https://colab.research.google.com/drive/1Te_yh1yOQ4DYHhWrPzbxE3Ttsu1fSgNQ#scrollTo=_2bgId5V1m5h

https://colab.research.google.com/drive/1PnombMcnWiAW2DHp14giJedbeiax8CD2

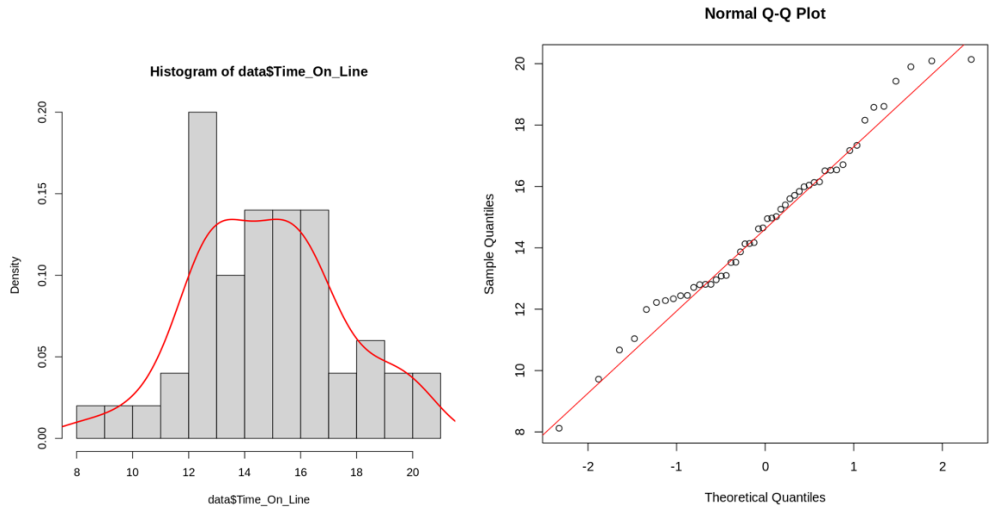
1–داده های زمان انتظار از چه توزیعی پیروی میکنند؟ نتایج آزمونهای موردنیاز را گزارش کنید.

```
data=read.csv("/content/Atlas_Khodro.csv",header = T)
head(data)
```

A data.frame: 6 × 6

	Time_On_Line	Number_Accepting_Insurance	Complaints_per_20_Customers	Wait_Time_for_Bus	Gas_To_Fill	Time_On_Line_Improved
	<dbl>	<int>	<int>	<dbl>	<dbl>	<dbl>
1	20.14	160	0	-4.36	65.25	11.160
2	14.97	174	1	-3.75	9.75	10.203
3	9.72	156	1	-1.75	32.25	11.127
4	14.65	146	3	-1.13	14.00	13.068
5	12.81	146	2	-1.36	12.25	11.004
6	19.90	170	2	-2.78	7.50	12.936

```
hist(data$Time_On_Line,freq = F,breaks = 15)
lines(density(data$Time_On_Line),col = "red",lwd = 2)

qqnorm(data$Time_On_Line)
qqline(data$Time_On_Line,col="red")
```



با توجه به شکل میتوان به طور تقریبی حدس زد داده های موجود در ستون زمان مورد انتظار میتوانند دارای توزیع نرمال باشند.

```
#shapiro test
shapiro.test(data$Time_On_Line)
```

```
        Shapiro-Wilk normality test

data:  data$Time_On_Line
W = 0.98199, p-value = 0.6386
```

```
# p-value > 0.05 --> normal distribution --> T
```

با توجه به مقدار p–value در تست shapiro میتوان نتیجه گرفت داده های مورد نظر میتوانند دارای توزیع نرمال باشند.

```r
# Skewness and Kurtosis test

install.packages("moments")
library(moments)

jarque.test(data$Time_On_Line)
```

```
        Jarque-Bera Normality Test

data:  data$Time_On_Line
JB = 0.095754, p-value = 0.9533
alternative hypothesis: greater
```

```r
# p-value > 0.05 --> normal distribution --> T
```

با توجه به مقدار p–value در تست jarque میتوان نتیجه گرفت داده های مورد نظر میتوانند دارای توزیع نرمال باشند.

```r
#Anscombe-Glynn test

anscombe.test(data$Time_On_Line)
```

```
        Anscombe-Glynn kurtosis test

data:  data$Time_On_Line
kurt = 2.83100, z = 0.12428, p-value = 0.9011
alternative hypothesis: kurtosis is not equal to 3
```

```r
# p-value > 0.05 --> normal distribution --> T
```

با توجه به مقدار p–value در تست anscombe میتوان نتیجه گرفت داده های مورد نظر میتوانند دارای توزیع نرمال باشند.

2–بهترین برآورد شما از زمان انتظار مشتریان چقدر است؟

```
#point estimation of mean
mean_t=mean(data$Time_On_Line)
cat("point estimation =",round(mean_t,4),"\n")

#interval estimation of Mean with Unkown variance
#confidence interval=0.95
n=length(data$Time_On_Line)
sd_t=sd(data$Time_On_Line)
alpha=0.05
t=qt(alpha/2,n-1)

lower_mean = mean_t - (abs(t) * sd_t / sqrt(n))
upper_mean = mean_t + (abs(t) * sd_t / sqrt(n))

cat("interval estimation =","[",round(lower_mean,4),",",round(upper_mean,4),"]")
```

به ترتیب برآورد نقطه ای و بازه ای از میانگین زمان انتظار مشتریان عبارتند از:

```
point estimation = 14.7788
interval estimation = [ 14.02024 , 15.53736 ]
```

3–چند درصد مشتریان بیش از 15 دقیقه منتظر می مانند تا بتوانند به مرحله عقد قرارداد برسند؟

```
count15=0
for (x in data$Time_On_Line){
  if(x>15){
    count15=count15+1
  }
}
cat("percentage of values more than 15 = ",count15/length(data$Time_On_Line)*100,"%")
```

```
percentage of values more than 15 =  46 %
```

با توجه به این که این ستون از توزیع نرمال پیروی میکند از روش زیر هم میتوان استفاده نمود که نتیجه‌ای یکسان در پی دارد:

```
pnorm(15, mean(data$Time_On_Line), sd(data$Time_On_Line), lower.tail = FALSE)*100
```

```
percentage of values more than 15 =  46.6976 %
```

4–چند درصد جمعیت مشتریان اطلس خودرو، بیمه تکمیلی را روی قرارداد خود دریافت می کنند؟

```
sum_Insurance=0
for (x in data$Number_Accepting_Insurance){
  if(!is.na(x)){
    sum_Insurance=sum_Insurance+x
  }
}
cat("percentage of people whose insurance is accepted ",(sum_Insurance/(250*30))*100)
```

```
percentage of people whose insurance is accepted =  61.12
```

5–فرض کنید دو نفر به دفتر اطلس خودرو در فرودگاه مراجعه می کنند. با چه احتمالی هردو خواستار بیمه تکمیلی هستند؟

```
p=sum_Insurance/(250*30)
nq5=2

#probablity of exactly 2 insurance requests
cat("P(X=2) =",round(dbinom(2,nq5,p),4),"\n")
```

```
P(X=2) = 0.3736
```

6–اگر به صورت تصادفی 10 نفر را که به دفتر اطلس خودرو در فرودگاه مراجعه کردند، انتخاب کنید با چه احتمالی حداقل 5 نفرشان خواستار بیمه تکمیلی هستند؟

```
nq6=10
cat("P(X>=5) =",round(pbinom(4,nq6,p,lower.tail = F),4),"\n")
```

```
P(X>=5) = 0.8519
```

7–برای پاسخ به پرسش های 5 و 6 از چه توزیعی استفاده کردید؟ نتایج آزمونهای موردنیاز را گزارش کنید .

متغیرهای مورد بررسی ما از نوع گسسته هستند و در حل هر دو سوال از توزیع دوجمله ای استفاده میکنیم زیرا

الف. آزمایش ها از هم مستقل اند.

ب. آزمایش ها دارای دو برآمد موفقیت و شکست هستند. ←– خواستار بیمه بودن یا نبودن

پ. احتمال برآمد ها ثابت است. ←– 0. 6112

8–فرض کنید شما 20 مراجعه کننده به دفتر اطلس خودرو در فرودگاه را به شکل تصادفی انتخاب کردید و از آنان درباره تجربه دریافت خدمت شان پرسش کردید. با چه احتمالی دقیقا 4 شکایت ثبت خواهید کرد؟

```
mean_Complaints=sum(data$Complaints_per_20_Customers[1:25])/ (25*20)
cat("P(X=4) =",dpois(4,mean_Complaints),"\n")
```

9–فرض کنید شما 20 مراجعه کننده به دفتر اطلس خودرو در فرودگاه را به شکل تصادفی انتخاب کردید و از آنان درباره تجربه دریافت خدمت شان پرسش کردید. با چه احتمالی بیش از 4 شکایت ثبت خواهید کرد؟

```
cat("P(X>4) = P(X>=5) =",round(ppois(4,mean_Complaints, lower.tail = F),4),"\n")
```

```
P(X=4) = 0
-----------------------------------------------------
P(X>4) = P(X>=5) = 0
```

10–برای پاسخ به پرسش های 8 و 9 از چه توزیعی استفاده کردید؟ نتایج آزمونهای موردنیاز را گزارش کنید .

متغیرهای مورد بررسی ما از نوع گسسته هستند در حل هر دو سوال از توزیع پوآسن استفاده میکنیم زیرا

الف. آزمایش ها از هم مستقل اند.

ب. رخداد ها در یک بازه بررسی میشوند. ←– انتخاب تصادفی ۲۰ مشتری

```
poisson.test(x=sum(data$Complaints_per_20_Customers[1:25]),T=25)
```

```
        Exact Poisson test

data:  sum(data$Complaints_per_20_Customers[1:25]) time base: 25
number of events = 24, time base = 25, p-value = 1
alternative hypothesis: true event rate is not equal to 1
95 percent confidence interval:
 0.6150901 1.4284039
sample estimates:
event rate
      0.96
```

11–بهترین برآورد شما از زمان انتظار مشتریان برای اتوبوس چقدر است؟

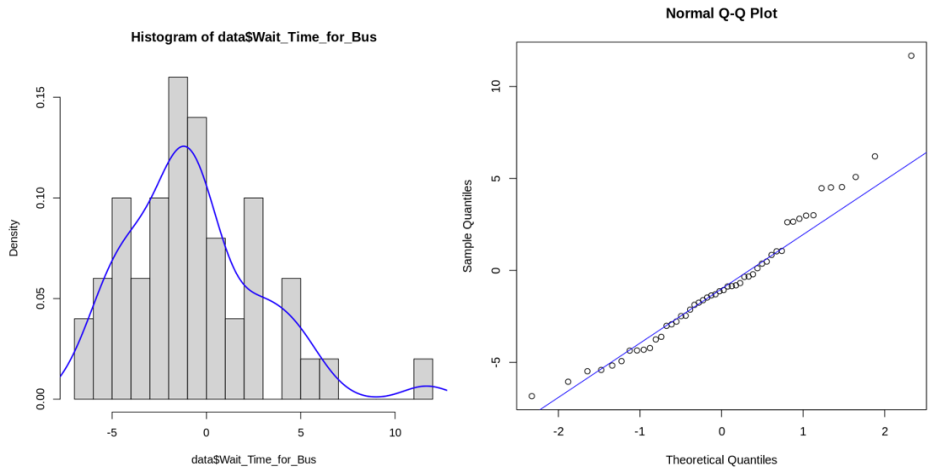
```
#point estimation of Mean
sum_tBus=0
for (x in data$Wait_Time_for_Bus){
  sum_tBus=sum_tBus+ (x+10)
}
mean_tBus=sum_tBus/50
cat("point estimation =",round(mean_tBus,4),"\n")

#interval estimation of Mean with Unkown variance
#confidence interval=0.95
n_tBus=length(data$Wait_Time_for_Bus)
sd_tBus=sd(data$Wait_Time_for_Bus)
alpha=0.05
t11=qt(alpha/2,n_tBus-1)

lower_mean11 = mean_tBus - (abs(t11) * sd_t / sqrt(n_tBus))
upper_mean11 = mean_tBus + (abs(t11) * sd_t / sqrt(n_tBus))

cat("interval estimation =","[",round(lower_mean11,4),",",round(upper_mean11,4),"]")
```

```
point estimation = 9.2898
interval estimation = [ 8.5312 , 10.0484 ]
```

12–چند درصد مشتریان اطلس خودرو در فرودگاه، بیش از 15 دقیقه منتظر اتوبوس باقی میمانند؟
(راهنمایی: به پرسش 13 مراجعه کنید.)

13–برای پاسخ به پرسش 12 از چه توزیعی استفاده کردید؟ نتایج آزمونهای موردنیاز را گزارش کنید
(راهنمایی: بررسی کنید آیا این شاخص از توزیع نرمال پیروی میکند. اگر نه، چه تغییر متغیری باید بدهید تا این بتوانید از توزیع نرمال استفاده کنید. این شاخص را به زمان انتظار مشتری تبدیل کنید، سپس با تغییر متغیر لگاریتمی آن را به توزیع نرمال برگردانید .)

برای سوال ۱۳ ابتدا تست های مربوطه و همچنین نمودار توزیع این ستون را برای پاسخ به پرسش نرمال بودن داده ها را بررسی میکنیم؛ با مشاهده نتایج به دست آمده میتوان عدم پیروی داده ها از توزیع نرمال را ثابت نمود.



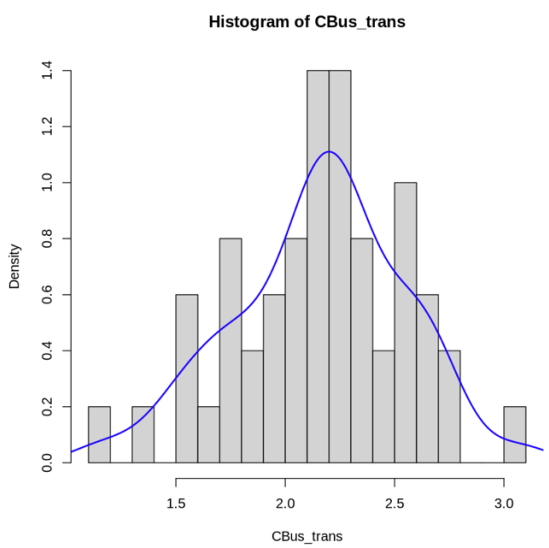
```r
#shapiro test
shapiro.test(data$Wait_Time_for_Bus)
cat("----------------------------------------")
# Skewness and Kurtosis test
jarque.test(data$Time_On_Line)
cat("----------------------------------------")
#Anscombe-Glynn test
anscombe.test(data$Time_On_Line)
```

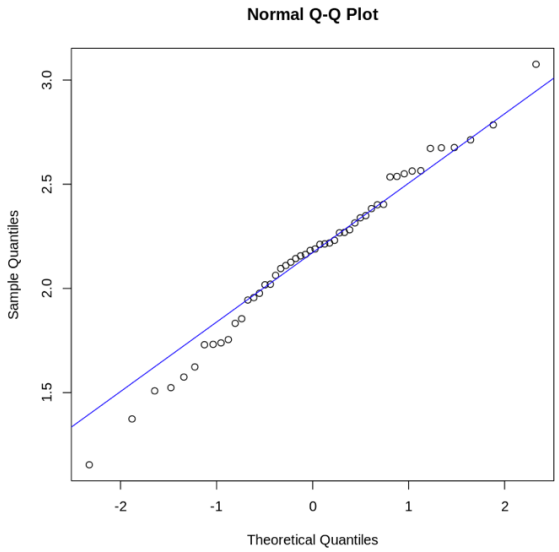
```
        Shapiro-Wilk normality test

data:  data$Wait_Time_for_Bus
W = 0.94992, p-value = 0.03387
----------------------------------------

        Jarque-Bera Normality Test

data:  data$Time_On_Line
JB = 0.095754, p-value = 0.9533
alternative hypothesis: greater
----------------------------------------

        Anscombe-Glynn kurtosis test

data:  data$Time_On_Line
kurt = 2.83100, z = 0.12428, p-value = 0.9011
alternative hypothesis: kurtosis is not equal to 3
```

```r
# p-value in Shapiro-Wilk normality test  < 0.05 --> normal distribution --> F
```

سپس با تبدیل متغیر های این ستون از شاخص به زمان اصلی خود و جایگزین کردن لگاریتم آن ها در برداری جدید دوباره نرمال بودن توزیع را بررسی میکنیم.

```r
CBus=data$Wait_Time_for_Bus
CBus_trans=CBus+10
CBus_trans=log(CBus_trans)
```





```
#shapiro test
shapiro.test(CBus_trans)
cat("----------------------------------------")
# Skewness and Kurtosis test
jarque.test(CBus_trans)
cat("----------------------------------------")
#Anscombe-Glynn test
anscombe.test(CBus_trans)
```

```
        Shapiro-Wilk normality test

data:  CBus_trans
W = 0.98814, p-value = 0.8935
----------------------------------------

        Jarque-Bera Normality Test

data:  CBus_trans
JB = 0.52775, p-value = 0.7681
alternative hypothesis: greater
----------------------------------------

        Anscombe-Glynn kurtosis test

data:  CBus_trans
kurt = 2.93832, z = 0.31746, p-value = 0.7509
alternative hypothesis: kurtosis is not equal to 3
```

```
# p-value in all of tests  > 0.05 --> normal distribution --> T
```

در نهایت برای سوال ۱۲ برای بررسی درصد مشتریانی که بیش از ۱۵ دقیقه منتظر اتوبوس بوده‌اند را از طریق زیر محاسبه میکنیم:

```
cat("percentage of people who wait more than 15 min for bus = ",
    round(pnorm(log(15), mean(CBus_trans), sd(CBus_trans),
    lower.tail =FALSE)*100,4),"%")
```

```
percentage of people who wait more than 15 min for bus =  8.0096 %
```

14–بهترین برآورد شما از حجم باقی مانده تا پرشدن باک بنزین بر حسب لیتر چقدر است؟

```
#point estimation of mean
mean_g=mean(data$Gas_To_Fill)
cat("point estimation =",round(mean_g,4),"\n")

#interval estimation of Mean with Unkown variance
#confidence interval=0.95
n14=length(data$Gas_To_Fill)
sd_g=sd(data$Gas_To_Fill)
alpha=0.05
t14=qt(alpha/2,n14-1)

lower_mean14 = mean_g - (abs(t14) * sd_t / sqrt(n14))
upper_mean14 = mean_g + (abs(t14) * sd_t / sqrt(n14))

cat("interval estimation =","[",round(lower_mean14,4),",",round(upper_mean14,4),"]")
```

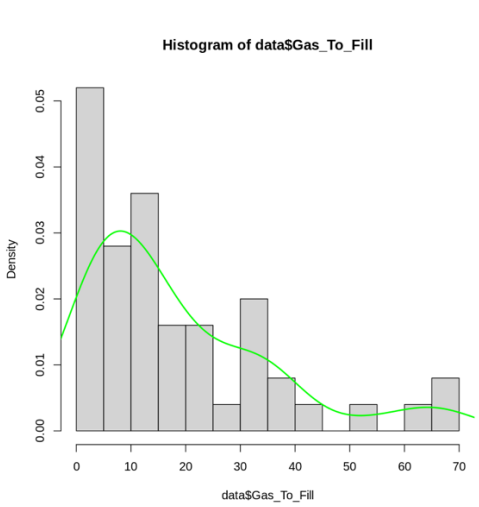
```
point estimation = 18.52
interval estimation = [ 17.7614 , 19.2786 ]
```

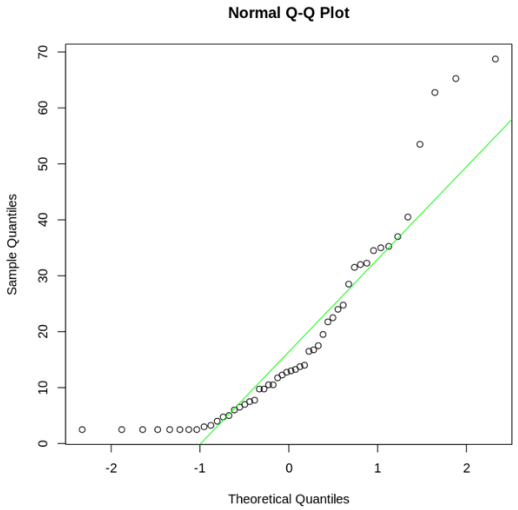
15–چند درصد خودروهای برگشتی 40 لیتر یا بیشتر بنزین نیاز دارند تا باکشان پر شود؟

16–با چه احتمالی دو خودروی برگشتی پشت سرهم نیاز دارند تا 40 لیتر یا بیشتر باکشان را پر کنند؟ (به این توجه کنید کدامیک از توزیعهای گسسته برای حل این مسئله مناسب است .)

17–برای پاسخ به پرسش های 15 و 16 چه فرضی درباره توزیع داده های ستون پنجم داشتید؟ نتایج آزمونهای موردنیاز را گزارش کنید .

متغیر حجم باقی مانده در باک متغیری از نوع پیوسته است که با توجه به نمودار و تست های انجام شده دارای توزیع نرمال نیست.





```r
#shapiro test
shapiro.test(data$Gas_To_Fil)
cat("----------------------------------------")
# Skewness and Kurtosis test
jarque.test(data$Gas_To_Fil)
cat("----------------------------------------")
#Anscombe-Glynn test
anscombe.test(data$Gas_To_Fil)
```

```
        Shapiro-Wilk normality test

data:  data$Gas_To_Fil
W = 0.83419, p-value = 5.932e-06
----------------------------------------

        Jarque-Bera Normality Test

data:  data$Gas_To_Fil
JB = 18.683, p-value = 8.773e-05
alternative hypothesis: greater
----------------------------------------

        Anscombe-Glynn kurtosis test

data:  data$Gas_To_Fil
kurt = 4.2560, z = 1.8711, p-value = 0.06133
alternative hypothesis: kurtosis is not equal to 3
```

اما با بررسی تست مربوط به توزیع نمایی و مقدار p_value میتوان نتیجه گرفت این داده ها دارای توزیع نمایی میباشند.

```r
# checking exponential dist

install.packages("exptest")
library("exptest")

shapiro.exp.test(data$Gas_To_Fil)
```

```
        Shapiro-Wilk test for exponentiality

 data:  data$Gas_To_Fil
 W = 0.018005, p-value = 0.266
```

درصد خودروهای بازگشتی که نیاز به ۴۰لیتر بنزین یا بیشتر دارند که در سوال ۱۵ خواسته شده:

```r
cat("percentage of cars which need >=40L gas = "
    round(pexp(40,rate=1/mean(data$Gas_To_Fil),lower.tail = F)*100,4),"%")
```

```
percentage of cars which need >=40L gas =  11.5345 %
```

برای سوال ۱۶ باید از توزیع هندسی استفاده کنیم زیرا اروی دادن رخداد پر کردن باک ۴۰لیتر یا بیش از آن را پس از دو موفقیت ، متغیری گسسته میباشد و احتمال وقوع آن خواسته شده است.

```
p16=pexp(40,rate=1/mean(data$Gas_To_Fil),lower.tail = F)
cat("P(X=2) =",round(dgeom(2, 1-p16),4))
```

```
P(X=2) = 0.0118
```

18–آیا براساس داده های موجود، میتوان ادعا کرد که بهبود در فرآیندها، وضعیت سیستم را از منظر زمان انتظار بهتر کرده است؟

چون باید درباره بهبود میانگین در شرایط دوم اظهار نظر کنیم (m2>m1) بنابراین H1 باید دارای هر شرطی جز این باشد پس سوال صورت فرضیه را برای این سوال به این صورت تعریف میکنیم:

```
# improved?
# H0: m1 = m2
# H1: m1 > m2
```

```
m1=mean(data$Time_On_Line)
m2=mean(data$Time_On_Line_Improved)
```

```
mu 1 = 14.7788
----------------------------------------
mu 2 = 11.37024
```

چون تعداد داده های ما از ۳۰ بزرگتر هستند نگرانی بابت داشتن توزیع نرمال در آن ها نیستیم.
هر چند در صورت بررسی تست ها و نمودار داده ها این امر اثبات میگردد.
همینطور چون واریانس جامعه برای ما مشخص نیست باید در آزمون فرضیه میانگین از تست t استفاده کنیم.
به دلیل این که از جامعه مورد نظر یک نمونه ۵۰ تایی دو بار در شرایط مختلف بررسی شده است تست t را به صورت زیر مینویسیم:

```
t.test(data$Time_On_Line,data$Time_On_Line_Improved,
       paired = T,alternative = "two.sided",conf.level = 0.95)
```

```
        Paired t-test

data:  data$Time_On_Line and data$Time_On_Line_Improved
t = 7.797, df = 49, p-value = 3.91e-10
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 2.530043 4.287077
sample estimates:
mean of the differences
                3.40856
```

میتوان نقطه بحرانی را نیز برای این سری از داده ها مشخص کرد:

```
round(qt(c(.025, .975), df=49),4)
```

```
-2.0096 · 2.0096
```

```
# p-value > 0.05 --> reject H0.  & t = 7.797 > 2.0096
```

آماره آزمون در ناحیه بحرانی قرار نگرفته است بنابراین میتوان با مقادیر به دست آمده برای p_vlaue و t ، فرضیه صفر را رد کرد و نتیجه گرفت میانگین در شرایط دوم، بهبود یافته است و مشتریان زمان کمتری منتظر میمانند.

19–از آنجایی که گرفتن بیمه تکمیلی برای اطلس خودرو سودآور است، مدیران یک بسته تشویقی برای نیروهای فروش در دفتر مستقر در فرودگاه تعریف می کنند تا آنان تشویق شوند مشتریان را بیشتر به سمت دریافت بیمه تکمیلی روی قرارداد کرایه خودرو ترغیب کنند. پس از اجرای این طرح تشویقی، نمونه گیری مجدد روی 250 قرارداد کرایه نشان داد که 67 درصد این قراردادها دارای بیمه تکمیلی بودند.
آیا میتوان ادعا کرد اعمال این طرح تشویقی برای کارکنان فروش تغییری در حجم قراردادهای دارای بیمه تکمیلی خودرو کرده است؟

چون درباره هرگونه تغییر که باعث شود m1 وm2 با هم برابر نباشند از ما سوال شده است صورت فرضیه را برای این سوال به صورت زیر تعریف میکنیم:

```
# Changed?
# H0: p1  = p2
# H1: p1 != p2
```

و برای پاسخ به این سوال باید از آزمون فرضیه استفاده کنیم که ابتدا مقادیر p1 وp2 را که درصد افرادی که بیمه دریافت کرده اند را از کل دو نمونه ای که در شرایط پس از تغییر نشان میدهند محاسبه میکنیم.

```
p1=sum(data$Number_Accepting_Insurance[1:30])/(250*30)
p2=0.67
prop.test(x = c(p1*7500,p2*250), n = c(7500,250), alternative = "two.sided",
          conf.level = 0.95)
```

```
        2-sample test for equality of proportions with continuity correction

data:  c(p1 * 7500, p2 * 250) out of c(7500, 250)
X-squared = 3.2828, df = 1, p-value = 0.07001
alternative hypothesis: two.sided
95 percent confidence interval:
 -0.12018868  0.00258868
sample estimates:
prop 1 prop 2
0.6112 0.6700
```

```
# p_value > 0.05 --> accept H0
```

که با بررسی مقدار p_value میتوان فرضیه صفر را اثبات کرد و نتیجه گرفت این طرح تشویقی باعث تغییر در حجم قراردادهای دارای بیمه شده است.