qbs120_ps4_gibran

Gibran Erlangga

10/7/2021

Question 1

(Based on Rice 6.3) Let $\bar{X}$ be the average of a sample of n independent standard normal RVs.

(a) Determine c such that $P(|\bar{X}| < c) = 0.5$. Solve for c as a function of n.

$$
\begin{aligned}
P(|\bar{X}| < c) &= 0.5 \\
P(-c < \bar{X} < c) &= 0.5 \\
P(\frac{-c-\mu}{1/\sqrt{n}} < \frac{\bar{X}-\mu}{1/\sqrt{n}} < \frac{c-\mu}{1/\sqrt{n}}) &= 0.5 \\
P(-c\sqrt{n} < Z < c\sqrt{n}) &= 0.5 \\
2\Phi(c\sqrt{n}) - 1 &= 0.5 \\
\Phi(c\sqrt{n}) &= 1.5/2 \\
\Phi(c\sqrt{n}) &= 0.75
\end{aligned}
$$

Using the Cumulative Normal Distribution table as reference, we get:

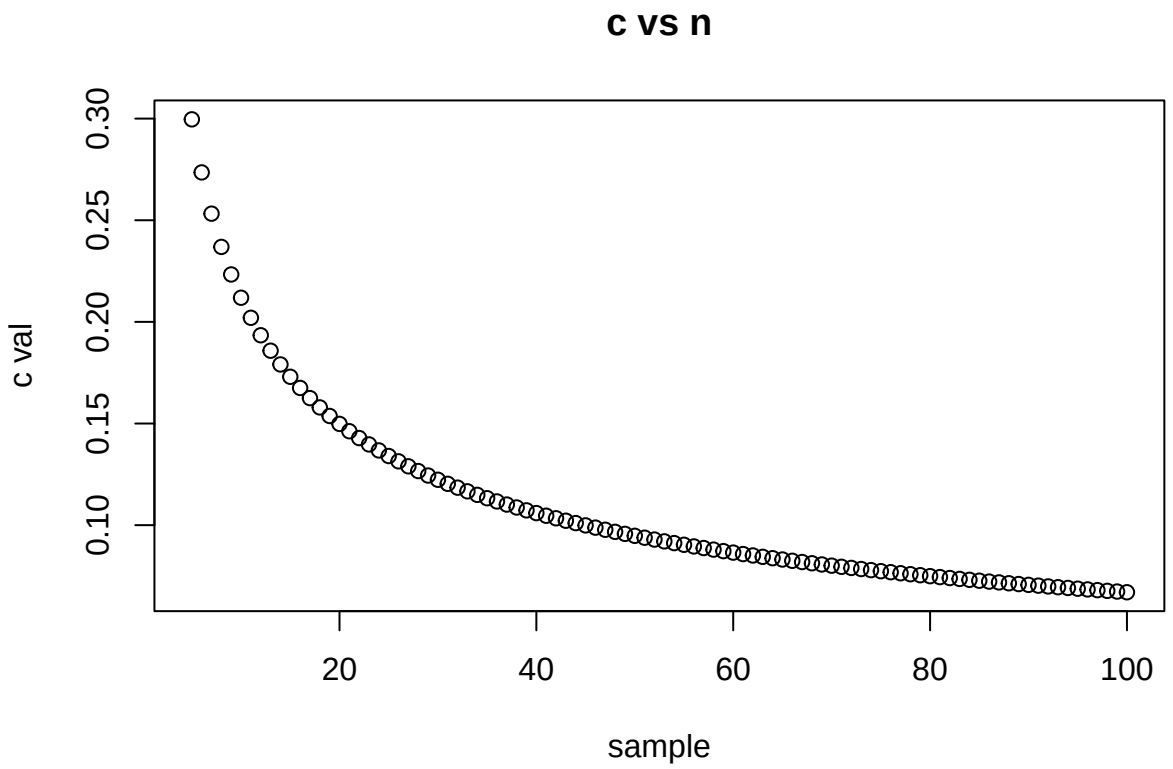
$$
\begin{aligned}
c\sqrt{n} &= 0.67 \\
c &= 0.67/\sqrt{n}
\end{aligned}
$$

(b) Using only R *norm() functions for the standard normal distribution, compute the exact value of c for n= 5,...100 and visualize as a plot of c vs. n.

```
sample = 5:100

get_c <- function(n_iter) {
  return(0.67/sqrt(n_iter))
}

plot(sample, get_c(sample), ylab = "c val", main="c vs n")
```



(c) If the variance was not known, how would you solve the problem and what additional piece of information would you need to get an exact answer?

We can find the sample variance using the sample and proceed to approximate $\bar{X}$ with t-distribution with n-1 degree of freedom.

(d) If the n RVs are independent and have the same distribution with expectation 0 and variance 1 but the exact distribution is not known, how would you approach the problem?

Approach the problem using CLT, with standard normal distribution ($\mu = 0, \sigma = 1$).

Question 2

(Based on Rice 6.6)

(a) Show that if $T \sim t_n$, then $T^2 \sim F_{1,n}$.

T is said to have a t distribution with n degrees freedom, and we know that:

$$T = \frac{Z}{\sqrt{V/n}}$$

with:
- $Z \sim N(0, 1)$
- $V \sim \chi^2$

So,

$$T^2 = \frac{Z^2}{V/n} = \frac{Z^2/1}{V/n}$$

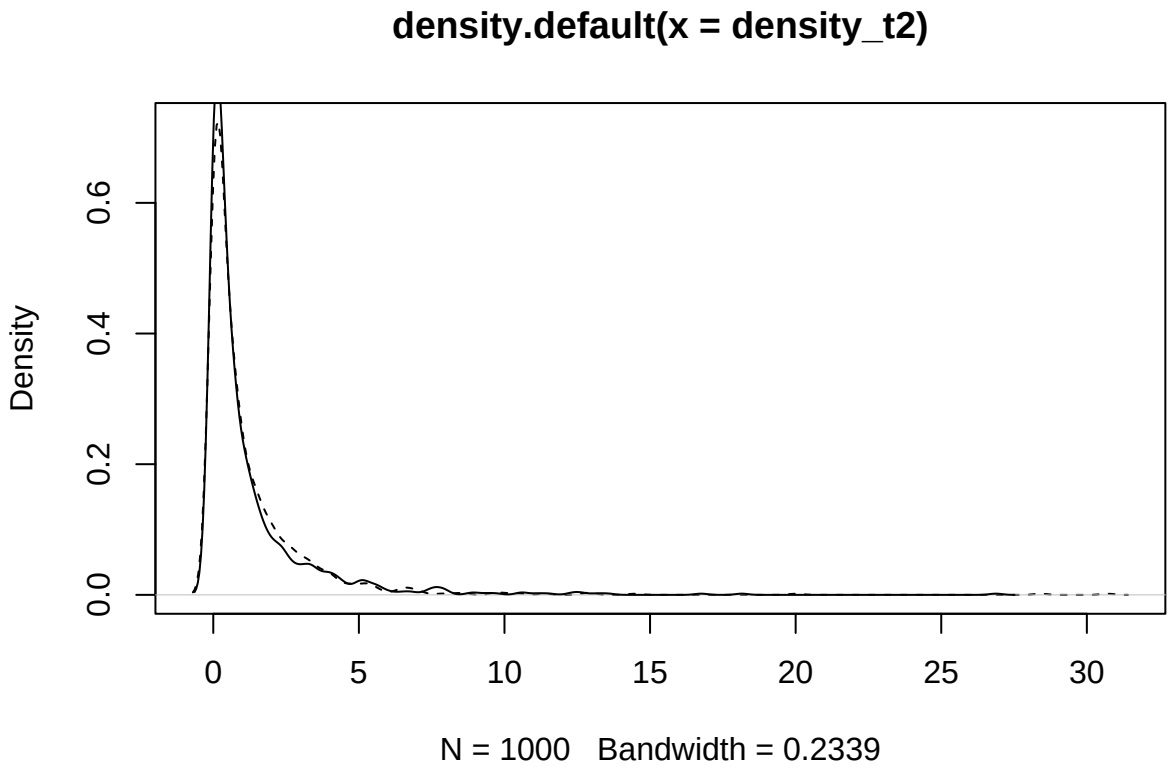
By definition, this is $F_{1,n}$ with $Z^2 = U$ with U as χ^2 distribution with 1 degree of freedom, and V as χ^2 distribution with n degree of freedom.

(b) For n=10, demonstrate this equivalence numerically by plotting the kernel density estimates for 1000 randomly generated T^2 values and 1000 randomly generated $F_{1,n}$ values.

```
n= 1000
deg = 10

density_t2 <- (rt(n, df=10))^2
density_f <- rf(n, df1=1, df2=deg)

plot(density(density_t2), type="l", lty="dashed")
lines(density(density_f))
```



Question 3

Optional

Question 4

(Based on Rice 7.3) Which of the following is a random variable? Justify your answers.
a) **The population mean is not a random variable**, because it is one of the parameters from the population. The value will always be the same given the same population.
b) **The population size (n) is not a random variable**. This is the number of data points in a given population.
c) **The sample size (n) is not a random variable**. This is the number of sample you choose to be taken from a particular population definition.
d) **The sample mean is a random variable**. Its value will be different for every iteration of the sampling, depending in the elements inside the sample result.
e) **The variance of the sample mean is not a random variable**, as this is one of the parameters of the sample mean.
f) **The largest value in the sample is a random variable**. This value differs depending on the elements inside the sample result.
g) **The population variance is not a random variable**. This value will remain constant given the same population.
h) **The estimated variance of the sample mean is a random variable**. The estimated variance is derived from an RV which will yield in a random value every time we do the estimation.

Question 5

(Based on Rice 7.4) Two populations are surveyed with simple random sampling. A sample of size n_1 is used for population I, which has a population standard deviation of σ_1; a sample of size $n_2 = 3n_1$ is used for population II, which has a population standard deviation of $\sigma_2 = 2\sigma_1$.

(a) Ignoring the finite population correction, in which of the two samples would you expect the estimate of the population mean to be more accurate (i.e., smallest variance)? Provide a mathematical justification for your answer.

We have population I with n_1, σ_1 and population II with $3n_1, 2\sigma_1$.

The comparison between standard error in first and second population can be viewed as:

$$\sigma\bar{X}_1 = \sigma 1/\sqrt{n_1}$$
$$\sigma\bar{X}_2 = 2\sigma 1/\sqrt{3n_1}$$
$$\sigma\bar{X}_2 = \frac{2}{\sqrt{3}}\sigma 1$$
$$\sigma\bar{X}_2 = 1.154\sigma 1$$

Therefore, the estimation of the population mean for first population will be more accurate than second population.

(b) For what ratio of n_2/n_1 would the estimates have equivalent accuracy (i.e., equivalent variances)?

$$\frac{\sigma 1}{\sqrt{n_1}} = \frac{\sigma 2}{\sqrt{n_2}}$$
$$\frac{n_2}{n_1} = \frac{\sigma_2^2}{\sigma_1^2}$$
$$\frac{n_2}{n_1} = \frac{4\sigma_1^2}{\sigma_1^2}$$
$$\frac{n_2}{n_1} = 4$$

(c) Verify this ratio via simulation,i.e.,create populations I and II by simulating 1000 normal RVs for each with $\mu = 1$ and $\sigma_1 = 1$ and generate 1000 estimates of the population mean μ using random samples with $n_1 = 100$ and n_2 set to give the ratio you found in b). Plot the distributions of these estimates using a kernel density estimate (the distributions should look similar). Why won't these empirical distributions look identical?

They are not identical due to different initial standard deviation in both populations, and we might need more sampling as 1000 is insufficient.

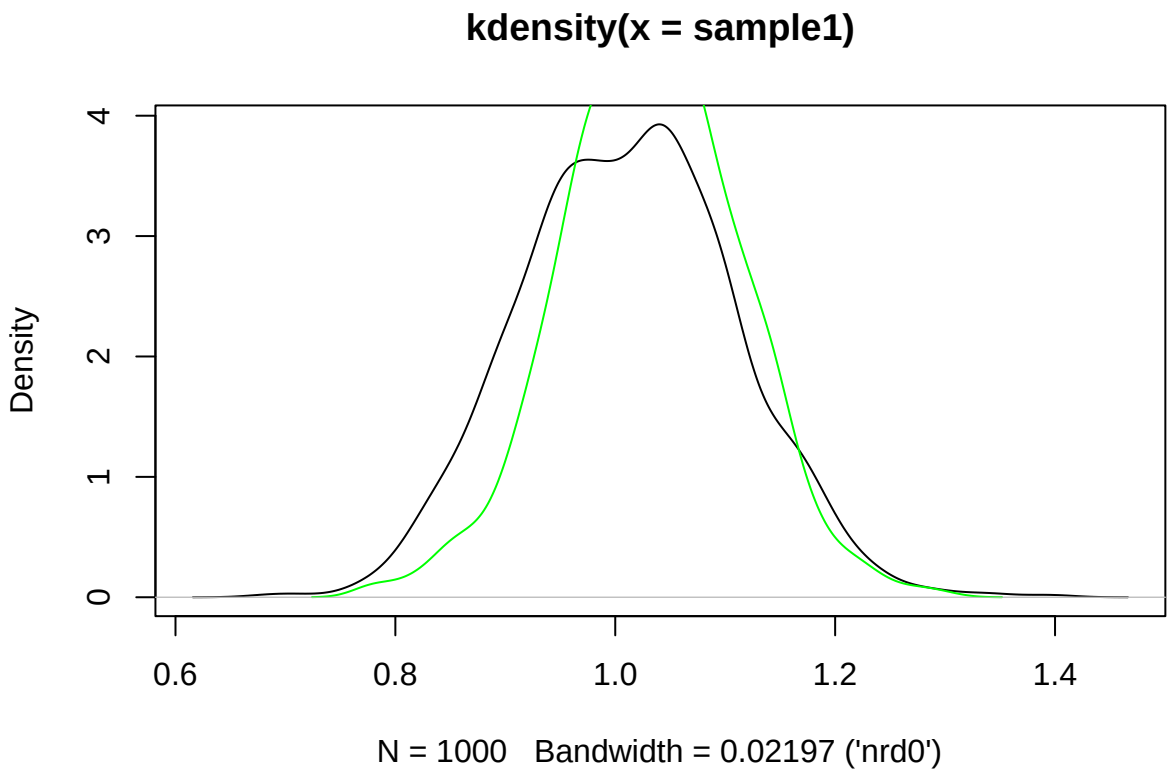
```
library(kdensity)

set.seed(10)
population1 = rnorm(1000, mean=1, sd=1)
population2 = rnorm(1000, mean=1, sd=2)

sample1 = c()
sample2 = c()

for (i in 1:1000) {
  sample1[i] = mean(sample(population1, 100))
  sample2[i] = mean(sample(population2, 400))
}

plot(kdensity(sample1))
lines(kdensity(sample2), col='green')
```



Question 6

(Based on Rice 7.10) True or false (and state why): If a sample from a population is large, a histogram of the values in the sample will be appropriately normal, even if the population is not normal? Verify your answer via simulation using a population whose elements have a U(0,1) distribution.

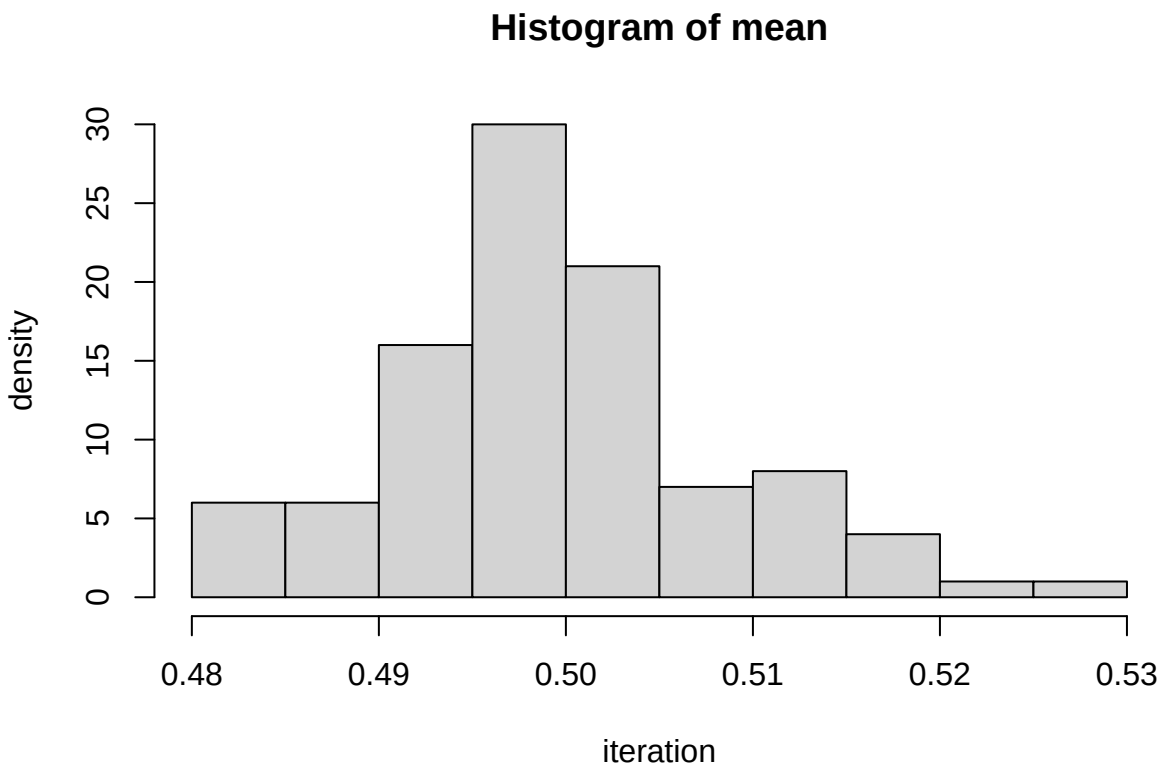
If we are sampling the a large number of data from a non-normal distribution, the sample distribution will approximate the population distribution. It would be a different scenario when you sample the average sample from a population, which will likely be looking similar to normal distribution.

Below is the code to take random sample from $U(0,1)$ with 1000 individual samples each iteration for 10 iterations. Here's the comparison of the distribution shape looks like between individual sample and mean sample:

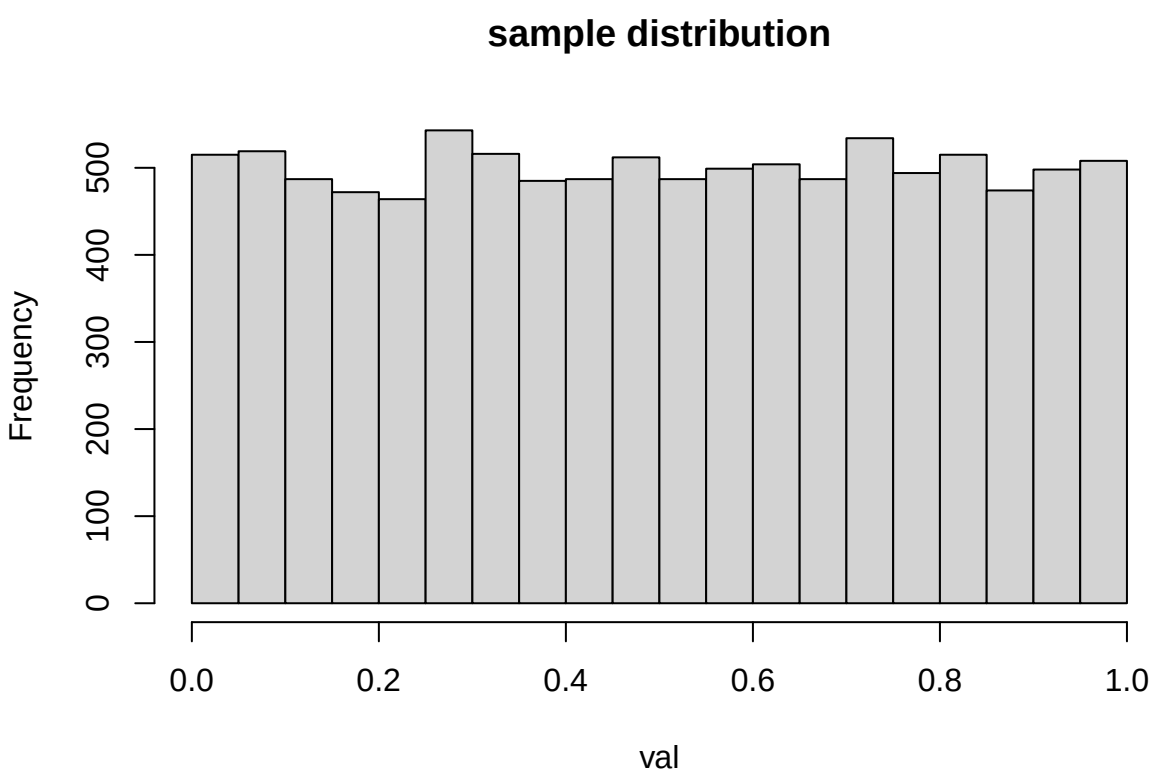
```
n=1000
mean = c()

for (i in 1:100) {
  x_val = runif(n, 0, 1)
  mean[i] = mean(x_val)
}

hist(mean,
     xlab='iteration',
     ylab='density')
```




```r
#individual sample
val = runif(10000, 0, 1)
hist(val, main="sample distribution")
```



Question 7

(Based on Rice 7.16) True or false? Justify your answers.

a) T. The center of a 95% confidence interval is the sample mean, with the confidence interval boundary as the lower bound and the upper bound.
b) F. 95% confidence interval does not tells us about the probability of sample mean or $/mu$.
c) F. 95% confidence interval does not tells us about the proportion of population included in the interval.
d) F. This is not the definition of 95% CI. 95% confidence interval means that with a large number of repeated samples, 95% of such calculated confidence intervals would include the true value of the parameter.

Question 8

Optional