <-  0.5

samples <- replicate(num_repetitions, runif(n, u_min, u_max))

xbar_10 <- sapply(as.data.frame(samples), mean)
# xbar_10 <- colMeans(samples)  # alternate function to calculate means for each column

hist(xbar_10, main = "Histogram of x_bar (n=10)", prob = TRUE, breaks = 10, xlim = c(u_min,u_
```



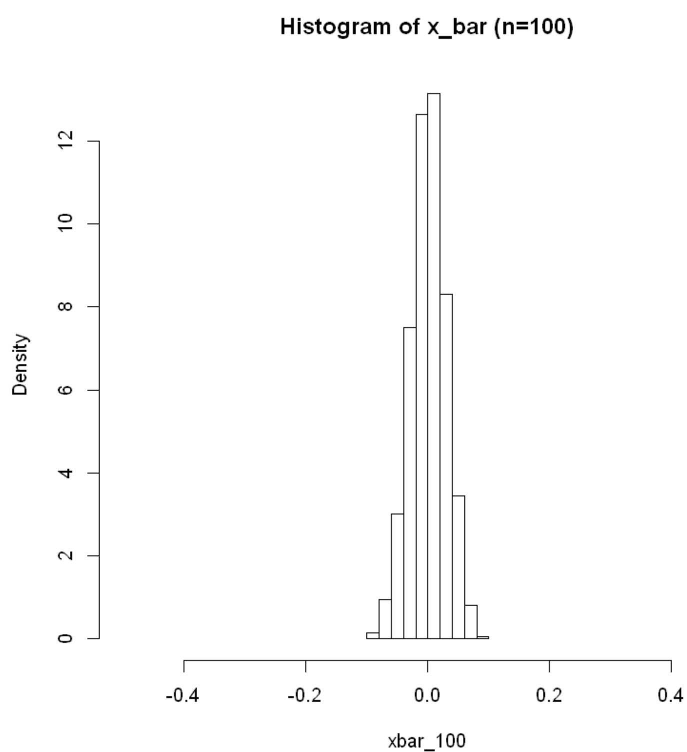
In [12]:

```
n <- 20

samples <- replicate(num_repetitions, runif(n, u_min, u_max))

xbar_20 <- sapply(as.data.frame(samples), mean)

hist(xbar_20, main = "Histogram of x_bar (n=20)",prob=TRUE, breaks = 10, xlim=c(u_min,u_max))
```

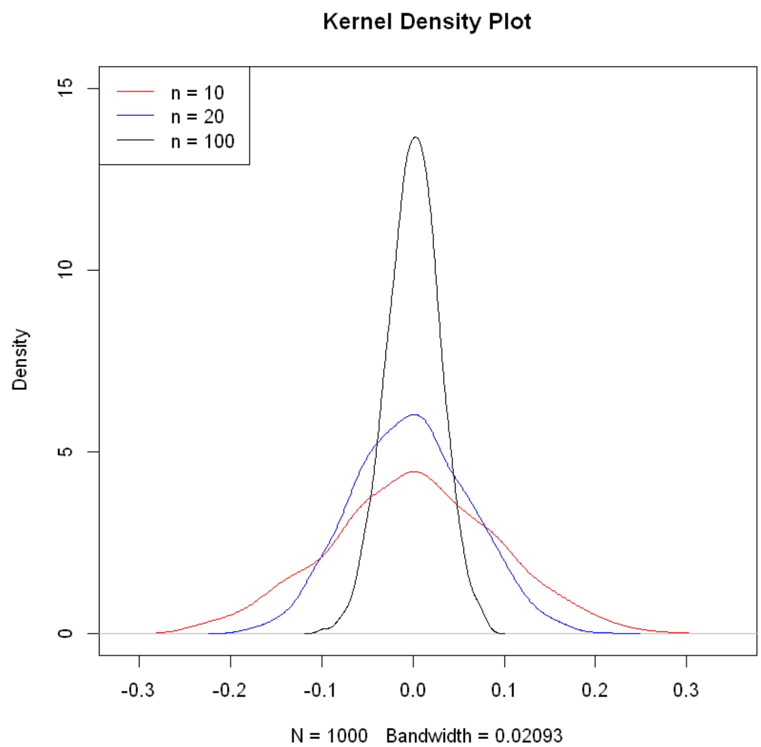


In [13]:

```
n <- 100

samples <- replicate(num_repetitions, runif(n, u_min, u_max))

xbar_100 <- sapply(as.data.frame(samples), mean)

hist(xbar_100, main = "Histogram of x_bar (n=100)", prob = TRUE, breaks = 10, xlim = c(u_min,
```



In [14]:
```
plot(density(xbar_10),col = "red",ylim = c(0,15), main = "Kernel Density Plot")
 lines(density(xbar_20), col = "blue")
 lines(density(xbar_100), col = "black")
 legend("topleft",legend=c("n = 10", "n = 20", "n = 100"),
   col=c("red", "blue", "black"), lty=c(1,1,1))
```



In [15]:
```
test <- as.data.frame(cbind(xbar_10,xbar_20,xbar_100))

names(test)[1] <- "xbar_10"
names(test)[2] <- "xbar_20"
names(test)[3] <- "xbar_100"

std.devs <- sapply(test,sd, na.rm=TRUE)
kable(std.devs, col.names="Std. Dev.",digits=3, align="c",caption="Summary Statistics","html'
```

Summary Statistics

	Std. Dev.
xbar_10	0.093
xbar_20	0.064
xbar_100	0.029