FSML_Part 2_YuHsuanTING

Yu-Hsuan TING

Sep. 16 2019

Exercise 1:

(a). $X \sim \mathcal{N}(-1, 0.01)$ 0.01 is variance. Compute:

1. $P(X \leq -0.98)$
2. $P(X \leq -1.02)$
3. $P(X \geq -0.82)$
4. $P(X \in [-1.22; -0.96])$

```
#1
pnorm(-0.98,mean = -1,sd = sqrt(0.01))
```

```
## [1] 0.5792597
```

```
#2.
pnorm(-1.02,mean = -1,sd = sqrt(0.01))
```

```
## [1] 0.4207403
```

```
#3.
1-pnorm(-0.82,mean = -1,sd = sqrt(0.01))
```

```
## [1] 0.03593032
```

```
#4.
pnorm(-0.96,mean = -1,sd = sqrt(0.01))-pnorm(-1.22,mean = -1,sd = sqrt(0.01))
```

```
## [1] 0.6415183
```

(b). $X \sim \mathcal{N}(0, 1)$ determine t such that:

1. $P(X \leq t) = 0.9$
2. $P(X \leq t) = 0.2$
3. $P(X \in [-t, t]) = 0.95$

```
#1.
qnorm(0.9)
```

```
## [1] 1.281552
```

```r
#2.
qnorm(0.2)
```

```
## [1] -0.8416212
```

```r
#3.
qnorm(1-(1-0.95)/2)
```

```
## [1] 1.959964
```

Exercise 2:

(a) Give the definition of a density function f_d

For continuous variable we use density function, we need to define first the number of class and the class range

table class	relative frequence	density
$[\rho_1, \rho_2[$	f_1	d_1
$[\rho_2, \rho_3[$	f_2	d_2
...	...	...
$[\rho_k, \rho_{k+1}[$	f_k	d_k

where $f_i = P(x \in [\rho_i, \rho_{i+1}])$ and $d_i = \frac{f_i}{\rho_{i+1} - \rho_i}$

Therefore density function is define as:

$$
\begin{aligned}
f_d(x) = d_i \quad &\text{if} \quad t \in [\rho_i, \rho_{i+1}[\\
0 \quad & \quad otherwise
\end{aligned}
$$

- $\forall x \in \mathbb{R} \quad f_d(x) \geq 0$
- $\int f_d(x)dx = 1$

(b) Let θ_n an estimator of a parameter θ. Give the definition of θ_n an unbiased estimator of θ.

we say that θ_n is an unbiased estimator of θ if $\mathbb{E}[\theta_n] = \theta$ (expectation of θ_n is θ)

(c) Let $X_1, ..., X_n$ a n-sample. We denote by μ the expectation of X_1 and σ^2 its variance. Let $\overline{X_n}$ the empirical mean associated. Compute the expectation and the variance of $\overline{X_n}$.

note that $\mathbb{E}[X_i] = \mu$ *and* $V[X_i] = \sigma^2$

$\overline{X_n} = \frac{1}{n}\sum_{i=1}^{n} X_i$ we see that $\overline{X_n}$ just depend on $X_1, ..., X_n$ so it is an estimator

$\mathbb{E}[\overline{X_n}] = \mathbb{E}[\frac{1}{n}\sum_{i=1}^{n} X_i] = \frac{1}{n}\sum_{i=1}^{n} \mathbb{E}[X_i] = \frac{1}{n} \times n\mu = \mu$ (here we understand that $\overline{X_n}$ is an unbiased estimator for μ)

$V[\overline{X_n}] = V[\frac{1}{n}\sum X_i] = \frac{1}{n^2}\sum V[X_i] = \frac{1}{n^2} \times n\sigma^2 = \frac{\sigma^2}{n}$

(d) Let $X_1, ..., X_n$ a n-sample with a $\mathcal{N}(\mu, \sigma^2)$ distribution. Give an unbiased estimator of σ^2 when we assume that μ is unknown. Prove the fact that it is unbiased.

note that $\mathbb{E}[X_i] = \mu$ *and* $V[X_i] = \sigma^2$

$\hat{\sigma_n^2}$ is an estimator because it's just a function of $X_1, ..., X_n$, and we can compute the expectation of $\hat{\sigma_n^2}$

we denote that:

$V[X_i] = \mathbb{E}[X_i^2] - (\mathbb{E}[X_i])^2 \quad \text{so} \quad \sum \mathbb{E}[X_i^2] = n(\sigma^2 + \mu)$

$V[\overline{X_n}] = \mathbb{E}[\overline{X_n}^2] - (\mathbb{E}[\overline{X_n}])^2 \quad \text{so} \quad \sum \mathbb{E}[\overline{X_n}^2] = n(\frac{\sigma^2}{n} + \mu) \quad \text{(refer to (c))}$

$\sum X_i = n\overline{X_n} \quad \text{so} \quad \sum 2X_i\overline{X_n} = 2n\overline{X_n}^2$

We can now compute the following:

$\mathbb{E}[\hat{\sigma_n^2}] = \mathbb{E}[\frac{1}{n}\sum_{i=1}^{n}(X_i - \overline{X_n})^2]$

$= \frac{1}{n}\mathbb{E}[\sum X_i^2 - \sum 2X_i\overline{X_n} + \sum \overline{X_n}^2]$

$= \frac{1}{n}(\mathbb{E}[\sum X_i^2] - \mathbb{E}[\sum 2X_i\overline{X_n}] + \mathbb{E}[\sum \overline{X_n}^2])$

$= \frac{1}{n}(\mathbb{E}[\sum X_i^2] - 2n\mathbb{E}[\sum \overline{X_n}^2] + \mathbb{E}[\sum \overline{X_n}^2])$

$= (\sigma^2 + \mu) - 2\mathbb{E}[\overline{X_n}^2] + \mathbb{E}[\overline{X_n}^2]$

$= (\sigma^2 + \mu) - (\frac{\sigma^2}{n} + \mu) = \frac{n-1}{n}\sigma^2$

we see from the equation $\mathbb{E}[\hat{\sigma_n^2}] = \frac{n-1}{n}\sigma^2 \neq \sigma^2$ so it is not an unbiased estimator, although $\frac{n-1}{1} \to 1 \quad \text{when} \quad n \to \infty$ we can say that $\hat{\sigma_n}^2$ is asymptotically an unbiased estimator of σ^2

we can do a linear transformation for our estimator (it would still be an estimator) $\mathbb{E}[\hat{\sigma_n^2}] = \frac{n-1}{n}\sigma^2 \quad \text{to} \quad \mathbb{E}[\frac{n}{n-1}\hat{\sigma_n^2}] = \sigma^2$ .Therefore we can say that $\frac{n}{n-1}\hat{\sigma_n^2}$ is an unbiased estimator of σ^2

Exercise 3:

- read table into T1

```
T1=read.table('dataexam.txt')
head(T1)
```

```
##           V1
## 1 0.00000000
## 2 0.03811531
## 3 0.20292690
## 4 0.31850700
## 5 0.89276500
## 6 1.04180300
```

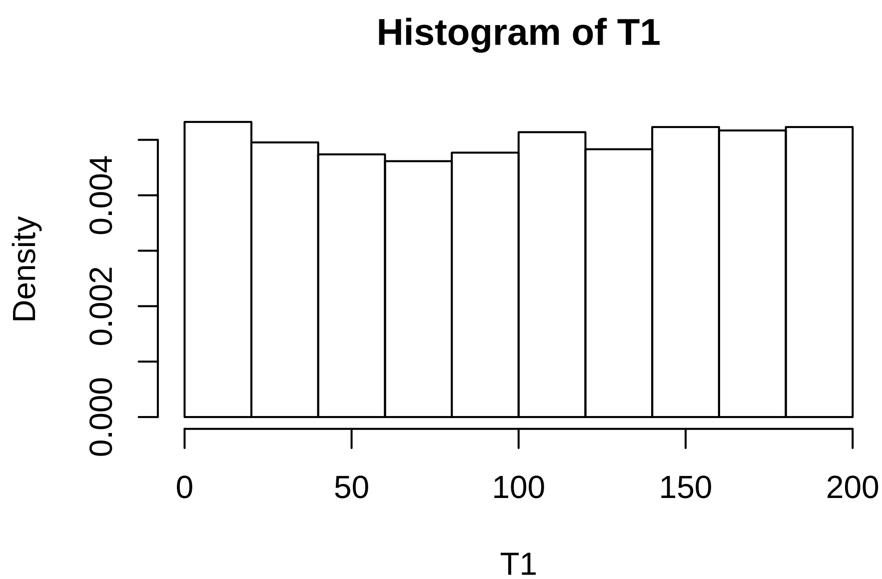
```
dim(T1)
```

```
## [1] 1625    1
```

(a) Make a test to show that those times are distributed according to a uniform distribution.

- draw a histogram

```
T1=as.matrix(T1)
hist(T1,freq=FALSE)
```



Uniform distrivurion $X \sim \mathrm{U}(a, b)$ where a is the lowest of x and b is the highest value of x

with density function $f(x) = \frac{1}{b-a}$ for $a \leq x \leq b$

theoretical mean and sd are $\mu = \frac{a+b}{2}$ and $\sigma = \sqrt{\frac{(b-a)^2}{12}}$

- all the value is between

```
maxT1=max(T1[,1])
minT1=min(T1[,1])
```

- now we compute the sample mean

```
sm=mean(T1[,1])
ssd=sd(T1[,1])
sm
```

```
## [1] 100.8505
```

```
ssd
```

```
## [1] 58.42301
```

- we check the theoretical mean and sd, it is very close to the sample true mean and sd

```
tm=(maxT1+minT1)/2
tsd=sqrt((maxT1-minT1)**2/12)
tm
```

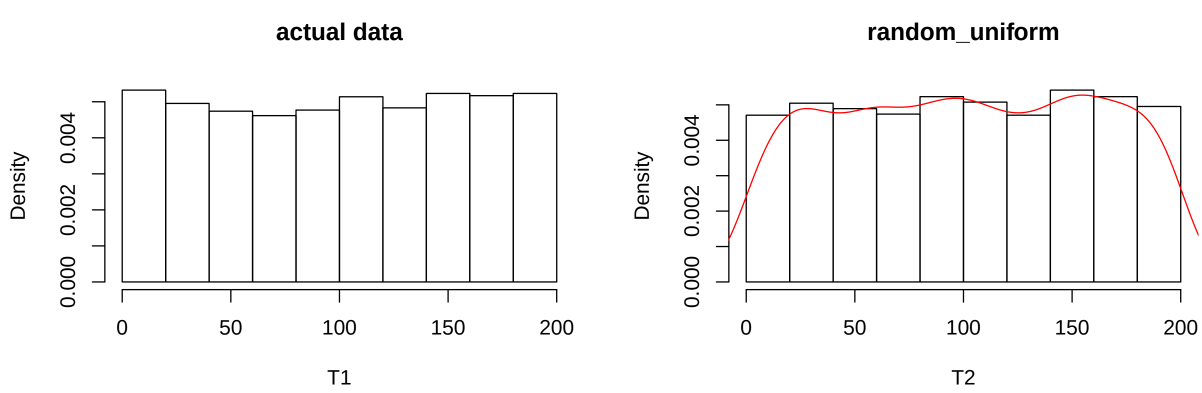
```
## [1] 99.977
```

```
tsd
```

```
## [1] 57.72175
```

We see from the graph generate random uniform distribution it looks similar as our dataset, that we can say our dataset is uniform distribution

```
T2=runif(1625, minT1, maxT1)
par(mfrow=c(1,2))
hist(T1,freq=FALSE,main="actual data")
hist(T2, freq = FALSE, main="random_uniform")
lines(density(T2),col='red')
```



```r
#density(T2)
```

We can also check by `ks.test` to perform a one- or two-sample Kolmogorov-Smirnov test.

Here the null hypothesis is that the data follow a uniform distribution. We see that the p-value is high so that we don't reject the null hypothesis.

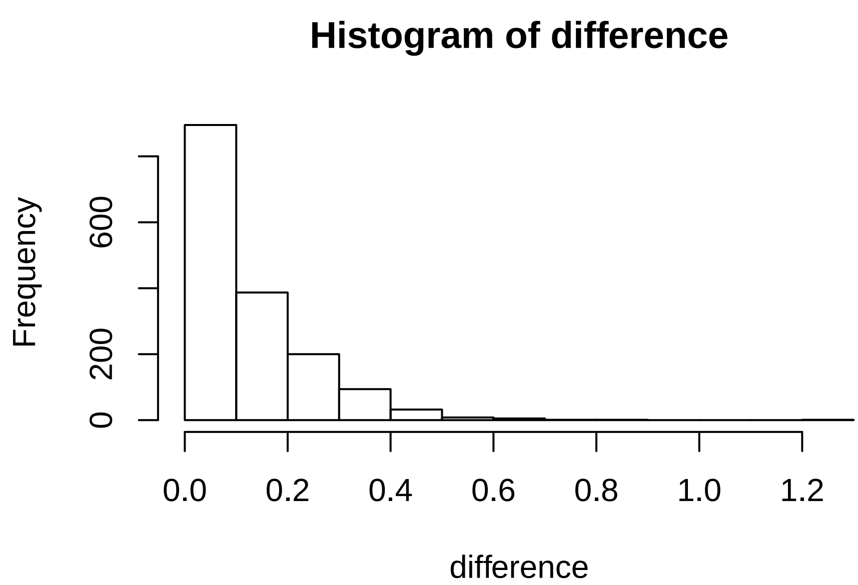
```r
ks.test(T1,T2)
```

```
##
##  Two-sample Kolmogorov-Smirnov test
##
## data:  T1 and T2
## D = 0.013538, p-value = 0.9984
## alternative hypothesis: two-sided
```

(b) Now if we consider the time between two events, how can you modelize this distribution?

draw the histogram of the difference between the data, it seems to be exponential distribution

```r
difference=diff(T1[,1])
hist(difference)
```



if $x \sim E(\lambda)$, $\mathbb{E}[x] = \frac{1}{\lambda}$

```r
#actual
m=mean(difference)
m
```

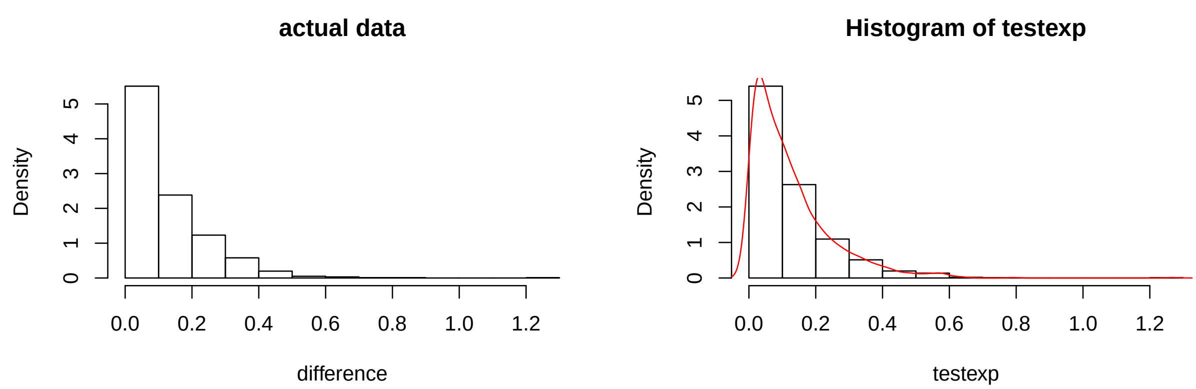
```
## [1] 0.1231244
```

```r
#theoretical
lambda=1/m
lambda
```

```
## [1] 8.121868
```

```r
#compare 2 histogram

par(mfrow=c(1,2))
hist(difference,freq=FALSE,main="actual data")
#curve(dexp(lambda))
testexp=rexp(n=length(difference),lambda)
hist(testexp,freq = FALSE)
#curve(dexp,xlim = c(0,1.2),add = TRUE)
lines(density(testexp),col='red')
```



We can also use `ks.test` to check the goodness of fit for exponential distribution. As the p-value is high we also don't reject the null hypothesis that is the difference is exponential distribution.

```r
ks.test(difference,testexp)
```

```
## Warning in ks.test(difference, testexp): p-value will be approximate in the
## presence of ties
```

```
##
##  Two-sample Kolmogorov-Smirnov test
##
## data:  difference and testexp
## D = 0.025246, p-value = 0.6787
## alternative hypothesis: two-sided
```

(c) Guess the value of the parameter and compute a confidence interval for it.

we assum that lambda is 8.121868, we are going to compute the confidence interval

- Thanks to Central Limit theorem we have $\sqrt{n}\frac{\overline{x_n}-\mu}{\sigma} \sim \mathcal{N}(0,1)$

- we denote that:
    1. $\mu = \mathbb{E}[x]$ which is $\frac{1}{\lambda}$ in exponential distribution
    2. $\sigma = \sqrt{V[x]}$ which is $\sqrt{\frac{1}{\lambda^2}}$
- so we can get: $\sqrt{n}(\lambda\overline{x_n} - 1) \sim \mathcal{N}(0,1)$
- we define S and T such that: $\mathrm{P}(\sqrt{n}(\lambda\overline{x_n} - 1) \in [S,T]) = 1-\alpha$
- we make a choice that S=-T, so T is the fractile of $1-\frac{\alpha}{2}$
- it means that $P(\mathcal{N}(0,1) \leq t) = 1-\frac{\alpha}{2}$
  $\mathrm{P}(\sqrt{n}(\lambda\overline{x_n} - 1) \in [-T,T]) = 1-\alpha$
  $\Leftrightarrow P(\frac{1}{\bar{x}}(1-\frac{t}{\sqrt{n}}) \leq \lambda \leq \frac{1}{\bar{x}}(1+\frac{t}{\sqrt{n}}))$ with $\alpha = 0.05$

```
t=qnorm(0.975) #1-0.05/2
lb=1/m*(1-t/sqrt(length(difference)))
ub=1/m*(1+t/sqrt(length(difference)))
cat("confidence interval:[",c(lb,ub),"]")
```

```
## confidence interval:[ 7.726855 8.516881 ]
```

Exercise 4:

Let X be a random variable whose distribution is an exponential with parameter $\lambda > 0$

(a) We define the conditional probability $\mathrm{P}(A \mid B)$ by:

$$\mathrm{P}(A \mid B) = \frac{\mathrm{P}(A \cap B)}{\mathrm{P}(B)}$$

if $P(B) \neq 0$ Prove that the exponential random variable is with no memory which means:

$$\forall s,t > 0, \quad \mathrm{P}(X > t+s \mid X > t) = P(X > s)$$

- for exponential cumulative probability $P(X < x) = 1 - e^{-\lambda x}$ that is $P(X > x) = e^{-\lambda x}$
- equations:
  $P(X > t+s \mid X > t) = \frac{P(X>t+s \quad \cap \quad X>t)}{P(X>t)} = \frac{P(X>t+s)}{P(x>t)}$
  $\Leftrightarrow \frac{e^{-\lambda(t+s)}}{e^{-\lambda t}} = \frac{e^{-\lambda t} \times e^{-\lambda s}}{e^{-\lambda t}} = e^{-\lambda s}$
  $\Leftrightarrow P(X > s)$

here since it is exponential distribution and s,t > 0, we can say that $P(X > t+s \cap X > t) = P(X > t+s)$

- let's try with the code, set lambda=1, random choose an integer between 1 to 5 for t and s

```r
#random choose t and s
set.seed(10)
t=sample(1:5,1)
s=sample(1:5,1)
t
```

```
## [1] 3
```

```r
s
```

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

```r
(1-pexp(t+s))/(1-pexp(t))
```

```
## [1] 0.3678794
```

```r
1-pexp(s)
```

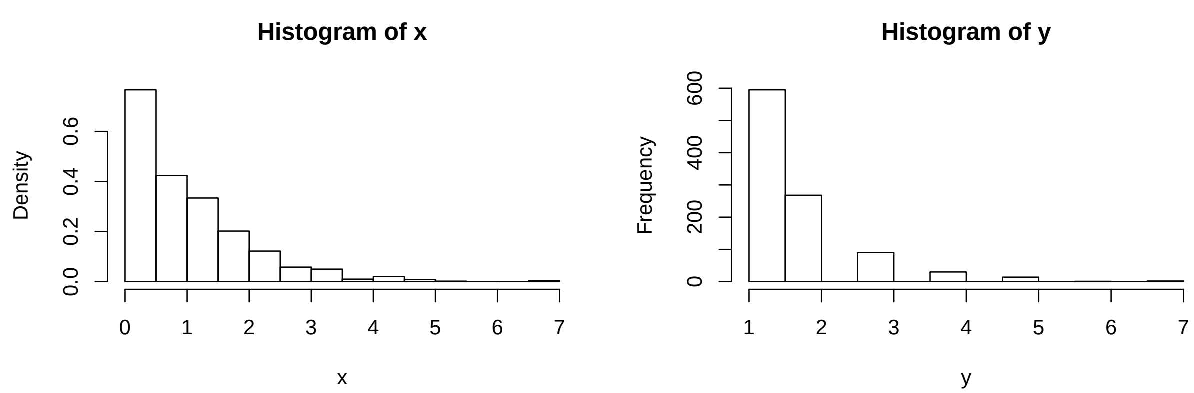
```
## [1] 0.3678794
```

(b) Let's consider $Y = E(x) + 1$ where E(x) is the biggest integer smaller or equal to x

let's try with 1000 random data, here we can see that our data is not anymore continuous. Moreover, y cannot be 0 the smallest is 1, we can assum it to be geometric distribution.

```r
set.seed(10)
par(mfrow=c(1,2))
x=rexp(1000)
hist(x, freq=FALSE)

y=floor(x)+1
hist(y)
```

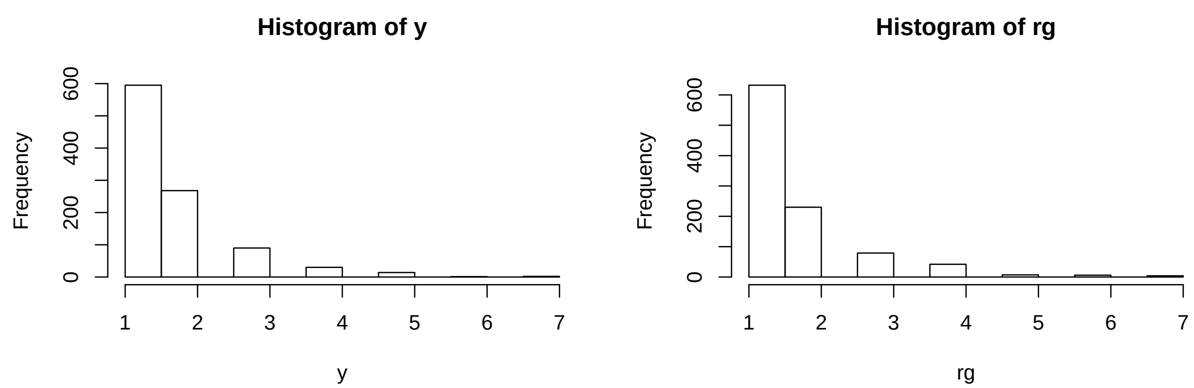


```r
table(y)
```

```
## y
##   1   2   3   4   5   6   7
## 595 268  90  30  14   1   2
```

with Geometric distribution, $\mu = \frac{1}{p}$ therefor we get p=0.6207325 and we do a random 1000 data from geometric distribution to see if it looks like our y. But `rgeom` start at 0 so we can add the value all to 1

```r
gp=1/mean(y)
par(mfrow=c(1,2))
hist(y)
rg=rgeom(1000,gp)+1
hist(rg)
```



We can then check the distribution from `ks.test`, but since ks.test is meant to use for continouse variable, therefore I use `chisq.test` instead and it suggest me that it is a geometric distribution.

```r
#ks.test(y,rg)
chisq.test(y,rg)
```

```
## Warning in chisq.test(y, rg): Chi-squared approximation may be incorrect
```

```
##
##  Pearson's Chi-squared test
##
## data:  y and rg
## X-squared = 25.789, df = 36, p-value = 0.8962
```