qbs120_final_gibran

Gibran Erlangga

11/15/2021

Question 1

(12 points) Properties of random variables.

(a) (1 point) What is a random variable? Define both discrete and continuous RVs in terms of the sample space.

Random variable is a function that maps sample space to real numbers. Discrete RVs are the type or RVs that have countable number of values. An example for this is the number of children in a family. Continuous RVs are the type of RVs that have continuous values. An example for this is the change in temperature throughout a single day.

(b) (1 point) One of the observed values in an experiment is 5.4, which is modeled as a RV during analysis. How can a fixed number be modeled as a RV? Where is the random component?

Fixed number is an example of the discrete random variable, as it said that 5.4 is one of the observed values from the sample space. The randomization component comes from the process of generating the numbers in the sample space itself, by following through the chosen distribution from the family of continuous random variables.

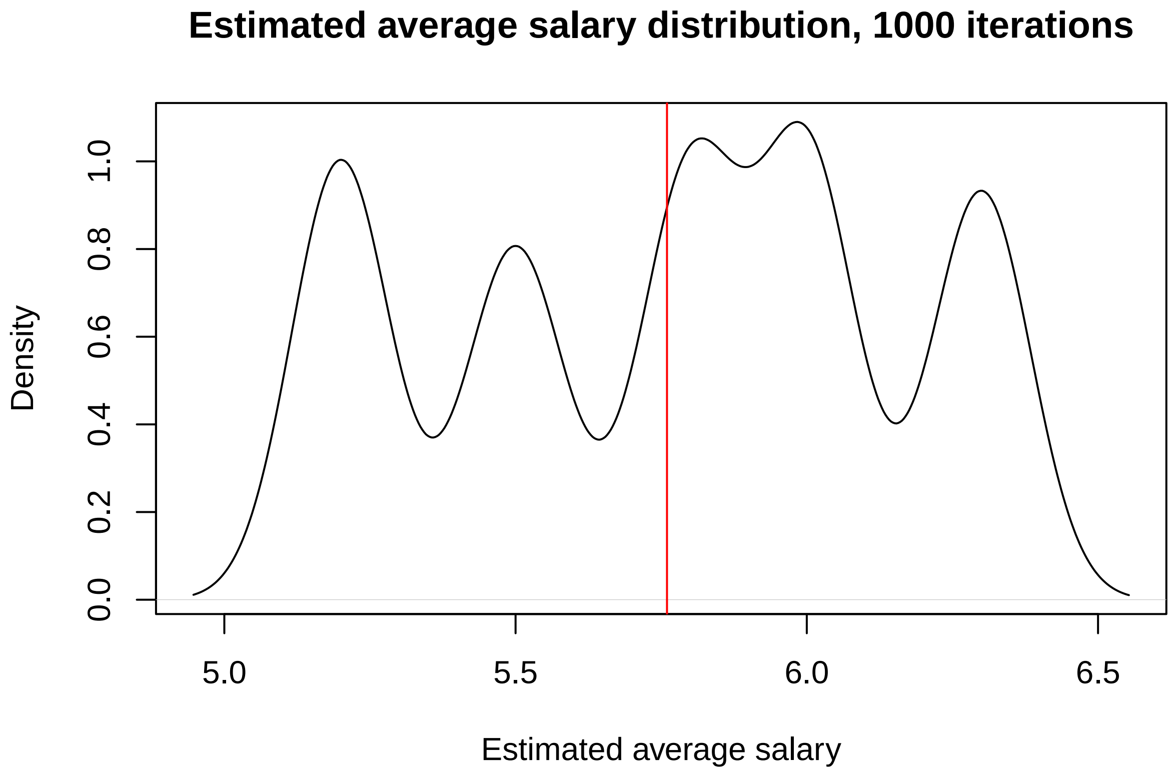
(c) (1 point) As part of a budget planning process, the average salary of Geisel employees during 2019-2020 fiscal year is computed. If the number of employees is fixed, is this average salary a RV? Explain.

It is not a random variable, because you will always get the same average salary value every time you compute it (assuming no salary changes in all employees during the stated period).

(d) (1 point) If the average salary is estimated using the the salary of one randomly selected employee, is the estimate a RV? Explain. Describe how you would model this value from a probabilistic perspective.

Yes, the estimated average salary will be a random variable as there will be a different person every time we compute the value. Consider following example: using the stated method, we estimate the average salary 1000 times. That can be expressed in R as:

```r
salaries <- c(5.2, 5.5, 5.8, 6.3, 6)

plot(density(sample(salaries, 1000, replace=T)),
             main="Estimated average salary distribution, 1000 iterations",
             xlab="Estimated average salary",
             ylab="Density")
abline(v=mean(salaries), col="red")
```



The red vertical line shows the true value of average salary, while the black line shows the distribution of estimates. We can see that the estimates are similar but not exactly the same. That shows the stochastic nature of the estimation generated by the given approach.

(e) (4 points) Let X_1 and X_2 be independent RVs with expectations μ_1 and μ_2, variances σ_1^2 and σ_2^2 and CDFs F_{X_1} and F_{X_2}. Find values for the following in terms of μ_1, μ_2, σ_1^2, and σ_2^2.

X_1 = Independent RV 1; $\mathrm{E}[X_1] = \mu_1$, $\mathrm{Var}(X_1) = \sigma_1^2$, CDF $= F_{X_1}$

X_2 = Independent RV 2; $\mathrm{E}[X_2] = \mu_2$, $\mathrm{Var}(X_2) = \sigma_2^2$, CDF $= F_{X_2}$

- $E[X_1 + X_2]$?

$$E[X_1 + X_2] = E[X_1] + E[X_2] = \mu_1 + \mu_2$$

- $E[X_1|X_2]$?

$$E[X_1|X_2] = E[X_1] = \mu_1$$

- $E[X_1^2]$?

$$\begin{aligned} E[X_1^2] &= Var(X_1) + [E[X_1]]^2 \\ &= \sigma_1^2 + \mu_1^2 \end{aligned}$$

- $Var(X_1 + X_2)$?

$$Var(X_1 + X_2) = Var(X_1) + Var(X_2) + 2Cov(X_1X_2)$$

$Cov(X_1X_2) = 0$ because it is independent, hence:

$$Var(X_1 + X_2) = \sigma_1^2 + \sigma_2^2$$

- $Var(X_1|X_2)$?

$$Var(X_1|X_2) = Var(X_1) = \sigma_1^2$$

- $F_{X_1}^{-1}(F_{X_1}(\mu_1))$?

$$F_{X_1}^{-1}(F_{X_1}(\mu_1)) = \mu_1$$

- $E[\mu_1\mu_2]$?

$$E[\mu_1\mu_2] = \mu_1\mu_2$$

- $Var(\sigma_1^2/\sigma_2^2)$?

$$Var(\sigma_1^2/\sigma_2^2) = 0$$

(f) (4 points) Let Y1, Y2 be dependent normally distributed RVs with expectations μ_1, μ_2, variances σ_1^2, σ_2^2, correlation ρ, and CDFs F_{X_1} and F_{X_2}. Find values for the following in terms of μ_1, μ_2, σ_1^2, σ_2^2, ρ and Y_2.

- $E[Y_1 + Y_2]$?

$$E[Y_1 + Y_2] = E[Y_1] + E[Y_2] = \mu_1 + \mu_2$$

- $E[Y_1|Y_2]$?

$$\begin{aligned} E[Y_1|Y_2] &= \sum_{Y_1} P_{Y_1|Y_2}(Y_1|Y_2) \\ &= (1-\rho)\cdot\mu_1 + \rho\cdot\mu_2 \end{aligned}$$

- $E[Y_1^2]$?

$$\begin{aligned} E[Y_1^2] &= Var(Y_1) + [E[Y_1]]^2 \\ &= \sigma_1^2 + \mu_1^2 \end{aligned}$$

- $Var(Y_1 + Y_2)$?

$$\begin{aligned} Var(Y_1 + Y_2) &= Var(Y_1) + Var(Y_2) + 2Cov(Y_1Y_2) \\ &= Var(Y_1) + Var(Y_2) + 2\cdot\rho\cdot\sqrt{Var(Y_1)\cdot Var(Y_2)} \\ &= \sigma_1^2 + \sigma_2^2 + 2\cdot\rho\cdot\sqrt{\sigma_1^2\cdot\sigma_2^2} \end{aligned}$$

- $Var(Y1|Y_2)$?

$$\begin{aligned} Var(Y_1|Y_2) &= Var(Y_1)\cdot(1-\rho^2) \\ &= \sigma_1^2\cdot(1-\rho^2) \end{aligned}$$

- $Var(Y_1/2 + \mu_2)$?

$$\begin{aligned}Var(Y_1/2 + \mu_2) &= (\frac{1}{2})^2 \cdot Var(Y_1) + Var(\mu_2)\\ &= \frac{1}{4} \cdot Var(Y_1) + 0\\ &= \frac{1}{4} \cdot \sigma_1^2\end{aligned}$$

- $1 - F_{Y_1}(\mu_1)$?

$$1 - F_{Y_1}(\mu_1) = 0.5$$

- $F_{Y_2}^{-1}(0.5)$?

$$F_{Y_2}^{-1}(0.5) = \mu_1$$

Question 2

(12 points) Properties of estimators. For parts (a) thru (h), consider an estimator $\hat{\theta}$ for distribution parameter θ.

(a) (0.5 point) Is $\hat{\theta}$ a random variable? Justify your answer.

$\hat{\theta}$ acts as an estimator, hence its a random variable. Its value will be depending on the sample we have at hand.

(b) (0.5 point) Is θ a random variable? Justify your answer.

θ is a parameter of a distribution, so it is not a random variable. For instance, our population is defined as all students in the QBS program for 2021-2022 cohort. From the data, we know that the median height of the population to be (for example) 6 feet.

(c) (1 point) What must be true for $\hat{\theta}$ to be an unbiased estimator of θ?

An estimator $\hat{\theta}$ is said to be unbiased if its expected value is equal to the true parameter θ, on average.

(d) (1 point) Under what conditions would a biased estimator be preferable to an unbiased estimator? What estimator property can be used to make this decision?

Ideally, an unbiased estimator is preferable to a biased estimator, although in practice, biased estimators (with generally small bias) are frequently used. Few reasons:

- an unbiased estimator does not exist without further assumptions about a population.
- an unbiased estimator is difficult to compute.
- an unbiased estimator is median-unbiased but not mean-unbiased (or the reverse).
- a biased estimator gives a lower value of some loss function (particularly mean squared error) compared with unbiased estimators.
- in some cases, being unbiased is too strong a condition, and the only unbiased estimators are not useful.

We can use consistency estimator property to support us in making the decision between one over the other.

(e) (1 point) What must be true for $\hat{\theta}$ to be a consistent estimator of θ?

The estimate has to converge to the true value of the parameter as the sample size increases.

(f) (1 point) Under what conditions would one consistent estimator for a parameter be preferable to another consistent estimator for the same parameter?

Few cases:

- a case when one consistent estimator converges faster to the true value of parameter than the other consistent estimator.
- a case when both consistent estimators are biased, but one is less biased than the other.
- a case when one consistent estimator is unbiased and the other estimator is biased. In this case, we choose the unbiased one.

(g) (1 point) What is the interpretation of the 100(1—α)% confidence interval for θ?

100(1—α)% confidence interval signifies the probability of population parameter θ true value is within the given confidence interval range. For instance, if *alpha* is 0.01, then there is a 99% probability that the true value of population parameter θ is within the given confidence interval range (lower bound - upper bound).

(h) (1 point) An experiment generates 10 numeric measurements, X1, ..., X10 . These measurements are modeled as i.i.d. random variables where the common distribution has CDF F, expectation μ, and variance σ^2. What is the preferred estimate of μ? Why is this estimate preferred? What can be said about the distribution of this estimate?

Preferred estimate of μ in this case would be based on the expected value formula, which is mean of the data. This is preferred because this is an unbiased estimator. The distribution of the estimate is approximately be a normal $N(\mu, \sigma^2)$.

(i) (1 point) For the 10 measurement experiment in the prior question, assume that just a single measurement, e.g., X1, is used to estimate μ. Is this estimate unbiased? Is it consistent? What can be said about the distribution of this estimate?

Using X_1 as an estimate for μ: this is a biased estimate, and this is a consistent one. We do not have any distribution for the estimate as it is a constant.

(j) (1 point) What is a maximum likelihood estimate? What can be said about the sampling distribution of MLEs as the size of the experimental sample grows towards infinity?

MLE is a way of estimating the distribution parameters from the given data, assuming the data we have is a good representative of the population data. When the experimental sample grows towards infinity, sequences of maximum likelihood estimators will converge to the true value.

(k) (1 point) Describe the general process for computing maximum likelihood estimates. Are there ever cases where there is ambiguity regarding the correct MLE?

Steps: First, set up the likelihood function ($P(data|p)$); Second, apply derivative to the likelihood function and setting it to be equal to 0. Third, Solve for p. the value of p will be the MLE. There are some cases where ambiguity is present regarding the MLE results.

(l) (2 points) Let X1,...,Xn be iid normal RVs with distribution N(μ, σ^2). What is the distribution of $S = c + \sum_{i=1}^{n} X_i$, where c is a constant? If the Xi are not normal but are independent with means μ and variance σ^2, what can be said about the distribution of S?

We will estimate the mean through its expected value, and the standard deviation through the estimated variance. Hence:

Expected value:

$$\begin{aligned} E(S) &= c + \sum_{i=1}^{n} X_i \\ &= E(c) + E(\sum_{i=1}^{n} X_i) \\ &= c + n \cdot \mu \end{aligned}$$

Variance:

$$\begin{aligned} Var(S) &= Var(c + \sum_{i=1}^{n} X_i) \\ &= Var(X_i) + 2 \cdot Cov(X_i, X_i) \\ &= Var(X_i) + 0 \\ &= n \cdot \sigma_2 \end{aligned}$$

Therefore, S ~ $N(c + n \cdot \mu, n \cdot \sigma_2)$

Question 3

(12 points) Properties of hypothesis testing. (a) (1 point) Describe the difference between a composite and simple hypothesis.

Simple hypothesis is a hypothesis which all parameters of the distribution are specified. An example of a simple hypothesis is X follows a uniform distribution on [0,1].

Composite hypothesis is a hypothesis which parameters of the distribution are not completely specified. An example of a composite hypothesis is X follows a normal distribution with mean μ= 0 (hence the value for σ^2 can be anything > 0).

(b) (1 point) Let μ_1 and μ_2 be the expected values of the birth weights for two different mouse strains. What are appropriate null and alternative hypotheses for comparing these mean weights?

$$\begin{aligned} H_0 &: \mu_1 = \mu_2 \\ H_A &: \mu_1 \neq \mu_2 \end{aligned}$$

In words, the null hypothesis is that there is no difference in the mean weights between μ_1 and μ_2. For the alternate hypothesis, there is a difference between μ_1 and μ_2.

(c) (1 point) What is the sampling distribution of a test statistic?

Sampling distribution of a test statistic is the distribution of a sample to which the test statistic is derived from.

(d) (1 point) Define the type I error rate α, the type II error rate β and the power for a hypothesis test.

Type I error rate (α) is probability of rejecting H_0 when H_0 is true (false positive).
Type II error rate (β) is probability of accepting H_0 when H_A is true (false negative).
Power for a hypothesis test is probability of rejecting H_0 when H_0 is false. Can also be expressed as 1 - β.

(e) (1 point) Can the power of a hypothesis test be determined using just the sampling distribution?

Yes you can. Assuming the sample is normally distributed, you can do hypothesis test of a single population mean μ, and use the sample standard deviation to approximate the population standard deviation. If the sample size is large enough, t-test will work even if the population is not normally distributed. The power of the hypothesis test will increase as the sample size increases.

(f) (1 point) What is a p-value?

p-value is the probability of encountering a test statistic equal to or larger than the observed statistic assuming H_0 is true.

(g) (1 point) How can one obtain perfect type I error control? How can one obtain perfect type II error control?

To obtain perfect type I error control, you just need to accept all null hypotheses. Conversely, to obtain perfect type II error control, you need to reject all null hypotheses.

(h) (1 point) Under what conditions is type I error control more important that type II error control? Under what conditions is type II error control more important that type I error control?

Type I error control more important that type II error control when the cost of getting a wrong prediction is higher than letting a right prediction goes away. The most relevant example to illustrate this is covid cases. Let's say a person is suspected for COVID by some symptoms showed in the past few days. Then he proceed to get a covid test. The null hypothesis setting will be:

H_0: test result is negative.

H_A: test result is positive.

Hence, the definition of type I and II error will be:

Type I error: The person tested negative, when in fact he actually has covid, so this will spread the virus to his friends and relatives.

Type II error: The person tested positive when he does not have covid, so this will incur some extra costs for him to isolate himself for a few days ahead.

In this case, the consequences of having type I error is higher than type II error because have the virus spread in any way will harm the community, especially when it gets spread into more vulnerable people, as opposed to mandating the person to isolate himself for a few days ahead.

Conversely, type II error control more important that type I error control when the cost of getting a wrong prediction is lower than letting a right prediction goes away. For example, the Dartmouth Alumni Gym regularly do a daily water quality test in the club's swimming pool. If the level of contaminants are too high, then they temporarily close the pool to perform a water treatment. In this case, the null hypothesis setting will be:

H_0: water quality is good, so it's safe for people to use it.

H_A: water quality is not good, so it's not safe for people to use it.

Hence, the definition of type I and II error will be:

Type I error: We close the pool when we do not need to (because the water quality is good)

Type II error: We do not close the pool when we need to (because the water quality is not good)

In this case, the consequences of having type II error is higher than type I error as it can incur serious illness to people who use the pool, as opposed to closing the pool for a day to fix the water quality.

(i) (1 point) If the observations in a sample are correlated but are assumed to be independent for a one sample t test, what are the implications for type I error control? Justify.

Type I error happens when we reject the H_0 when H_0 is true. If the null is false and the test statistic is in the rejection region, no error was made (it's classified as true positive).

(j) (1 point) Define the family-wise error rate (FWER) and the false discovery rate (FDR) for the scenario in which m total hypotheses are tested, the null hypothesis is rejected for a of those hypotheses, and the null was true for b of the a rejected hypotheses.

Family-wise Error Rate (FWER) is the probability of making at least one type I error in the multiple testing context. FWER also known as cumulative type I error. Given the above scenario, FWER will be:

$$Pr(b > 0) = Pr(b/a > 0)$$

False Discovery Rate (FDR) is the expected proportion of false positives among all significant tests. Given the above scenario, FDR will be:

$$E[b/a]$$

(k) (1 point) Under what conditions are the FWER and FDR equal?

If all null hypotheses are true.

(l) (1 point) Under what conditions is FWER control preferable to FDR control? Under what conditions is FDR control preferable to FWER control?

Often the control of the FWER is not needed. The control of the FWER is important when a conclusion from the various individual inferences is likely to be erroneous (when at least one of them is). This may be the case when several new treatments are competing against a standard, and a single treatment is chosen from the set of treatments which are declared significantly better than the standard.

Question 4

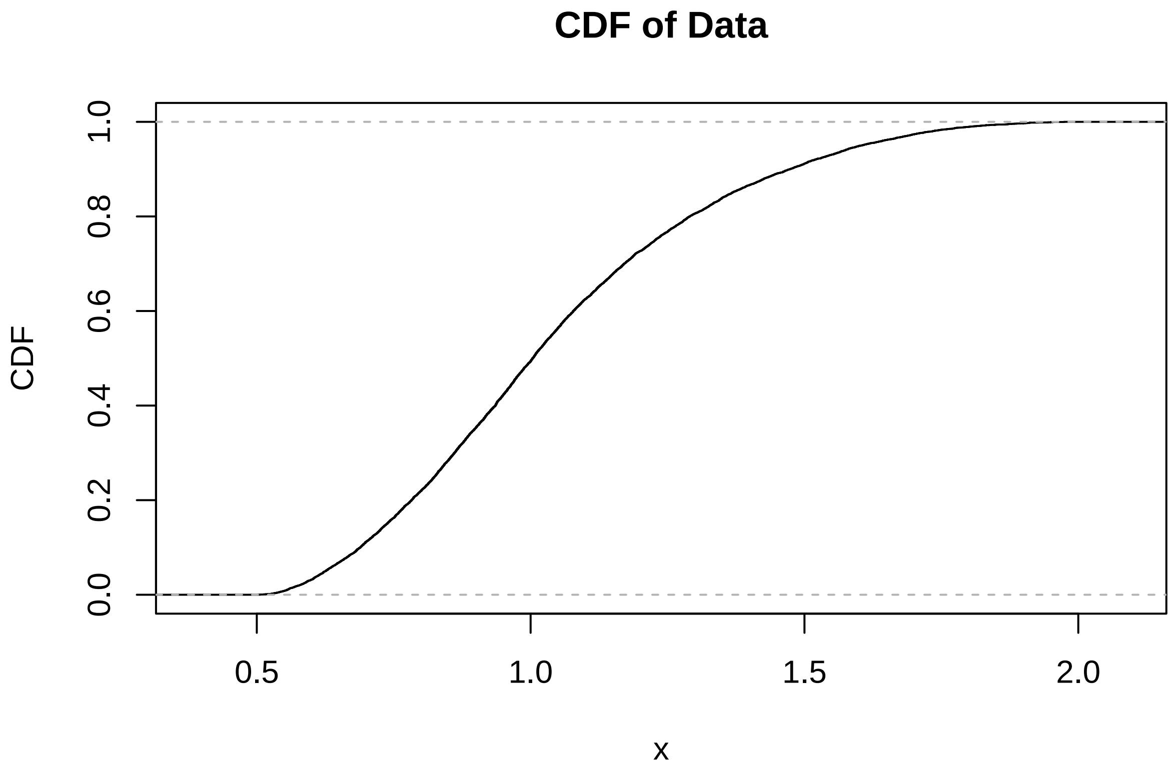
(7 points) Let X1,...Xn be iid RVs with Xi U(2,4). Let T = X(n)/X(1), i.e., the ratio of the last and first order statistics.

(a) (2 points) Use simulation to approximate the distribution of T for n = 100 (generate at least 1000 simulated data sets).

```r
n_sample=10000
n=100

sim.data = matrix(runif(n_sample*n, 2, 4), nrow=n_sample, ncol=n)
sim.T = sim.data[,n] / sim.data[,1]
```

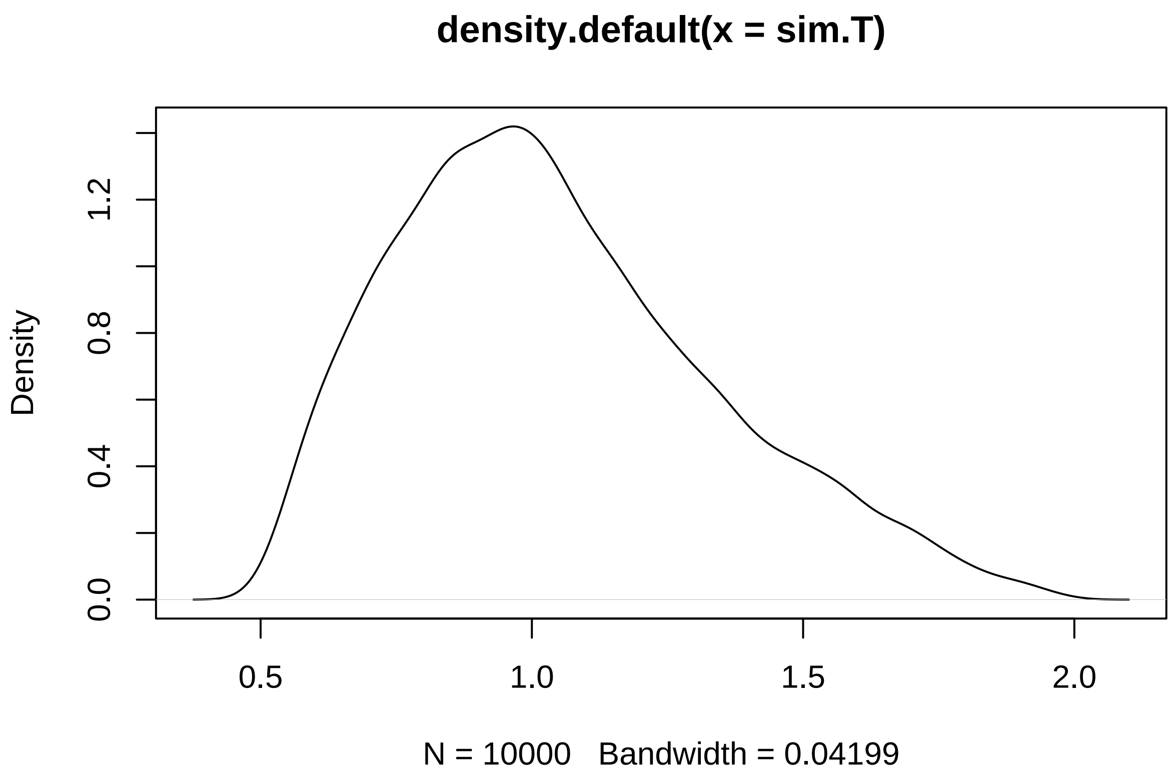
(b) (1 point) Visualize the approximate CDF for this distribution.

```r
cdf_plot <- ecdf(sim.T)

plot(cdf_plot, xlab='x', ylab='CDF', main='CDF of Data')
```



(c) (1 point) Visualize the approximate density for this distribution.

```
plot(density(sim.T))
```



(d) (1 point) Estimate $P(T \geq 1.99|n = 100)$.

```
n_iter = 10000
n = 100
T_val <- rep(0, n_iter)

for (i in 1:n_iter) {
  temp <- runif(n, 2, 4)
  temp <- sort(temp)
  T_val[i] <- temp[n]/ temp[1]
}

num_T_val_passed_threshold <- length(which(T_val >= 1.99))

# all simulated values more than threshold over all simulated values
num_T_val_passed_threshold/length(T_val)
```

```
## [1] 0.1538
```

(e) (2 points) Estimate E[T] using the simulation results. Is your estimate unbiased and/or consistent? Justify.

```
mean(T_val)
```

```
## [1] 1.970349
```

This estimate is unbiased and consistent.

Question 5

An experiment generates the following independent measurements:

```
data <- c(11,9,15,10,13,11,13,10,8,13,6,10,10,9,11,9,12,14,10,15,11,6,11,9,19,
          9,8,5,12,8,8,9,8,8,9,7,11,9,9,10,9,20,14,14,8,12,13,10,10,16,
          9,7,7,12,6,10,17,7,7,13,9,6,11,13,11,11,8,10,9,11,13,13,12,12,12,
          14,10,8,13,11,8,6,10,18,9,4,10,12,8,11,4,13,17,8,13,15,6,9,7,12)
```

(a) (2 points) If the observations are assumed to be iid, what is the 95% CI for the mean of the common distribution according to the asymptotic normal distribution of the sample average? Compute using just the following R functions: mean(), sd(), length(), sqrt() and qnorm().

```
mean <- mean(data)
sd <- sd(data)
n <- length(data)
se <- sd/sqrt(n)

# asymptotic normal
error <- qnorm(0.975)*se
left <- mean-error
right <- mean+error

cat(left, right)
```

```
## 9.820778 11.03922
```

(b) (2 points) Compute the 95% CI for the mean using the basic bootstrap method (use of bootstrap-specific R packages not allowed).

```
bootstrap_data = c()

for (i in 1:10000) {
  sample_bootstrap <- sample(data, 1000, replace=T)
  bootstrap_data<- c(bootstrap_data, mean(sample_bootstrap))
}

bt_mean <- mean(bootstrap_data)
bt_sd <- sd(bootstrap_data)
bt_n <- length(bootstrap_data)
bt_se <- bt_sd/sqrt(bt_n)
bt_error <- qnorm(0.975)*bt_se

bt_left <- bt_mean-error
bt_right <- bt_mean+error

cat(bt_left, bt_right)
```

```
## 9.819711 11.03815
```

(c) (2 points) Compute the 95% CI for the mean using the percentile bootstrap method (use of bootstrap-specific R packages not allowed).

```
perct_bootstrap <- quantile(bootstrap_data, c(0.025, 0.975))
perct_bootstrap_left <- perct_bootstrap[1][[1]]
perct_bootstrap_right <- perct_bootstrap[-1][[1]]
perct_bootstrap
```

```
##   2.5%  97.5%
## 10.240 10.623
```

```
cat(perct_bootstrap_left, perct_bootstrap_right)
```

```
## 10.24 10.623
```

(d) (1 point) How do the asymptotic normal, basic bootstrap and percentile bootstrap 95% CIs compare? Do the relative widths of these three CIs match your expectations based on the observed values and statistical theory? Explain.

```
# asymptotic normal
cat(left, right)
```

```
## 9.820778 11.03922
```

```
# basic bootstrap
cat(bt_left, bt_right)
```

```
## 9.819711 11.03815
```

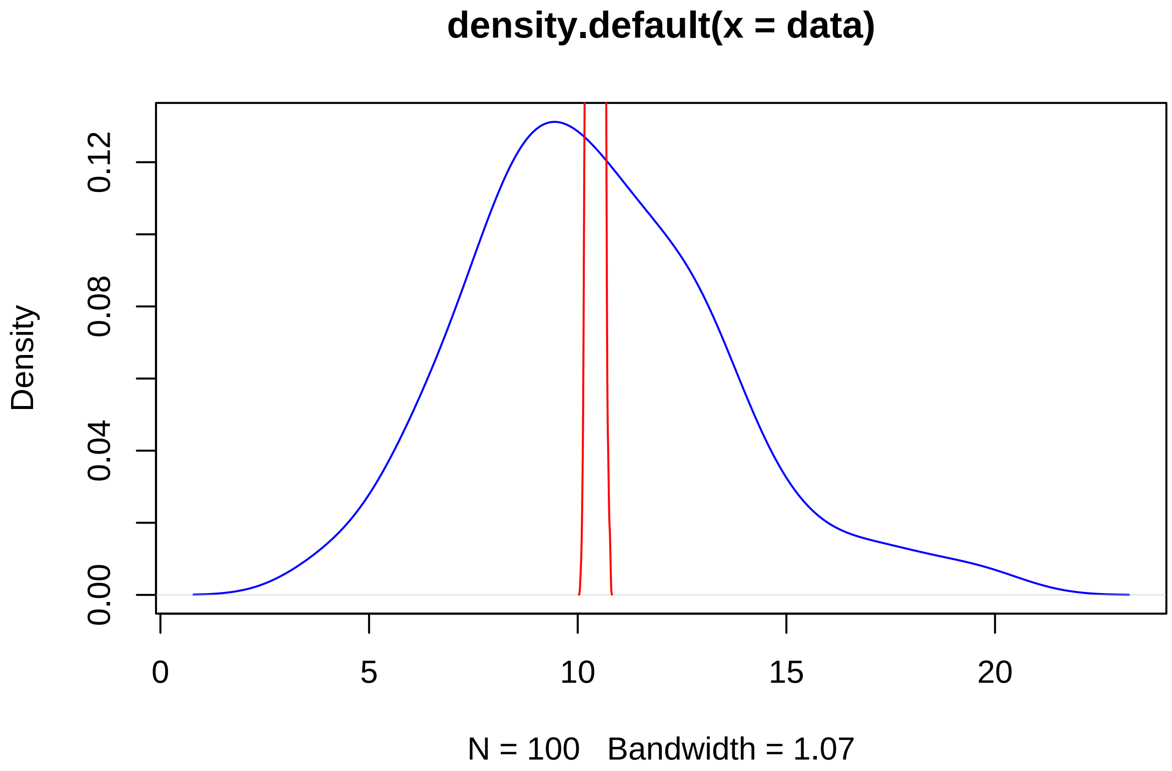
```
# percentile bootstrap
cat(perct_bootstrap_left, perct_bootstrap_right)
```

```
## 10.24 10.623
```

CI of asymptotic normal and basic bootstrap looks similar, which match with my expectation as the value for bootstrap is taken from the original data. CI from percentile bootstrap method seems to be narrow than the others, which might be the effect of our bootstrap distribution of means not being a good enough approximation to the true distribution of the mean.

(e) (2 points) Compare the asymptotic normal distribution of the mean to the bootstrap distribution using both density and probability plots. Does the bootstrap distribution visually match the asymptotic normal distribution?

```
plot(density(data), col='blue')
lines(density(bootstrap_data), col='red')
```



Bootstrap distribution does not visually match with the asymptotic normal distribution.

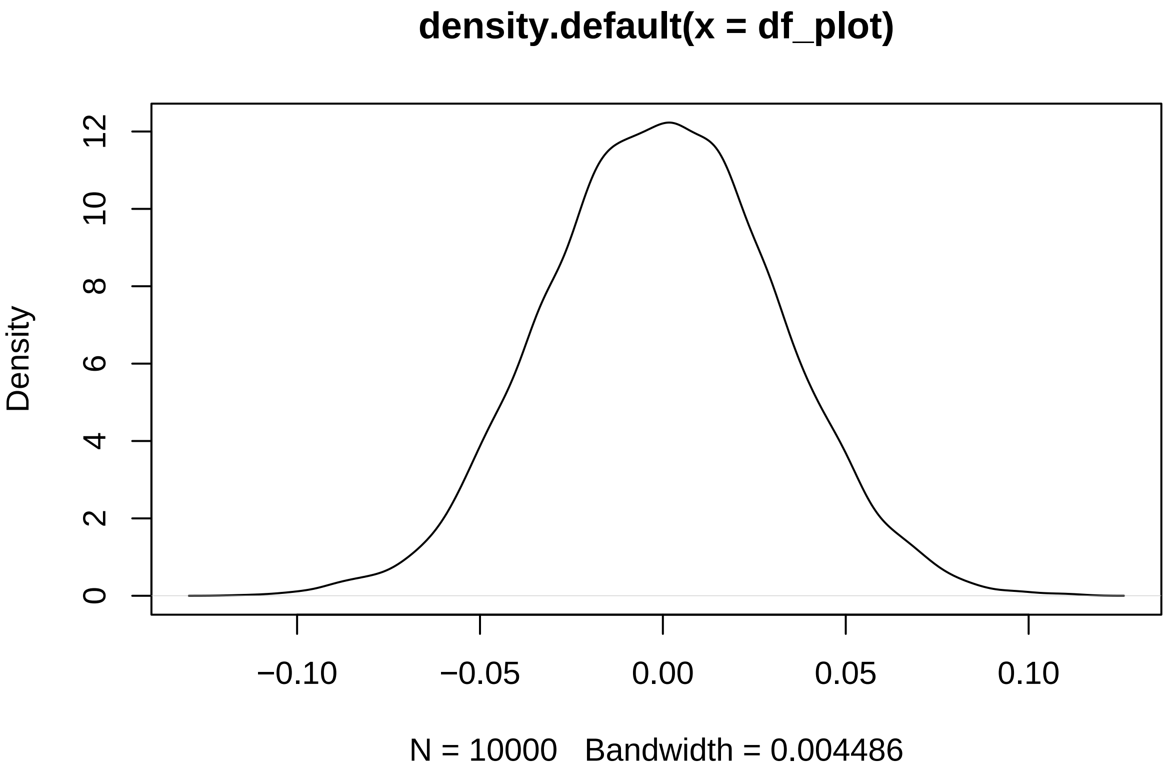
(f) (1 point) Test the goodness-of-fit of the asymptotic normal distribution to the bootstrap distribution using a two-sided Kolmogorov-Smirnov (KS) test (you can use the ks.test() method setting y="pnorm"; hint: this expects the x value to be standard normal by default so you'll need to standardize the bootstrap means).

```
std_mean <- c()

for (i in 1:length(bootstrap_data)) {
  # standardize bootstrap means
  temp_std_mean <- (bootstrap_data[i] - mean)/sd
  std_mean <- c(std_mean, temp_std_mean)
}

# run two-sided ks test
ks.test(std_mean, y="pnorm")
```

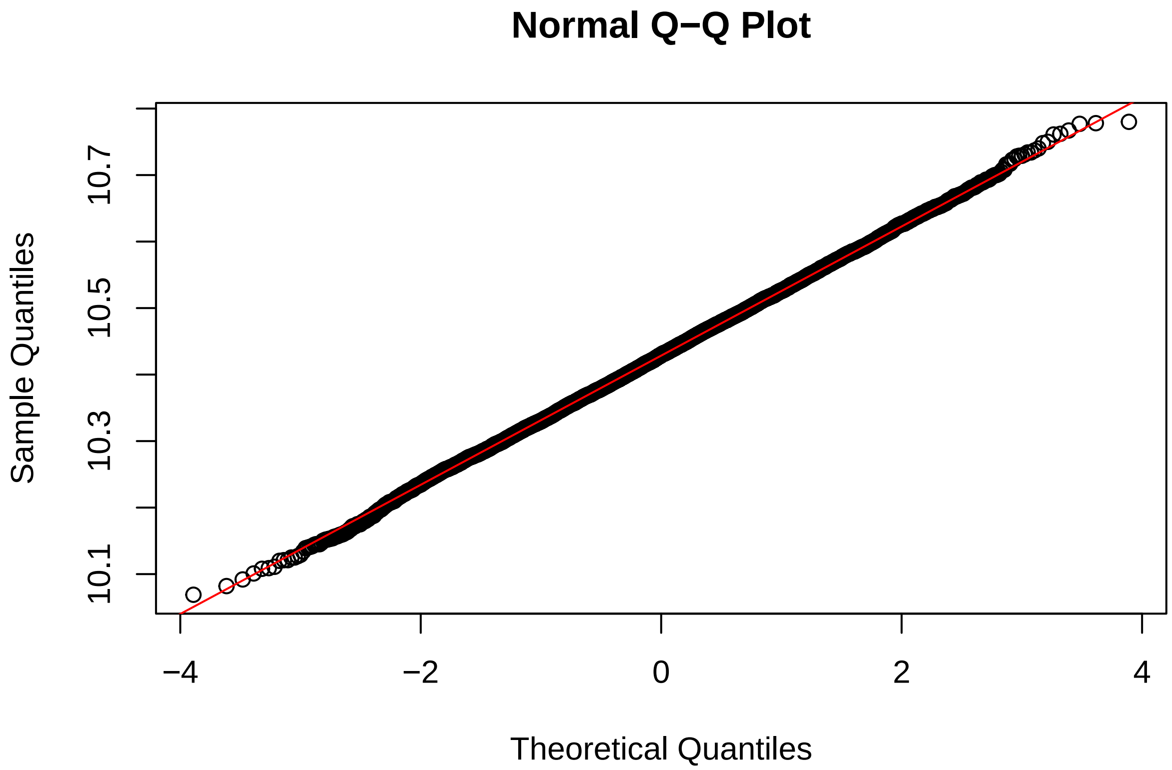
```
## Warning in ks.test(std_mean, y = "pnorm"): ties should not be present for the
## Kolmogorov-Smirnov test

##
##  One-sample Kolmogorov-Smirnov test
##
## data:  std_mean
## D = 0.46319, p-value < 2.2e-16
## alternative hypothesis: two-sided
```

(g) (2 points) What is your best estimate for the true distribution of the data? Evaluate visually using both a probability plot and statistically using a KS test.

```
# density plot
df_plot <- (bootstrap_data - mean)/sd
plot(density(df_plot))
```



```
# qq-plot
qqnorm(bootstrap_data)
qqline(bootstrap_data, col='red')
```



```
# ks test
ks.test(df_plot, y="pnorm")
```

```
## Warning in ks.test(df_plot, y = "pnorm"): ties should not be present for the
## Kolmogorov-Smirnov test
```

```
##
##  One-sample Kolmogorov-Smirnov test
##
## data:  df_plot
## D = 0.46319, p-value < 2.2e-16
## alternative hypothesis: two-sided
```

Question 6

(10 points) Exploring the Central Limit Theorem (CLT): In this problem, you will explore how well the CLT applies to the sum of RVs that are dependent and not identically distributed. Specifically, we will look at the distribution of the sum of eigenvalues of sample covariance matrices computed on data matrices that hold independent standard normal values. The theoretical joint distribution of these eigenvalues is provided by random matrix theory.

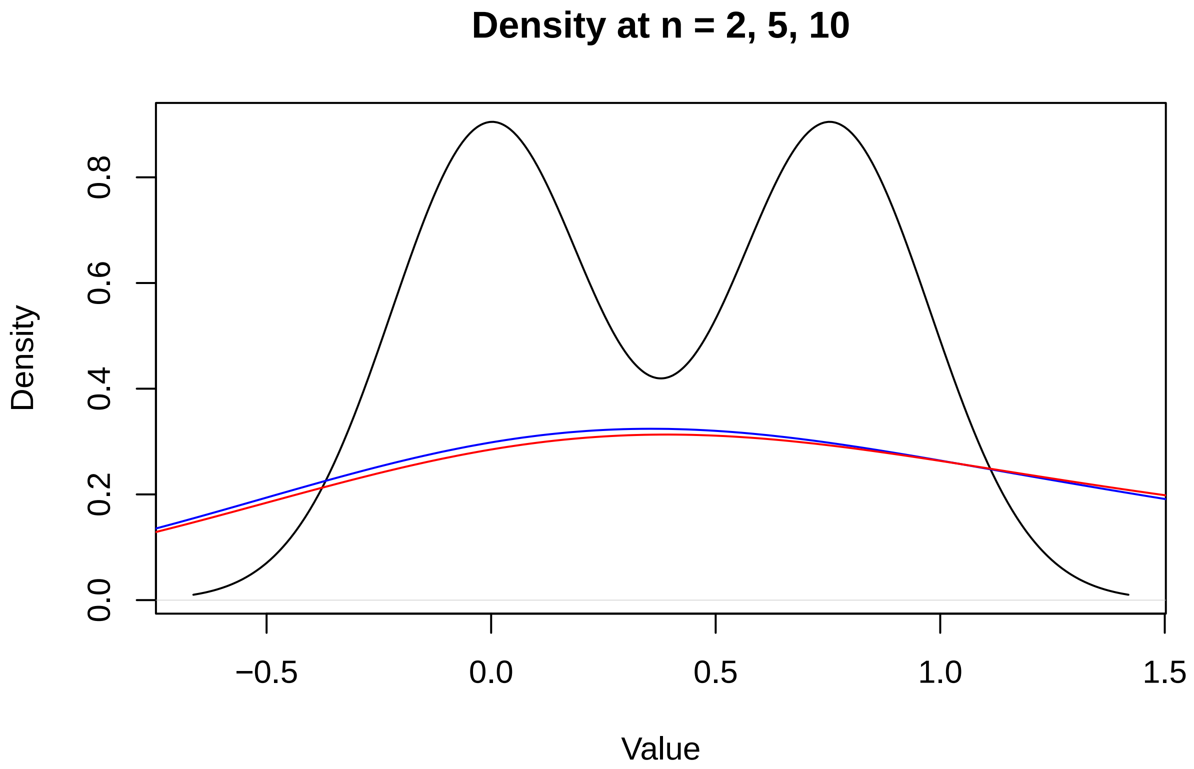
(a) (2 points) For n=2,...,10, simulate 1,000 n-by-n matrices of independent N(0,1) RVs. For each matrix, find the eigenvalues of the sample covariance matrix (can use the cov() and eigen() R functions). Plot the estimated density of the eigenvalue sums for n=2, 5 and 10 (put

all three estimated densities on a single plot). Do these estimated density plots match your expectations? Explain.

```
library(hash)
```

```
## hash-2.2.6.1 provided by Decision Patterns
```

```
n_dim = 2:10
n = 1000

# create an empty hash to store all values
storage <- hash()

for (i in n_dim) {
  temp_simulation <- list()
  for (j in 1:n) {
    temp_simulation[[j]] <- matrix(rnorm(i*i, 0, 1), ncol = i, nrow=i)
  }
  # store each list of 1000 datasets in hash
  storage[[as.character(i)]] <- temp_simulation
}

# use 1st list as sample
density_n_2 <- density(eigen(cov(storage[["2"]][[1]]), symmetric=T)$values)
density_n_5 <- density(eigen(cov(storage[["5"]][[1]]), symmetric=T)$values)
density_n_10 <- density(eigen(cov(storage[["10"]][[1]]), symmetric=T)$values)

plot(density_n_2,
     main="Density at n = 2, 5, 10",
             xlab="Value",
             ylab="Density")
lines(density_n_5, col='blue')
lines(density_n_10, col='red')
```



(b) (2 point) Explore the dependency structure between the eigenvalues by finding and print the sample correlation matrix for n=2,3,4, and 5, i.e., for each n, compute the correlation matrix for a 1,000 by n matrix holding all simulated eigenvalues from a) for that n value. What does this empirical correlation structure suggest about the dependence of the eigenvalues?

```
n_corr = c(2, 3, 4, 5)

cor_results = list()
random_list_index <- sample(1:1000, 1)

for (i in n_corr) {
  cor_results[[i]] <- cor(storage[[as.character(i)]][[random_list_index]])
}

cor_results[[2]]
```

```
##      [,1] [,2]
## [1,]    1    1
## [2,]    1    1
```

```
cor_results[[3]]
```

```
##            [,1]       [,2]       [,3]
## [1,]  1.0000000  0.9918385 -0.7028211
## [2,]  0.9918385  1.0000000 -0.7877845
## [3,] -0.7028211 -0.7877845  1.0000000
```

```
cor_results[[4]]
```

```
##            [,1]       [,2]       [,3]       [,4]
## [1,]  1.0000000  0.1276071  0.4054453 -0.3825246
## [2,]  0.1276071  1.0000000 -0.1188976 -0.3567640
## [3,]  0.4054453 -0.1188976  1.0000000 -0.8829390
## [4,] -0.3825246 -0.3567640 -0.8829390  1.0000000
```

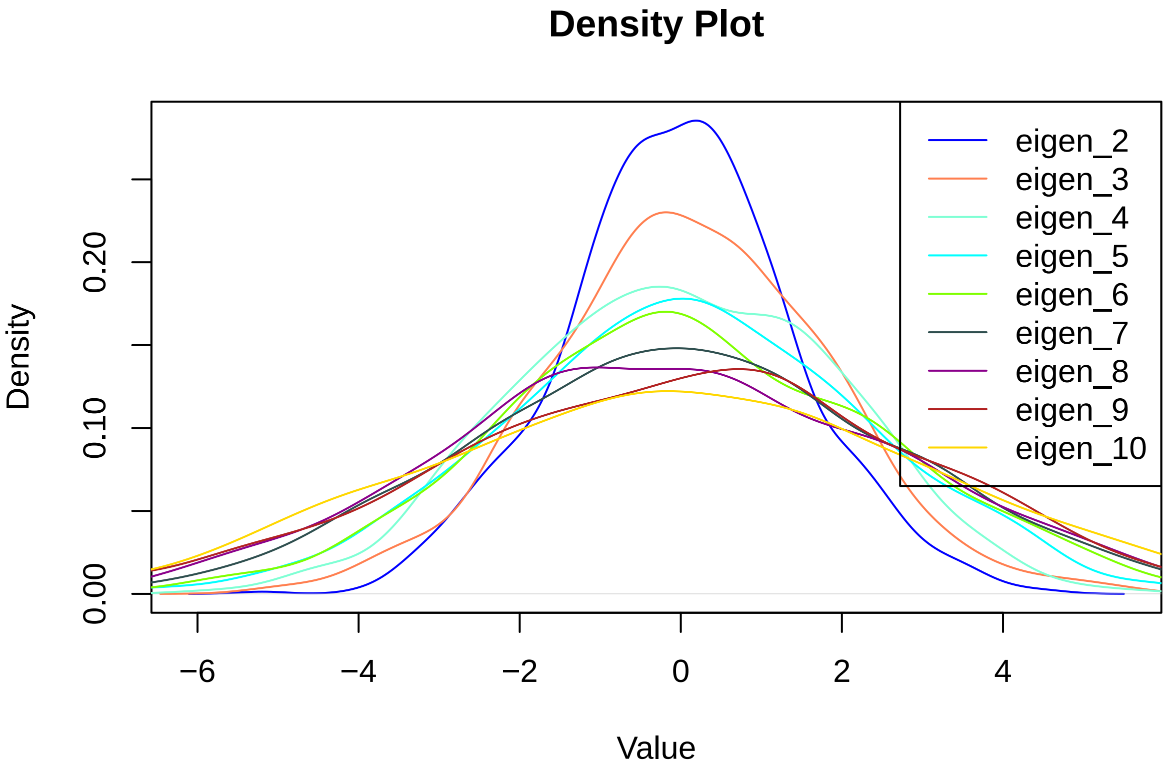
```
cor_results[[5]]
```

```
##            [,1]       [,2]       [,3]       [,4]       [,5]
## [1,]  1.0000000 -0.5771356  0.8114337  0.4352544  0.1368971
## [2,] -0.5771356  1.0000000 -0.3341682 -0.1859078 -0.7069886
## [3,]  0.8114337 -0.3341682  1.0000000 -0.1151327 -0.1813206
## [4,]  0.4352544 -0.1859078 -0.1151327  1.0000000  0.1095885
## [5,]  0.1368971 -0.7069886 -0.1813206  0.1095885  1.0000000
```

From the correlation structure in each defined n value, no matter which sample we use, the correlation values between combinations of grids stay the same.

(c) (1 points) Assuming the CLT holds for the sum of eigenvalues, what is the approximate distribution of a standardized version of this sum?

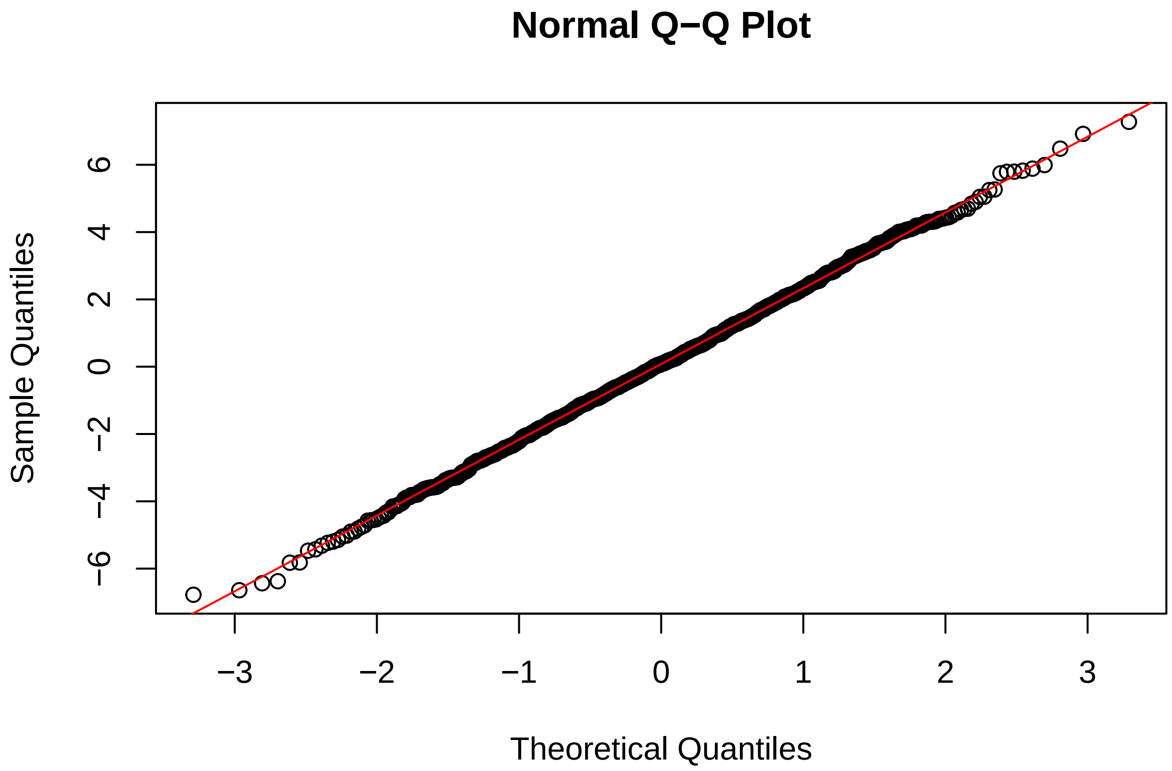
```
eigen_total_sums <- hash()

for (i in n_dim) {
  temp_eigen <- 0
  for (j in 1:n) {
    temp_eigen_sum <- sum(eigen(storage[[as.character(i)]][[j]], symmetric = T)$values)
    temp_eigen <- c(temp_eigen, temp_eigen_sum)
  }
  eigen_total_sums[[as.character(i)]] <- temp_eigen
}

# color options
colors = c("blue", "coral", "aquamarine", "cyan", "chartreuse",
           "darkslategrey", "darkmagenta", "firebrick", "gold")

names = c("eigen_2", "eigen_3", "eigen_4", "eigen_5",
          "eigen_6", "eigen_7", "eigen_8", "eigen_9",
          "eigen_10")

plot(density(eigen_total_sums[["2"]]), col=colors[1],
     main="Density Plot",
             xlab="Value",
             ylab="Density")
lines(density(eigen_total_sums[["3"]]), col=colors[2])
lines(density(eigen_total_sums[["4"]]), col=colors[3])
lines(density(eigen_total_sums[["5"]]), col=colors[4])
lines(density(eigen_total_sums[["6"]]), col=colors[5])
lines(density(eigen_total_sums[["7"]]), col=colors[6])
lines(density(eigen_total_sums[["8"]]), col=colors[7])
lines(density(eigen_total_sums[["9"]]), col=colors[8])
lines(density(eigen_total_sums[["10"]]), col=colors[9])

legend("topright",
       legend=names,
       col=colors,
       lty=1)
```



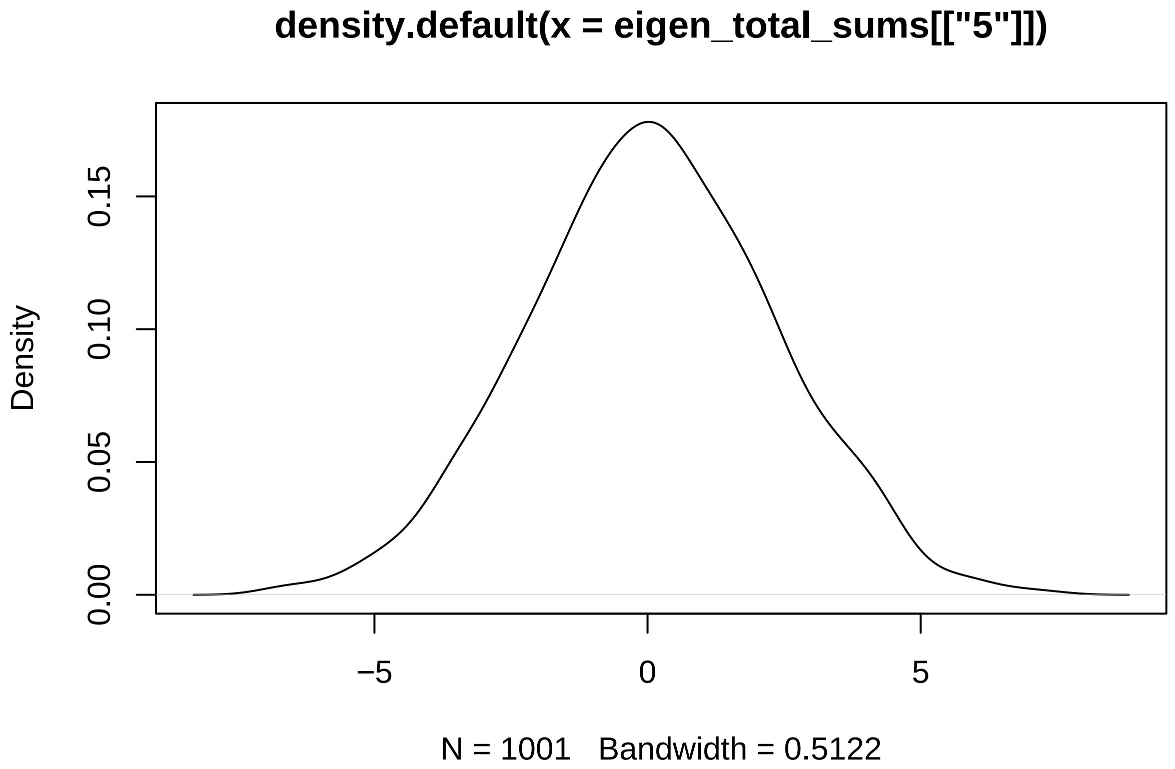
We can see from a) that the chosen sample distribution is not normally distributed. However, when CLT holds, as the sample size increases, the distribution of sample means approaches a normal distribution (in this case, it is the distribution of sum of the eigenvalues).

(d) (2 points) For n=5, use a probability plot to visualize how well the empirical distribution of the standardized sum matches the theoretical distribution according to the CLT. Does this plot match your expectations?

```
qqnorm(eigen_total_sums[["5"]])
qqline(eigen_total_sums[["5"]], col="red")
```



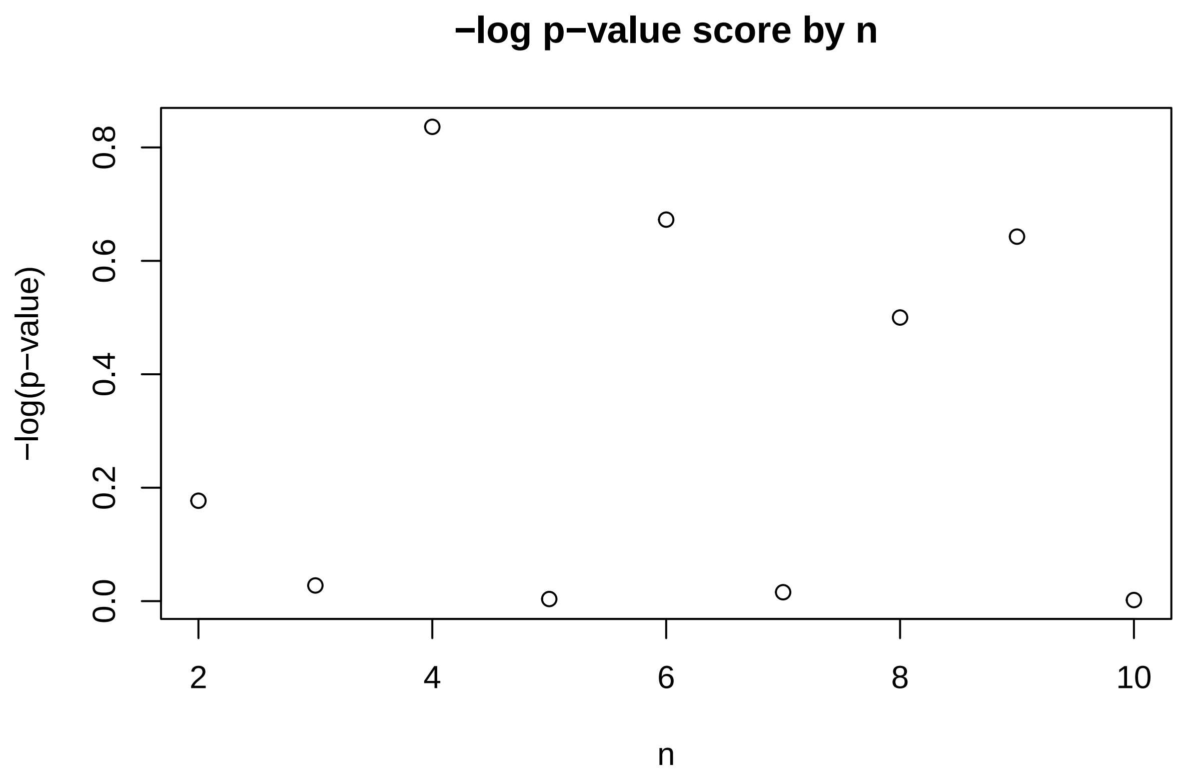
```
plot(density(eigen_total_sums[["5"]]))
```



This match with my expectation as explained in c), the distribution will approaches a normal distribution as the sample size grows.

(e) (3 points) Use a two-sided Kolmogorov-Smirnov test to assess the goodness-of-fit of the approximate distribution of a standardized eigenvalue sum according to the CLT (what you specified in part c)) for the empirical distribution (note above hint regarding use of ks.test() with y="pnorm"). Plot the -log(p-value) from these tests relative to the value of n. Do these results match your expectations?

```
eigen_total_sums_std <- hash()

for (i in n_dim) {
  temp_mean <- mean(eigen_total_sums[[as.character(i)]])
  temp_sd <- sd(eigen_total_sums[[as.character(i)]])
  eigen_total_sums_std[[as.character(i)]] <- (eigen_total_sums[[as.character(i)]] - temp_mean)/temp_sd
}

ks_pvalues <- c()

for (i in n_dim) {
  #temp_pval <- ks.test(eigen_total_sums[[as.character(i)]], y="pnorm")$p.value
  temp_pval <- ks.test(eigen_total_sums_std[[as.character(i)]], y="pnorm")$p.value
  ks_pvalues <- c(ks_pvalues, -log(temp_pval))
}

plot(n_dim, ks_pvalues,
     main="-log p-value score by n",
     xlab="n",
     ylab="-log(p-value)")
```


−log p−value score by n
−log(p−value)
n