به نام خدا

دانشگاه تهران

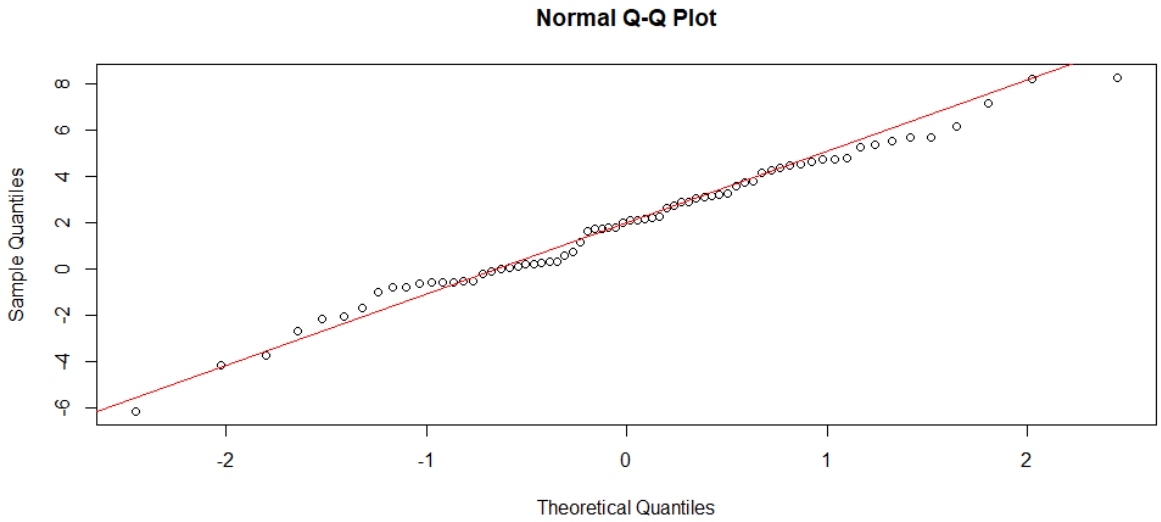
دانشکده‌ی مهندسی برق و کامپیوتر

تمرین کامپیوتری دوم استنباط آماری

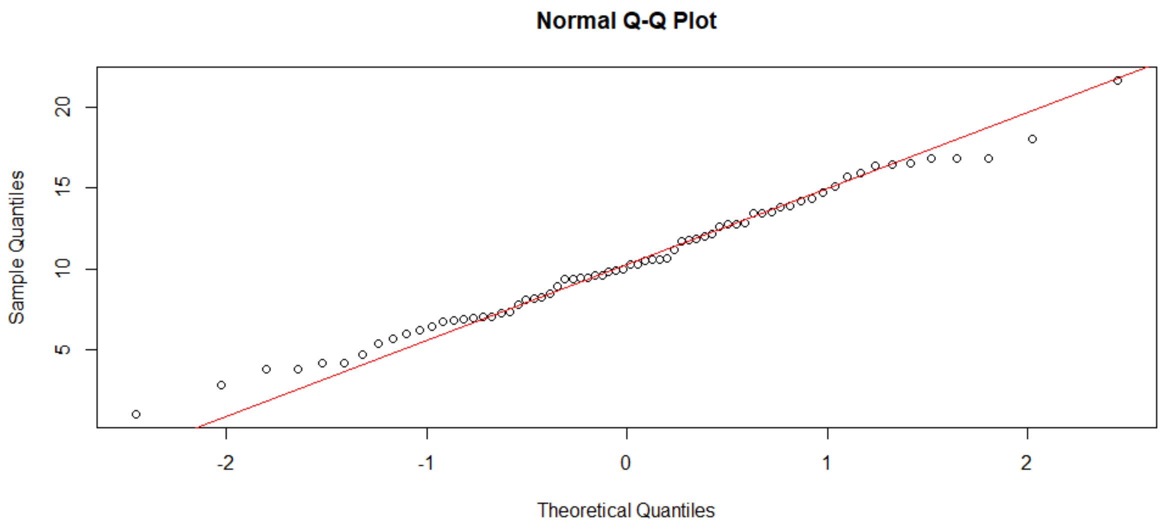
نام و نام خانوادگی	علیرضا فداکار
شماره‌ی دانشجویی	810195555

سوال 1)

با استفاده از تابع qqnorm نمودار qq plot را برای دو توزیع نرمال $N(2,3)$ و $N(10,4)$ نسبت به نرمال استاندارد به ترتیب در دو شکل 1 و 2 رسم می کنیم.



شکل 1-1



شکل 2-1

سوال ۲)

مشکل این آماره این است که به ازای $i = n$ داریم:

$$F^{-1}\left(\frac{i}{n}\right) = F^{-1}(1) = \infty$$

بنابرین آماره D_n گفته شده در صورت سوال برای هر n بینهایت می شود. این مشکل در آماره $kolmogorov\ smirnov$ با فرض اینکه F پیوسته باشد وجود ندارد. برای اینکه موارد مذکور را با استفاده از شبیه سازی ببینیم کافیست کد شکل 2-1 را پیاده سازی کنیم:

```r
n0 <- 10
size <- 1000
Dn <- matrix(0, size-n0+1)
Dn_kol <- matrix(0, size-n0+1)
for(n in c(n0:size)) {
  X <- rnorm(n)
  Fn <- ecdf(X)
  Y <- sort(X)
  Xn <- pnorm(Y)
  D <- abs(Y-qnorm(Fn(Y)))
  Dn[n-n0] <- max(D)
  D_kol <- abs(Xn-Fn(Y))
  Dn_kol[n-n0] <- max(D_kol)
}
```

شکل 1-2: پیاده سازی کد در r

در کد شکل 2-1 آماره ذکر شده در صورت سوال و آماره Kolmogorov smirnov را با استفاده از تابع توزیع گوسی پیاده سازی کرده ایم که آماره kolmogorov به صورت زیر است:

$$kolmogorov\ statistic \Longrightarrow Dn = \sup\ |F(x_i) - \frac{i}{n}|$$

آماره گفته شده در صورت سوال را برای n های مختلف در بردار Dn و آماره Kolmogorov را در بردار Dn_kov ذخیره کرده ایم. پس از اجرای کد مشاهده می شود که تمام عناصر Dn ، inf هستند که در شکل 2-2 چند عنصر اول آن قابل مشاهده است:

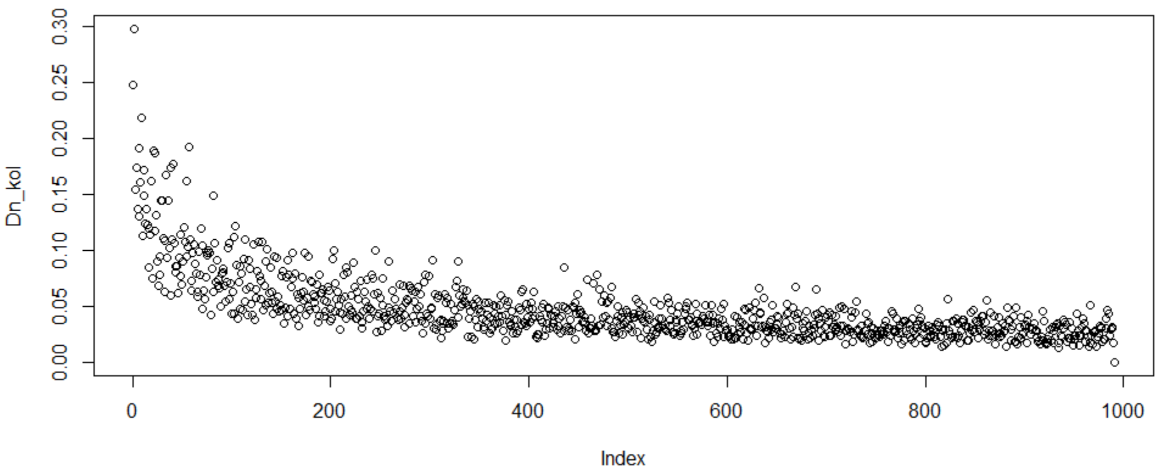
```
> Dn
      [,1]
[1,]   Inf
[2,]   Inf
[3,]   Inf
[4,]   Inf
[5,]   Inf
[6,]   Inf
[7,]   Inf
```

شکل 2-2

اما آماره Kolmogorov که در شکل 2-3 چند عنصر اول آن قابل مشاهده است هیچ مشکلی نداشته و به خوبی همگرا می شود که نمودار آن به ازای $10 \leq n \leq 1000$ در شکل 2-4 نشان داده شده است.

```
> Dn_kol
            [,1]
[1,] 0.24756922
[2,] 0.29741176
[3,] 0.15448170
[4,] 0.17378357
[5,] 0.13680588
[6,] 0.19138723
[7,] 0.12972518
[8,] 0.16051469
[9,] 0.21784195
```

شکل 2-3



شکل 2-4: نمودار Dn_kol

سوال 3)

$$\begin{cases} H_0: \mu_1 - \mu_2 = 0 \\ H_1: \mu_1 - \mu_2 \neq 0 \end{cases}$$

فرض میکنیم طول هر دو سمپل X و Y برابر n باشد. در این صورت:

$$Var(\bar{X} - \bar{Y}) = \sigma^2 \left(\frac{1}{n} + \frac{1}{n}\right) = \frac{2\sigma^2}{n}$$

آماره به صورت زیر است:

$$Z = \frac{\bar{X} - \bar{Y}}{\sigma\sqrt{\frac{2}{n}}}$$

بنابرین ناحیه rejection region به صورت زیر است:

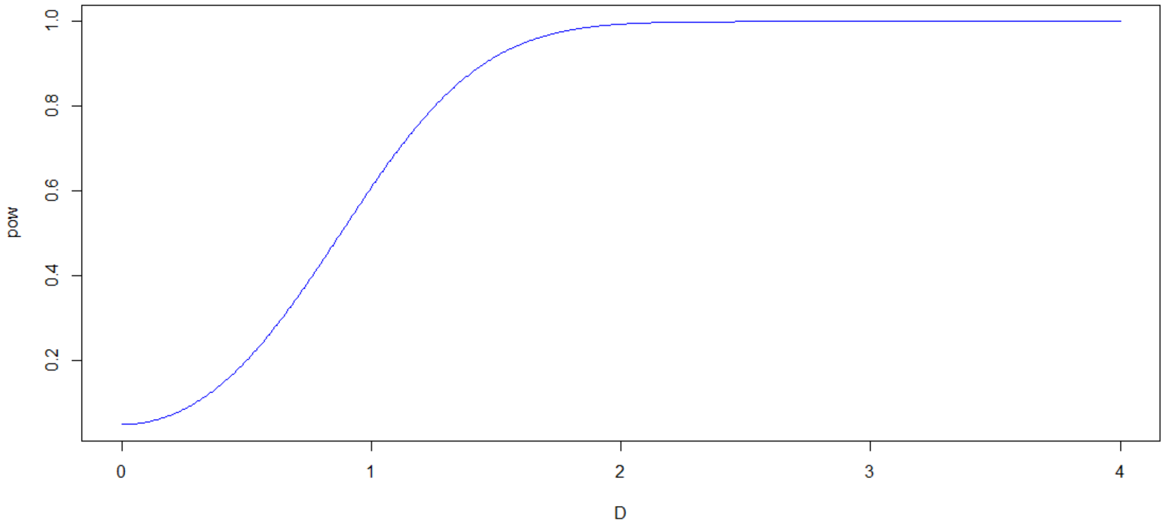
$$|\bar{X} - \bar{Y}| > z\left(\frac{\alpha}{2}\right)\sigma\sqrt{\frac{2}{n}}$$

مقدار *power* اگر $\mu_1 - \mu_2 = \Delta$ به صورت زیر بدست می آید:

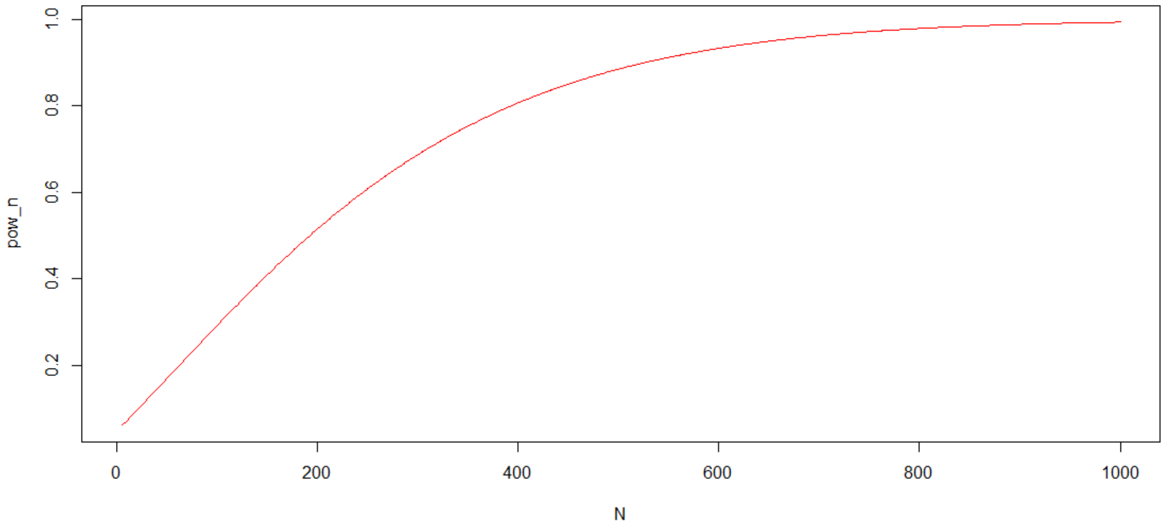
$$power = p\left[|\bar{X} - \bar{Y}| > z\left(\frac{\alpha}{2}\right)\sigma\sqrt{\frac{2}{n}}\right]$$

$$= 1 - \phi\left[z\left(\frac{\alpha}{2}\right) - \frac{\Delta}{\sigma}\sqrt{\frac{n}{2}}\right] + \phi\left[-z\left(\frac{\alpha}{2}\right) - \frac{\Delta}{\sigma}\sqrt{\frac{n}{2}}\right]$$

طبق فرض سوال $\sigma = 10$ بنابرین به کمک rstudio نمودار power بر اساس رابطه بالا بر حسب Δ و n به ترتیب به صورت شکل 3-1 و 3-2 است.

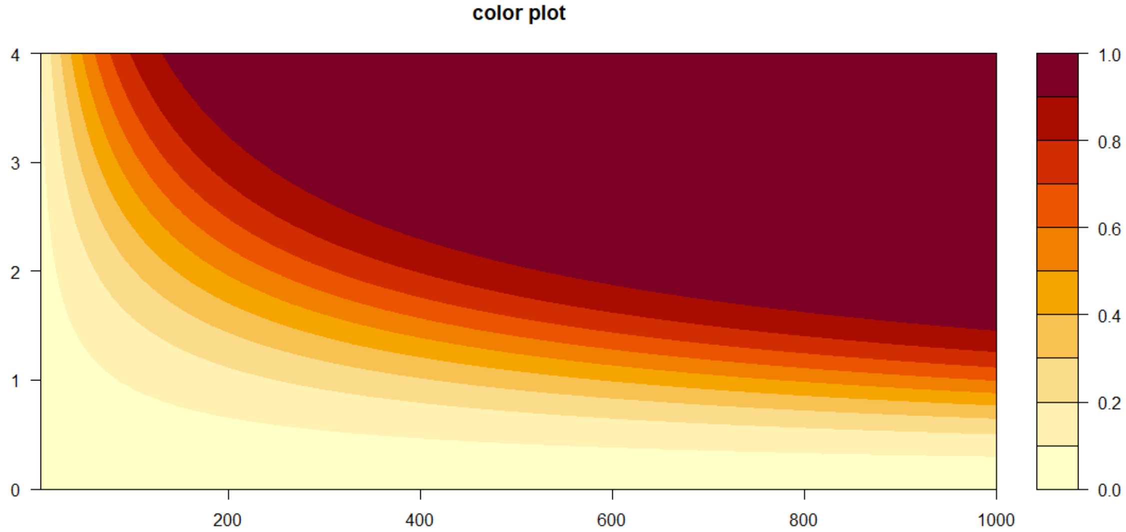


شکل 3-1: نمودار power بر حسب Δ



شکل 3-2 : نمودار power بر حسب n با فرض $\Delta = 2$

همچنین شکل 3-3 نمودار power را بر حسب هر دو متغیر Δ و n نشان می دهد که محور عمودی Δ و محور افقی n می باشد. در نواحی که رنگ پر رنگ تر است مقدار power بیشتر است. از این نمودار نیز می توان نتیجه گرفت با افزایش Δ و n مقدار $power$ افزایش می یابد.



شکل 3-3: نمودار power بر حسب Δ و n

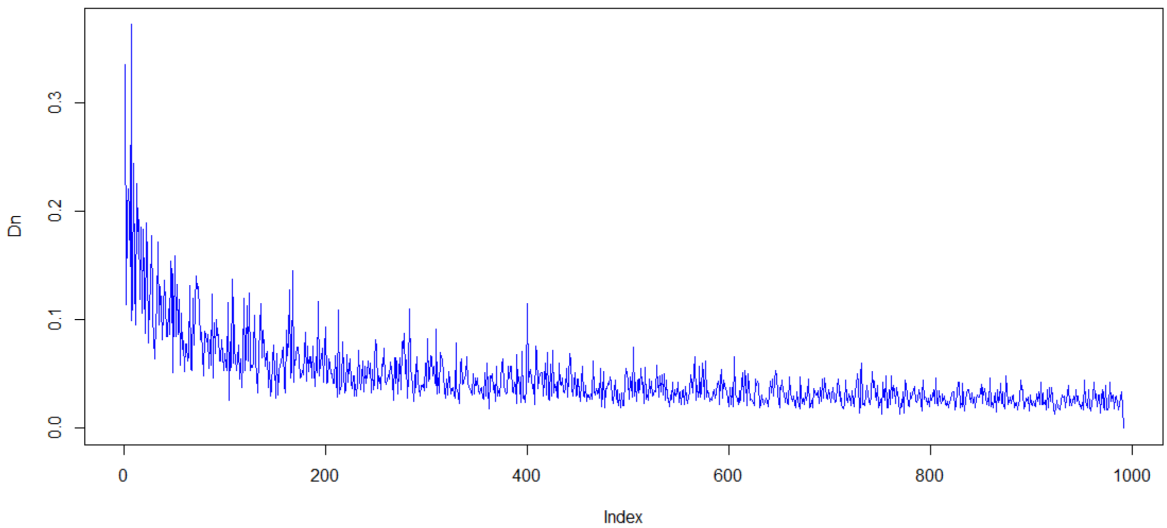
سوال 4)

کافیست طبق شکل 4-1 از توزیع نرمال نمونه با n های مختلف نمونه برداری کرده و نمودار Dn را بر حسب n رسم کنیم. این کارها را مطابق کد شکل 4-1 انجام می دهیم. نمودار در شکل 4-2 نشان داده شده است.

```r
n0 <- 10
size <- 1000
Dn <- matrix(0, size-n0+1)
for(n in c(n0:size)) {
  X <- rnorm(n)
  Fn <- ecdf(X)
  Y <- sort(X)
  Xn <- pnorm(Y)
  D <- abs(Xn-Fn(Y))
  Dn[n-n0] <- max(D)
}

plot(Dn, type = 'l', col = 'blue')
```

شکل 4-1



شکل 4-2: نمودار D_n بر حسب n

سوال 5)

الف)

از آماره χ^2 pearson می توان استفاده کرد.

ب)

فرض کنید f_i مقادیر مشاهده شده در جدول و e_i مقادیر $expected$ باشد. در این صورت تحت فرض صفر برای $0 \leq i \leq 12$ داریم $e_i = n \times p_i$ که در آن $n = 6115$ تعداد کل خانواده ها و:

$$p_i = \binom{12}{i}(0.5)^{12}$$

به کمک rstudio که کد آن در شکل 1-5 قابل مشاهده است ، مقادیر e_i را محاسبه می کنیم که در جدول 1-5 نشان داده شده است.

```
e <- matrix(0, 13)
n <- 6115
for(i in c(1:13)){
  e[i] <- n*dbinom(i-1, size = 12, prob = 0.5)
}
```

شکل 1-5

تعداد فرزندان (i)	f_i	e_i
0	7	1.49292
1	45	17.91504
2	181	98.53271
3	478	328.44238
4	829	738.99536
5	1112	1182.39258
6	1343	1379.45801
7	1033	1182.39258
8	670	738.99536
9	286	328.44238
10	104	98.53271
11	24	17.91504
12	3	1.49292

جدول 1-5

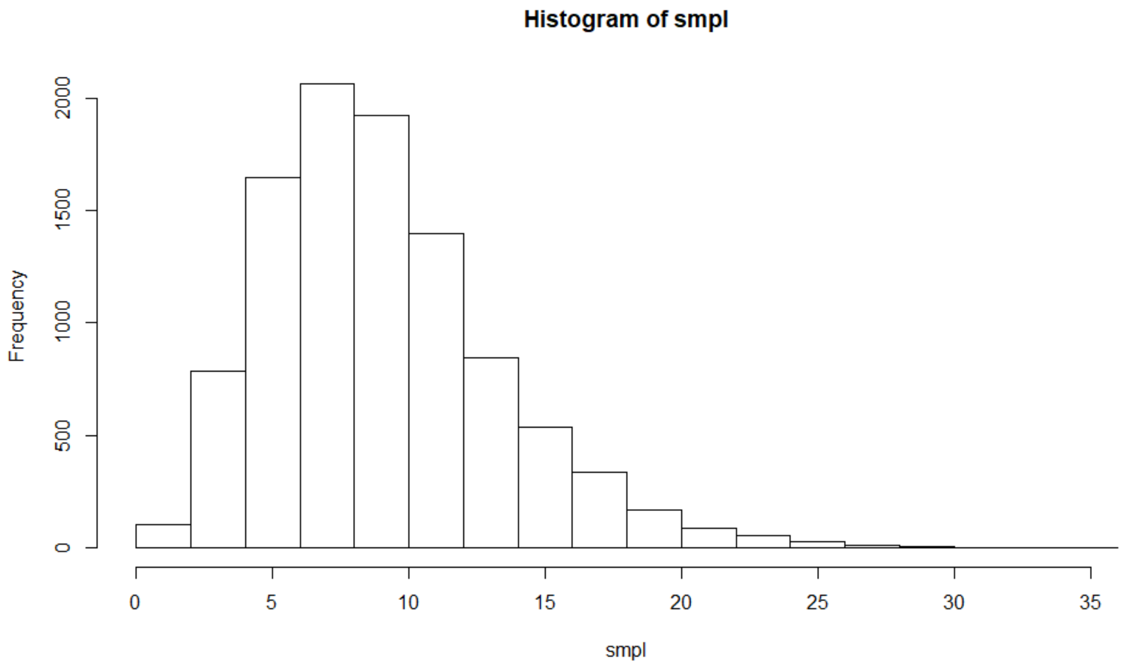
چون e_0 و e_{12} کمتر از 5 هستند سطر 1 را با سطر 2 و سطر 13 را با سطر 12 ، merge می کنیم.
جدول به صورت جدول 2-5 تغییر می کند:

تعداد فرزندان (i)	f_i	e_i
0,1	52	19.40796
2	181	98.53271
3	478	328.44238
4	829	738.99536
5	1112	1182.39258
6	1343	1379.45801
7	1033	1182.39258
8	670	738.99536
9	286	328.44238
10	104	98.53271
12,11	27	19.40796

پس به کمک نرم افزار rstudio آماره χ^2 به صورت زیر محاسبه می شود:

$$T = \sum_{i=0}^{10} \frac{(e_i - f_i)^2}{e_i} = 242.0463$$

می دانیم T تحت فرض H_0 توزیع chi square با درجه آزادی $df = 11 - 1 - 1 = 9$ دارد.
هیستوگرام این توزیع در شکل 2-5 رسم شده است:



شکل 2-5

ج)

به کمک دستور $1-pchisq(T, df=9)$ مقدار p_value برابر 0 بدست می آید. پس به ازای هر سطح معناداری α فرض H_0 رد می شود.

سوال 6)

الف)

می توانیم از f test برای مقایسه واریانس دو توزیع استفاده کنیم.

ب)

مطابق شکل 6-1 ، دو سمپل X و Y به طول 10000 به ترتیب با واریانس 3 و 10 نمونه برداری می کنیم. سپس $f\ test$ را روی این دو سمپل با سطح معناداری 0.05 و فرض H_1 دو طرفه اعمال می کنیم.

```r
#Generating two normal samples
N <- 10000
X <- rnorm(N, mean = 0, sd = sqrt(3))
Y <- rnorm(N, mean = 0, sd = sqrt(10))

#F test
var.test(X, Y, ratio = 1, alternative = c("two.sided"),
         conf.level = 0.95)
```

شکل 6-1

نتیجه تست پس از اجرای کد به صورت شکل 6-2 می باشد:

```
        F test to compare two variances

data:  X and Y
F = 0.30299, num df = 9999, denom df = 9999,
p-value < 2.2e-16
alternative hypothesis: true ratio of variances is not equal to 1
95 percent confidence interval:
 0.2913403 0.3151028
sample estimates:
ratio of variances
         0.3029887
```

شکل 6-2

سوال 7)

الف)

از روش پارامتریک two sample t-test و روش غیر پارامتری mann-whitney test استفاده می کنیم.

قبل از شروع با استفاده از تابع read.csv مطابق شکل 7-1 دیتاست را در یک ماتریس ذخیره می کنیم و چون دیتاست دو ستون دارد ، دو ستون آن را در بردار های X و Y ذخیره می کنیم.

```r
f <- file.choose()

#Importing csv data
grades <- read.csv(file = f, head = FALSE)
grades <- as.matrix(grades)
x <- grades[,1]
Y <- grades[,2]
```

شکل 7-1

ابتدا روش two sample t-test را بررسی می کنیم. برای این کار کافیست دستور $t.test(X,Y)$ را اجرا کنیم. خروجی این تابع در شکل 7-2 نشان داده شده است:

```
        Welch Two Sample t-test

data:  X and Y
t = 0.50875, df = 95.863, p-value = 0.6121
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -4.323031  7.302623
sample estimates:
mean of x mean of y
 67.61224  66.12245
```

شکل 7-2

همان طور که در شکل 7-2 مشاهده می شود چون مقدار p-value از سطح معناداری $\alpha = 5\ \%$ بیشتر است فرض H_0 را نمی توان رد کرد و بنابرین فرض H_0 را می پذیریم و میانگین دو داده برابر است.

حال از روش غیر پارامتری mann-whitney test استفاده می کنیم. کافیست دستور wilcox.test را وارد کنیم. نتیجه در شکل 7-3 نشان داده شده است:

```
        Wilcoxon rank sum test with continuity correction

data:  X and Y
W = 1235, p-value = 0.809
alternative hypothesis: true location shift is not equal to 0
```

شکل 3-7 : نتیجه تست mann whitney test

همان طور که در شکل 3-7 مشاهده میشود چون p-value بزرگتر از سطح معناداری است بیشتر است نمیتوان فرض صفر را رد کرد و می توان گفت میانگین دو داده برابر است.

همانطور که مشاهده می شود در هر دو تست پارامتری و غیر پارامتری به نتیجه یکسانی رسیدیم.

ب)

سوال ۸)

ابتدا به صورت شکل 8-1 دیتاست را import می کنیم در این سوال ما فقط از ستون دوم دیتاست که مربوط به بچه هاست استفاده می کنیم.

```r
f <- file.choose()

#Importing csv data
heights <- read.csv(file = f)
heights <- as.matrix(heights)
childs <- heights[,2]
```

شکل 8-1

الف)

چون طبق فرض انحراف معیار جامعه نامشخص است از توزیع t برای بدست آوردن بازه اطمینان استفاده می کنیم. بازه اطمینان $100(1-\alpha)\%$ برای یک sample با طول n از جامعه به صورت زیر است:

$$\left(\bar{X} - s_{\bar{X}} t_{n-1}\left(\frac{\alpha}{2}\right)\ , \bar{X} - s_{\bar{X}} t_{n-1}\left(\frac{\alpha}{2}\right)\right)$$

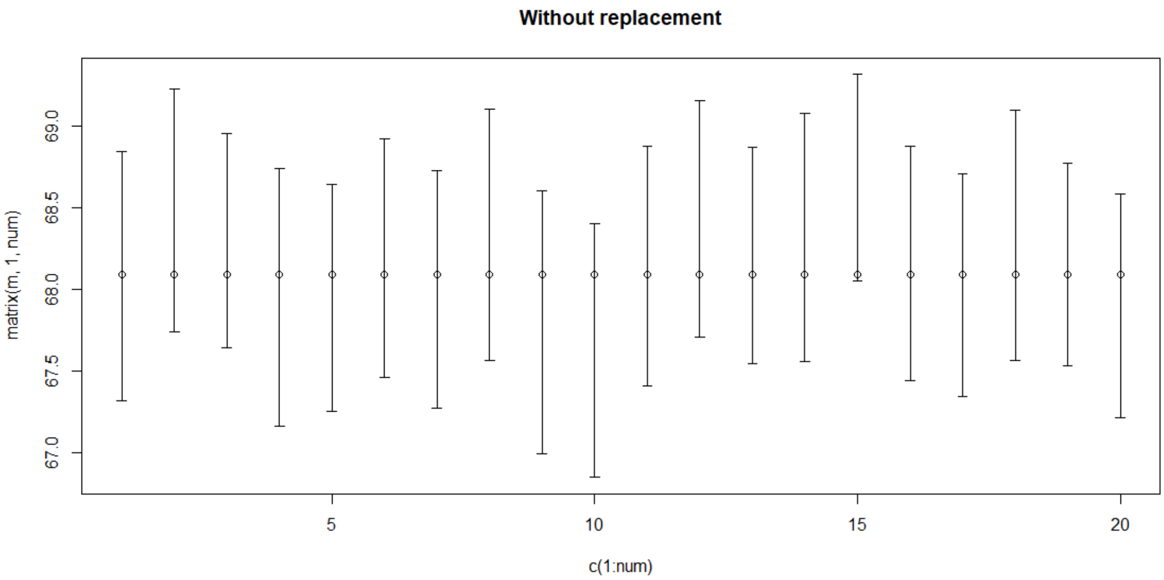
در نتیجه به صورت کد شکل 8-2 و دستور t.test (متد conf.int) بازه اطمینان ها را بدست می آوریم متغیر $output$ تعداد بازه هایی که میانگین واقعی جامعه را دارند نشان می دهد:

```r
size <- 1e4
u <- matrix(0, 1, size)
l <- matrix(0, 1, size)
a <- 1e-1/2
m <- mean(childs)
smpl_size <- 10
d_f <- smpl_size - 1
output <- 0

#Without replacement
for (i in c(1:size)) {
  smpl <- sample(childs, smpl_size, replace = FALSE, prob = NULL)
  test <- t.test(smpl, alternative = "two.sided", conf.level = 0.9)
  confidence <- test$conf.int
  l[i] <- confidence[1]
  if (u[i]>m && m>l[i]) {
    output <- output + 1
  }
```

شکل 8-2 : بدست آوردن بازه اطمینان

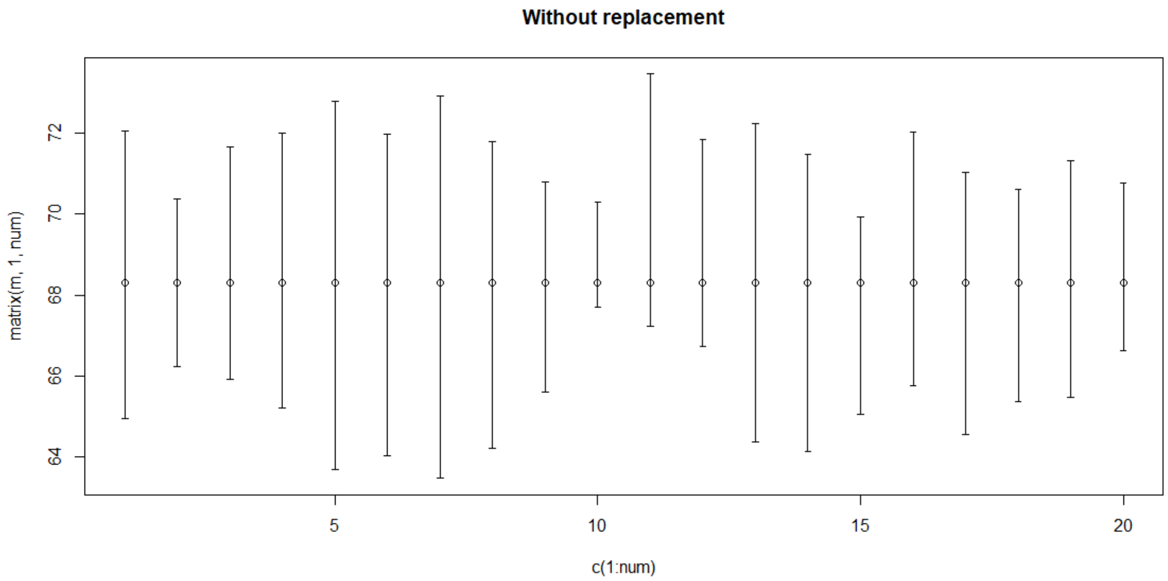
پس از اجرای کد مشاهده می کنیم که مقدار $output$ برابر 19479 یعنی تعداد کل نمونه هاست بنابرین 97.395% بازه ها میانگین واقعی جامعه را شامل می شوند. همچنین تعداد 20 بازه اول را رسم کرده ایم که در نمودار شکل 3-8 قابل مشاهده است:



شکل 3-8 : بازه اطمینان 20 سمپل

ب)

به طور مشابه با الف پس از تغییر تعداد سمپل و سطح معناداری و طول هر سمپل و سپس اجرای کد ، مقدار $output$ برابر 9021 می شود که این بار 90.21% بازه ها شمل میانگین واقعی هستند. شکل 4-8 تعداد 20 بازه اطمینان را نشان می دهد.



شکل 4-8 : تعداد 20 بازه اطمینان

از انجام این دو آزمایش نتیجه می گیریم که اگر n سمپل داشته باشیم و بازه اطمینان هر یک را بدست آوریم ، تقریبا $100(1-\alpha)\%$ بازه ها شامل میانگین واقعی جامعه هستند.

سوال ۹)

الف و ب)

ابتدا سمپل را ایجاد کرد و سپس تست one sample signed rank را اعمال می کنیم و سپس $power$ را بدست می آوریم که در کد شکل 1-9 قابل مشاهده است:

```
1  #sample from beta dist
2  size <- 50
3  a <- 2
4  b <- 5
5  m0 <- 0.4
6  alpha <- 0.05
7  X <- rbeta(size, a, b)
8  m <- sum(X > m0)
9  p_value <- pbinom(m, size, 0.5, lower.tail = FALSE)
10
11 # Calculating p-value using wilcox.test
12 wilcox.test(X, mu = m0, alternative = 'greater')
13
14 # Calculating power of the test
15 p <- 1 - pbeta(m0, a, b)
16 decision <- qbinom(1 - alpha, size, 0.5)
17 power <- pbinom(decision, size, p, lower.tail = FALSE)
```

شکل 1-9

توجه داشته باشید مقدار p_value را به دو روش محاسبه می کنیم. در روش اول ابتدا test statistic را که برابر تعداد داده های بزرگتر از 0.4 هست محاسبه کرده ایم و در متغیر m ذخیره می کنیم. سپس ب استفاده از دستور pbinom مقدار p_value را (خط 9 شکل 1-9) محاسبه می کنیم که این مقدار برابر است با:

$$p\ value = \ 0.9999881$$

در روش دوم می توانیم از دستور wilcox.test استفاده کنیم (خط 12 کد) که البته این روش کمی خطا داشته و مقدار 1 را به عنوان p-value خروجی می دهد.

در نهایت چون مقدار p-value نزدیک یک است در هر سطح معناداری نمی توانیم H_0 را رد کنیم.

مقدار power را نیز در خط 15,16,17 کد شکل 1-9 محاسبه کرده ایم که برابر است با:

$$power = 1.070415 \times 10^{-9}$$

که قابل انتظار نیز بود.

ج)

در این قسمت پس از تغییر $m0 = 0.6$ (شکل 1-9) کد را دوباره اجرا می کنیم. مقدار **p-value** و **power** به صورت زیر بدست می آید:

$$p - value = 1$$

$$power = 3.430504 \times 10^{-32}$$

د)

سوال 10)

ابتدا دیتاست را به صورت شکل 10-1 ، import می کنیم و سمپلی به طول 70 از آن در نظر می گیریم.

```
f <- file.choose()

#Importing csv data
heights <- read.csv(file = f)
heights <- as.matrix(heights)
fathers <- heights[,2]
n1 = 5
n2 = 500
pow <- matrix(0, n2-n1+1)

smpl_size <- 10
pop_mean <- mean(fathers)
pop_sd <- sd(fathers)

smpl <- sample(fathers, smpl_size, replace = FALSE, prob = NULL)
```

شکل 10-1

سپس t-test را اعمال کرده و بازه اطمینان را بدست آورده و در نهایت power را با دو روش بدست می آوریم در روش اول به صورت دقیق با استفاده از توزیع t-student و روش دوم با استفاده از تقریب نرمال بدست می آوریم. شکل 10-2 کدهای مربوطه را نشان می دهد:

```r
d_f <- smpl_size - 1

test <- t.test(smpl, mu = 60, alternative = "two.sided")

confidence <- test$conf.int
u <- confidence[2]
l <- confidence[1]

# Exact Power using normal estimation
power1 <- 0
u1 <- (u-pop_mean)/pop_sd
l1 <- (l-pop_mean)/pop_sd
power1 <- pt(u1, df = d_f, lower.tail = FALSE)
power1 <- power1 + pt(l1, df = d_f)

# Power using normal estimation
power <- 0
power <- pnorm(u, pop_mean, pop_sd, lower.tail = FALSE)
power <- power + pnorm(l, pop_mean, pop_sd)
```

شکل 2-10

بازه اطمینان به صورت زیر است:

$$confidence\ interval \Longrightarrow (67.82396\,, 68.86175)$$

از طرفی p-value نیز برابر است با:

$$p - value = \ 2.2 \times 10^{-16}$$

بنابرین فرض H_0 رد می شود و ناحیه $rejection$ خارج بازه اطمینان است. مقدار power نیز به صورت زیر است:

$$power = 0.7724859$$

ب)

در این حالت که تعداد نمونه 10 باشد نتایج به صورت زیر است:

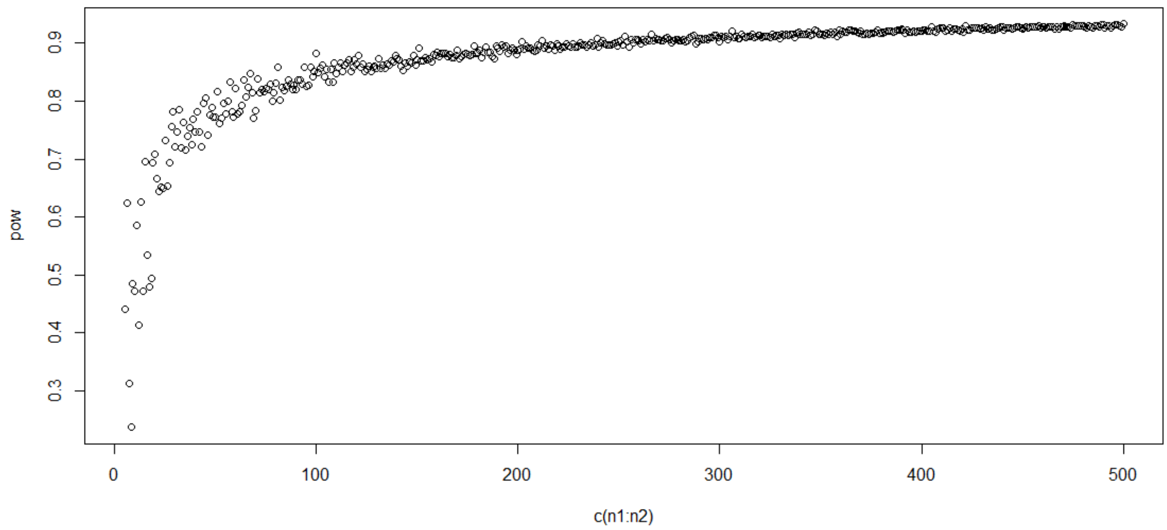
$$p - value = 2.583 \times 10^{-8}$$

$$confidence\ interval \Longrightarrow (67.59177\,, 69.80823)$$

$$power = 0.5449347$$

ج)

با مقایسه دو قسمت قبل مشاهده می کنیم با کاهش اندازه سمپل مقدار power کمتر شده است. برای درک بهتر این موضوع نمودار power بر حسب اندازه سمپل را در یک نمودار رسم می کنیم که در شکل 10-3 نشان داده شده است. همان طور که مشاهده می شود مقدار power در نهایت همگرا می شود.



شکل 10-3 : نمودار power بر حسب n