第二次作业

赵炜

目录

1 Question #1: **4**

1.1 a 5

1.2 b 5

1.3 c 6

1.4 d 6

2 Question #2: **7**

2.1 a 8

2.2 b 9

2.3 c 10

2.4 d 11

2.5 e 12

2.6 f 12

3 Question #3 **13**

3.1 a 13

3.2 b 13

4 Question #4 **14**

4.1 a 16

4.2 b 16

4.3 c 17

4.4 d 18
4.5 e 18
4.6 f 18

5 Question #5: 19

5.1 a 20
5.2 b 22
5.3 c 23
5.4 d 23

6 Question #6: 23

6.1 a 24
6.2 b 24
6.3 c 25

7 Question #7 26

7.1 a 27
7.2 b 28
7.3 c 28
7.4 d 30
7.5 e 30

8 Question #8 30

8.1 a 31
8.2 b 32
8.3 c 32
8.4 e 35
8.5 f 35

9 Question #9: 36

9.1 a 37

9.2 b 39

9.3 c 40

9.4 d 42

```
knitr::opts_chunk$set(
  message = FALSE,
  warning = FALSE,
  error = FALSE,
  out.width = "100%",
  fig.showtext = TRUE,
  fig.align = "center",
  comment = "#>",
  df_print = "tibble",
  paged.print = FALSE,
  split = FALSE,
  cache = TRUE
)

library(tidyverse)
```

```
## -- Attaching core tidyverse packages ------------------------ tidyverse 2.0.0 --
## v dplyr     1.1.4     v readr     2.1.5
## v forcats   1.0.0     v stringr   1.5.1
## v ggplot2   3.5.1     v tibble    3.2.1
## v lubridate 1.9.3     v tidyr     1.3.1
## v purrr     1.0.2
## -- Conflicts ---------------------------------------- tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
## i Use the conflicted package (<http://conflicted.r-lib.org/>) to force all conflicts to become err
```

```
library(showtext)
```

```
## 载入需要的程序包: sysfonts
## 载入需要的程序包: showtextdb
```

```r
showtext_auto()

# 添加微软雅黑字体
font_add("Microsoft YaHei", "C:/Windows/Fonts/msyh.ttc")
```

```r
options(scipen = 1, digits = 4)
# library(tidyverse)
# 这里 family 设置成你系统中的中文字体名。
old <- theme_set(theme(text=element_text(family="Microsoft YaHei",size=14))+
                   theme_minimal())
# 还原默认主题
# theme_set(old)
```

1 Question #1:

BigBangTheory. (Attached Data: BigBangTheory)

The Big Bang Theory, a situation comedy featuring Johnny Galecki, Jim Parsons, and Kaley Cuoco-Sweeting, is one of the most-watched programs on network television. The first two episodes for the 2011–2012 season premiered on September 22, 2011; the first episode attracted 14.1 million viewers and the second episode attracted 14.7 million viewers. The attached data file BigBangTheory shows the number of viewers in millions for the first 21 episodes of the 2011–2012 season (*the Big Bang theory* website, April 17, 2012).

```r
# read data
data_q1 <- read_csv("data/BigBangTheory.csv") %>%
  rename(viewers = `Viewers (millions)`, air_date = `Air Date`) %>%
  mutate(air_date = mdy(air_date))
data_q1
```

```
#> # A tibble: 21 x 2
#>    air_date   viewers
#>    <date>       <dbl>
#>  1 2011-09-22    14.1
#>  2 2011-09-22    14.7
#>  3 2011-09-29    14.6
#>  4 2011-10-06    13.6
#>  5 2011-10-13    13.6
#>  6 2011-10-20    14.9
```

```
#>  7 2011-10-27    14.5
#>  8 2011-11-03    16
#>  9 2011-11-10    15.9
#> 10 2011-11-17    15.1
#> # i 11 more rows
```

1.1 a

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

```
data_q1 %>%
  summarise(min_viewrs = min(viewers),
            max_viewers = max(viewers))
```

```
#> # A tibble: 1 x 2
#>   min_viewrs max_viewers
#>        <dbl>       <dbl>
#> 1       13.3        16.5
```

number of viewers 最小值为 13.3, 最大值为 16.5。

1.2 b

b. Compute the mean, median, and mode.

```
# 平均值 和中位数
data_q1 %>%
  summarise(mean_viewrs = mean(viewers),
            median_viewers = median(viewers)
  )
```

```
#> # A tibble: 1 x 2
#>   mean_viewrs median_viewers
#>         <dbl>          <dbl>
#> 1        15.0             15
```

```
# 众数
data_q1 %>%
 count(viewers) %>%
  slice_max(order_by = n)
```

```
#> # A tibble: 4 x 2
#>   viewers     n
#>     <dbl> <int>
#> 1    13.6     2
#> 2    14       2
#> 3    16.1     2
#> 4    16.2     2
```

平均值为 15.0, 中位数为 15 众数有 3 个，分别是 13.6，14，16.1，16.2 出现的次数均为 2。

1.3 c

c. Compute the first and third quartiles.

```
data_q1 %>%
  summarise(q1 = quantile(viewers, probs = 0.25),
            q3 = quantile(viewers, probs = 0.75))
```

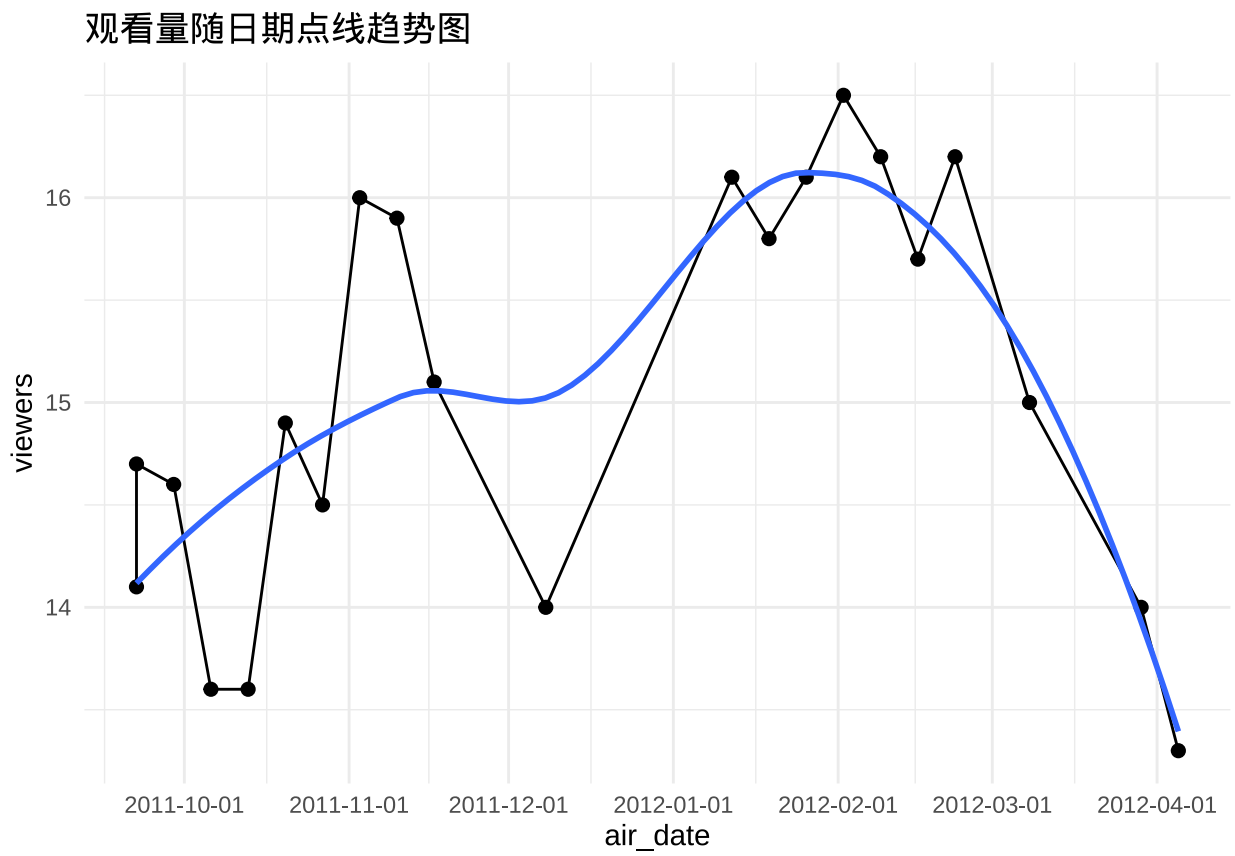
```
#> # A tibble: 1 x 2
#>      q1    q3
#>   <dbl> <dbl>
#> 1  14.1    16
```

1/4 分位数为 14.1, 3/4 分位数为 16。

1.4 d

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

```
ggplot(data_q1,aes(air_date,viewers)) +
         geom_point(size=2) +
         geom_line() +
  scale_x_date(date_breaks = "1 months", date_labels = "%Y-%m-%d") +
  labs(title = " 观看量随日期点线趋势图") +
  geom_smooth(se=FALSE)
```



总体来看，第二季的收视率比第一季高，有上升趋势；但是结尾部分收视率急速下滑，可能是大家对结尾不感兴趣了。

2 Question #2:

NBAPlayerPts. (Attached Data: NBAPlayerPts)

CbSSports.com developed the Total Player Rating system to rate players in the National Basketball Association (NBA) based on various offensive and defensive statistics. The attached data file NBAPlayerPts shows the average number of points scored per game (PPG) for 50 players with the highest ratings for a portion of the 2012–2013 NBA season (CbSSports.com website, February 25, 2013). Use classes starting at 10 and ending at 30 in increments of 2 for PPG in the following.

```
data_q2 <- read_csv("data/NBAPlayerPts.csv")
data_q2 <- data_q2 %>%
  mutate(class = cut(PPG, breaks = seq(
    from = 10, to = 30, by = 2
```

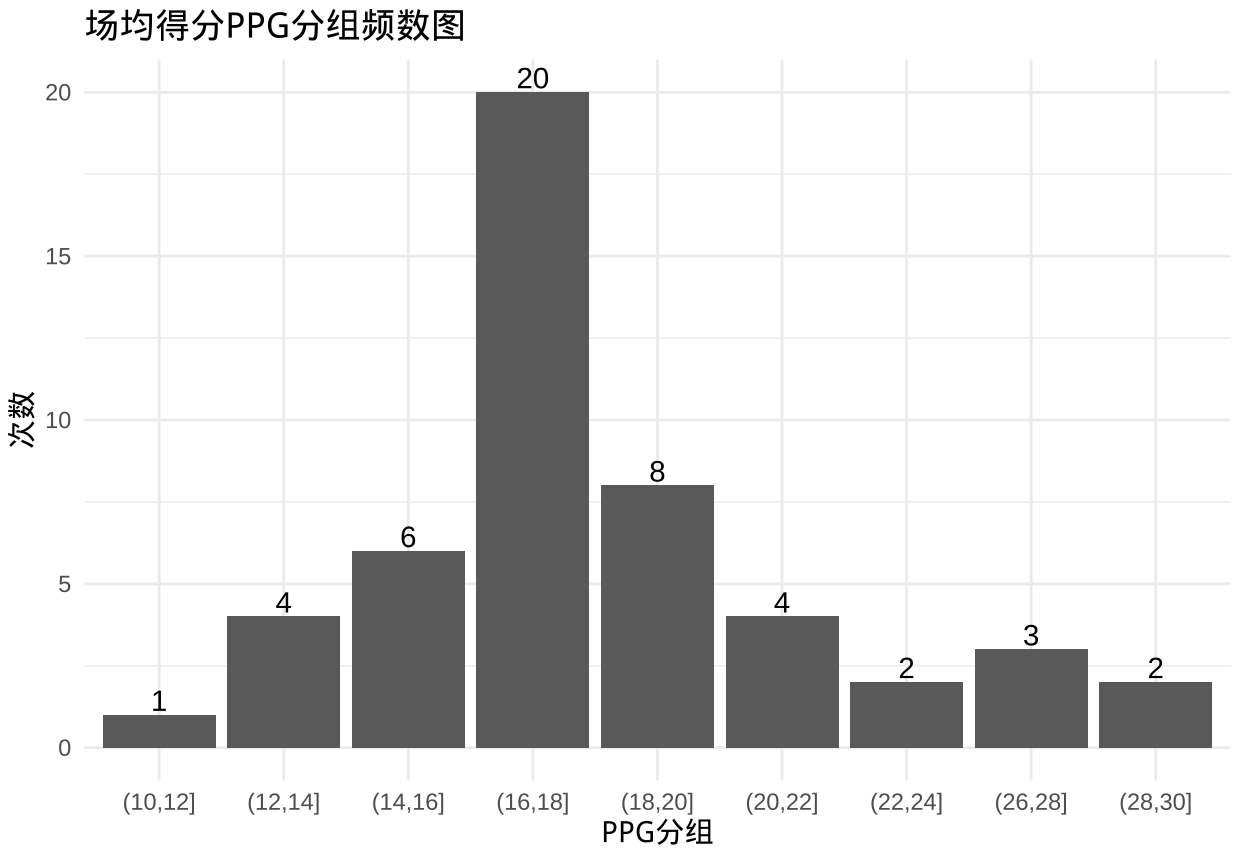
```
  )))
data_q2
```

```
#> # A tibble: 50 x 4
#>     Rank Player                   PPG class
#>    <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
```

2.1 a

a. Show the frequency distribution.

```
ggplot(data_q2 %>% count(class),aes(x=class,y=n)) +
  geom_col()+
  geom_text(aes(label = n),vjust = -0.2)+
  labs(title = " 场均得分 PPG 分组频数图") +
  xlab("PPG 分组") +
  ylab(" 次数")
```



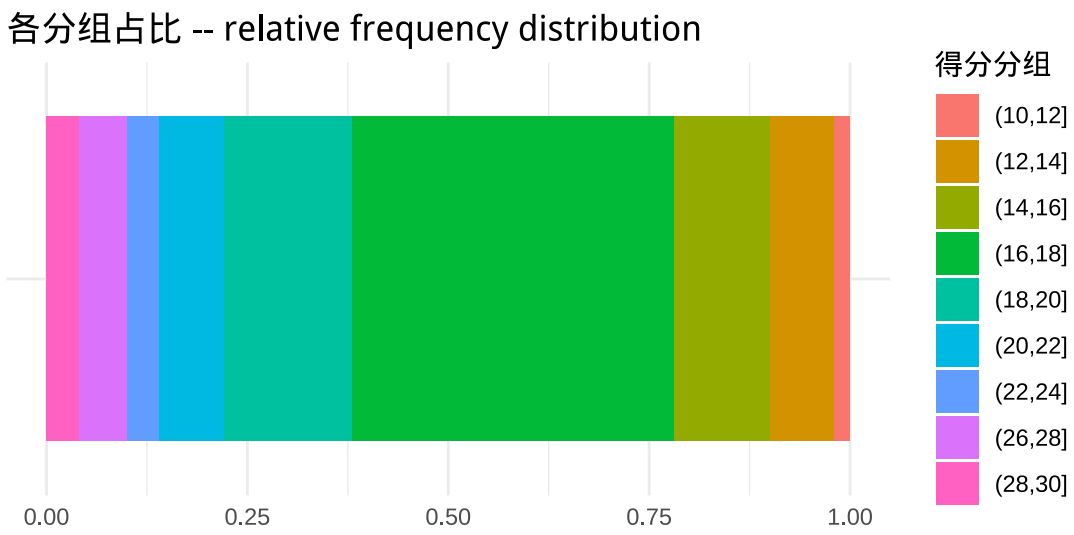
2.2 b

b. Show the relative frequency distribution.

```
table(data_q2$class) / nrow(data_q2)
```

```
#>
#> (10,12] (12,14] (14,16] (16,18] (18,20] (20,22] (22,24] (24,26] (26,28] (28,30]
#>    0.02    0.08    0.12    0.40    0.16    0.08    0.04    0.00    0.06    0.04
```

```
ggplot(data_q2 %>% count(class), aes(x = "", group = class, y = n)) +
  geom_col(aes(fill = class),position = "fill") +
  coord_flip() +
  labs(title = " 各分组占比 -- relative frequency distribution",
       fill = " 得分分组") +
  xlab("") +
  ylab("")
```



2.3 c

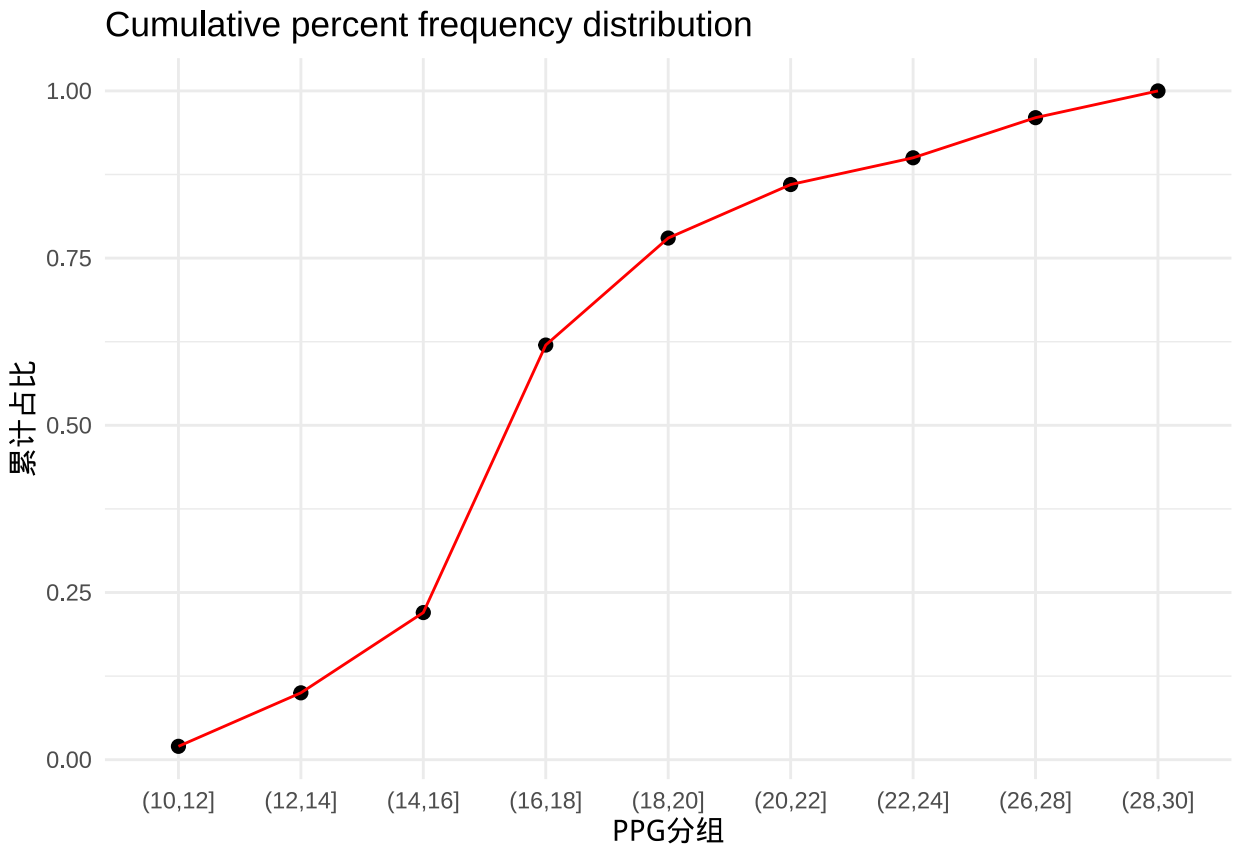
c. Show the cumulative percent frequency distribution.

```
data_q2 %>%
  count(class) %>%
  mutate(prop_n = n / sum(n), cusum_prop = cumsum(n) / sum(n))
```

```
#> # A tibble: 9 x 4
#>   class       n prop_n cusum_prop
#>   <fct>   <int>  <dbl>      <dbl>
#> 1 (10,12]     1   0.02       0.02
#> 2 (12,14]     4   0.08       0.1
#> 3 (14,16]     6   0.12       0.22
#> 4 (16,18]    20   0.4        0.62
#> 5 (18,20]     8   0.16       0.78
#> 6 (20,22]     4   0.08       0.86
#> 7 (22,24]     2   0.04       0.9
#> 8 (26,28]     3   0.06       0.96
#> 9 (28,30]     2   0.04       1
```

```
data_q2 %>%
  count(class) %>%
  mutate(prop_n = n / sum(n), cusum_prop = cumsum(n) / sum(n),
```

```
        x=row_number()) %>%
  ggplot(aes(x=class, y = cusum_prop))+
  geom_point(size = 2)+
  geom_line(aes(x = x, y = cusum_prop),color = "red")+
  xlab("PPG 分组") +
  ylab(" 累计占比")+
  ggtitle("Cumulative percent frequency distribution")
```

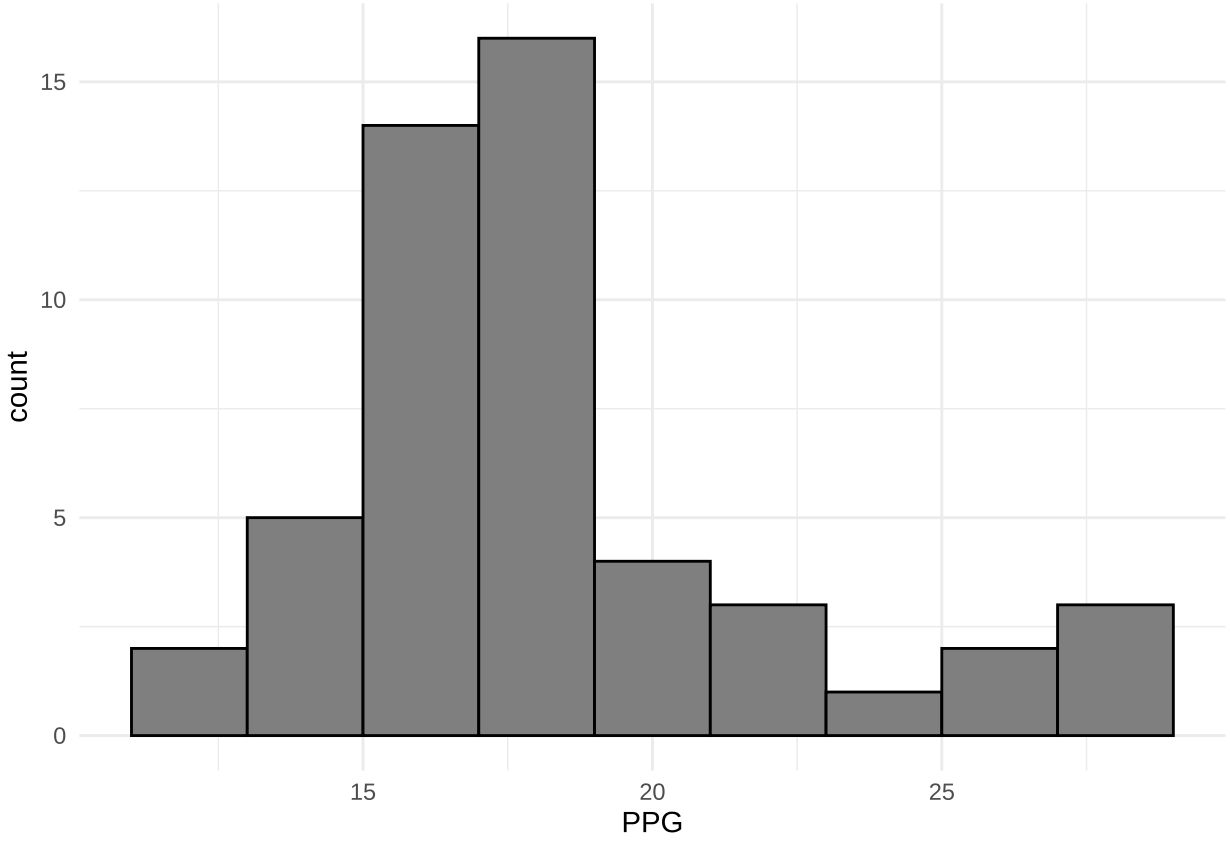


累计占比，请见表格的 `cusum_prop` 列，分区情况请见图。

2.4 d

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

```
ggplot(data_q2,aes(PPG)) +
  geom_histogram(binwidth = 2,fill="grey50",color = "black")
```



2.5 e

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

```
e1071::skewness(data_q2$PPG)
```

```
#> [1] 1.124
```

通过计算偏度得到值为为 1.1, 结合长尾直方图，可以判断 PPG 的数据是右偏的。

2.6 f

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

```
data_q2 %>%
  filter(PPG>=20) %>%
  tally() / nrow(data_q2)
```

```
#>      n
#> 1 0.22
```

PPG 大于等于 20 分的球员占比为 22%。

3 Question #3

A researcher reports survey results by stating that the standard error of the mean is 20. The population standard deviation is 500.

$$se = \sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}}$$
$$n = \frac{\sigma^2}{se^2}$$

3.1 a

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

```
se = 20
sigma = 500

sigma^2 / se^2
```

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

至少需要 625 个样本。

3.2 b

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

样本量足够大，可以认为均值服从正态分布 $N\ (\bar{x}, \sigma/\sqrt{n})$

$$P(x \leq \bar{x} + 25) = \Phi(\frac{25}{\sigma_{\bar{x}}})$$
$$P(x \leq \bar{x} - 25) = \Phi(\frac{-25}{\sigma_{\bar{x}}})$$
$$P(\bar{x} - 25 \leq x \leq \bar{x} + 25) = \Phi(\frac{25}{\sigma_{\bar{x}}}) - \Phi(\frac{-25}{\sigma_{\bar{x}}})$$

```
pnorm(25/20) - pnorm(-25/20)
```

```
#> [1] 0.7887
```

总体均值的点估计在 ±25 之间的概率为 0.7887005。

4 Question #4

Young Professional Magazine (Attached Data: Professional)

Young Professional magazine was developed for a target audience of recent college graduates who are in their first 10 years in a business/professional career. In its two years of publication, the magazine has been fairly successful. Now the publisher is interested in expanding the magazine's advertising base. Potential advertisers continually ask about the demographics and interests of subscribers to *young Professionals*. To collect this information, the magazine commissioned a survey to develop a profile of its subscribers. The survey results will be used to help the magazine choose articles of interest and provide advertisers with a profile of subscribers. As a new employee of the magazine, you have been asked to help analyze the survey results.

Some of the survey questions follow:

1. What is your age?
2. Are you: Male_____ Female_____
3. Do you plan to make any real estate purchases in the next two years?

   Yes_____ No_____
4. What is the approximate total value of financial investments, exclusive of your home, owned by you or members of your household?
5. How many stock/bond/mutual fund transactions have you made in the past year?
6. Do you have broadband access to the Internet at home? Yes_____ No_____
7. Please indicate your total household income last year. _____
8. Do you have children? Yes_____ No_____

The file entitled Professional contains the responses to these questions.

Managerial Report:

Prepare a managerial report summarizing the results of the survey. In addition to statistical summaries, discuss how the magazine might use these results to attract advertisers. You might also comment on how the survey results could be used by the magazine's editors to identify topics that would be of interest to readers. Your report should address the following issues, but do not limit your analysis to just these areas.

```
data_q4 <- read_csv("data/Professional.csv",col_select = 1:8) %>%
  rename(
    age = Age,
    gender = Gender,
    buy_estate = `Real Estate Purchases?`,
    value_investment = `Value of Investments ($)`,
    num_transaction = `Number of Transactions`,
    access_internet = `Broadband Access?`,
    ttl_income = `Household Income ($)`,
    have_child = `Have Children?`
  ) %>%
  select(1:8) %>%
  mutate(across(is.character, as_factor))

data_q4
```

```
#> # A tibble: 410 x 8
#>      age gender buy_estate value_investment num_transaction access_internet
#>    <dbl> <fct>  <fct>                <dbl>           <dbl> <fct>
#>  1    38 Female No                   12200               4 Yes
#>  2    30 Male   No                   12400               4 Yes
#>  3    41 Female No                   26800               5 Yes
#>  4    28 Female Yes                  19600               6 No
#>  5    31 Female Yes                  15100               5 No
#>  6    32 Male   No                   39700               3 Yes
#>  7    32 Male   Yes                  21900               2 Yes
#>  8    26 Female Yes                  41900               2 Yes
#>  9    26 Male   Yes                  16100               4 Yes
#> 10    34 Female Yes                  18400              11 Yes
#> # i 400 more rows
#> # i 2 more variables: ttl_income <dbl>, have_child <fct>
```

4.1 a

a. Develop appropriate descriptive statistics to summarize the data.

```
summary(data_q4)
```

```
#>       age          gender    buy_estate value_investment num_transaction
#>  Min.   :19.0   Female:181   No :229    Min.   :     0   Min.   : 0.00
#>  1st Qu.:28.0   Male  :229   Yes:181    1st Qu.: 18300   1st Qu.: 4.00
#>  Median :30.0                           Median : 24800   Median : 6.00
#>  Mean   :30.1                           Mean   : 28538   Mean   : 5.97
#>  3rd Qu.:33.0                           3rd Qu.: 34275   3rd Qu.: 7.00
#>  Max.   :42.0                           Max.   :133400   Max.   :21.00
#>  access_internet   ttl_income     have_child
#>  Yes:256         Min.   : 16200   Yes:219
#>  No :154         1st Qu.: 51625   No :191
#>                  Median : 66050
#>                  Mean   : 74460
#>                  3rd Qu.: 88775
#>                  Max.   :322500
```

4.2 b

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

```
age_sigma = sd(data_q4$age)
(age_bar = mean(data_q4$age))
```

```
#> [1] 30.11
```

```
income_sigma = sd(data_q4$ttl_income)
(income_bar = mean(data_q4$ttl_income))
```

```
#> [1] 74460
```

```
# 总体方差未知, 使用 t 分布
n = nrow(data_q4)
n
```

```
#> [1] 410
```

```
t_alpha = qt(0.975, df = n - 1)
t_alpha
```

```
#> [1] 1.966
```

总体方差未知，样本均值的区间估计公式为：

$$\bar{x} \pm t_{\alpha/2} \times \frac{s}{\sqrt{n}}$$

t_alpha * age_sigma / sqrt(n) = 0.3906642
t_alpha * income_sigma / sqrt(n) = 3380.2565944

年龄均值的 95% 的置信区间为 [29.7215309, 30.5028594] 。

家庭总收入 ttl_income 的均值的 95% 的置信区间为 [7.1079256×10^4, 7.7839769×10^4] 。

4.3 c

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

```
p_hat_1 = mean(data_q4$access_internet == "Yes")
p_hat_1
```

```
#> [1] 0.6244
```

```
p_hat_2 = mean(data_q4$have_child == "Yes")
p_hat_2
```

```
#> [1] 0.5341
```

```
z_alpha = qnorm(0.975)
z_alpha
```

```
#> [1] 1.96
```

样本比例的区间估计公式为：

$$\hat{p} \pm Z_{\alpha/2} \times \sqrt{\frac{\hat{p} * (1 - \hat{p})}{n}}$$

家中有宽带比例的 95% 的置信区间为 [0.577514 , 0.6712665]

有孩子的比例的 95% 的置信区间为 [0.4858615 , 0.5824312]

4.4 d

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

对在线经纪商（通过互联网提供金融服务）来说，这个杂志是很好的广告渠道。可以通过第 1 问的数据汇总结果看出:

1. 3/4 及以上的人都有投资，投资标的物价值的均值为 28539，中位数为 24800，占他们全年收入的 1/3 以上了。
2. 一年平均交易 6 次左右，说明交易的需求还是有的。
3. 62% 人的家里都有宽带，可以访问互联网。

基于以上的汇总数据，这个 *Young Professional* 杂志是一个比较的好的渠道推广金融服务，潜在客户多，而且投资的金额也不低。

4.5 e

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

对售卖教育软件跟儿童电脑游戏的公司来说，这个杂志也是适合的，儿童的比例 219/(219+192) = 0.53 , 超过一半了; 而且家庭的年总收入的均值为 74k，应该算是中产家庭了，在孩子教育跟娱乐方面还是很愿意消费的。

4.6 f

f. Comment on the types of articles you believe would be of interest to readers of *Young Professional*.

针对《Young Professional Magazine》的特定目标受众——近期大学毕业、处于商业或专业职业生涯前 10 年的年轻人，以及他们的特点（53% 有孩子，年收入均值为 7.5 万美元，其中 1/3 的资产用于投资），以下是一些可能引起他们兴趣的文章类型：

1. **职业发展与晋升**：文章可以提供如何在职场快速成长、获得晋升机会的策略。
2. **财务管理与投资**：鉴于他们有相当一部分资产用于投资，可以提供股票市场分析、投资策略、退休规划和税务规划的文章。
3. **家庭与工作平衡**：探讨如何在繁忙的工作和育儿责任之间找到平衡，提供时间管理和家庭规划的实用建议。
4. **教育与育儿**：提供关于如何为孩子提供良好教育和成长环境的文章，包括教育储蓄和家庭教育技巧。
5. **健康与福利**：鉴于他们处于职业生涯的早期阶段，可能对健康保险、健康生活方式和工作场所健康福利感兴趣。
6. **房地产与住房**：提供购房指南、抵押贷款建议和房地产市场分析，帮助他们做出明智的住房决策。
7. **科技与效率工具**：介绍可以提高工作效率和生活质量的最新科技工具和应用程序。
8. **法律与合规**：提供关于劳动合同、知识产权和商业法律的基础知识，帮助他们在职场中避免法律风险。

这些文章类型能够满足《Young Professional Magazine》读者在职业发展、家庭生活、财务规划和个人兴趣等方面的多元化需求。

5 Question #5:

Quality Associate, Inc. (Attached Data: Quality)

Quality associates, inc., a consulting firm, advises its clients about sampling and statistical procedures that can be used to control their manufacturing processes. in one particular application, a client gave Quality associates a sample of 800 observations taken during a time in which that client's process was operating satisfactorily. the sample standard deviation for these data was .21; hence, with so much data, the population standard deviation was assumed to be .21. Quality associates then suggested that random samples of size 30 be taken periodically to monitor the process on an ongoing basis. by analyzing the new samples, the client could quickly learn whether the process was operating satisfactorily. when the process was not operating satisfactorily, corrective action could be taken to eliminate the problem. the design specification indicated the mean for the process should be 12. the hypothesis test suggested by Quality associates follows.

$$H_0 : \mu = 12 \; H_a : \mu \neq 12$$

Corrective action will be taken any time H_0 is rejected.

Data are available in the data set Quality.

Managerial Report

```
data_q5 <- read_csv("data/Quality.csv") %>%
  rename(s1 = `Sample 1`,
         s2 = `Sample 2`,
         s3 = `Sample 3`,
         s4 = `Sample 4`)


# cal_p <- function(vec,miu,sigma,n){
#   a <- mean(vec) - miu
#   if(a >=0) {return(2*(1-pnorm(a/(sigma/sqrt(n)))))}
#     else
#       return(2*pnorm(a/(sigma/sqrt(n))))
# }
```

5.1 a

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.

```
lapply(data_q5, t.test, mu = 12, alternative = "two.sided", conf.level = 0.99)
```

```
#> $s1
#>
#>  One Sample t-test
#>
#> data:  X[[i]]
#> t = -1, df = 29, p-value = 0.3
#> alternative hypothesis: true mean is not equal to 12
#> 99 percent confidence interval:
#>  11.85 12.07
#> sample estimates:
#> mean of x
#>     11.96
#>
#>
#> $s2
#>
#>  One Sample t-test
```

```
#>
#> data:  X[[i]]
#> t = 0.71, df = 29, p-value = 0.5
#> alternative hypothesis: true mean is not equal to 12
#> 99 percent confidence interval:
#>  11.92 12.14
#> sample estimates:
#> mean of x
#>     12.03
#>
#>
#> $s3
#>
#>  One Sample t-test
#>
#> data:  X[[i]]
#> t = -2.9, df = 29, p-value = 0.006
#> alternative hypothesis: true mean is not equal to 12
#> 99 percent confidence interval:
#>  11.78 11.99
#> sample estimates:
#> mean of x
#>     11.89
#>
#>
#> $s4
#>
#>  One Sample t-test
#>
#> data:  X[[i]]
#> t = 2.2, df = 29, p-value = 0.04
#> alternative hypothesis: true mean is not equal to 12
#> 99 percent confidence interval:
#>  11.98 12.19
#> sample estimates:
#> mean of x
#>     12.08
```

可以看到，第一个样本的 p-value $= 0.3127$，第二个样本的 p-value $= 0.4818$，第三个样本的 p-value $= 0.006469$，

第四个样本的 p-value = 0.03906。

5.2 b

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

```
lapply(data_q5, sd)
```

```
#> $s1
#> [1] 0.2204
#>
#> $s2
#> [1] 0.2204
#>
#> $s3
#> [1] 0.2072
#>
#> $s4
#> [1] 0.2061
```

```
# 差异
lapply(data_q5, FUN = function(x) sd(x) - 0.21)
```

```
#> $s1
#> [1] 0.01036
#>
#> $s2
#> [1] 0.01036
#>
#> $s3
#> [1] -0.002829
#>
#> $s4
#> [1] -0.003891
```

假设总体的标准差为 0.21，这个假设看起来合理，样本的标准差偏差不大。

5.3 c

c. compute limits for the sample mean $\overline{x}$ around $\mu = 12$ such that, as long as a new sample mean is within those limits, the process will be considered to be operating satisfactorily. if $\overline{x}$ exceeds the upper limit or if $\overline{x}$ 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.

样本量超过 30，可以认为样本均值服从正态分布 $\bar{x} \sim N(\bar{x}, \frac{\sigma}{\sqrt{n}})$

```
# 区间估计, 显著性水平: 0.01

z_alpha = qnorm(0.995)


# 均值 =12 的置信区间:
c(12 - z_alpha *  0.21/sqrt(30), 12 + z_alpha *  0.21/sqrt(30))
```

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

5.4 d

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?

```
# 区间估计, 显著性水平: 0.1

z_alpha = qnorm(0.95)


# 均值 =12 的置信区间:
c(12 - z_alpha *  0.21/sqrt(30), 12 + z_alpha *  0.21/sqrt(30))
```

```
#> [1] 11.94 12.06
```

增大显著性水平，置信区间会变小，第一类错误 Type I Error 的概率增大。

6 Question #6:

Vacation occupancy rates were expected to be up during March 2008 in Myrtle Beach, South Carolina (*the sun news,* February 29, 2008). Data in the file Occupancy (Attached file **Occupancy**) will allow you to

replicate the findings presented in the newspaper. The data show units rented and not rented for a random sample of vacation properties during the first week of March 2007 and March 2008.

```
data_q6 <- read_csv("data/Occupancy.csv", skip = 1) %>%
  rename(march_2007 = `March 2007`, march_2008 = `March 2008`)
  #  %>% mutate(across(is.character,as.factor))
data_q6
```

```
#> # A tibble: 200 x 2
#>    march_2007 march_2008
#>    <chr>      <chr>
#>  1 Yes        No
#>  2 No         Yes
#>  3 Yes        Yes
#>  4 No         No
#>  5 No         Yes
#>  6 Yes        No
#>  7 No         No
#>  8 No         Yes
#>  9 No         Yes
#> 10 Yes        Yes
#> # i 190 more rows
```

6.1 a

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

```
(p_2007 = mean(data_q6$march_2007 == "Yes", na.rm = TRUE))
```

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

```
(p_2008 = mean(data_q6$march_2008 == "Yes", na.rm = TRUE))
```

```
#> [1] 0.4667
```

6.2 b

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

两比例之差的区间估计:

$$\sigma_{\hat{p_1}-\hat{p_2}} = \sqrt{\frac{p_1 * (1-p_1)}{n_1} + \frac{p_2 * (1-p_2)}{n_2}}$$

$$ME = Z_{\alpha/2} \times \sigma_{\hat{p_1}-\hat{p_2}}$$

```
p = p_2008 - p_2007
n1 = sum(!is.na(data_q6$march_2007))
n2 = sum(!is.na(data_q6$march_2008))
me = qnorm(1 - 0.05 / 2) * sqrt(p_2007 * (1 - p_2007) / n1 +
                                p_2008 * (1 - p_2008) / n2
                                )

c(p - me, p + me)
```

```
#> [1] 0.01302 0.22032
```

2008 年与 2007 年之差的比例，95% 的置信区间为 0.0130152, 0.2203182 。

6.3 c

c. On the basis of your findings, does it appear March rental rates for 2008 will be up from those a year earlier?

假设检验：

$$H_0 : P_{2008} - P_{2007} >= 0; H_a : P_{2008} - P_{2007} < 0$$

原假设为真，$P_{2008} = P_{2007} = P$,

$$\bar{p} = \frac{n_1\bar{p_1} + n_2\bar{p_2}}{n_1 + n_2}$$

$$\sigma_{\bar{p}} = \sqrt{\bar{p}(1-\bar{p})(\frac{1}{n_1} + \frac{1}{n_2})}$$

检验统计量为：

$$z = \frac{\bar{p_1} - \bar{p_2}}{\sqrt{\bar{p}(1-\bar{p})(\frac{1}{n_1} + \frac{1}{n_2})}}$$

```
p_bar = (n1 * p_2007 + n2*p_2008) /(n1 + n2)
sigma_bar = sqrt(p_bar * (1- p_bar) * (1/n1 + 1/n2))

z = (p_2008 - p_2007) / sigma_bar
z
```

```
#> [1] 2.205
```

```
pnorm(z)
```

```
#> [1] 0.9863
```

```
# 右侧曲线面积
1-pnorm(z)
```

```
#> [1] 0.01373
```

对于单侧检验 p 值为 $0.0137343 < 0.05$ ，在 0.05 的显著性水平下，拒绝 H0，认为 2008 年 3 月的入住比例显著比 2007 年 3 月的高。入住率提高，那么租金会上涨。

7 Question #7

Air Force Training Program (data file: Training)

An air force introductory course in electronics uses a personalized system of instruction whereby each student views a videotaped lecture and then is given a programmed instruc-tion text. the students work independently with the text until they have completed the training and passed a test. Of concern is the varying pace at which the students complete this portion of their training program. Some students are able to cover the programmed instruction text relatively quickly, whereas other students work much longer with the text and require additional time to complete the course. The fast students wait until the slow students complete the introductory course before the entire group proceeds together with other aspects of their training.

A proposed alternative system involves use of computer-assisted instruction. In this method, all students view the same videotaped lecture and then each is assigned to a computer terminal for further instruction. The computer guides the student, working independently, through the self-training portion of the course.

To compare the proposed and current methods of instruction, an entering class of 122 students was assigned randomly to one of the two methods. one group of 61 students used the current programmed-text method and the other group of 61 students used the proposed computer-assisted method. The time in hours was recorded for each student in the study. Data are provided in the data set training (see Attached file).

Managerial Report

读入数据：

```
tr <- read_csv("data/Training.csv")
tr
```

```
#> # A tibble: 61 x 2
#>    Current Proposed
#>      <dbl>    <dbl>
#>  1      76       74
#>  2      76       75
#>  3      77       77
#>  4      74       78
#>  5      76       74
#>  6      74       80
#>  7      74       73
#>  8      77       73
#>  9      72       78
#> 10      78       76
#> # i 51 more rows
```

7.1 a

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?

```
summary(tr)
```

```
#>     Current        Proposed
#>  Min.   :65.0   Min.   :69.0
#>  1st Qu.:72.0   1st Qu.:74.0
#>  Median :76.0   Median :76.0
#>  Mean   :75.1   Mean   :75.4
#>  3rd Qu.:78.0   3rd Qu.:77.0
#>  Max.   :84.0   Max.   :82.0
```

```
sd(tr$Current)
```

```
#> [1] 3.945
```

```r
sd(tr$Proposed)
```

```
#> [1] 2.506
```

中位数相同，均值接近，标准差有差异。

7.2 b

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

两种方法的样本均值差异不大，说明效果不明显。下面采用统计学的假设检验的方法来验证一下：

总体方差未知，总体的均值假设检验，采用 t 统计量。

```r
t.test(tr$Current,tr$Proposed)
```

```
#>
#>  Welch Two Sample t-test
#>
#> data:  tr$Current and tr$Proposed
#> t = -0.6, df = 102, p-value = 0.5
#> alternative hypothesis: true difference in means is not equal to 0
#> 95 percent confidence interval:
#>  -1.5477  0.8263
#> sample estimates:
#> mean of x mean of y
#>     75.07     75.43
```

检验结果表面，95% 的置信水平下，不能拒绝原假设，认为两种方法没有显著的差异。

7.3 c

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.

```r
# 方差
lapply(tr, var)
```

```
#> $Current
#> [1] 15.56
#>
#> $Proposed
#> [1] 6.282
```

```
# 标准差
lapply(tr, sd)
```

```
#> $Current
#> [1] 3.945
#>
#> $Proposed
#> [1] 2.506
```

两个总体方差的统计推断：

总体方差的假设检验的统计量为

$$F = \frac{s_1}{s_2}$$

```
f_stats = var(tr$Current) / var(tr$Proposed)
f_stats
```

```
#> [1] 2.477
```

```
# 临界值统计量 F0.05 , 单侧
qf(0.95, df1 = nrow(tr)-1, df2 = nrow(tr) -1)
```

```
#> [1] 1.534
```

```
# p 值 单侧
1- pf(f_stats,df1 = nrow(tr)-1, df2 = nrow(tr) -1)
```

```
#> [1] 0.000289
```

```
# p 值双侧
# (1- pf(f_stats,df1 = nrow(tr)-1, df2 = nrow(tr) -1)) * 2
#
# F Test to Compare Two Variances
# var.test(tr$Proposed, tr$Current)
# var.test(tr$Current, tr$Proposed)
```

在显著性水平为 0.05 条件下，拒绝 h0，认为两个总体的方差有显著差异。

7.4 d

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

总体均值没有差异，但是方差有差异，实验组的方差减小，说明有第二组大部分人完成时间的差异在缩小，第二种方法有一定的作用。

7.5 e

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?

我觉得还需要再对组里的每个人组织一次测试，看看两组最终的平均测试成绩是否有明显差异。

8 Question #8

The Toyota Camry is one of the best-selling cars in North America. The cost of a previously owned Camry depends upon many factors, including the model year, mileage, and condition. To investigate the relationship between the car's mileage and the sales price for a 2007 model year Camry, Attached data file Camry show the mileage and sale price for 19 sales (Pricehub website, February 24, 2012).

```
camry <- read_csv("data/Camry.csv")

camry <- camry %>%
  rename(miles = `Miles (1000s)`, price = `Price ($1000s)`)

camry
```

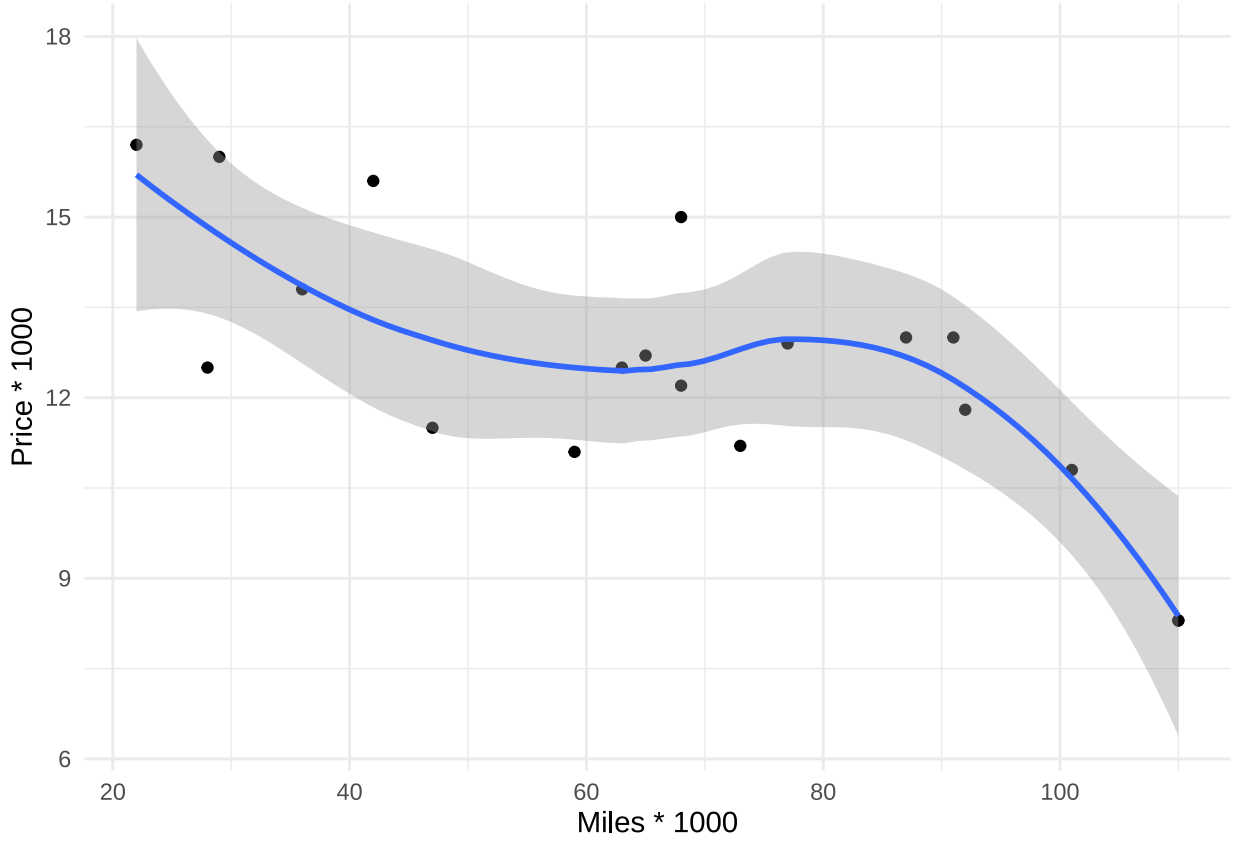
```
#> # A tibble: 19 x 2
#>    miles price
#>    <dbl> <dbl>
#>  1    22  16.2
#>  2    29  16
#>  3    36  13.8
#>  4    47  11.5
```

```
#>  5    63  12.5
#>  6    77  12.9
#>  7    73  11.2
#>  8    87  13
#>  9    92  11.8
#> 10   101  10.8
#> 11   110   8.3
#> 12    28  12.5
#> 13    59  11.1
#> 14    68  15
#> 15    68  12.2
#> 16    91  13
#> 17    42  15.6
#> 18    65  12.7
#> 19   110   8.3
```

8.1 a

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

```
ggplot(camry, aes(x = miles, y = price)) +
  geom_point() +
  geom_smooth() +
  xlab("Miles * 1000") +
  ylab("Price * 1000")
```



8.2 b

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

行驶里程越长，价格有降低的趋势

8.3 c

c. Develop the estimated regression equation that could be used to predict the price ($1000s) given the miles (1000s).

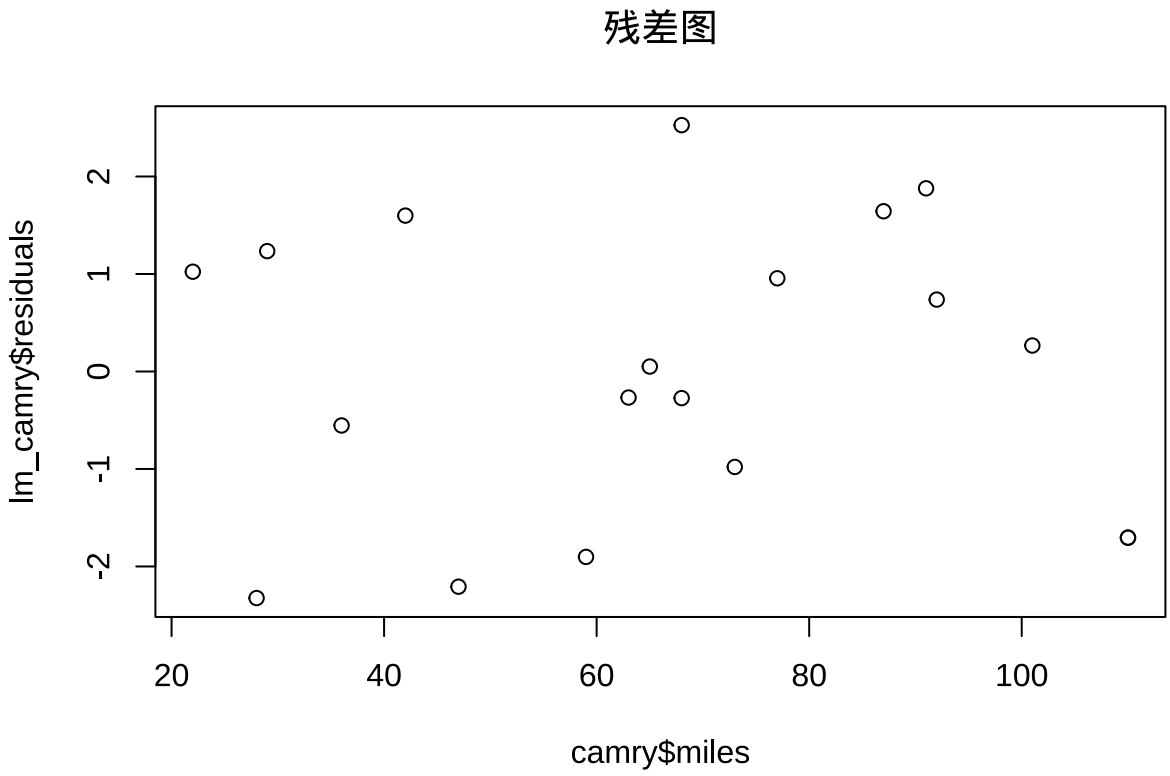
```
lm_camry <- lm(price ~ miles, data = camry)

summary(lm_camry)
```

```
#>
```

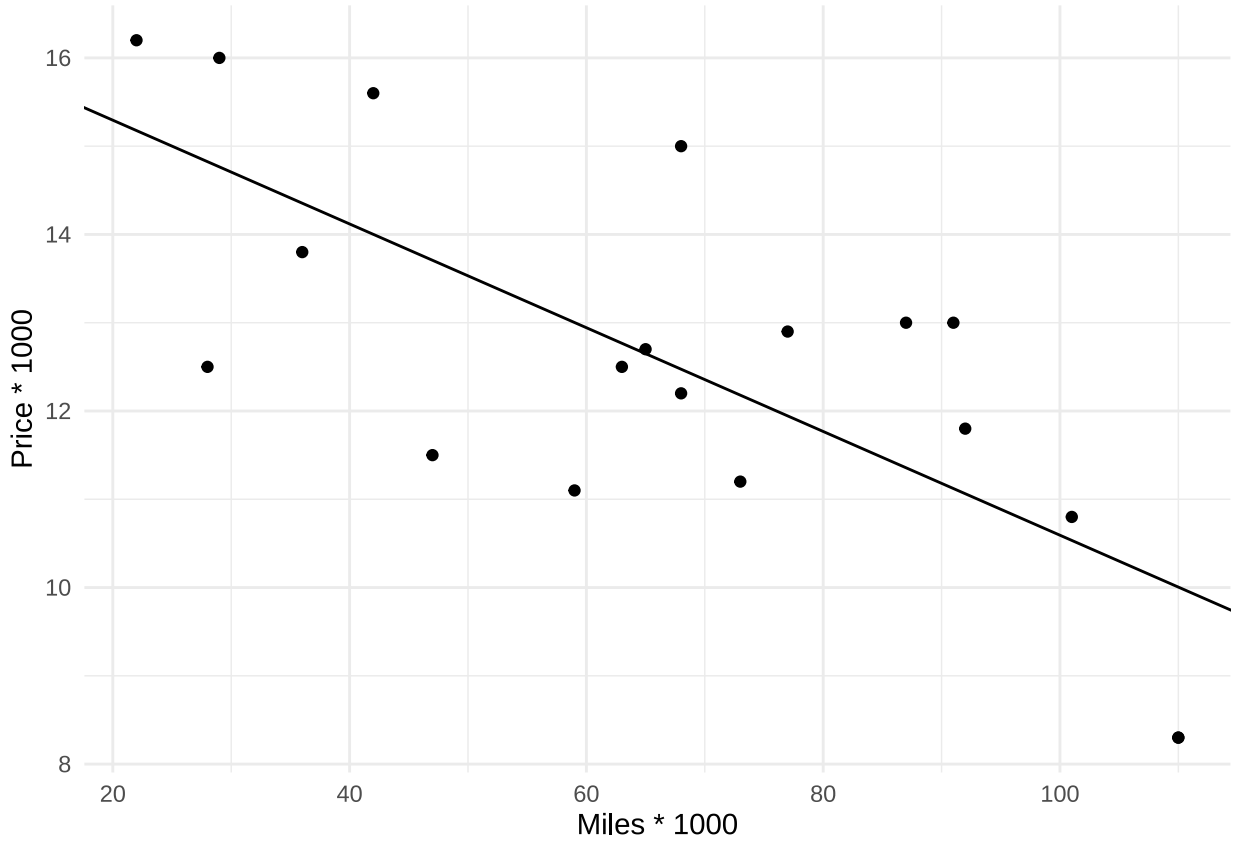
```
#> Call:
#> lm(formula = price ~ miles, data = camry)
#>
#> Residuals:
#>     Min      1Q  Median      3Q     Max
#> -2.3241 -1.3419  0.0506  1.1290  2.5269
#>
#> Coefficients:
#>             Estimate Std. Error t value Pr(>|t|)
#> (Intercept)  16.4698     0.9488   17.36    3e-12 ***
#> miles        -0.0588     0.0132   -4.46  0.00035 ***
#> ---
#> Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
#>
#> Residual standard error: 1.54 on 17 degrees of freedom
#> Multiple R-squared:  0.539,  Adjusted R-squared:  0.512
#> F-statistic: 19.8 on 1 and 17 DF,  p-value: 0.000348
```

```
plot(camry$miles, lm_camry$residuals,main = " 残差图")
```




```
# std.Error of b1
# sqrt(sum((lm_camry$fitted.values - camry$price)^2) / (nrow(camry)-2) )
#  / sqrt(sum((camry$miles - mean(camry$miles))^2))
#  t statistic  =  b1/std.Error
```

```
ggplot(camry, aes(x = miles, y = price)) +
  geom_point() +
  geom_abline(slope = -0.05877, intercept = 16.46976)+
  xlab("Miles * 1000") +
  ylab("Price * 1000")
```



$$price = 16.46976 - 0.05877 * miles$$

d d. Test for a significant relationship at the .05 level of significance.

`miles` 的系数假设检验的 p 值为 0.000348，小于 0.05，拒绝原假设，认为系数不为 0，说明 miles 与 pirce 存在一个显著的关系。

8.4 e

e. Did the estimated regression equation provide a good fit? Explain.

通过判定系数 R-squared = 0.5387, 说明该模型的拟合效果不太行，miles 变量只解释了 53.8% 的 price 的变异性。

8.5 f

f. Provide an interpretation for the slope of the estimated regression equation.

里程每增加 1000 英里，车辆的价格下降 58.77 美元。

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.

```
predict(lm_camry,data.frame(miles=60))
```

```
#>     1
#> 12.94
```

模型预测价格是 12943 美元。这个价格只能拿来参考，还要考虑车况，保养，有没有事故等情况。

9 Question #9:

附件 WE.xlsx 是某提供网站服务的 Internet 服务商的客户数据。数据包含了 6347 名客户在 11 个指标上的表现。其中" 流失 "指标中 0 表示流失，"1"表示不流失，其他指标含义看变量命名。

读入数据:

```
we <- readxl::read_xlsx("data/WE.xlsx")

we <- we %>%
  rename(id = `客户 ID`,
         is_lost = 流失,
         happy_index = 当月客户幸福指数,
         happy_index_chg = 客户幸福指数相比上月变化,
         cust_support = 当月客户支持,
         cust_support_chg = 客户支持相比上月的变化,
         ser_prior = 当月服务优先级,
         ser_prior_chg = 服务优先级相比上月的变化,
         login_cnt = 当月登录次数,
         blog_cnt_chg = 博客数相比上月的变化,
         visit_add = 访问次数相比上月的增加,
         cust_expired = 客户使用期限,
         visit_interval= 访问间隔变化)
we
```

```
#> # A tibble: 6,347 x 13
```

```
#>       id is_lost happy_index happy_index_chg cust_support cust_support_chg
#>    <dbl>   <dbl>       <dbl>           <dbl>        <dbl>            <dbl>
#>  1     1       0           0               0            0                0
#>  2     2       0          62               4            0                0
#>  3     3       0           0               0            0                0
#>  4     4       0         231               1            1               -1
#>  5     5       0          43              -1            0                0
#>  6     6       0         138             -10            0                0
#>  7     7       0         180              -5            1                1
#>  8     8       0         116             -11            0                0
#>  9     9       0          78              -7            1               -2
#> 10    10       0          78             -37            0                0
#> # i 6,337 more rows
#> # i 7 more variables: ser_prior <dbl>, ser_prior_chg <dbl>, login_cnt <dbl>,
#> #   blog_cnt_chg <dbl>, visit_add <dbl>, cust_expired <dbl>,
#> #   visit_interval <dbl>
```

9.1 a

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

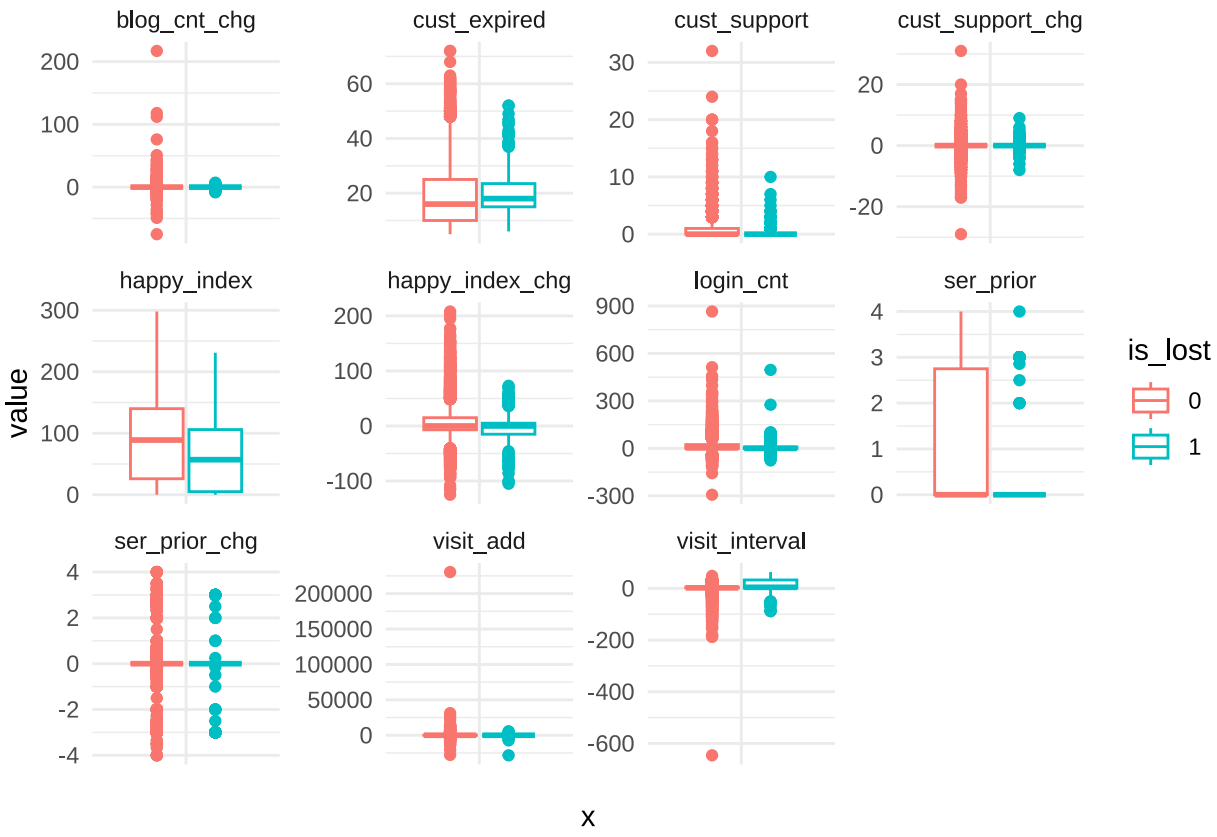
```
we %>% group_by(is_lost) %>%
  summarise(across(happy_index:visit_interval, .fns= mean, na.rm =TRUE)) %>%
  pivot_longer(happy_index:visit_interval,values_to = "mean") %>%
  arrange(name,is_lost) %>%
  select(name,is_lost,mean) %>%
  print(n = 30)
```

```
#> # A tibble: 22 x 3
#>    name             is_lost      mean
#>    <chr>              <dbl>     <dbl>
#>  1 blog_cnt_chg           0   0.171
#>  2 blog_cnt_chg           1  -0.102
#>  3 cust_expired           0  18.8
#>  4 cust_expired           1  20.4
#>  5 cust_support           0   0.724
#>  6 cust_support           1   0.372
#>  7 cust_support_chg       0  -0.00930
```

```
#>  8 cust_support_chg       1   0.0372
#>  9 happy_index            0  88.6
#> 10 happy_index            1  63.3
#> 11 happy_index_chg        0   5.53
#> 12 happy_index_chg        1  -3.74
#> 13 login_cnt              0  16.1
#> 14 login_cnt              1   8.06
#> 15 ser_prior              0   0.830
#> 16 ser_prior              1   0.500
#> 17 ser_prior_chg          0   0.0327
#> 18 ser_prior_chg          1  -0.0167
#> 19 visit_add              0 107.
#> 20 visit_add              1 -95.8
#> 21 visit_interval         0   3.51
#> 22 visit_interval         1   8.49
```

画图看一下：

```
we %>% select(is_lost:visit_interval) %>%
  pivot_longer(happy_index:visit_interval) %>%
  ggplot(aes(x="",y=value,colour=as.factor(is_lost))) +
  geom_boxplot() +
  facet_wrap(vars(name),scales = "free_y")+
  labs(colour = "is_lost")
```



可以看出 11 个变量均可能存在显著不同。

9.2 b

b. 通过均值比较的方式验证上述不同是否显著。

```
we %>% select(is_lost:visit_interval) %>%
  pivot_longer(happy_index:visit_interval) %>%
  group_by(name) %>%
  group_modify(~broom::tidy(t.test(value~is_lost, data=.x)), .keep = TRUE) %>%
  mutate(p.value = round(p.value,4))
```

```
#> # A tibble: 11 x 11
#> # Groups:   name [11]
#>    name         estimate estimate1 estimate2 statistic p.value parameter conf.low
#>    <chr>           <dbl>     <dbl>     <dbl>     <dbl>   <dbl>     <dbl>    <dbl>
#>  1 blog_cnt_c~    0.273     0.171     -0.102      2.53   0.0116      696.   0.0613
#>  2 cust_expir~   -1.53     18.8       20.4       -2.98   0.0031      380.  -2.55
```

```
#>  3 cust_suppo~   0.353    0.724      0.372      5.51   0           419.   0.227
#>  4 cust_suppo~  -0.0464  -0.00930    0.0372    -0.632  0.528       407.  -0.191
#>  5 happy_index  25.3     88.6       63.3        7.62   0           369.  18.8
#>  6 happy_inde~   9.27     5.53      -3.74       5.78   0           366.   6.12
#>  7 login_cnt     8.08    16.1        8.06       3.57   0.0004      363.   3.63
#>  8 ser_prior     0.330    0.830      0.500      5.14   0           373.   0.204
#>  9 ser_prior_~   0.0494   0.0327    -0.0167     0.641  0.522       364.  -0.102
#> 10 visit_add   202.     107.       -95.8        1.91   0.0563      448.  -5.46
#> 11 visit_inte~  -4.97     3.51       8.49      -4.10   0.0001      346.  -7.36
#> # i 3 more variables: conf.high <dbl>, method <chr>, alternative <chr>
```

可以看到 客户支持相比上月的变化 (cust_support_chg), 服务优先级相比上月的变化 (ser_prior_chg) 这两个变量的均值在是否流失，没有显著性的差异。

9.3 c

c. 以" 流失 "为因变量，其他你认为重要的变量为自变量（提示：a、b 两步的发现），建立回归方程对是否流失进行预测。

```
glm_we <- glm(
  is_lost ~ happy_index + happy_index_chg + cust_support +
    ser_prior + login_cnt + blog_cnt_chg + visit_add +
    cust_expired + visit_interval,
  family = binomial(link = "logit"),
  data = we
)

summary(glm_we)
```

```
#>
#> Call:
#> glm(formula = is_lost ~ happy_index + happy_index_chg + cust_support +
#>     ser_prior + login_cnt + blog_cnt_chg + visit_add + cust_expired +
#>     visit_interval, family = binomial(link = "logit"), data = we)
#>
#> Coefficients:
#>                   Estimate Std. Error z value Pr(>|z|)
#> (Intercept)     -2.8744187  0.1214855  -23.66  < 2e-16 ***
#> happy_index     -0.0052250  0.0011610   -4.50  6.8e-06 ***
```

```
#> happy_index_chg -0.0095009  0.0024239   -3.92  8.9e-05 ***
#> cust_support    -0.0352242  0.0743801   -0.47    0.636
#> ser_prior       -0.0372737  0.0751390   -0.50    0.620
#> login_cnt        0.0009104  0.0019523    0.47    0.641
#> blog_cnt_chg    -0.0000236  0.0207961    0.00    0.999
#> visit_add       -0.0001171  0.0000407   -2.88    0.004 **
#> cust_expired     0.0141828  0.0052602    2.70    0.007 **
#> visit_interval   0.0170011  0.0042767    3.98  7.0e-05 ***
#> ---
#> 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: 2466
#>
#> Number of Fisher Scoring iterations: 6
```

可以看到，有的变量不显著，可以通过逐步回归的方式，剔除不显著的自变量。

```
step_glm <- step(glm_we, trace = 0)
```

```
summary(step_glm)
```

```
#>
#> Call:
#> glm(formula = is_lost ~ happy_index + happy_index_chg + visit_add +
#>     cust_expired + visit_interval, family = binomial(link = "logit"),
#>     data = we)
#>
#> Coefficients:
#>                   Estimate Std. Error z value Pr(>|z|)
#> (Intercept)     -2.8937138  0.1200232  -24.11  < 2e-16 ***
#> happy_index     -0.0055843  0.0010716   -5.21  1.9e-07 ***
#> happy_index_chg -0.0094253  0.0022500   -4.19  2.8e-05 ***
#> visit_add       -0.0001133  0.0000401   -2.83   0.0047 **
#> cust_expired     0.0149233  0.0051896    2.88   0.0040 **
#> visit_interval   0.0171150  0.0042775    4.00  6.3e-05 ***
```

```
#> ---
#> 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: 2447.4  on 6341  degrees of freedom
#> AIC: 2459
#>
#> Number of Fisher Scoring iterations: 6
```

通过逐步回归，最终选取了 5 个变量，当月客户幸福指数 (happy_index)，客户幸福指数相比上月变化 (happy_index_chg)，访问次数相比上月的增加 (visit_add)，客户使用期限 (cust_expired)，访问间隔变化 (visit_interval) 。

9.4 d

d. 根据上一步预测的结果，对尚未流失（流失 =0）的客户进行流失可能性排序，并给出流失可能性最大的前 100 名用户 ID 列表。

```
we %>% add_column(pred_prob = predict(step_glm, type = "response")) %>%
  # select(id,is_lost,pred_prob,everything()) %>%
  filter(is_lost == 0) %>%
  arrange(-pred_prob) %>%
  slice(1:100) %>%
  pull(id)
```

```
#>   [1] 2287  109 1971    1 2025 2076   14   76   18    3  929   21 2244 1929 1287
#>  [16] 1459   51  128   59  183   55  121 2240 1520  137 2599 1236  154   68 1862
#>  [31] 2080 1143  146  119  171  190   89   42    5    2  123  101 2546 1286   95
#>  [46]   61 2680 1616 2838 2289 1438 1392 2924  106 1393 2481  203   69 3671 1574
#>  [61] 2255 1395 1478 2235 1204   62 1141  139  798 2830 1151 2739 1693   12 3042
#>  [76]  142 4245   57 1908   10 2286 1951 3076 2242 2062 1110 3124 2047  868  104
#>  [91] 1953 1019 2903  680 2656 2744 1446 2306  163 1155
```