定量分析二次作业

蔡子豪

目录


```
#load library
library(tidyverse)
library(lubridate)
library(dplyr)
library(readxl)
library(gridExtra)
```

1 Question #1: BigBangTheory

```
bbt<- read_csv("./data/BigBangTheory.csv")
view(bbt)
```

```
bbt %>% summarise(
  min=min(`Viewers (millions)`),
  max=max(`Viewers (millions)`)
)
```

1.0.0.1 a. Compute the minimum and the maximum number of viewers.

```
#> # A tibble: 1 x 2
#>     min   max
#>   <dbl> <dbl>
#> 1  13.3  16.5
```

```
 mean(bbt$`Viewers (millions)`)
```

1.0.0.2 b. Compute the mean, median, and mode.

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

```
  median(bbt$`Viewers (millions)`)
```

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

```
  # which.max(table(bbt$`Viewers (millions)`))

  freq <- table(bbt$`Viewers (millions)`)
  max_freq <- max(freq)
  modes <- as.numeric(names(freq[freq == max_freq]))
  # 输出众数
  print(modes)
```

```
#> [1] 13.6 14.0 16.1 16.2
```

```
  quantile(bbt$`Viewers (millions)`,0.25)
```

1.0.0.3 c. Compute the first and third quartiles.

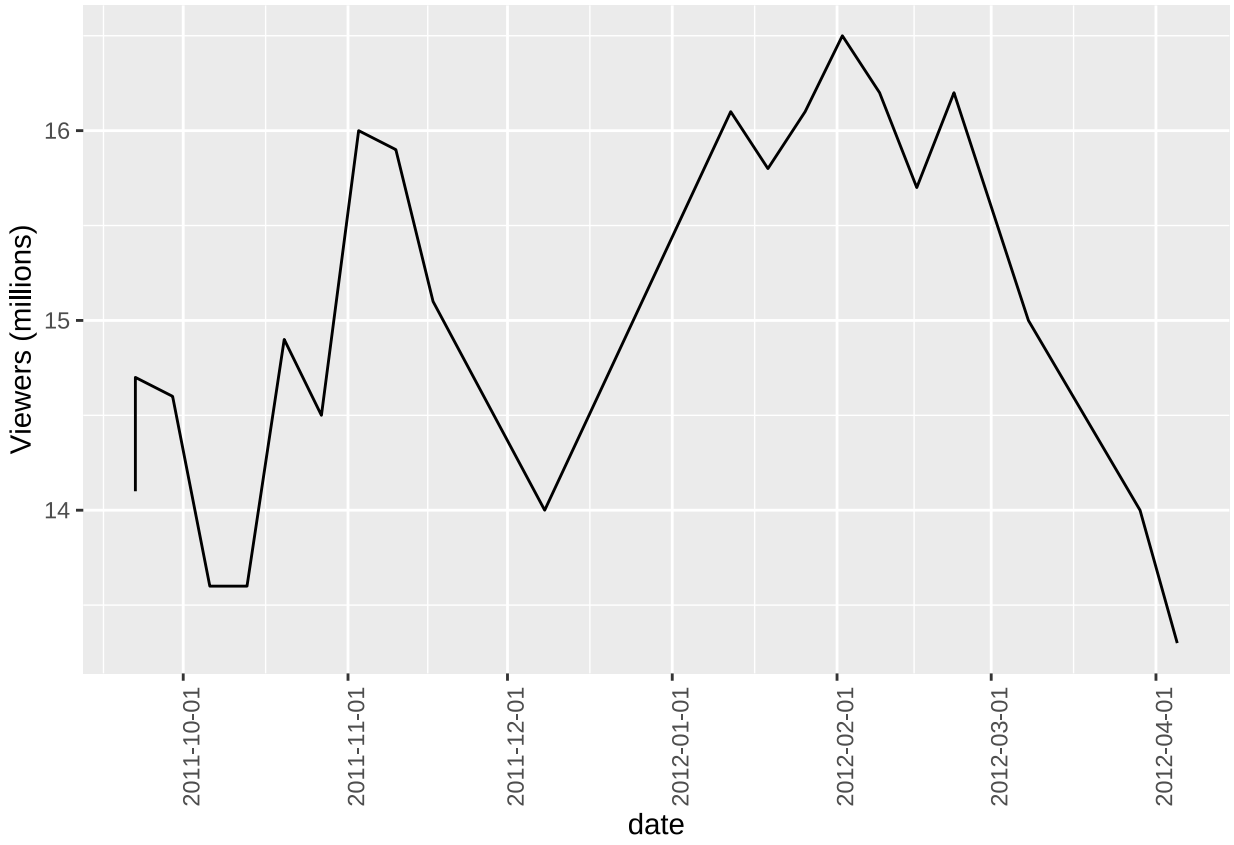
```
#>  25%
#> 14.1
```

```
  quantile(bbt$`Viewers (millions)`,0.75)
```

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

```
bbt$date <- mdy(bbt$`Air Date`)
 bbt %>%
   arrange(desc(date)) %>%
   ggplot(aes(x = date, y = `Viewers (millions)`)) +
   geom_line() +
   scale_x_date(date_labels = "%Y-%m-%d", date_breaks = "1 month") +
   theme(axis.text.x = element_text(angle = 90))
```

1.0.0.4 d. has viewership grown or declined over the 2011–2012 season? Discuss.



2 Question #2: NBAPlayerPts.

```
nba<- read_csv("./data/NBAPlayerPts.csv")
view(nba)
```

```
breaks <- seq(10, 30, by = 2)  # 从 10 到 30, 每 2 分一个区间

# 使用 cut() 函数将 PPG 列分组
nba$PPG_Group <- cut(nba$PPG, breaks = breaks, right = FALSE)
nba
```

2.0.0.1 a. Show the frequency distribution.

```
#> # A tibble: 50 x 4
```

```
#>     Rank Player                   PPG PPG_Group
#>    <dbl> <chr>                  <dbl> <fct>
#>  1     1 LeBron James, MIA       27   [26,28)
#>  2     2 Kevin Durant, OKC       28.8 [28,30)
#>  3     3 James Harden, HOU       26.4 [26,28)
#>  4     4 Kobe Bryant, LAL        27.1 [26,28)
#>  5     5 Russell Westbrook, OKC  22.9 [22,24)
#>  6     6 Carmelo Anthony, NY     28.4 [28,30)
#>  7     7 David Lee, GS           19.2 [18,20)
#>  8     8 Stephen Curry, GS       21   [20,22)
#>  9     9 LaMarcus Aldridge, POR  20.8 [20,22)
#> 10    10 Paul George, IND        17.6 [16,18)
#> # i 40 more rows
```

```
  frequency_ds <- table(nba$PPG_Group)
  frequency_ds
```

```
#>
#> [10,12) [12,14) [14,16) [16,18) [18,20) [20,22) [22,24) [24,26) [26,28) [28,30)
#>       1       3       7      19       9       4       2       0       3       2
```

```
relative_frequency_ds <- prop.table(frequency_ds)
  relative_frequency_ds
```

2.0.0.2 b. Show the relative frequency distribution.

```
#>
#> [10,12) [12,14) [14,16) [16,18) [18,20) [20,22) [22,24) [24,26) [26,28) [28,30)
#>    0.02    0.06    0.14    0.38    0.18    0.08    0.04    0.00    0.06    0.04
```

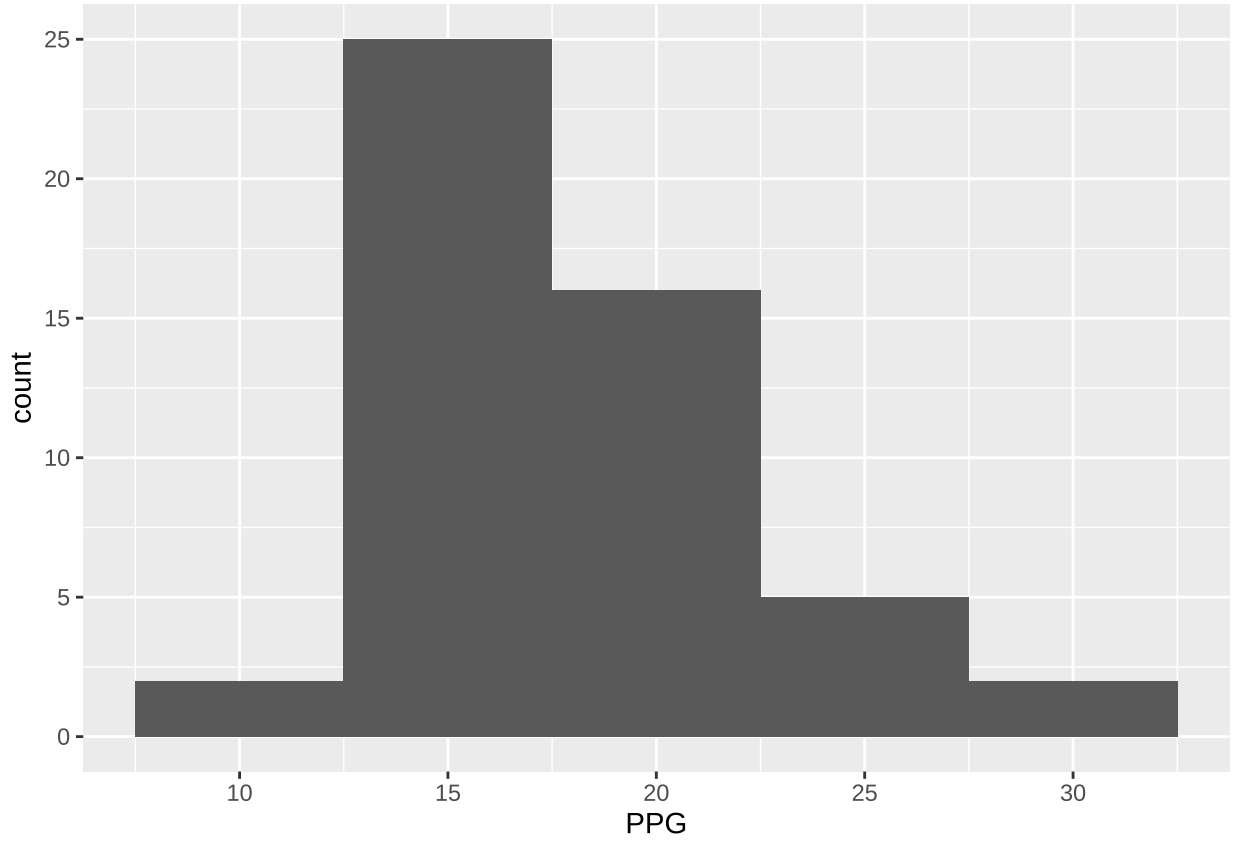
```
  cumulative_frequency <- cumsum(relative_frequency_ds)
  cumulative_frequency
```

2.0.0.3 c. Show the cumulative percent frequency distribution.

```
#> [10,12) [12,14) [14,16) [16,18) [18,20) [20,22) [22,24) [24,26) [26,28) [28,30)
#>    0.02    0.08    0.22    0.60    0.78    0.86    0.90    0.90    0.96    1.00
```

```
ggplot(nba,aes(x=PPG))+geom_histogram(binwidth = 5)
```

2.0.0.4 d. Develop a histogram for the average number of points scored per game.




```
# 通过直方图和频率曲线可以看到，数据右偏，使用 skewness 验证
  e1071::skewness(nba$PPG)
```

2.0.0.5 e. Do the data appear to be skewed? Explain.

```
#> [1] 1.124025
```

```
  # 结果验证 skewness 右偏
```

```
goodPlayer <- filter(nba,PPG>=20)
rate <-  nrow (goodPlayer)/nrow (nba)
rate
```

2.0.0.6 f. What percentage of the players averaged at least 20 points per game?

```
#> [1] 0.22
```

3 Question 3:

```
SE=20
sd=500
n=(sd/SE)^2
n
```

3.0.0.1 a. How large was the sample used in this survey?

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

```
# n>30 所有抽样符合正态分布

zu <- 25/SE
zl <- -25/SE
pnorm(zu)-pnorm(zl)
```

3.0.0.2 b. What is the probability that the point estimate was within ±25 of the population mean?

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

4 Question #4: Young Professional Magazine

```
pd<- read_csv("./data/Professional.csv")
 view(pd)
 mean_age <- mean(pd$Age)  # 读者年龄均值
 mean_houseIncome <- mean(pd$`Household Income ($)`) # 读者家庭收入均值
 sd_age <- sd(pd$Age)
 sd_houseIncome <- sd(pd$houseIncome)
 male_prop=sum(pd$Gender=='Male',is.na=TRUE)/sum(!is.na(pd$Gender)) # 样本中男性读者占比为 56.1%
```

4.0.0.1 a. Develop appropriate descriptive statistics to summarize the data.

```
 # 手动计算数据
 n <- length(pd$Age)  # 样本大小
 # 计算标准误差
 se <- sd_age / sqrt(n)
 ?qnorm
 # 使用 qnorm 计算置信区间
 confidence_interval <- qnorm(c(0.025, 0.975), mean_age, se)
 confidence_interval
```

4.0.0.2 b. Develop 95% confidence intervals for the mean age and household income of subscribers.

```
#> [1] 29.72269 30.50170
```

```
 # 使用置信极值计算计算区间
 z_value <- qnorm(0.975)  # 0.975 用于双尾检验（95% 置信区间），z 分布 0 对称
 z_value
```

```
#> [1] 1.959964
```

```
 error_margin <- z_value * se
 confidence_interval <- c(mean_age - error_margin, mean_age + error_margin)
 confidence_interval
```

```
#> [1] 29.72269 30.50170
```

```r
# t 检验代码直接获取置信区间
result_age <- t.test(pd$Age, conf.level = 0.95)
result_houseIncome <- t.test(pd$`Household Income ($)`, conf.level = 0.95)
result_age
```

```
#>
#>  One Sample t-test
#>
#> data:  pd$Age
#> t = 151.52, df = 409, p-value < 2.2e-16
#> alternative hypothesis: true mean is not equal to 0
#> 95 percent confidence interval:
#>  29.72153 30.50286
#> sample estimates:
#> mean of x
#>   30.1122
```

```r
result_houseIncome
```

```
#>
#>  One Sample t-test
#>
#> data:  pd$`Household Income ($)`
#> t = 43.302, df = 409, p-value < 2.2e-16
#> alternative hypothesis: true mean is not equal to 0
#> 95 percent confidence interval:
#>  71079.26 77839.77
#> sample estimates:
#> mean of x
#>  74459.51
```

```r
# 我们有 95% 的信心确定读者的平均年龄在 30 到 31 岁，读者的平均家庭收入在 71078 到 77840
```

```r
hasKid_num=sum(pd$`Have Children?`=='Yes',na.rm = TRUE)
broadband_num=sum(pd$`Broadband Access?`=='Yes',na.rm = TRUE)
total_num=sum(!is.na(pd$`Have Children?`))
```

```
result_hasKid=prop.test(hasKid_num,total_num,conf.level = 0.95)
result_broadband=prop.test(broadband_num,total_num,conf.level = 0.95)

# 我们有 95% 的信心有孩子的读者占总数的 48.46% 到 58.31%, 家庭中装了宽带的占比为 57.53% 到 67.11%
```

4.0.0.3 c. Develop 95% confidence intervals for the proportion of subscribers who have broadband access at home and the proportion of subscribers who have children.

```
networkpd <- pd %>% filter(
   `Broadband Access?`=='Yes'
 ) %>% summarise(
   buy_num=sum(`Real Estate Purchases?`=='Yes',na.rm = TRUE),
   total_num=sum(!is.na(`Real Estate Purchases?`))
 )
 result_propInvest=prop.test(networkpd$buy_num,networkpd$total_num,conf.level = 0.95)
 result_propInvest
```

4.0.0.4 d. Would Young Professional be a good advertising outlet for online brokers? Justify your conclusion with statistical data.

```
#>
#>  1-sample proportions test with continuity correction
#>
#> data:  networkpd$buy_num out of networkpd$total_num, null probability 0.5
#> X-squared = 2.4414, df = 1, p-value = 0.1182
#> alternative hypothesis: true p is not equal to 0.5
#> 95 percent confidence interval:
#>  0.3875848 0.5124019
#> sample estimates:
#>         p
#> 0.4492188
```

```
# 我们有 95% 的信心家中装了宽带的读者投资占总收入的 38.76% 到 51.24% , 想要购买房地产的占比还是比较高的
```

```
kidpd <- pd %>% filter(
  `Have Children?`=='Yes'
) %>% summarise(
  mean_invest=mean(`Value of Investments ($)`),
  mean_income=mean(`Household Income ($)`),
  has_pc=sum(`Real Estate Purchases?`=='Yes',na.rm = TRUE),
)
result_propInvest=prop.test(kidpd$mean_invest,kidpd$mean_income,conf.level = 0.95)
result_propInvest
```

4.0.0.5 e. Would this magazine be a good place to advertise for companies selling educational software and computer games for young children?

```
#>
#>  1-sample proportions test with continuity correction
#>
#> data:  kidpd$mean_invest out of kidpd$mean_income, null probability 0.5
#> X-squared = 3992.8, df = 1, p-value < 2.2e-16
#> alternative hypothesis: true p is not equal to 0.5
#> 95 percent confidence interval:
#>  0.3803511 0.3873720
#> sample estimates:
#>         p
#> 0.3838555
```

```
result_kidwithInternet=prop.test(kidpd$has_pc,hasKid_num,conf.level = 0.95)
result_kidwithInternet
```

```
#>
#>  1-sample proportions test with continuity correction
#>
#> data:  kidpd$has_pc out of hasKid_num, null probability 0.5
#> X-squared = 4.1096, df = 1, p-value = 0.04264
#> alternative hypothesis: true p is not equal to 0.5
#> 95 percent confidence interval:
#>  0.3632447 0.4977272
#> sample estimates:
#>         p
```

```
#> 0.4292237
```

```
  # 我们有 95% 的信心家中有孩子的读者投资占总收入的 38.04% 到 38.74% , 投资占比还是比较高的，而家中有孩
```

4.0.0.6 f. Comment on the types of articles you believe would be of interest to readers of Young Professional. 鉴于读者的平均年纪在 30-31 岁，且读者大概率的家中接入了宽带，同时经济的投资占比较高，所以杂志中刊登一些经济股票之类的版块会较受欢迎

5 Question #5: Quality Associate, Inc

```
qua<- read_csv("./data/Quality.csv")
  view(qua)
```

```
sampleTest <-  function (sampleName) {
    # 在已知方差的时候要使用 z 检验
    se=0.21/sqrt(length(qua[[sampleName]]))
    mean=mean(qua[[sampleName]])
    result <- (1-pnorm(abs(mean-12)/se))*2
    return (result)
  }
  result_df <- qua %>%
    summarise(across(everything(), ~sampleTest(cur_column())))
  result_df
```

5.0.0.1 a. Conduct a hypothesis test for each sample at the .01 level of significance and determine what action, if any, should be taken. Provide the p-value for each test.

```
#> # A tibble: 1 x 4
#>   `Sample 1` `Sample 2` `Sample 3` `Sample 4`
#>        <dbl>      <dbl>      <dbl>      <dbl>
#> 1      0.281      0.455    0.00379     0.0339
```

```
sdCalc <- function(sampleName){
  sd <- sd(qua[[sampleName]])
  return (sd)
}
result_sd <- qua %>%
  summarise(across(everything(), ~sdCalc(cur_column())))
result_sd[1,]
```

5.0.0.2 b. compute the standard deviation for each of the four samples. does the assumption of .21 for the population standard deviation appear reasonable?

```
#> # A tibble: 1 x 4
#>   `Sample 1` `Sample 2` `Sample 3` `Sample 4`
#>        <dbl>      <dbl>      <dbl>      <dbl>
#> 1      0.220      0.220      0.207      0.206
```

```
t.test( result_sd[1,],mu=0.21)
```

```
#>
#>  One Sample t-test
#>
#> data:  result_sd[1, ]
#> t = 0.8821, df = 3, p-value = 0.4427
#> alternative hypothesis: true mean is not equal to 0.21
#> 95 percent confidence interval:
#>  0.2008780 0.2261178
#> sample estimates:
#> mean of x
#> 0.2134979
```

```
# p 值为 0.4427 在 0.5 的显著水平下 总体方差当作.21 是合理的
```

```
qnorm(c(0.005,0.995),mean=12,sd=.21)
```

5.0.0.3 c. compute limits for the sample mean x around = 12 such that, as long as a new sample mean is within those limits, the process will be considered to be operating

satisfactorily. if x exceeds the upper limit or if x is below the lower limit, corrective action will be taken. these limits are referred to as upper and lower control limits for quality control purposes.

```
#> [1] 11.45908 12.54092
```

5.0.0.4 d. discuss the implications of changing the level of significance to a larger value. what mistake or error could increase if the level of significance is increased? # 显著水平提升之后，检测到不合格的次数会变多，会增加客户很多不必要的检测错误和改正的行为

6 Question #6:

```
ocp<- read_csv("./data/Occupancy.csv")
  view(ocp)
```

```
yes_mar07 <- sum(ocp$`Mar-07` == "Yes", na.rm = TRUE)
  yes_mar08 <- sum(ocp$`Mar-08` == "Yes", na.rm = TRUE)
  total_mar07 <- sum(!is.na(ocp$`Mar-07`))
  total_mar08 <- sum(!is.na(ocp$`Mar-08`))
  prop_mar07=yes_mar07/total_mar07
  prop_mar08=yes_mar08/total_mar08
  prop_mar07
```

6.0.0.1 a. Estimate the proportion of units rented during the first week of March 2007 and the first week of March 2008.

```
#> [1] 0.35
```

```
  prop_mar08
```

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

```
prop.test(c(yes_mar07,yes_mar08), c(total_mar07,total_mar08), conf.level = 0.95,correct = FALSE)
```

6.0.0.2 b. Provide a 95% confidence interval for the difference in proportions.

```
#>
#>  2-sample test for equality of proportions without continuity correction
#>
#> data:  c(yes_mar07, yes_mar08) out of c(total_mar07, total_mar08)
#> X-squared = 4.8611, df = 1, p-value = 0.02747
#> alternative hypothesis: two.sided
#> 95 percent confidence interval:
#>  -0.22031818 -0.01301516
#> sample estimates:
#>    prop 1    prop 2
#> 0.3500000 0.4666667
```

```
  # 置信区间是-0.22 -0.01
```

6.0.0.3 c. On the basis of your findings, does it appear March rental rates for 2008 will be up from those a year earlier? 95% 的置信区间没有包括 0，置信区间为负，所以可以推断租房比例是逐年下降的

7 Question #7: Air Force Training Program

```
aft <- read.csv("./data/Training.csv")
 View(aft)
```

```
mean(aft$Current)
```

7.0.0.1 a. use appropriate descriptive statistics to summarize the training time data for each method.what similarities or differences do you observe from the sample data?

```
#> [1] 75.06557
```

```
  sd(aft$Current)
```

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

```
  mean(aft$Proposed)
```

```
#> [1] 75.42623
```

```
  sd(aft$Proposed)
```

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

```

 t.test(aft$Current,aft$Proposed)
```

7.0.0.2 b. Comment on any difference between the population means for the two methods. Discuss your findings

```
#>
#>  Welch Two Sample t-test
#>
#> data:  aft$Current and aft$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
```

```
  # p 值为 0.5 在 0.05 的显著水平下，拒绝原假设-两个方法是不同的，所以两个方法是相同的
```

```

s1 <- sd(aft$Current)^2
  s2 <- sd(aft$Proposed)^2
  s1
```

7.0.0.3 c. compute the standard deviation and variance for each training method. conduct a hypothesis test about the equality of population variances for the two training methods. Discuss your findings.

```
#> [1] 15.5623
```

```
  s2
```

```
#> [1] 6.281967
```

```
  # H0: 假设两个方法的总体方差是不等的
  var.test(aft$Current,aft$Proposed)
```

```
#>
#>  F test to compare two variances
#>
#> data:  aft$Current and aft$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
```

```
  # p 值为 0.000578 无法拒绝原假设 两个方法的总方差是不等的
```

7.0.0.4 d. what conclusion can you reach about any differences between the two methods? what is your # recommendation? explain.

拟用的替代方案能更好的消除了学员之间的差距，倘若 60 分是及格分，那么均分差不多的情况下，替代方案能够保证更多的人通过测试，所以替代方案更好。

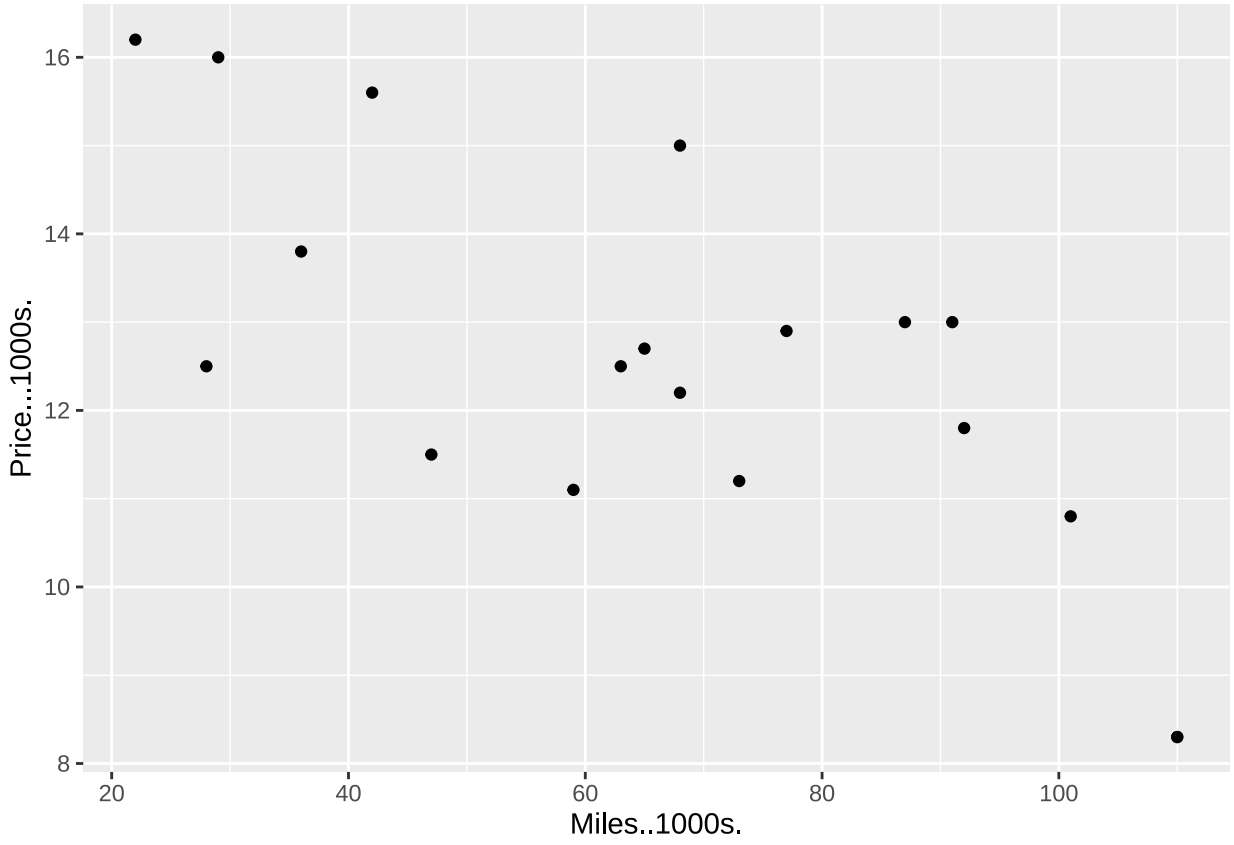
7.0.0.5 e. can you suggest other data or testing that might be desirable before making a final decision on the training program to be used in the future? 需要计算最高值，最低值，数据的左右偏移，来判断方案的极端情况，当遇到只考虑两端人才的时候会有用

8 Question #8

```
cmy <- read.csv("./data/Camry.csv")
 View(cmy)
```

```
ggplot(cmy,aes(x=Miles..1000s.,y=Price...1000s.))+geom_point()
```

8.0.0.1 a. Develop a scatter diagram with the car mileage on the horizontal axis and the price on the vertical axis.




```
# 从图中可以看出，随着里程数的增加，价格有下降趋势
# c. Develop the estimated regression equation that could be used to predict the price ($1000s)
lm_cmy <- lm(Price...1000s. ~ Miles..1000s., data = cmy)
summary(lm_cmy)
```

8.0.0.2 b. what does the scatter diagram developed in part (a) indicate about the relationship between the two variables?

```
#>
#> Call:
```

```
#> lm(formula = Price...1000s. ~ Miles..1000s., data = cmy)
#>
#> 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..1000s. -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
```

```
  # 回归方程为 y=-0.0588x+16.47 （x 为里程数 y 为价格）
```

8.0.0.3 d. Test for a significant relationship at the .05 level of significance. 5% 显著性水平下，p 值为 0.000348

8.0.0.4 e. Did the estimated regression equation provide a good fit? Explain. Multiple R-squared: 0.5387 具有一定的解释效果

8.0.0.5 f. Provide an interpretation for the slope of the estimated regression equation. 回归方程的斜率是-0.0588 意味着没增加 1000 里程数，价格下降 58.8 美元

```
price <- (-0.0588*60+16.47)*1000
  price
```

8.0.0.6 g. Suppose that you are considering purchasing a previously owned 2007 Camry that has been driven 60,000 miles. Using the estimated regression equation developed in part (c), predict the price for this car. Is this the price you would offer the seller.

```
#> [1] 12942
```

9 Question #9:

```r
we <-  read_excel('./data/WE.xlsx') %>%
  set_names(
    "id",
    "lost",
    "happy",
    "happy_c",
    "support",
    "support_c",
    "priority",
    "priority_c",
    "login",
    "blog_c",
    "visit_c",
    "limit",
    "gap"
  )
view(we)
```

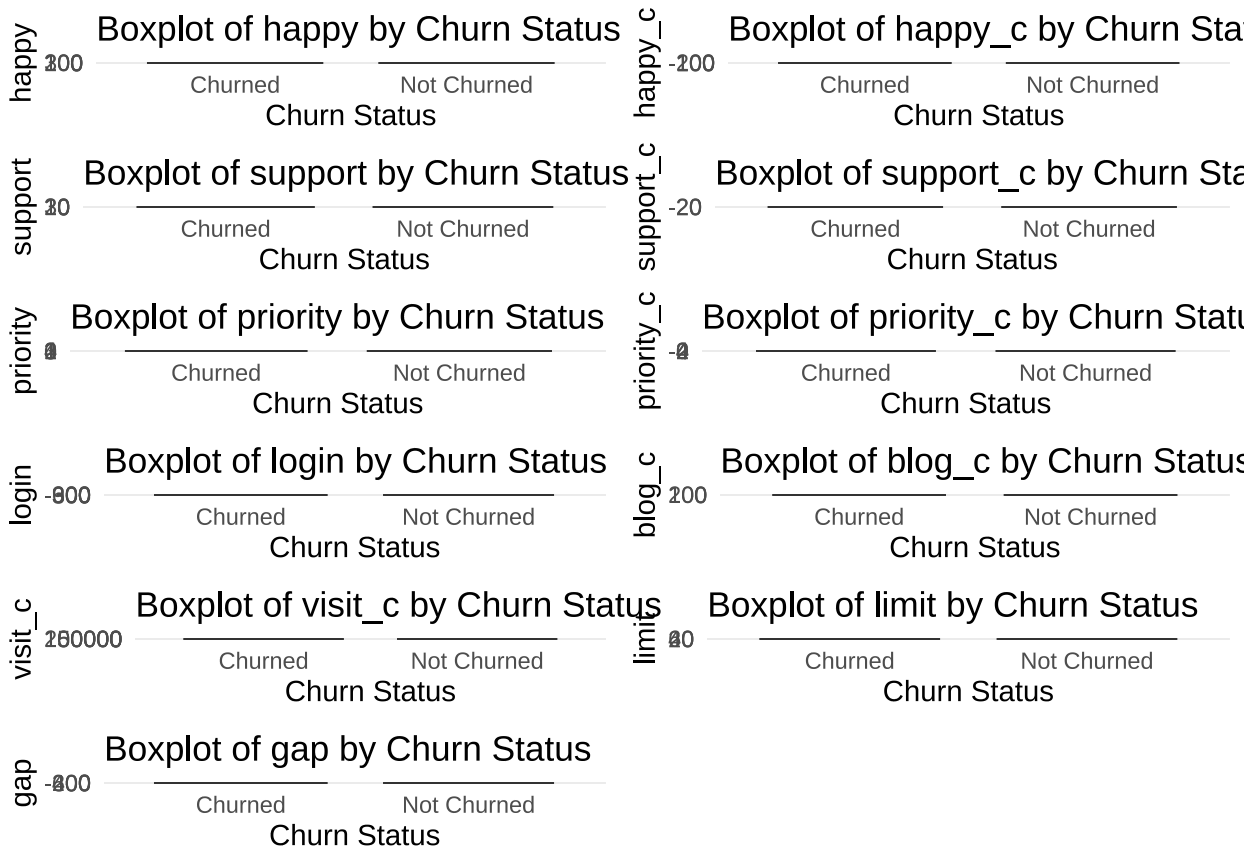
```r
# 添加流失状态标签
we <- we %>%
  mutate(Churn = ifelse(lost == 1, "Churned", "Not Churned"))

# 计算描述性统计
desc_stats <- we %>%
  group_by(Churn) %>%
  summarise(
    count = n(),
    happy_mean = mean(happy, na.rm = TRUE),
    support_mean = mean(support, na.rm = TRUE),
    priority_mean = mean(priority, na.rm = TRUE),
    login_mean = mean(login, na.rm = TRUE),
    blog_c_mean = mean(blog_c, na.rm = TRUE),
    visit_c_mean = mean(visit_c, na.rm = TRUE),
    limit_mean = mean(limit, na.rm = TRUE),
```

```
    gap_mean = mean(gap, na.rm = TRUE)
  )
print(desc_stats)
```

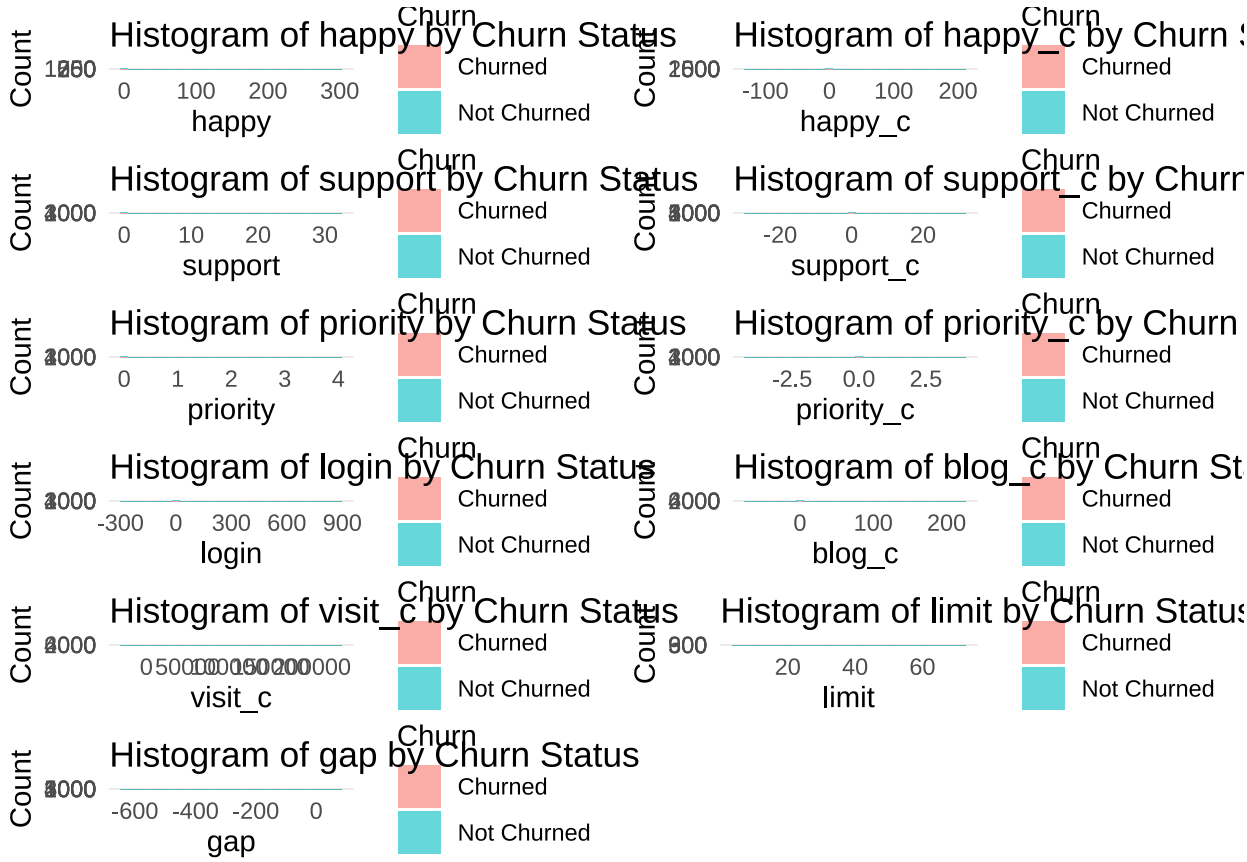
9.0.0.1 a. 通过可视化探索流失客户与□流失客户的□为特点（或特点对□），你能发现流失与□流失客户□为在哪些指标有可能存在显著不同？

```
#> # A tibble: 2 x 10
#>   Churn       count happy_mean support_mean priority_mean login_mean blog_c_mean
#>   <chr>       <int>      <dbl>        <dbl>         <dbl>      <dbl>       <dbl>
#> 1 Churned       323       63.3        0.372         0.500       8.06      -0.102
#> 2 Not Churned  6024       88.6        0.724         0.830      16.1        0.171
#> # i 3 more variables: visit_c_mean <dbl>, limit_mean <dbl>, gap_mean <dbl>
```

```
# 选择需要比较的指标
metrics <- c("happy", "happy_c","support","support_c", "priority","priority_c", "login", "blog_c

# 创建箱线图
plots <- list()
for (metric in metrics) {
  p <- ggplot(we, aes(x = Churn, y = .data[[metric]], fill = Churn)) +
    geom_boxplot() +
    labs(title = paste("Boxplot of", metric, "by Churn Status"), x = "Churn Status", y = metric)
    theme_minimal() +
    theme(legend.position = "none")
  plots[[metric]] <- p
}

# 显示所有箱线图
do.call(grid.arrange, c(plots, ncol = 2))
```




```r
# 创建直方图
plots_hist <- list()
for (metric in metrics) {
  p <- ggplot(we, aes(x = .data[[metric]], fill = Churn)) +
    geom_histogram(alpha = 0.6, position = "identity", bins = 30) +
    labs(title = paste("Histogram of", metric, "by Churn Status"), x = metric, y = "Count") +
    theme_minimal()
  plots_hist[[metric]] <- p
}

# 显示所有直方图
do.call(grid.arrange, c(plots_hist, ncol = 2))
```




```
# t 检验方式测试是否显著
  # 定义函数进行 t 检验
  perform_t_test <- function(data, metric) {
    churned <- data %>% filter(Churn == "Churned") %>% pull(metric)
    not_churned <- data %>% filter(Churn == "Not Churned") %>% pull(metric)
    test <- t.test(churned, not_churned)
    return(test$p.value)
  }

  # 计算各指标的 p 值
  p_values <- sapply(metrics, function(x) perform_t_test(we, x))
  # 结果输出
  p_values_df <- data.frame(
    Metric = metrics,
    P_Value = p_values
  )
```

```
print(p_values_df)
```

9.0.0.2 b. 通过均值□较的□式验证上述不同是否显著。

```
#>                Metric      P_Value
#> happy           happy 2.096694e-13
#> happy_c       happy_c 1.571085e-08
#> support       support 6.280509e-08
#> support_c   support_c 5.277532e-01
#> priority     priority 4.380969e-07
#> priority_c priority_c 5.218233e-01
#> login           login 4.037446e-04
#> blog_c         blog_c 1.157611e-02
#> visit_c       visit_c 5.630696e-02
#> limit           limit 3.056793e-03
#> gap               gap 5.215207e-05
```

```
# 结果发现 ‘客户流失’与 ‘服务优先级相比上月的变化’和 ‘客户支持相比上月的变化’无关
```

```
# 针对数值和分类数据的比较 使用 logit 回归
set.seed(1234)
we_logit<-glm(lost ~ happy + support + priority+login + visit_c + limit
              + gap + happy_c   +blog_c+visit_c,
              data = we,
              family = binomial(link = "logit"))
summary(we_logit)
```

9.0.0.3 c. 以” 流失 “为因变量，其他你认为重要的变量为□变量（提示：a、b 两步的发现），建□回归□程对是否流失进□预测。

```
#>
#> Call:
#> glm(formula = lost ~ happy + support + priority + login + visit_c +
#>     limit + gap + happy_c + blog_c + visit_c, family = binomial(link = "logit"),
#>     data = we)
#>
#> Coefficients:
```

```
#>               Estimate Std. Error z value Pr(>|z|)
#> (Intercept) -2.874e+00  1.215e-01 -23.661  < 2e-16 ***
#> happy       -5.225e-03  1.161e-03  -4.500 6.78e-06 ***
#> support     -3.522e-02  7.438e-02  -0.474  0.63581
#> priority    -3.727e-02  7.514e-02  -0.496  0.61985
#> login        9.104e-04  1.952e-03   0.466  0.64098
#> visit_c     -1.170e-04  4.069e-05  -2.877  0.00401 **
#> limit        1.418e-02  5.260e-03   2.696  0.00701 **
#> gap          1.700e-02  4.277e-03   3.975 7.03e-05 ***
#> happy_c     -9.501e-03  2.424e-03  -3.920 8.87e-05 ***
#> blog_c      -2.357e-05  2.080e-02  -0.001  0.99910
#> ---
#> 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
```

```
library(car)
vif(we_logit)
```

```
#>    happy  support priority    login  visit_c    limit      gap  happy_c
#> 1.513596 2.166698 2.128518 1.293839 1.034792 1.247978 1.197948 1.240227
#>   blog_c
#> 1.068660
```

```
we$churn_prob <- predict(we_logit, newdata = we %>% select(-id, -lost), type = "response")

top_100_churn_prob <- we %>%
  filter(lost == 0) %>%
  select(id, churn_prob)%>%
  arrange(desc(churn_prob)) %>%
  slice(1:100)
```

```
print(top_100_churn_prob)
```

9.0.0.4 d. 根据上□步预测的结果，对尚未流失（流失 =0）的客户进□流失可能性排序，并给出流失可能性最□的前 100 名□户 ID 列表。

```
#> # A tibble: 100 x 2
#>       id churn_prob
#>    <dbl>      <dbl>
#>  1  2287      0.364
#>  2   109      0.290
#>  3  1971      0.235
#>  4  2025      0.222
#>  5     1      0.215
#>  6   929      0.213
#>  7  2076      0.209
#>  8    76      0.194
#>  9    14      0.192
#> 10    18      0.188
#> # i 90 more rows
```