assignment2

wanghaochong

2024-11-15

目录

```
#Question1 A
#Compute the minimum and the maximum number of viewers.
q1_data = read.csv("BigBangTheory.csv")
q1_data
```

```
##              Air.Date Viewers..millions.
## 1  September 22, 2011                14.1
## 2  September 22, 2011                14.7
## 3  September 29, 2011                14.6
## 4     October 6, 2011                13.6
## 5    October 13, 2011                13.6
## 6    October 20, 2011                14.9
## 7    October 27, 2011                14.5
## 8    November 3, 2011                16.0
## 9   November 10, 2011                15.9
## 10  November 17, 2011                15.1
## 11   December 8, 2011                14.0
## 12   January 12, 2012                16.1
## 13   January 19, 2012                15.8
## 14   January 26, 2012                16.1
## 15   February 2, 2012                16.5
## 16   February 9, 2012                16.2
## 17  February 16, 2012                15.7
## 18  February 23, 2012                16.2
## 19      March 8, 2012                15.0
## 20     March 29, 2012                14.0
## 21      April 5, 2012                13.3
```

```r
min_viewer = min(q1_data$Viewers..millions.)
max_viewer = max(q1_data$Viewers..millions.)

print("The maximum number of viewers is:")
```

```
## [1] "The maximum number of viewers is:"
```

```r
max_viewer
```

```
## [1] 16.5
```

```r
print("The minimum number of viewers is:")
```

```
## [1] "The minimum number of viewers is:"
```

```r
min_viewer
```

```
## [1] 13.3
```

```r
#Question1 B
#Compute the mean, median, and mode.
mean_viewer = mean(q1_data$Viewers..millions.)
median_viewer = median(q1_data$Viewers..millions.)
mode_viewer = which.max(q1_data$Viewers..millions.)

print("The mean number of viewers is:")
```

```
## [1] "The mean number of viewers is:"
```

```r
mean_viewer
```

```
## [1] 15.04286
```

```r
print("The median number of viewers is:")
```

```
## [1] "The median number of viewers is:"
```

```r
median_viewer
```

```
## [1] 15
```

```r
print("The mode number of viewers is:")
```

```
## [1] "The mode number of viewers is:"
```

```r
mode_viewer
```

```
## [1] 15
```

```r
#Question1 C
#Compute the first and third quartiles.
first_quartiles = quantile(q1_data$Viewers..millions., probs = 0.25)
third_quartiles = quantile(q1_data$Viewers..millions., probs = 0.75)

print("The first quartiles of viewers is:")
```

```
## [1] "The first quartiles of viewers is:"
```

```r
first_quartiles
```

```
##  25%
## 14.1
```

```r
print("The third quartiles of viewers is:")
```

```
## [1] "The third quartiles of viewers is:"
```

```r
third_quartiles
```

```
## 75%
##  16
```

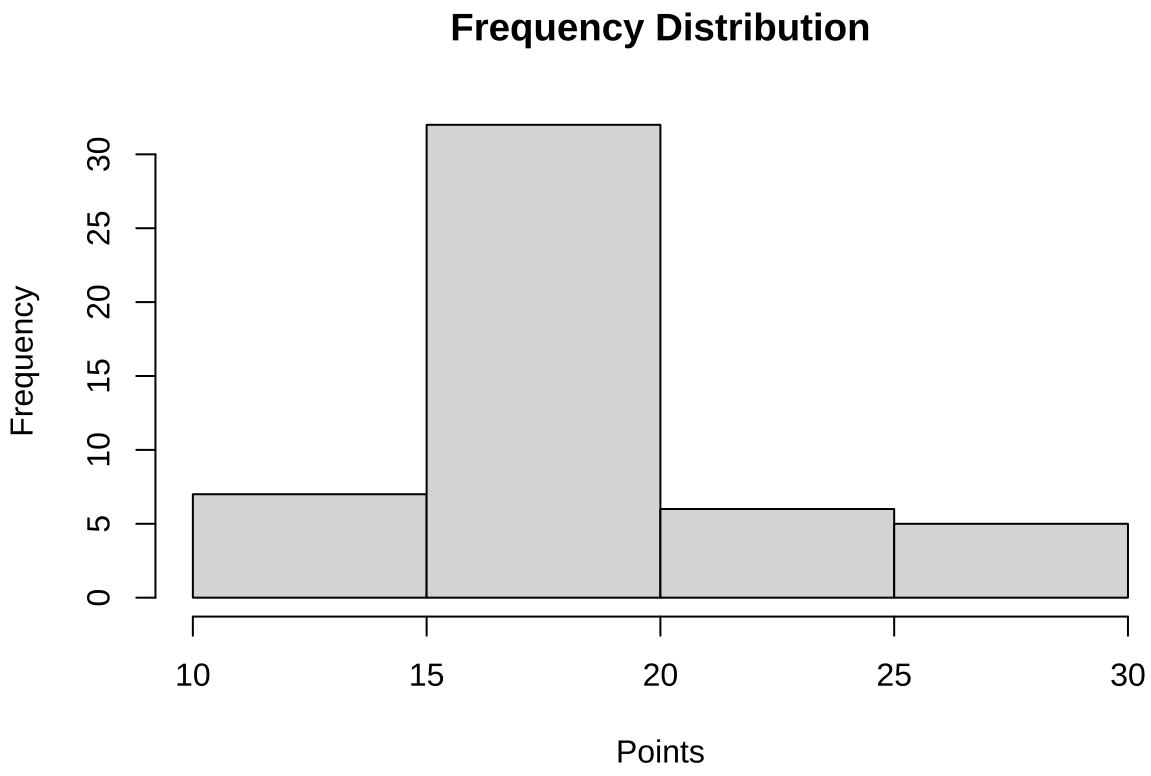
```r
#Question1 D
#has viewership grown or declined over the 2011-2012 season? Discuss.

#Question2 A
#Show the frequency distribution.
q2_data = read.csv("NBAPlayerPts.csv")

frq = table(q2_data$PPG)
frq
```

```
##
## 11.7 12.3 13.6   14 14.6 15.1 15.5 15.7 16.2 16.3 16.4 16.5 16.7 16.9   17 17.1
##    1    1    2    1    2    1    1    2    1    2    1    1    2    1    2    1
## 17.2 17.3 17.5 17.6 17.7   18 18.2 18.3 18.5 18.7 18.9 19.2 20.8   21 21.1 21.2
##    2    1    2    2    1    1    1    2    1    1    1    2    1    1    1    1
## 22.9 23.3 26.4   27 27.1 28.4 28.8
##    1    1    1    1    1    1    1
```

```r
# 绘制直方图
hist(q2_data$PPG, breaks = 5, main = "Frequency Distribution", xlab = "Points", ylab = "Frequency")
```




```r
#Question2 B
#Show the relative frequency distribution.
relative_freq = prop.table(frq)
relative_freq
```

```
##
## 11.7 12.3 13.6   14 14.6 15.1 15.5 15.7 16.2 16.3 16.4 16.5 16.7 16.9   17 17.1
## 0.02 0.02 0.04 0.02 0.04 0.02 0.02 0.04 0.02 0.04 0.02 0.02 0.04 0.02 0.04 0.02
## 17.2 17.3 17.5 17.6 17.7   18 18.2 18.3 18.5 18.7 18.9 19.2 20.8   21 21.1 21.2
## 0.04 0.02 0.04 0.04 0.02 0.02 0.02 0.04 0.02 0.02 0.02 0.04 0.02 0.02 0.02 0.02
```

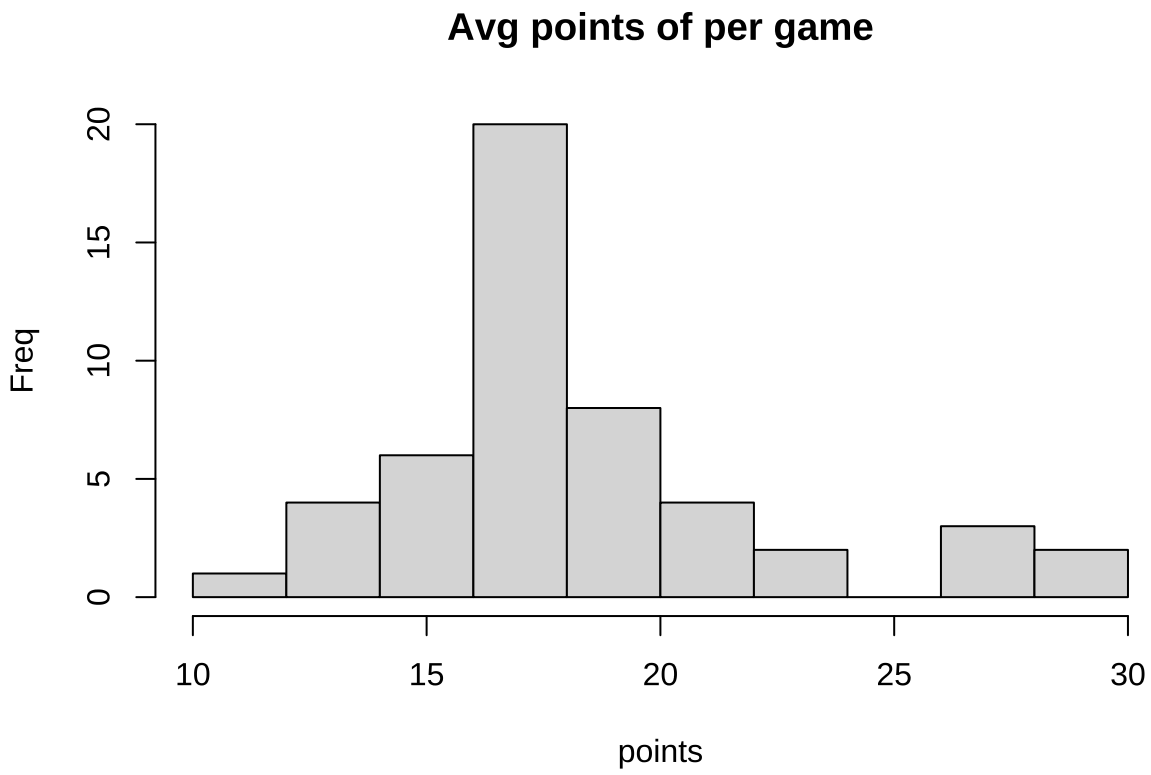
```
## 22.9 23.3 26.4   27 27.1 28.4 28.8
## 0.02 0.02 0.02 0.02 0.02 0.02 0.02
```

```r
#Question2 C
#Show the cumulative percent frequency distribution.
# 计算累积相对频率
cumulative_freq <- cumsum(relative_freq)

print(cumulative_freq)
```

```
## 11.7 12.3 13.6   14 14.6 15.1 15.5 15.7 16.2 16.3 16.4 16.5 16.7 16.9   17 17.1
## 0.02 0.04 0.08 0.10 0.14 0.16 0.18 0.22 0.24 0.28 0.30 0.32 0.36 0.38 0.42 0.44
## 17.2 17.3 17.5 17.6 17.7   18 18.2 18.3 18.5 18.7 18.9 19.2 20.8   21 21.1 21.2
## 0.48 0.50 0.54 0.58 0.60 0.62 0.64 0.68 0.70 0.72 0.74 0.78 0.80 0.82 0.84 0.86
## 22.9 23.3 26.4   27 27.1 28.4 28.8
## 0.88 0.90 0.92 0.94 0.96 0.98 1.00
```

```r
#Question2 D
#Develop a histogram for the average number of points scored per game.
hist(q2_data$PPG, main = "Avg points of per game", xlab = "points", ylab = "Freq")
```



```r
#Question2 E
#Do the data appear to be skewed
# 它看起来是向右倾斜的，因为它有一条向右的长尾巴。
```

```r
#Question2 F
#What percentage of the players averaged at least 20 points per game
total_player = length(q2_data$Player)

at_least_20 = sum(q2_data$PPG>=20)

percent_of_20 = at_least_20/total_player*100
print(paste0("The percentage of the players averaged at least 20 points per game is : ",  percent_of_
```

```
## [1] "The percentage of the players averaged at least 20 points per game is : 22%"
```

```r
#Question3 A
#How large was the sample used in this survey
#SE = 20,  population standard deviation = 500
# n = 500/20^2 = 625
SE = 20
sigma = 500
n = (sigma / SE)^2
n
```

```
## [1] 625
```

```r
#Question3 B
#What is the probability that the point estimate was within ±25 of the population mean?
p = pnorm(1.25) - pnorm(-1.25)
p
```

```
## [1] 0.7887005
```

```r
#Question4 A
#Develop appropriate descriptive statistics to summarize the data.
q4_data = read.csv("Professional.csv")
summary(q4_data$Age)
```

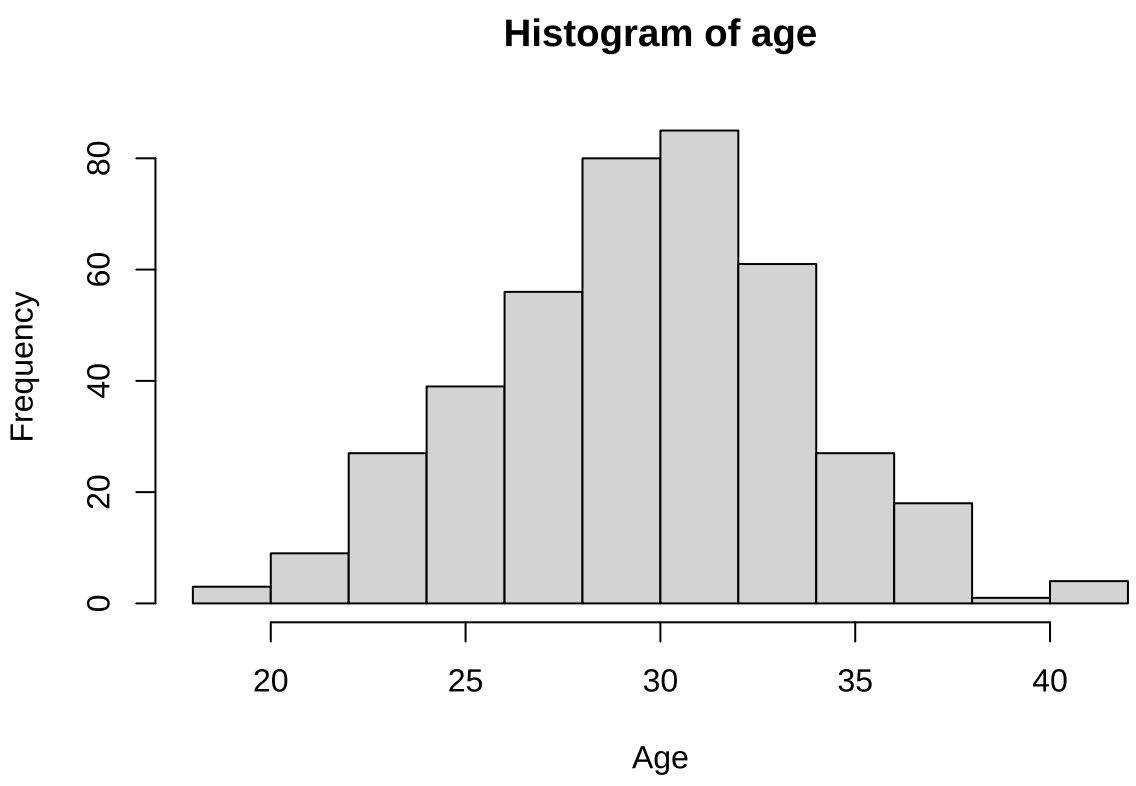
```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   19.00   28.00   30.00   30.11   33.00   42.00
```

```r
mean_age = mean(q4_data$Age)
min_age = min(q4_data$Age)
max_age = max(q4_data$Age)
sd_age = sd(q4_data$Age)

table_gender = table(q4_data$Gender)
props_gender = prop.table(table_gender)
props_gender
```

```
##
##    Female      Male
## 0.4414634 0.5585366
```

```r
hist(q4_data$Age, main = "Histogram of age", xlab = "Age")
```




```r
#Question4 B
se_age = sd_age/sqrt(length(q4_data$Age))
lower_age = mean_age - 1.96*se_age
upper_age = mean_age + 1.96*se_age
print(paste0("The 95% confidence interval for age is(",  lower_age,",",  upper_age,")" ))
```

```
## [1] "The 95% confidence interval for age is(29.7226797615731,30.5017104823294)"
```

```
mean_income = mean(q4_data$Household.Income)
se_income = sd(q4_data$Household.Income)/sqrt(length(q4_data$Household.Income))
lower_income = mean_income - 1.96*se_income
upper_income = mean_income + 1.96*se_income
print(paste0("The 95% confidence interval for household income is(",  lower_income,",", upper_income,
```

```
## [1] "The 95% confidence interval for household income is(71089.1964312007,77829.8279590432)"
```

```
#Question4 C
n_broadband = sum(q4_data$Broadband.Access. == "Yes")
prop_broad = n_broadband/length(q4_data$Broadband.Access.)
prop_broad
```

```
## [1] 0.6243902
```

```
n_children = sum(q4_data$Have.Children. == "Yes")
prop_children = n_children/length(q4_data$Have.Children.)
prop_children
```

```
## [1] 0.5341463
```

```
se_broadband = sqrt(prop_broad * (1 - prop_broad)/length(q4_data$Broadband.Access.))
se_broadband
```

```
## [1] 0.02391688
```

```
se_children = sqrt(prop_children * (1 - prop_children)/length(q4_data$Have.Children.))
se_children
```

```
## [1] 0.02463559
```

```
lower_broadband = prop_broad - 1.96 * se_broadband
upper_broadband = prop_broad + 1.96 * se_broadband
print(paste0("The 95% confidence interval for the proportion of subscribers with broadband access is
```

```
## [1] "The 95% confidence interval for the proportion of subscribers with broadband access is (0.577
```

```r
lower_children <- prop_children - 1.96 * se_children
upper_children <- prop_children + 1.96 * se_children
print(paste0("The 95% confidence interval for the proportion of subscribers  with children is (",  lo
```

```
## [1] "The 95% confidence interval for the proportion of subscribers  with children is (0.4858605863
```

```r
#Question4 D
mean_invest = mean(q4_data$Value.of.Investments....)
if(mean_invest > mean_income & mean_income > 70000){
  print("Young Professional would be a good advertising outlet")
}else{
  print("Young Professional is not a good advertising outlet")
}
```

```
## [1] "Young Professional is not a good advertising outlet"
```

```r
#Question4 E
if(prop_children > 0.5) {
  print(" 对于销售幼儿教育软件和电脑游戏的公司来说，这本杂志可能是一个做广告的好地方，因为有超过一半的订阅者
} else {
  print(" 对于销售幼儿教育软件和电脑游戏的公司来说，这本杂志可能不是一个理想的广告投放地方，因为有子女的订阅
}
```

```
## [1] "对于销售幼儿教育软件和电脑游戏的公司来说，这本杂志可能是一个做广告的好地方，因为有超过一半的订
```

```r
#Question4 F
mean_age
```

```
## [1] 30.1122
```

```r
mean_invest
```

```
## [1] 28538.29
```

```r
mean_income
```

```
## [1] 74459.51
```

```r
# 从以上数据中可以看出《青年专业人士》杂志读者大多在 30 岁左右，
# 大部分事业有成，支出和收入客观
# 他们可能会对投资与理财，生活方式与健康的文章内容更感兴趣
```

```r
#Question5 A
#Conduct a hypothesis test for each sample at the .01 level of significance and determine what action

# 原假设 H0:The process mean is equal to the specified mean of 12,
#The alternative hypothes is H1:The process mean is not equal to 12
q5_data = read.csv("Quality.csv")
sigma = 0.21
n = 30
alpha = 0.01
mu = 12
mean_1 = mean(q5_data$Sample.1)
mean_2 = mean(q5_data$Sample.2)
mean_3 = mean(q5_data$Sample.3)
mean_4 = mean(q5_data$Sample.4)
z1 = (mean_1 - mu)/(sigma/sqrt(n))
z2 = (mean_2 - mu)/(sigma/sqrt(n))
z3 = (mean_3 - mu)/(sigma/sqrt(n))
z4 = (mean_4 - mu)/(sigma/sqrt(n))
p1 = 2*(1 - pnorm(abs(z1)))
p2 = 2*(1 - pnorm(abs(z2)))
p3 = 2*(1 - pnorm(abs(z3)))
p4 = 2*(1 - pnorm(abs(z4)))

print(paste0("The sample 1 p-value is ", p1, ". Large than sg, Fail to reject H0" ))
```

```
## [1] "The sample 1 p-value is 0.281008276157385. Large than sg, Fail to reject H0"
```

```r
print(paste0("The sample 2 p-value is ", p2, ". Large than sg, Fail to reject H0" ))
```

```
## [1] "The sample 2 p-value is 0.454650325085948. Large than sg, Fail to reject H0"
```

```r
print(paste0("The sample 3 p-value is ", p3, ". Small than sg, reject H0" ))
```

```
## [1] "The sample 3 p-value is 0.00379031788780271. Small than sg, reject H0"
```

```r
print(paste0("The sample 4 p-value is ", p4, ". Small than sg, reject H0" ))
```

```
## [1] "The sample 4 p-value is 0.0338933553193166. Small than sg, reject H0"
```

```r
#Question5 B
sample_sd = apply(q5_data, 2, sd)
sample_sd
```

```
##  Sample.1  Sample.2  Sample.3  Sample.4
## 0.2203560 0.2203560 0.2071706 0.2061090
```

```r
# 四个样本的标准差接近于 0.21，假设值合理
```

```r
#Question5 C
z = qnorm(1 - alpha/2)
LCL = round(mu - z * sigma / sqrt(n), 2)
UCL = round(mu + z * sigma / sqrt(n), 2)

control_limits = c(LCL, UCL)
control_limits
```

```
## [1] 11.9 12.1
```

```r
#Question5 D
# 讨论将显著性水平改为更大值的影响。如果提高显著性水平，可能会增加什么错误或失误?
# 提高显著性水平可能会增加拒绝原假设的概率，从而导致不必要的行动措施
```

```r
#Question 6 A
# 估计 2007 年 3 月第一周和 2008 年 3 月第一周出租单元的比例。
q6_data = read.csv("Occupancy.csv", header = FALSE)
summary(q6_data)
```

```
##       V1                  V2
##  Length:202         Length:202
##  Class :character   Class :character
##  Mode  :character   Mode  :character
```

```r
prop_2007 = sum(q6_data$V1 == "Yes")/ length(q6_data$V1)
prop_2008 = sum(q6_data$V2 == "Yes")/ 150
prop_2007
```

```
## [1] 0.3465347
```

```r
prop_2008
```

```
## [1] 0.4666667
```

```r
#Question 6 B
# 为比例差异提供一个 95% 的置信区间
prop.test(x = c(sum(q6_data$V1 == "Yes"), sum(q6_data$V2 == "Yes")), n = c(200, 150), conf.level = 0.
```

```
##
##  2-sample test for equality of proportions with continuity correction
##
## data:  c(sum(q6_data$V1 == "Yes"), sum(q6_data$V2 == "Yes")) out of c(200, 150)
## X-squared = 4.3872, df = 1, p-value = 0.03621
## alternative hypothesis: two.sided
## 95 percent confidence interval:
##  -0.226151510 -0.007181823
## sample estimates:
##    prop 1    prop 2
## 0.3500000 0.4666667
```

```r
#Question 6 C
# 基于发现评估 2008 年 3 月租金上涨情况
# 置信区间的上限约为-0.007，且区间范围内不包括 0，所以 2008 年的租金会比 2007 年有所上升
```

```r
#Question 7 A
q7_data = read.csv("Training.csv")
current = q7_data$Current
proposed = q7_data$Proposed
summary(current)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   65.00   72.00   76.00   75.07   78.00   84.00
```

```r
summary(proposed)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   69.00   74.00   76.00   75.43   77.00   82.00
```

```r
# 从均值来看两种方式差异不大，提议的方法率高于当前方法
# 从标准差来看，提议的方法标准误差较小，其数据分布更为集中
# 两种方法在训练上表现出一定的相似性，但提议方法的数据分布更为集中，可能具有更高的稳定性。
```

```r
#Question 7 B
t_test_result = t.test(current, proposed)
t_test_result
```

```
##
##  Welch Two Sample t-test
##
## data:  current and proposed
## t = -0.60268, df = 101.65, p-value = 0.5481
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -1.5476613  0.8263498
## sample estimates:
## mean of x mean of y
##  75.06557  75.42623
```

```r
# 通过数据分析，可以观察到两种方法在训练时间上的总体均值略有不同，但差异并不大。
# 这种差异不足以成为唯一依据来证明哪种方式更优于另一种，因此需要考虑更多方面的因素
```

```r
#Question 7 c
sd_current = sd(current)
sd_proposed = sd(proposed)
sd_proposed
```

```
## [1] 2.506385
```

```r
sd_current
```

```
## [1] 3.944907
```

```r
var.test(current, proposed)
```

```
##
##  F test to compare two variances
##
## data:  current and proposed
## F = 2.4773, num df = 60, denom df = 60, p-value = 0.000578
## alternative hypothesis: true ratio of variances is not equal to 1
## 95 percent confidence interval:
##  1.486267 4.129135
## sample estimates:
```

```
## ratio of variances
##           2.477296
```

```
var.test
```

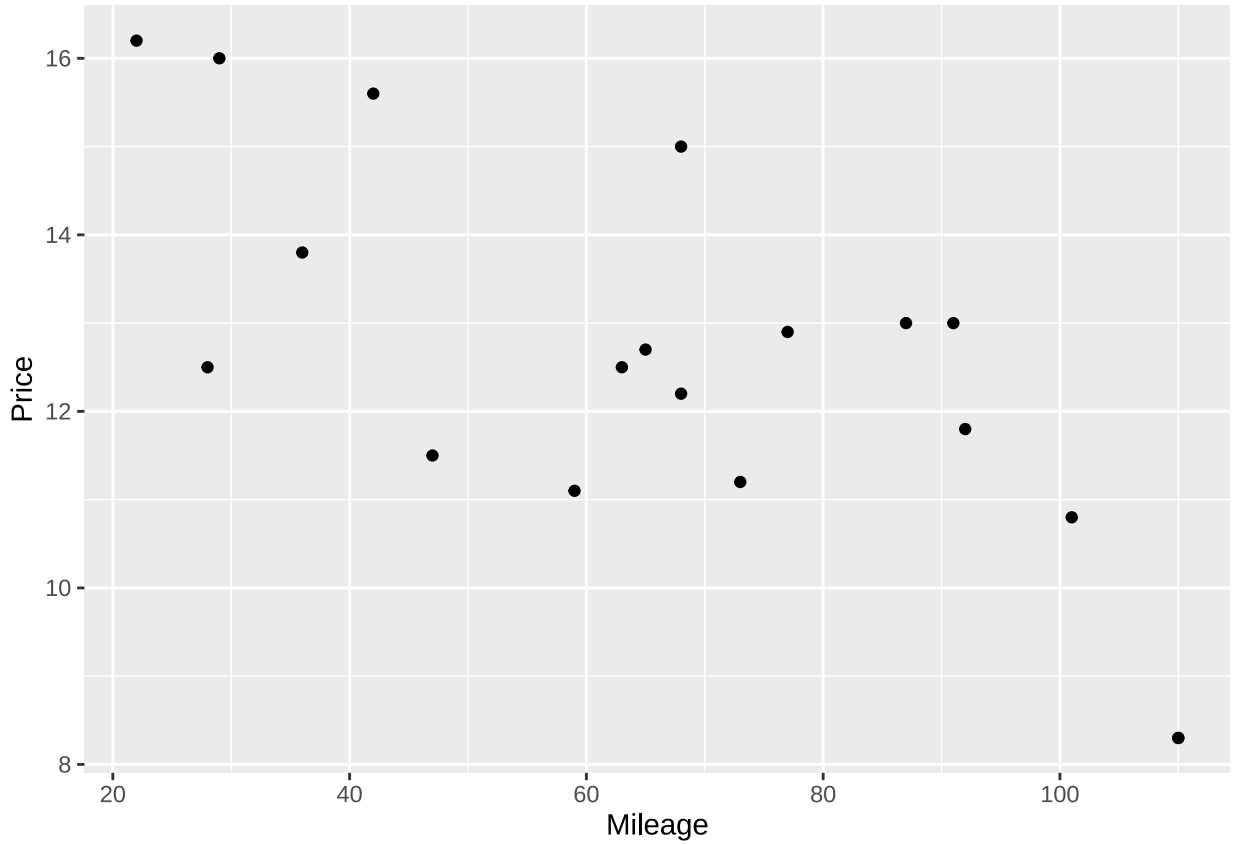
```
## function (x, ...)
## UseMethod("var.test")
## <bytecode: 0x7f99bba7e380>
## <environment: namespace:stats>
```

```
# 样本 current 的方差  大约是样本 proposed 方差的 2.477296 倍
```

```
#Question 7 D
# 关于这两种方法之间的任何差异，你能得出什么结论？你的建议是什么？请解释。
# 两种方法在训练上表现出一定的相似性，但提议方法的数据分布更为集中，可能具有更高的稳定性。
```

```
#Question 7 E
# 建议其他可能需要的数据或测试
# 实际教学效果还需考虑其他因素，如学生的掌握程度、教学资源的利用效率等
# 测试两种方法对学生学习的影响
# 比较两种教学方式的教学成本
```

```
#Question 8 A
library(ggplot2)
q8_data = read.csv("Camry.csv")
miles = q8_data$Miles..1000s.
prices = q8_data$Price...1000s.
ggplot(data = q8_data, aes(x = miles, y = prices)) +
  geom_point() +
  labs(x = "Mileage", y = "Price")
```




```
#Question 8 B
# 第（a）部分中绘制的散点图表明了两个变量的关系
# 随着公里数的增长，车的价格随之有下降的趋势
```

```
#Question 8 C
# 开发可用于预测价格（1000 美元）的估计回归方程给定英里（1000）。
summary(q8_data)
```

```
##  Miles..1000s.    Price...1000s.
##  Min.   : 22.00   Min.   : 8.30
##  1st Qu.: 44.50   1st Qu.:11.35
##  Median : 68.00   Median :12.50
##  Mean   : 66.74   Mean   :12.55
##  3rd Qu.: 89.00   3rd Qu.:13.40
##  Max.   :110.00   Max.   :16.20
```

```
model_q8 = lm(prices ~ miles, data = q8_data)
summary(model_q8)
```

```
##
```

```
## Call:
## lm(formula = prices ~ miles, data = q8_data)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -2.32408 -1.34194  0.05055  1.12898  2.52687
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 16.46976    0.94876  17.359 2.99e-12 ***
## miles       -0.05877    0.01319  -4.455 0.000348 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.541 on 17 degrees of freedom
## Multiple R-squared:  0.5387, Adjusted R-squared:  0.5115
## F-statistic: 19.85 on 1 and 17 DF,  p-value: 0.0003475
```

```
predicted_price <- predict(model_q8, newdata = q8_data)
print(predicted_price)
```

```
##        1        2        3        4        5        6        7        8
## 15.17673 14.76531 14.35389 13.70738 12.76700 11.94416 12.17926 11.35642
##        9       10       11       12       13       14       15       16
## 11.06255 10.53359 10.00462 14.82408 13.00209 12.47313 12.47313 11.12133
##       17       18       19
## 14.00125 12.64945 10.00462
```

```
#Prices = 16.470 - 0.059*Mileages
```

```
#Question 8 D
# 在 0.05 的显著性水平上检验显著性关系。
result = cor.test(miles, prices, method = 'pearson', conf.level = 0.95)
result
```

```
##
##  Pearson's product-moment correlation
##
## data:  miles and prices
## t = -4.4552, df = 17, p-value = 0.0003475
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
```

```
##  -0.8910894 -0.4196015
## sample estimates:
##        cor
## -0.7339328
```

```
#p-value = 0.0003475 < 0.05
```

```
#Question 8 E
# 估计的回归方程是否提供了很好的拟合? 解释一下。
#Multiple R-squared:  0.5387,
# 因为汽车行驶公里数是一个很好决定汽车价格的因素，所以它提供了一个很好的拟合
```

```
#Question 8 F
# 对估计回归方程的斜率进行解释。
# 汽车每多行驶 1000 英里数，车价就会下降 59 美金
```

```
#Question 8 G
price_1 = 16.470 - 0.059*60
price_1
```

```
## [1] 12.93
```

```
# 根据公式的预测价格应该是在 12.93 左右，这个价格不一定是你的卖价，但是是一个参考的依据
```

```
#Question 9 A
library(readxl)
q9_data = readxl::read_xlsx("WE.xlsx") %>%
  set_names("id","churn","happy_index","chg_hi","support","chg_supprt",
            "priority","chg_priority","log_in_fre","chg_blog_fre","chg_vis","y_age","chg_interval")
glimpse(q9_data)
```

```
## Rows: 6,347
## Columns: 13
## $ id           <dbl> 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17~
## $ churn        <dbl> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, ~
## $ happy_index  <dbl> 0, 62, 0, 231, 43, 138, 180, 116, 78, 78, 91, 40, 215, 0,~
## $ chg_hi       <dbl> 0, 4, 0, 1, -1, -10, -5, -11, -7, -37, -1, 14, 15, 0, 63,~
## $ support      <dbl> 0, 0, 0, 1, 0, 0, 1, 0, 1, 0, 0, 0, 0, 0, 0, 0, 0, 0, 1, ~
## $ chg_supprt   <dbl> 0, 0, 0, -1, 0, 0, 1, 0, -2, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0~
## $ priority     <dbl> 0, 0, 0, 3, 0, 0, 3, 0, 3, 0, 0, 0, 0, 0, 0, 0, 0, 0, 3, ~
## $ chg_priority <dbl> 0, 0, 0, 0, 0, 0, 3, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, ~
## $ log_in_fre   <dbl> 0, 0, 0, 167, 0, 43, 13, 0, -9, -7, 14, 0, 71, 0, 5, 0, 4~
```

```
## $ chg_blog_fre <dbl> 0, 0, 0, -8, 0, 0, -1, 0, 1, 0, 3, 0, 9, 0, 1, 0, 0, 0, 6~
## $ chg_vis      <dbl> 0, -16, 0, 21996, 9, -33, 907, 38, 0, 30, 0, 15, 8658, 0,~
## $ y_age        <dbl> 72, 72, 60, 68, 62, 63, 62, 51, 61, 61, 58, 61, 62, 62, 6~
## $ chg_interval <dbl> 33, 33, 33, 2, 33, 2, 2, 8, 9, 16, 2, 33, 2, 33, 2, 33, 3~
```

```
#Question 9 B
q9_data %>%
  select(-id) %>%
  group_by(churn)
```

```
## # A tibble: 6,347 x 12
## # Groups:   churn [2]
##    churn happy_index chg_hi support chg_supprt priority chg_priority log_in_fre
##    <dbl>       <dbl>  <dbl>   <dbl>      <dbl>    <dbl>        <dbl>      <dbl>
##  1     0           0      0       0          0        0            0          0
##  2     0          62      4       0          0        0            0          0
##  3     0           0      0       0          0        0            0          0
##  4     0         231      1       1         -1        3            0        167
##  5     0          43     -1       0          0        0            0          0
##  6     0         138    -10       0          0        0            0         43
##  7     0         180     -5       1          1        3            3         13
##  8     0         116    -11       0          0        0            0          0
##  9     0          78     -7       1         -2        3            0         -9
## 10     0          78    -37       0          0        0            0         -7
## # i 6,337 more rows
## # i 4 more variables: chg_blog_fre <dbl>, chg_vis <dbl>, y_age <dbl>,
## #   chg_interval <dbl>
```

```
#Question 9 C
q9_model = glm(churn ~ chg_blog_fre + chg_hi + chg_interval + chg_vis + happy_index
               + log_in_fre + priority  + support + y_age,
              data = q9_data,
              family = binomial(link = "logit"))
```

```
## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred
```

```
summary(q9_model)
```

```
##
## Call:
## glm(formula = churn ~ chg_blog_fre + chg_hi + chg_interval +
```

```
##     chg_vis + happy_index + log_in_fre + priority + support +
##     y_age, family = binomial(link = "logit"), data = q9_data)
##
## Coefficients:
##                Estimate Std. Error z value Pr(>|z|)
## (Intercept)  -2.874e+00  1.215e-01 -23.661  < 2e-16 ***
## chg_blog_fre -2.357e-05  2.080e-02  -0.001  0.99910
## chg_hi       -9.501e-03  2.424e-03  -3.920 8.87e-05 ***
## chg_interval  1.700e-02  4.277e-03   3.975 7.03e-05 ***
## chg_vis      -1.170e-04  4.069e-05  -2.877  0.00401 **
## happy_index  -5.225e-03  1.161e-03  -4.500 6.78e-06 ***
## log_in_fre    9.104e-04  1.952e-03   0.466  0.64098
## priority     -3.727e-02  7.514e-02  -0.496  0.61985
## support      -3.522e-02  7.438e-02  -0.474  0.63581
## y_age         1.418e-02  5.260e-03   2.696  0.00701 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 2553.1  on 6346  degrees of freedom
## Residual deviance: 2445.9  on 6337  degrees of freedom
## AIC: 2465.9
##
## Number of Fisher Scoring iterations: 6
```

```
vif(q9_model)
```

```
## chg_blog_fre       chg_hi chg_interval      chg_vis  happy_index   log_in_fre
##     1.068660     1.240227     1.197948     1.034792     1.513596     1.293839
##     priority      support        y_age
##     2.128518     2.166698     1.247978
```

```
#Question 9 D
q9_data %>%
  add_predictions(q9_model,type = "response") %>%
  arrange(desc(pred)) %>%
  filter(churn == 1) %>%
  slice_head(n=30)
```

```
## # A tibble: 30 x 14
##       id churn happy_index chg_hi support chg_supprt priority chg_priority
```

```
##    <dbl> <dbl>       <dbl>  <dbl>   <dbl>      <dbl>    <dbl>        <dbl>
##  1   357     1         203     25       7          6     2.86       -0.143
##  2  1363     1           0    -34       0          0     0           0
##  3  1672     1           2      1       0          0     0           0
##  4   299     1          14   -101       0          0     0           0
##  5  2951     1          20    -39       0          0     0           0
##  6  2922     1          13    -52       0          0     0           0
##  7  1021     1          12    -73       0          0     0           0
##  8   335     1           0    -64       0          0     0           0
##  9   156     1           8      0       0          0     0           0
## 10  3604     1           0    -78       0          0     0           0
## # i 20 more rows
## # i 6 more variables: log_in_fre <dbl>, chg_blog_fre <dbl>, chg_vis <dbl>,
## #   y_age <dbl>, chg_interval <dbl>, pred <dbl>
```