Statistical Computing - Homework 1

Wentao Huang, 1303160224, Statistics 1602

2019/03/09

Problem 1

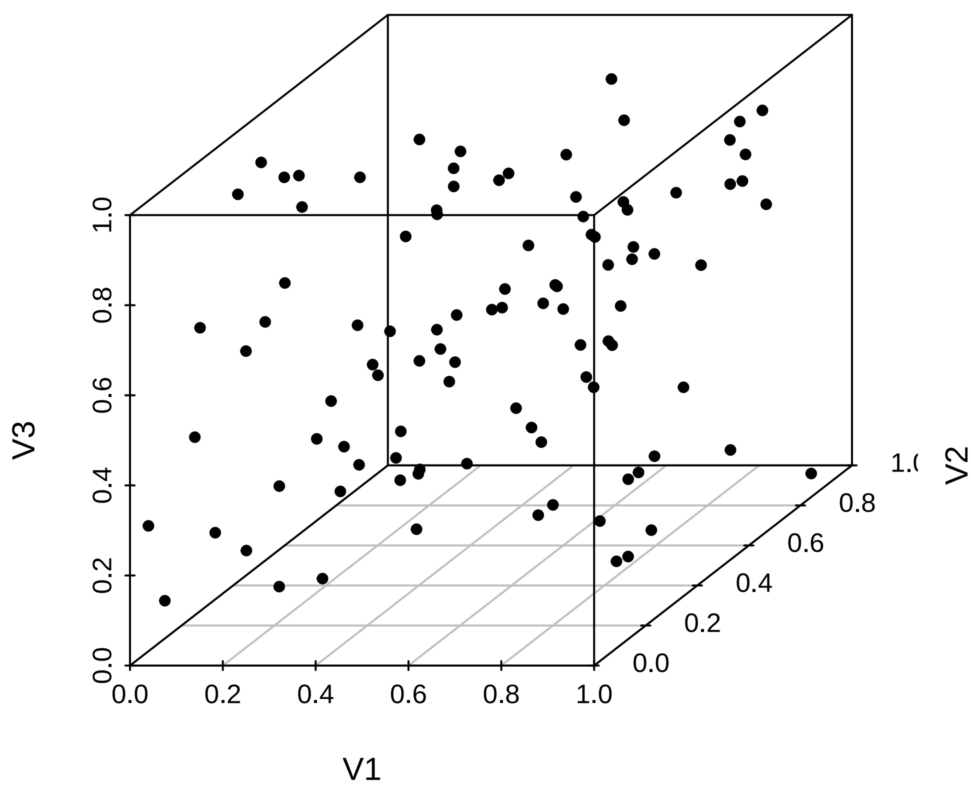
Simulate a sequence of uniform random numbers with R function `runif()`. Make 3-tuples with every 3 consecutive numbers from the sequence. Visualize the tuples in the cube $[0,1]^3$ and test the uniformity of these tuples with Pearson's χ^2 test.

Solution:

We can solve the problem with R, here are codes with comments below:

```r
# generate 3-tuples dataframe
set.seed(666)
dat1 = as.data.frame(array(runif(300),c(100,3)))

# plot in 3d
## install.packages("scatterplot3d")
library(scatterplot3d)
scatterplot3d(dat1,xlim=c(0,1),ylim=c(0,1),zlim=c(0,1),pch=20)
```



```r
# Pearson's chisq-test
chisq.test(dat1)
```

```
## Warning in chisq.test(dat1): Chi-squared approximation may be incorrect
```

```
##
##  Pearson's Chi-squared test
##
## data:  dat1
## X-squared = 36.87, df = 198, p-value = 1
```

The p-value of Pearson's test is almost equal to 1, which means the generated data follows the Uniform distribution.

Problem 2

Prove that $F(X) \sim U(0,1)$ for a continuous random variable X with CDF $F(\cdot)$. And explain why this is not true if X follows a discrete distribution.

Proof:

We suppose the random variable $X \sim F(\cdot)$ is continuous and it's CDF is arbitrary, it's PDF $f(t)$ must exist. We denote it's CDF as:

$$F(X) = P\{T \leq t\} = \int_{-\infty}^{X} f(t)dt$$

$$P\{T \leq t\} = P\{F(X) \leq t\} = P\{X \leq F^{-1}(t)\} = F[F^{-1}(t)] = t$$

which is monotone increasing and continuous with domain $(-\infty, \infty)$ and range $[0,1]$. Hence the CDF's inverse function F^{-1} must exist. $F(x)$ must be with domain $[0,1]$ and range $[0,1]$.

Therefore $F(x) \sim U(0,1)$

In the discrete case, there is no continuous PDF and inverse function, therefore this is not true if X follows a discrete distribution.

Q.E.D.

Problem 3

A fault-tolerant memory bank is built with 5 memory units. These units have independent random failure times, each with a distribution F. The memory bank is operable as long as 4 or more units are still working, so it fails when the second unit fails. Let $G(\cdot|\alpha, \beta)$ be the CDF of the $Beta(\alpha, \beta)$.

a)

Express the failure time of the memory bank as a function of a single variable $U \sim U(0,1)$ via $F(\cdot)$ and $G(\cdot|\alpha, \beta)$ and/or their inverses.

Solution:

From prob.3.a, we denote T as the failure times of bank, and X_i as the failure times of each unit. It is obviously that $X_i \sim F$.

When the second unit breaks down, the bank fails with $T = X_{(2)}$. Since $X_i \sim U(0,1)$, $X_{(2)} \sim Beta(2,4)$ and $T = F^{-1}(U_{(2)})$, $U_2 \sim G^{-1}(U)$.

As a result, $T = F^{-1}(G^{-1}(U))$, where $\alpha = 2,\ \beta = 4$.

b)

Suppose that F is the exponential distribution with mean $500,000$ (hours). Sample $10,000$ memory bank lifetimes using the expression from part a). Estimate the mean memory bank lifetime and give a 99% confidence interval for it.

Solution:

We can obtain that the parameters of Beta distribution are $\alpha = 2, \beta = 4$.

```r
# generate 10000 exponential variables via inverse transformation
set.seed(666)
n.unif = rbeta(10000,2,4)
n.exp = -(500000)*log(1-n.unif)
n.exp.bar = sum(n.exp)/10000
n.exp.bar
```

```
## [1] 227933.3
```

```r
# 99% confidence interval
chi.1 = qchisq(0.005, 10000, lower.tail = F)
chi.2 = qchisq(0.995, 10000, lower.tail = F)
10000*n.exp.bar/chi.1
```

```
## [1] 219842.3
```

```r
10000*n.exp.bar/chi.2
```

```
## [1] 236458.1
```

After computation, we conclude that the mean memory bank lifetime is 227933.3, and it's 99% confidence interval is $[219842.3, 236458.1]$.

c)

If we didn't use a redundant system, we might instead use 4 memory units and the bank would fail when the first of those failed. Find the expected lifetime of such a system without redundancy and give a 99% confidence interval. The redundant system takes 1.25 times as many memory units. Does it last at least 1.25 times as long on average?

Solution:

If we didn't use a redundant system, the paramaters of Beta distribution will turn to $\alpha = 1,\ \beta = 4$.

```r
# generate 10000 exponential variables via inverse transformation
set.seed(666)
n.unif = rbeta(10000,1,4)
n.exp2 = -(500000)*log(1-n.unif)
n.exp.bar2 = sum(n.exp2)/10000
n.exp.bar2
```

```
## [1] 122847.8
```

```r
# 99% confident interval
chi.1 = qchisq(0.005, 10000, lower.tail = F)
chi.2 = qchisq(0.995, 10000, lower.tail = F)
10000*n.exp.bar2/chi.1
```

```
## [1] 118487.1
```

```r
10000*n.exp.bar2/chi.2
```

```
## [1] 127442.4
```

```r
rate = n.exp.bar/n.exp.bar2
rate
```

```
## [1] 1.855411
```

After computation, we conclude that the mean memory bank lifetime without redundant system is 122847.8, and it's 99% confidence interval is $[118487.1, 127442.4]$.

And the rate between a) and b) is $1.855411 > 1.25$ therefore it doesn't last at least 1.25 times as long on average.

d)

Compare the speed of sampling by inversion with that of generating 5 exponential variables, sorting them and picking out the second smallest.

Solution:

```r
set.seed(666)
# case 1
t1.1 = Sys.time()
n.unif = rbeta(10000,2,4)
n.exp = -(500000)*log(1-n.unif)
n.exp.bar = sum(n.exp)/10000
t1.2 = Sys.time()
t1 = t1.2 - t1.1
t1
```

```
## Time difference of 0.006236792 secs
```

```r
# case 2
t2.1 = Sys.time()
n = 1
repeat {
   sort(rexp(5,1/500000))[2]
   n = n+1
   if(n==10000) {
```

```r
        break
    }
}
t2.2 = Sys.time()
t2 = t2.2 - t2.1
t2
```

```
## Time difference of 0.5268168 secs
```

The speed of inverse function method is about 0.01458049 seconds.

The speed of tranditional method is about 0.5428333 seconds.

Therefore, the first method is performing better than the second one.

Problem 4

For acceptance-rejection sampling,

a)

Show that the ratio $f(x)/g(x)$ of $N(0,1)$ to Cauchy densities is maximized at $x = \pm 1$.

Proof:

Since $f(x) \sim N(0,1), \quad g(x) \sim C(1,0)$, we can obtain their PDF that

$$f(x) = \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}}, \quad g(x) = \frac{1}{\pi(1+x^2)}$$

Then we denote that

$$h(x) = \frac{f(x)}{g(x)} = \sqrt{\frac{\pi}{2}}(1+x^2)e^{-\frac{x^2}{2}}$$

In order to find the maximum of $h(x)$, we should find the root of $h'(x)$

$$h'(x) = \sqrt{\frac{\pi}{2}} e^{-\frac{x^2}{2}} x\,(x+1)\,(x-1) = 0 \quad \Rightarrow \quad x = 0, \pm 1$$

$$\Rightarrow \quad h(0) = \sqrt{\frac{\pi}{2}} \quad < \quad h(\pm 1) = \sqrt{2\pi} e^{\frac{1}{2}}$$

Therefore, $x = \pm 1$ is the maximum of $h(x) = f(x)/g(x)$.

Q.E.D.

b)

Can we use acceptance-rejection with proposals from $N(0, 1)$ to sample from Cauchy distribution? Explain why or why not.

Answer:

No, we cannot use acceptance-rejection with proposals from $N(0, 1)$ to sample from Cauchy distribution.

Because Normal distribution has thin tails with the finite expectation μ and variance σ, on the contrary, Cauchy distribution has thick tails which it's expectation and variance are approaching to infinity. If we use AR method to pick the sample from $N(0, 1)$ out of the interval 3σ, in which appearently there are almostly no samples, but Cauchy distribution still has lots of samples there.

Therefore, we cannot use AR method with proposals from Normal distribution to sample from Cauchy distribution.