Delft University of Technology
Faculty EEMCS, Department DIAM

van Mourik Broekmanweg 6
2628 XE Delft

Midterm TW2080 2018-2019 **October 3rd, 2018**

This written midterm exam contains 5 questions, each question counts for 20% of the final grade of the written test. You are only allowed to use the two sheets with information on probability distributions and R-commands. You are not allowed to use any books or notes.

1. The standard exponential distribution has distribution function $F(x) = 1-e^{-x}$ on $[0, \infty)$.

   (a) Does the exponential distribution with parameter λ belong to a location-scale family $\{F_{a,b} : a \in \mathbb{R}, b > 0\}$ associated with the standard exponential distribution? If yes, then specify the parameters a and b. If no, then explain why.

   (b) Derive the expression of the quantile function corresponding to the exponential distribution with parameter λ.

2. Let $X_1, \ldots X_n$ be independent and U$[0, \theta]$-distributed, with $\theta > 0$ unknown. Consider estimators of the type $cX_{(n)}$, for $c \in \mathbb{R}$. You may use that

$$\mathrm{E}_\theta X_{(n)} = \frac{n\theta}{n+1} \quad \text{and} \quad \mathrm{var}(X_{(n)}) = \frac{n\theta^2}{(n+2)(n+1)^2}.$$

   (a) For which value of c is the estimator $cX_{(n)}$ unbiased for θ?

   (b) Determine the mean square errors of the estimators $cX_{(n)}$ for θ, for every value of $c > 0$.

   (c) Which value for c gives the best estimator?

3. Let $X_1, \ldots, X_n$ be a sample from a distribution with probability density

$$p_\theta(x) = \theta x^{\theta-1} \quad \text{for } x \in (0, 1)$$

   and 0 otherwise. Here $\theta > 0$ is an unknown parameter.

   (a) Determine the maximum likelihood estimator for θ.

   (b) Determine the method of moments estimator for θ.

4. Let $X = (X_1, \ldots, X_n)$ be a sample from the geometric distribution with parameter θ,

$$P_\theta(X_1 = k) = (1-\theta)^{k-1}\theta, \qquad k = 1, 2, \ldots,$$

where $0 \leq \theta \leq 1$ is unknown. As prior distribution for θ, we choose probability density

$$\pi(\theta) = 6(1-\theta)\theta, \qquad \theta \in (0, 1).$$

   (a) Compute the posterior density and report to which distribution it corresponds.

   (b) Determine the Bayes estimator for θ and specify to which estimator it converges as $n \to \infty$.

5. Let $X_1, \ldots, X_n$ be a sample from the $N(\mu, 1)$ distribution. We test $H_0 : \mu \leq 0$ against $H_1 : \mu > 0$ at significance level α_0. We take $T = \overline{X}$ as test statistic.

   (a) Explain that the critical region of the test is of the form $K_T = [c_{\alpha_0}, \infty)$.

   (b) Show that the size of the test is given by $1 - \Phi(\sqrt{n}c_{\alpha_0})$, where Φ denotes the distribution function of the standard normal distribution.

   (c) Show that $c_{\alpha_0} = \xi_{1-\alpha_0}/\sqrt{n}$, where $\xi_{1-\alpha_0}$ denotes the $(1-\alpha_0)$-quantile of the standard normal distribution.

Delft University of Technology
Faculty EEMCS, Department DIAM

van Mourik Broekmanweg 6
2628 XE Delft

R Midterm TW2080 **October 3rd, 2018**

Inloggen.

1. Log in on the desktop PC with

2. Log in at your personal exam environment
   - Netid: your personal `netid`
   - Passwd: your personal `password`

Your personal exam environment. In your personal exam environment you will have access to two folders `P:\` and `S:\`. In the folder `S:\` you will find the two sheets with information on probability distributions and R-commands as well as the following files

- `template-student.R` - a template for an R-script;
- `template-student.Rmd` - a template for an R-markdown file;
- `template-student.docx` - a template for a word-document, which can be used to report your answers

Before you start.

1. First copy all these files to the folder `P:\`. In this folder you are able to save all files you will produce during the R-exam. You cannot write in folder `S:\`.
2. Open R-studio (note that there is no correct matching icon). Your workspace should be the folder `P:\`.
3. When you open the template of the R-script or the .Rmd-file, ignore the error message.
4. **NOTE:** If you want to use the package `knitr`, note that it only produces a Word-file! Ignore the error message about choosing an internet browser and remove it with `ESC`.

How to answer the R-questions. You can answer the questions in two ways

1. Setup a R-markdown file which contains your comments, the R-commands used, and the numbers and figures required, and use `knitr` to produce a word file. Put your netid and study number in the filenames, e.g., `netid1234567.Rmd` and and `netid1234567.docx`.

2. Put your R-commands in a R-script file and put your comments, and the numbers and figures required in a word file. Put your netid and study number in the filenames, e.g., `netid1234567.R` and `netid1234567.docx`.

This R-midterm exam contains 3 questions, each question counts for 1/3 of the final grade of the R-test. You are only allowed to use the two sheets with information on probability distributions and R-commands. You are not allowed to use any books or notes.

Before you start, first load the libraries `car` and `MASS` in your `R` workspace.

1. Use the dataframe `States`.

   (a) Report the minimum and maximum percentage of graduating high-school students in a state that took the SAT exam (`percent`). For how many states is this percentage below 60%?

   (b) To investigate normality of the variable `SATV` in dataframe `States`, produce a figure with two plots next to each other: (i) a histogram and (ii) a normal QQ-plot.

   (c) Assuming that the $N(\mu, \sigma^2)$ distribution is the right model compute estimates $\overline{x}$ and s_x^2 for μ and σ^2 on the basis of `SATV` values.

   (d) Produce a boxplot of the amount of dollars spend on public education by the state (`dollars`) split over the different U.S. Census regions (`region`). Which region corresponds to the largest amount of dollars spend?

   (e) Investigate by means of a scatterplot whether there is a correlation between the amount of dollars spend on public education by the state (`dollars`) and the average teacher's salary in the state (`pay`), with the amount of dollars spend on public education by the state as x-variable. Compute the sample correlation coefficient.

2. Use the dataframe `States`.

   (a) Consider the variable `pop` in the dataframe `States`, which contains the population of each state in 1 000s. Make a new variable `popmilion`, which represents the population of each state in 1 000 000s and produce a histogram of `popmilion`.

   (b) We want to fit a gamma distribution to the observations `popmilion`. Compute the maximum likelihood estimates for the parameters of this distribution on the basis the observations `popmilion`. Report their values.

(c) Construct a histogram of the observations `popmilion` and compare the fit with the gamma distribution with the parameters estimated by maximum likelihood.

3. We want to compare the performance of the maximum likelihood estimator with the method of moment estimator for the parameter θ of a uniform distribution on $[0, \theta]$. Take a uniform distribution with $\theta = 1$. Setup a simulation consisting of 1000 repetitions, where in each repetition:

   (i) Generate a sample of 50 observations from a uniform distribution with parameters $\theta = 1$.

   (ii) Pretend you do not know θ and, on the basis of the generated observations, compute the maximum likelihood estimator

   $$T_1 = x_{(50)} = \max\{x_1, \ldots, x_{50}\}$$

   and the method of moments estimator $T_2 = 2\overline{x}$ for θ.

   Put the values of each of the two estimators in numerical vectors of length 1000.

   Use the results of the simulation to compare the performance of the estimators T_1 and T_2 for θ. For this comparison use a boxplot and compute the empirical mean squared error.

   Which estimator performs better and why?

Before you logout.

1. Make sure that you have saved a R-markdown file which contains your comments, the R-commands used, and the numbers and figures required and a corresponding word document containing the answers. Make sure that the filenames are your netid and study number, e.g., `netid1234567.Rmd` and `netid1234567.docx`.

   or

   Make sure you have saved an R-script file containing all R-commands and a word document containing your comments, and the numbers and figures required. Make sure that the filenames are your netid and study number, e.g., `netid1234567.R` and `netid1234567.docx`.

2. When you are finished, you can close the session by clicking EXIT.

Solutions to Midterm TW2080

Rik Lopuhaä

1. (a) This is Exercise 2.3.
The distribution function of the exponential distribution with parameter $\lambda > 0$ is

$$F_\lambda(y) = 1 - \mathrm{e}^{-\lambda y} = F_1(-\lambda y), \quad y \geq 0$$

where

$$F_1(x) = 1 - \mathrm{e}^{-x}, \quad x \geq 0$$

is the distribution function of the standard exponential distribution with $\lambda = 1$. This means that $F_\lambda(y)$ is of the form

$$F_{a,b}(y) = F_1\left(\frac{y-a}{b}\right),$$

with $a = 0$ and $b = 1/\lambda$. Hence, the exponential distribution with parameter $\lambda > 0$ belongs to the location-scale family $\{F_{a,b} : a \in \mathbb{R}, b > 0\}$ associated with the standard exponential distribution.

(b) This Example 2.8.
Let Q denote the quantile function corresponding to F_λ. Because F_λ is strictly increasing on $[0, \infty)$ and continuous, for $\alpha \in (0, 1)$, $Q(\alpha)$ is equal to the value x such that $F_\lambda(x) = \alpha$. For $\alpha \in (0, 1)$, we have

$$F_\lambda(x) = \alpha \Leftrightarrow 1 - \mathrm{e}^{-\lambda x} = \alpha \Leftrightarrow \mathrm{e}^{-\lambda x} = 1 - \alpha \Leftrightarrow x = -\frac{1}{\lambda}\log(1-\alpha).$$

This means that the quantile function corresponding to F_λ is given by

$$Q(\alpha) = -\frac{1}{\lambda}\log(1-\alpha), \quad \alpha \in (0, 1).$$

2. (a) This is Exercise 3.3 and feedback assignment Week 3.
The estimator $cX_{(n)}$ is unbiased for θ, if

$$\mathrm{E}_\theta[cX_{(n)}] = \theta, \quad \text{for all } \theta > 0.$$

This means that the estimator $cX_{(n)}$ is unbiased for θ, if

$$\theta = \mathrm{E}_\theta[cX_{(n)}] = c\mathrm{E}_\theta[X_{(n)}] = \frac{cn}{n+1}\theta, \quad \text{for all } \theta > 0.$$

Therefore, the estimator $cX_{(n)}$ is unbiased for θ, if $c = (n+1)/n$.

(b) We have

$$\begin{aligned}\mathrm{MSE}(\theta; cX_{(n)}) &= \mathrm{var}(cX_{(n)}) + \left(\mathrm{E}_\theta[cX_{(n)}] - \theta\right)^2 \\ &= c^2\mathrm{var}(X_{(n)}) + \left(c\mathrm{E}_\theta[X_{(n)}] - \theta\right)^2.\end{aligned}$$

When we insert the expressions for $\mathrm{var}(X_{(n)})$ and $\mathrm{E}_\theta[X_{(n)}]$, we find

$$\begin{aligned}\mathrm{MSE}(\theta; cX_{(n)}) &= c^2\frac{n\theta^2}{(n+2)(n+1)^2} + \left(c\frac{n\theta}{n+1} - \theta\right)^2 \\ &= \theta^2\frac{nc^2}{(n+2)(n+1)^2} + \theta^2\left(\frac{cn}{n+1} - 1\right)^2\end{aligned}$$

(c) Hence, we must minimize

$$c \mapsto \frac{nc^2}{(n+2)(n+1)^2} + \left(\frac{cn}{n+1} - 1\right)^2$$

Differentiation with respect to c gives

$$\frac{2cn}{(n+2)(n+1)^2} + 2\left(\frac{cn}{n+1} - 1\right)\frac{n}{n+1} = \frac{2cn}{(n+2)(n+1)^2} + \frac{2cn^2}{(n+1)^2} - \frac{2n}{n+1}.$$

Setting this equal to zero and solve for c, is equivalent with solving

$$\begin{aligned}\frac{c}{(n+2)(n+1)} + \frac{cn}{(n+1)} = 1 &\Leftrightarrow c(1+n(n+2)) = (n+2)(n+1) \\ &\Leftrightarrow c(n+1)^2 = (n+2)(n+1) \\ &\Leftrightarrow c = \frac{n+2}{n+1}\end{aligned}$$

To see that this is a minimum, compute the second derivative with respect to c:

$$\frac{2n}{(n+2)(n+1)^2} + \frac{2n^2}{(n+1)^2} > 0,$$

so that $c = (n+2)/(n+1)$ minimizes the mean squared error of $cX_{(n)}$.

3. (a) This is Exercise 3.10.
We set-up the loglikelihood

$$\log L(\theta; X_1, \ldots, X_n) = \log\left(\prod_{i=1}^n \theta X_i^{\theta-1}\right) = n\log\theta + (\theta-1)\sum_{i=1}^n \log X_i$$

To find the maximizer, we differentiate with respect to θ and set it equal to zero:

$$\frac{\mathrm{d}L}{\mathrm{d}\theta} = \frac{n}{\theta} + \sum_{i=1}^n \log X_i = 0 \Leftrightarrow \theta = -\frac{n}{\sum_{i=1}^n \log X_i}$$

To see whether this is a maximum, take the second derivative:

$$\frac{\mathrm{d}^2L}{\mathrm{d}\theta^2} = -\frac{n}{\theta^2} < 0,$$

which means that it is a maximum, so that the maximum likelihood estimator for θ is given by

$$\widehat{\theta} = -\frac{n}{\sum_{i=1}^{n} \log X_i}.$$

(b) This is Exercise 3.10(i) and 3.33(i).
We first determine the expectation of the X_i:

$$\mathrm{E}_\theta X_i = \int_0^1 x\theta x^{\theta-1}\,\mathrm{d}x = \theta \int_0^1 x^\theta\,\mathrm{d}x = \theta\left[\frac{x^{\theta+1}}{\theta+1}\right]_{x=0}^{x=1} = \frac{\theta}{\theta+1}$$

Next, we find the method of moments estimator, by setting the expectation equal to the sample mean and solve for θ:

$$\overline{X} = \mathrm{E}_\theta X_1 = \frac{\theta}{\theta+1} \Leftrightarrow (\theta+1)\overline{X} = \theta \Leftrightarrow \overline{X} = \theta(1-\overline{X}) \Leftrightarrow \theta = \frac{\overline{X}}{1-\overline{X}}.$$

The method of moments estimator is given by $\overline{X}/(1-\overline{X})$.

4. This is Example 3.39.

(a) The posterior density is given by

$$p_{\overline{\Theta}=\theta|X=x}(\theta) = \frac{p_\theta(x)\pi(\theta)}{\int_0^1 p_\vartheta(x)\pi(\vartheta)\,\mathrm{d}\vartheta}$$

where $\pi(\theta) = 6\theta(1-\theta)$, for $\theta \in [0,1]$, and

$$p_\theta(x) = \prod_{i=1}^{n} \theta(1-\theta)^{x_i-1} = \theta^n(1-\theta)^{\sum_{i=1}^n x_i - n} = \theta^n(1-\theta)^{n(\overline{x}-1)}$$

This leads to

$$p_{\overline{\Theta}=\theta|X=x}(\theta) = \frac{\theta^n(1-\theta)^{n(\overline{x}-1)}\cdot 6\theta(1-\theta)}{\int_0^1 p_\vartheta(x)\pi(\vartheta)\,\mathrm{d}\vartheta} = \frac{\theta^{n+1}(1-\theta)^{n(\overline{x}-1)+1}}{C(x_1,\ldots,x_n)}$$

This density is proportional to $\theta^{\alpha-1}(1-\theta)^{\beta-1}$, with $\alpha = n+2$ and $\beta = n(\overline{x}-1)+2$, and must therefore be a beta(α,β) distribution, with $\alpha = n+2$ and $\beta = n(\overline{x}-1)+2$.

(b) The Bayes estimator is the expectation of the posterior distribution (Theorem 3.36). If follows directly from part (a) and the distribution-sheet that the Bayes estimator is given by

$$T(X) = \frac{\alpha}{\alpha+\beta} = \frac{n+2}{n\overline{X}+4}$$

When $n \to \infty$, then

$$T(X) = \frac{n+2}{n\overline{X}+4} = \frac{1+2/n}{\overline{X}+4/n} \to \frac{1}{\overline{X}}.$$

5. (a) This is Example 4.6.

According to the law of large numbers, $\overline{X}$ will be close to the expectation of the X_i's, which is μ. Therefore, small values of $T = \overline{X}$ are in favor of $H_0 : \mu \leq 0$, and large values of $T = \overline{X}$ are in favor of $H_1 : \mu > 0$. This means that we reject $H_0 : \mu \leq 0$ for large values of $T = \overline{X}$, so that critical region is of the form $K_T = [c_{\alpha_0}, \infty)$.

(b) This is similar to Example 4.12; see also Exercise 4.10.

The size of the test is given by

$$\alpha = \sup_{\mu \leq 0} P_\mu \left(T \in K_T\right) = \sup_{\mu \leq 0} P_\mu \left(\overline{X} \geq c_{\alpha_0}\right).$$

Because $\overline{X} \sim N(\mu, 1/n)$, it follows that $\sqrt{n}(\overline{X} - \mu) \sim N(0,1)$. Therefore

$$\begin{aligned} P_\mu \left(\overline{X} \geq c_{\alpha_0}\right) &= P_\mu \left(\sqrt{n}(\overline{X} - \mu) \geq \sqrt{n}(c_{\alpha_0} - \mu)\right) \\ &= P\left(Z \geq \sqrt{n}(c_{\alpha_0} - \mu)\right), \quad Z \sim N(0,1) \\ &= 1 - \Phi(\sqrt{n}(c_{\alpha_0} - \mu)) \end{aligned}$$

which is increasing in μ. As a consequence

$$\begin{aligned} \alpha &= \sup_{\mu \leq 0} P_\mu \left(T \in K_T\right) = P_{\mu=0} \left(\overline{X} \geq c_{\alpha_0}\right) \\ &= P_{\mu=0} \left(\sqrt{n}\overline{X} \geq \sqrt{n} c_{\alpha_0}\right) = P\left(Z \geq \sqrt{n} c_{\alpha_0}\right), \quad Z \sim N(0,1) \\ &= 1 - \Phi(\sqrt{n} c_{\alpha_0}). \end{aligned}$$

(c) This is similar to Example 4.12; see also Exercise 4.10.

We must choose c_{α_0} such that the size of the test is less than α_0, i.e.,

$$1 - \Phi(\sqrt{n} c_{\alpha_0}) \leq \alpha_0 \Leftrightarrow \Phi(\sqrt{n} c_{\alpha_0}) \geq 1 - \alpha_0.$$

A simple picture of the density of the $N(0,1)$ distribution shows that

$$\Phi(\sqrt{n} c_{\alpha_0}) \geq 1 - \alpha_0 \Leftrightarrow \sqrt{n} c_{\alpha_0} \geq \xi_{1-\alpha_0} \Leftrightarrow c_{\alpha_0} \geq \xi_{1-\alpha_0}/\sqrt{n}.$$

Since, the critical region is supposed to be chosen as large as possible, this means that $c_{\alpha_0} = \xi_{1-\alpha_0}/\sqrt{n}$.

Uitwerkingen Midterm TW2080

Rik Lopuhaa

October 3, 2018

QUESTION 1

```
library(car)
library(MASS)
```

1A

```
summary(States$percent)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    4.00   11.50   25.00   33.75   57.50   74.00
```

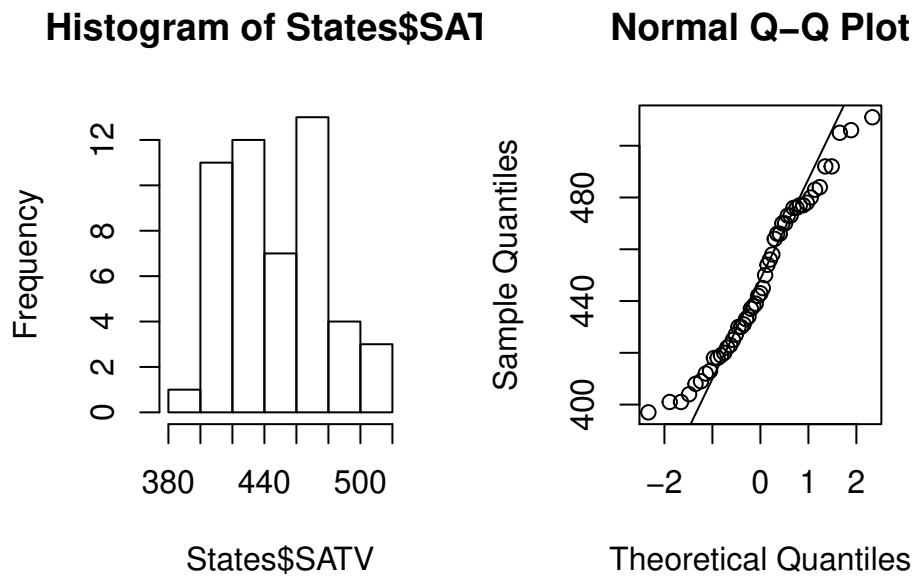
```
length(States$percent[States$percent<=60])
```

```
## [1] 42
```

The minimum percentage is 4 and the maximum percentage is 74. There are 42 states with a percentage less than 60%

1B

```
par(mfrow=c(1,2))
hist(States$SATV)
qqnorm(States$SATV)
qqline(States$SATV)
```




```
par(mfrow=c(1,1))
```

1C

```
mean(States$SATV)
```

```
## [1] 448.1569
```

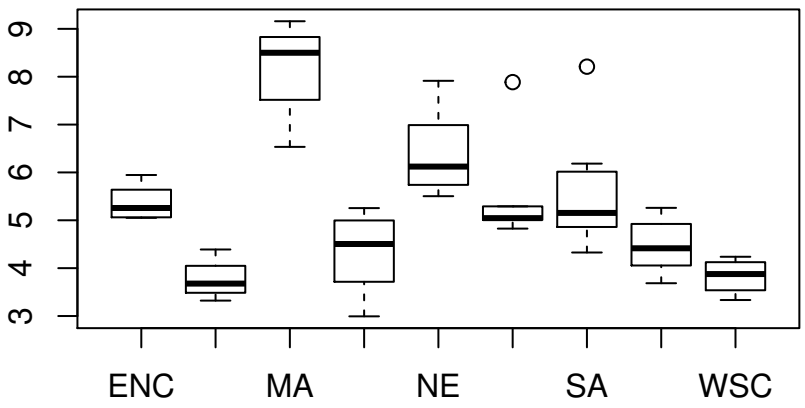
```
var(States$SATV)
```

```
## [1] 949.9349
```

The sample mean of SATV is 448.1569 The sample variance of SATV is 949.9349

(D)

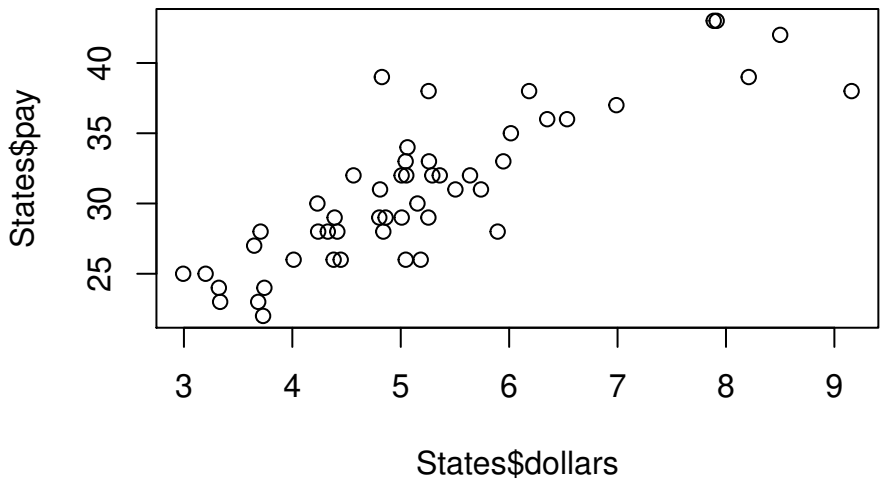
```
boxplot(split(States$dollars,States$region))
```



MA has the highest amount of dollars spend. In the help-file of States, one can see that MA stands for Mid-Atlantic.

(E)

```
plot(States$dollars,States$pay)
```




```
cor(States$dollars,States$pay)
```

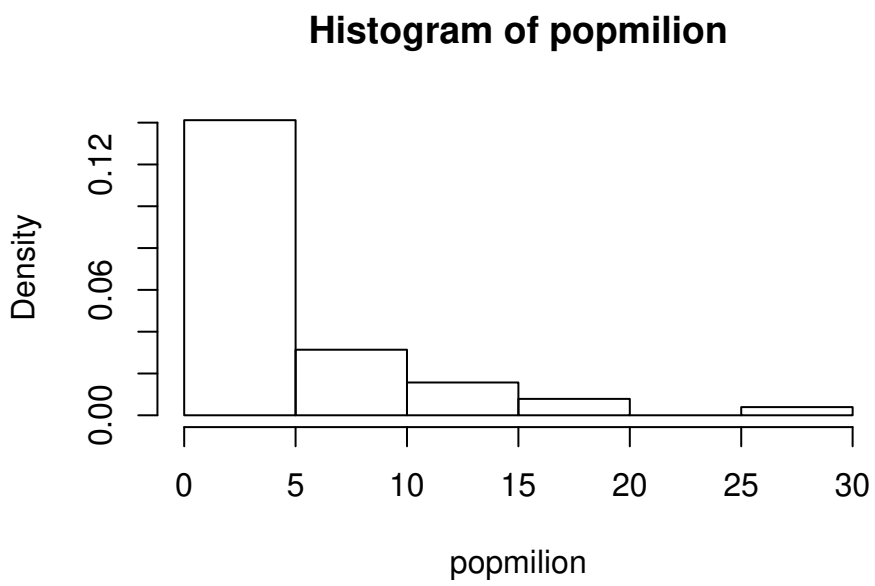
```
## [1] 0.8476737
```

The sample correlation coefficient is 0.8476737

QUESTION 2

(2A)

```
popmilion=States$pop/1000
hist(popmilion,prob=TRUE)
```



(2B)

```
ML=fitdistr(popmilion,densfun="gamma")$estimate
ML
```

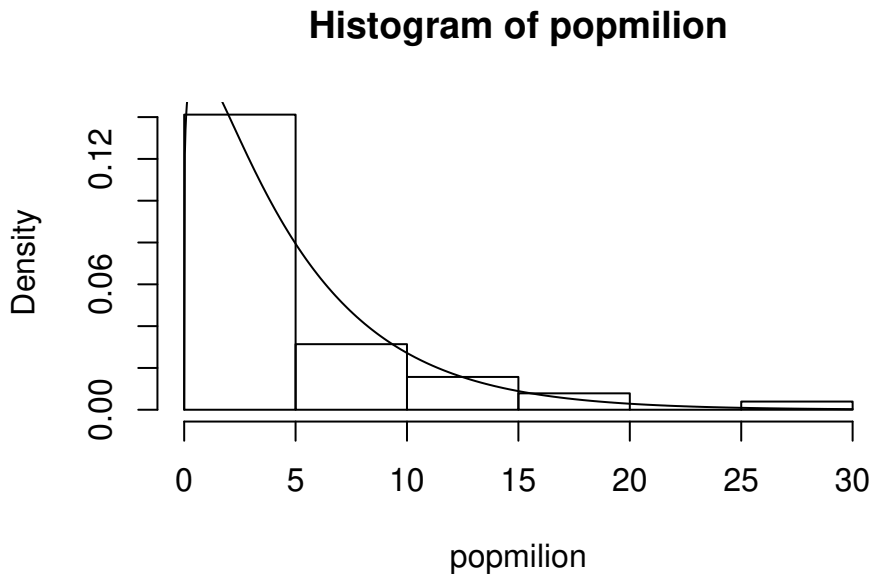
```
##     shape      rate
## 1.1435853 0.2345035
```

The maximum likelihood estimator for the shape parameter is 1.1435853 The maximum likelihood estimator for the rate parameter is 0.2345035

(2C)

```
hist(popmilion,prob=TRUE)
n=length(popmilion)
```

```
xas=seq(0,30,length=1000)
lines(xas,dgamma(xas,shape=ML[1],rate=ML[2]))
```

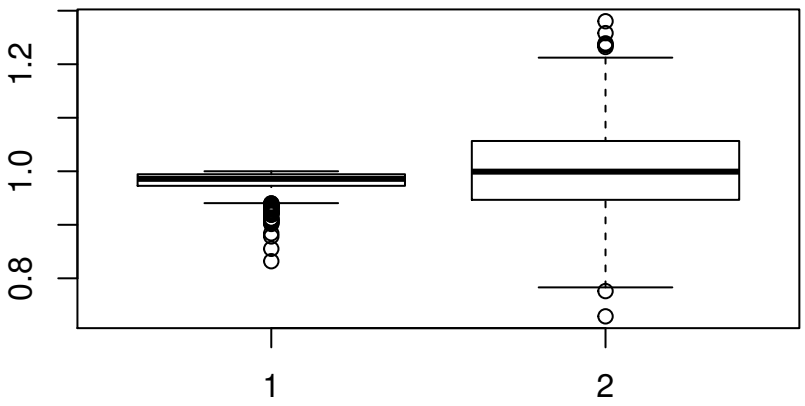


QUESTION 3

Preparation for the simulation

```
est1=numeric(1000)
est2=numeric(1000)
for (i in 1:1000){
  xdata=runif(50)
  est1[i]=max(xdata)
  est2[i]=2*mean(xdata)
}
```

```
boxplot(est1,est2)
```




```
var(est1)+(mean(est1)-1)^2
```

```
## [1] 0.0007516543
```

```
var(est2)+(mean(est2)-1)^2
```

```
## [1] 0.006682353
```

The mean squared error of the maximum is 0.0008336496 The mean squared error of two times the sample mean is 0.006996427. The MSE of the second estimator is a factor 8 times higher, so one should prefer the first estimator.