第二次作业

雷婷

目录

```r
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
)
```

```
library(showtext)
```

```
## Warning: 程序包'showtext'是用R版本4.4.2 来建造的
```

```
## 载入需要的程序包：sysfonts
```

```
## 载入需要的程序包：showtextdb
```

```
#library(showtextdb)
showtext_auto()

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

```
# setwd(choose.dir())
library(tidyverse)
```

1 Q1：BigBangTheory 的（附数据：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).

# 解答前的准备工作

```
# 导入数据，以逗号作为分隔符，并对列进行命名
bbt<-read_csv("BigBangTheory.csv") %>%
  rename(air_date = `Air Date`,viewers = `Viewers (millions)`) %>%
  mutate(air_date = mdy(air_date))
```

2 Q1 a. Compute the minimum and the maximum number of viewers.

```r
min(bbt$viewers)
```

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

```r
max(bbt$viewers)
```

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

3 Q1 b. Compute the mean, median, and mode.

```r
mean(bbt$viewers)
```

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

```r
median(bbt$viewers)
```

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

```r
# 就一个众数时，可用此法，即计数向量次数，找到最大的以后，返回行名
modes <- names(which.max(table(bbt$viewers)))

# 以下适合多个众数时
#table 函数对 data 向量进行计数。table 函数会统计 data 中每个不同元素出现的次数，并返回一个表格形式的结
counts <- table(bbt$viewers)
# 找到 counts 表格中的最大值，也就是数据中出现次数最多的值的出现次数。
max_count <- max(counts)
# 首先使用 counts == max_count 生成一个逻辑向量，该向量指示 counts 中哪些元素等于最大出现次数。然后 whi
modes <- names(counts)[which(counts == max_count)]
modes
```

```
#> [1] "13.6" "14"   "16.1" "16.2"
```

4 Q1 c. Compute the first and third quartiles.

```
# Quantile 计算分位数， probs = c(分位数)
quartiles <- quantile(bbt$viewers, probs = c(0.25,0.75))
quartiles
```

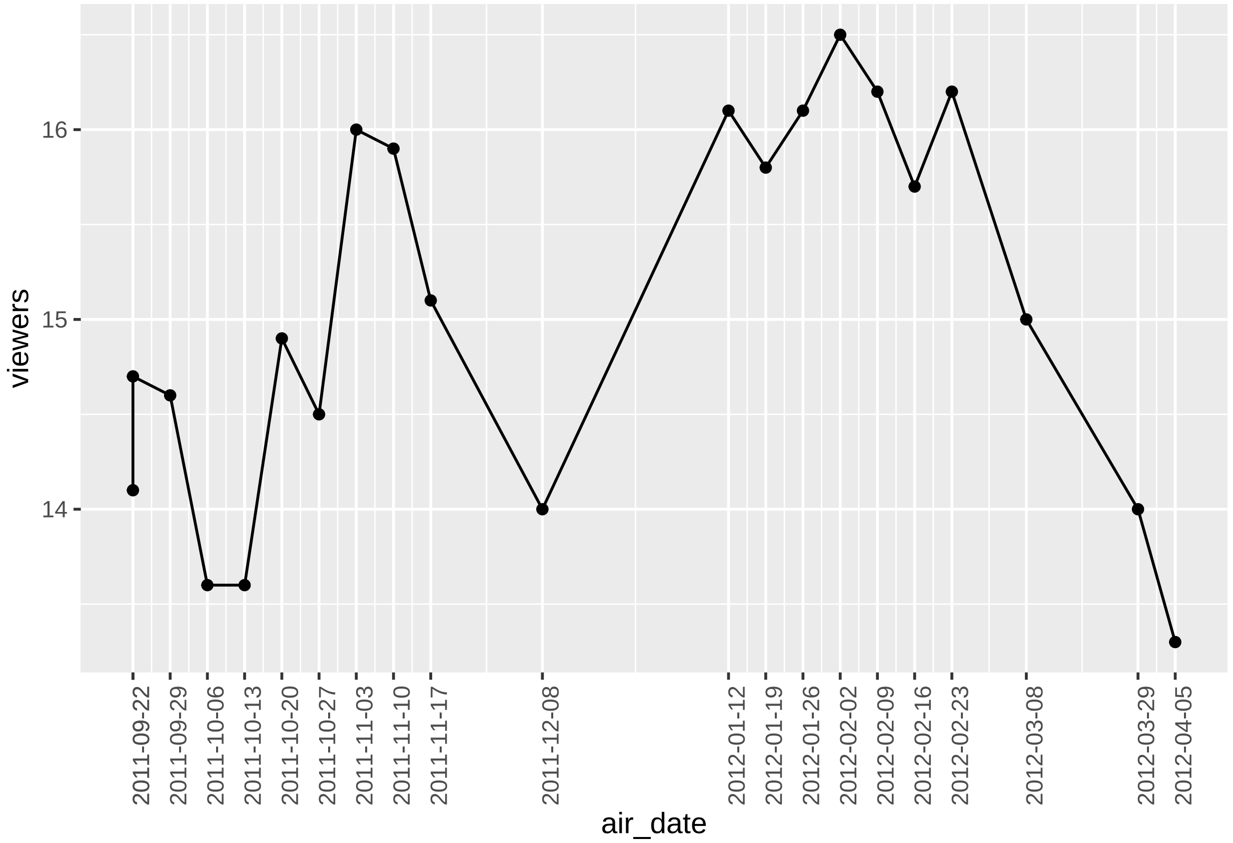
```
#>  25%  75%
#> 14.1 16.0
```

5 Q1 d. has viewership grown or declined over the 2011–2012 season? Discuss.

回答：收视率在下降观测方式一：线图看数据波动情况，发现有起伏，不利于直观得出结论观测方式二：计算每周较上周收视率的差值，统计后发现 7 次为正，12 次为负，说明整体呈下降观测方式三：使用线性回归模型来拟合收视率与日期数值之间的关系，计算出来后的线性斜率为负，说明呈下降趋势综上得出收视率在下降的结论

library(ggplot2) library(lubridate)

```
ggplot(data = bbt, aes(x = air_date , y = viewers)) +
  geom_line()+
  geom_point()+
#x 轴的间隔 = 实际数据值
  scale_x_date(breaks = bbt$air_date)+
# 调整 x 轴间距
  theme(axis.text.x = element_text(angle = 90))
```



```
  # scale_x_date(date_breaks = "1 month")
```

```
# 给数据表增加序列号
bbt <- bbt[order(bbt$air_date), ]
# 创建一个新的列 diff，用于存储相邻日期收视率的差值
bbt$diff <- c(NA, diff(bbt$viewers))
# 统计差值的正负情况。如果正差值的数量多于负差值，可能说明收视率总体上在增长；反之则可能在下降。
diff_zheng <- sum(bbt$diff > 0, na.rm = TRUE)
diff_fu <- sum(bbt$diff < 0, na.rm = TRUE)
diff_zheng
```

```
#> [1] 7
```

```
diff_fu
```

```
#> [1] 12
```

```
# 将日期转换为数值格式，例如可以使用 as.numeric 函数将日期转换为自某个起始日期以来的天数。假设日期存储在
dates_num <- as.numeric(bbt$air_date)
# 使用线性回归模型来拟合收视率与日期数值之间的关系。在 R 中，可以使用 lm 函数进行线性回归，例如
model <- lm(viewers ~ dates_num, data = bbt)
model
```

```
#> 
#> Call:
#> lm(formula = viewers ~ dates_num, data = bbt)
#> 
#> Coefficients:
#> (Intercept)    dates_num  
#>  -47.503649     0.004081
```

6 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.

```
# 读取文件
nba<-read.csv2("NBAPlayerPts.csv",sep = ",")

# 将 ppg 转换为数值型，便于后续分析
nba$PPG <- as.numeric(nba$PPG)

# PPG 使用从 10 开始到 30 结束的类，增量为 2
# 创建数据框
nba <- data.frame(
  Rank = c(nba$Rank),
  Player = c(nba$Player),
  PPG = c(nba$PPG)
)
# 使用 cut 函数进行分类并添加为新列
nba$PPG_category <- cut(nba$PPG, breaks = seq(10, 30, by = 2))
```

```
# 查看处理后的数据
nba
```

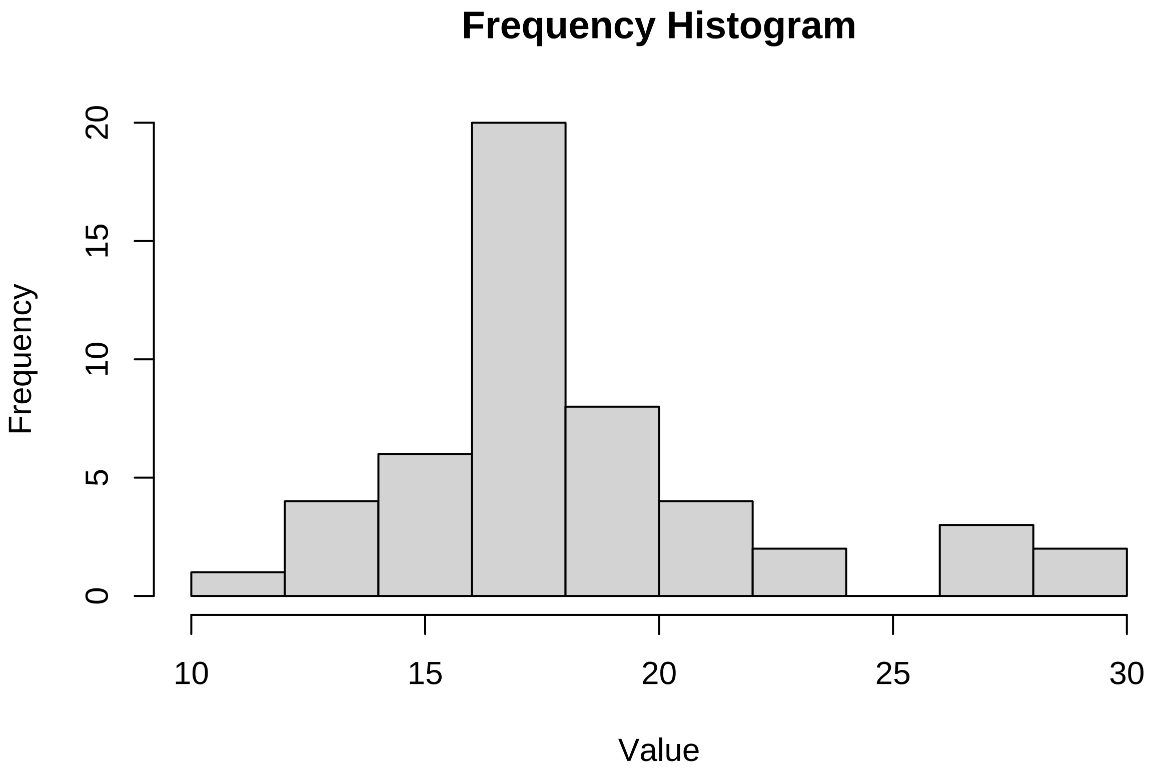
```
#>    Rank                       Player  PPG PPG_category
#> 1     1            LeBron James, MIA 27.0      (26,28]
#> 2     2            Kevin Durant, OKC 28.8      (28,30]
#> 3     3            James Harden, HOU 26.4      (26,28]
#> 4     4             Kobe Bryant, LAL 27.1      (26,28]
#> 5     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 21.0      (20,22]
#> 9     9       LaMarcus Aldridge, POR 20.8      (20,22]
#> 10   10             Paul George, IND 17.6      (16,18]
#> 11   11              Tony Parker, SA 21.1      (20,22]
#> 12   12            Jrue Holiday, PHI 19.2      (18,20]
#> 13   13             Dwyane Wade, MIA 21.2      (20,22]
#> 14   14           Nicolas Batum, POR 15.5      (14,16]
#> 15   15              Josh Smith, ATL 17.2      (16,18]
#> 16   16              Al Horford, ATL 16.7      (16,18]
#> 17   17            Al Jefferson, UTA 17.6      (16,18]
#> 18   18           Blake Griffin, LAC 18.5      (18,20]
#> 19   19             Paul Pierce, BOS 18.3      (18,20]
#> 20   20 Damian Lillard, POR (Rookie) 18.3      (18,20]
#> 21   21            Kyrie Irving, CLE 23.3      (22,24]
#> 22   22           Dwight Howard, LAL 16.4      (16,18]
#> 23   23        Brandon Jennings, MIL 18.9      (18,20]
#> 24   24               Luol Deng, CHI 16.5      (16,18]
#> 25   25          Deron Williams, BKN 17.0      (16,18]
#> 26   26             Joakim Noah, CHI 11.7      (10,12]
#> 27   27           Zach Randolph, MEM 15.7      (14,16]
#> 28   28                Rudy Gay, TOR 18.0      (16,18]
#> 29   29            Kemba Walker, CHA 17.7      (16,18]
#> 30   30        Chandler Parsons, HOU 14.6      (14,16]
#> 31   31             Greg Monroe, DET 15.7      (14,16]
#> 32   32              David West, IND 17.2      (16,18]
#> 33   33             Monta Ellis, MIL 18.2      (18,20]
#> 34   34               O.J. Mayo, DAL 17.5      (16,18]
#> 35   35              Marc Gasol, MEM 13.6      (12,14]
#> 36   36               Ty Lawson, DEN 16.3      (16,18]
#> 37   37              Chris Paul, LAC 16.2      (16,18]
#> 38   38          Greivis Vasquez, NO 13.6      (12,14]
#> 39   39              Chris Bosh, MIA 17.1      (16,18]
#> 40   40               Tim Duncan, SA 16.7      (16,18]
#> 41   41             Joe Johnson, BKN 17.0      (16,18]
#> 42   42        DeMarcus Cousins, SAC 17.3      (16,18]
#> 43   43           DeMar DeRozan, TOR 17.5      (16,18]
#> 44   44             Evan Turner, PHI 14.0      (12,14]
#> 45   45        Danilo Gallinari, DEN 16.9      (16,18]
#> 46   46            Klay Thompson, GS 16.3      (16,18]
#> 47   47            Paul Millsap, UTA 15.1      (14,16]
#> 48   48          Nikola Vucevic, ORL 12.3      (12,14]
#> 49   49             Brook Lopez, BKN 18.7      (18,20]
#> 50   50             George Hill, IND 14.6      (14,16]
```

7 Q2 a. Show the frequency distribution.

```r
table(nba$PPG_category)
```

```
#> 
#> (10,12] (12,14] (14,16] (16,18] (18,20] (20,22] (22,24] (24,26] (26,28] (28,30] 
#>       1       4       6      20       8       4       2       0       3       2
```

```r
# 直方图
nba_ppg <- c(nba$PPG)
#main = " 图名", xlab = "x 轴名", ylab = "y 轴名")
hist(nba_ppg, main = "Frequency Histogram", xlab = "Value", ylab = "Frequency")
```

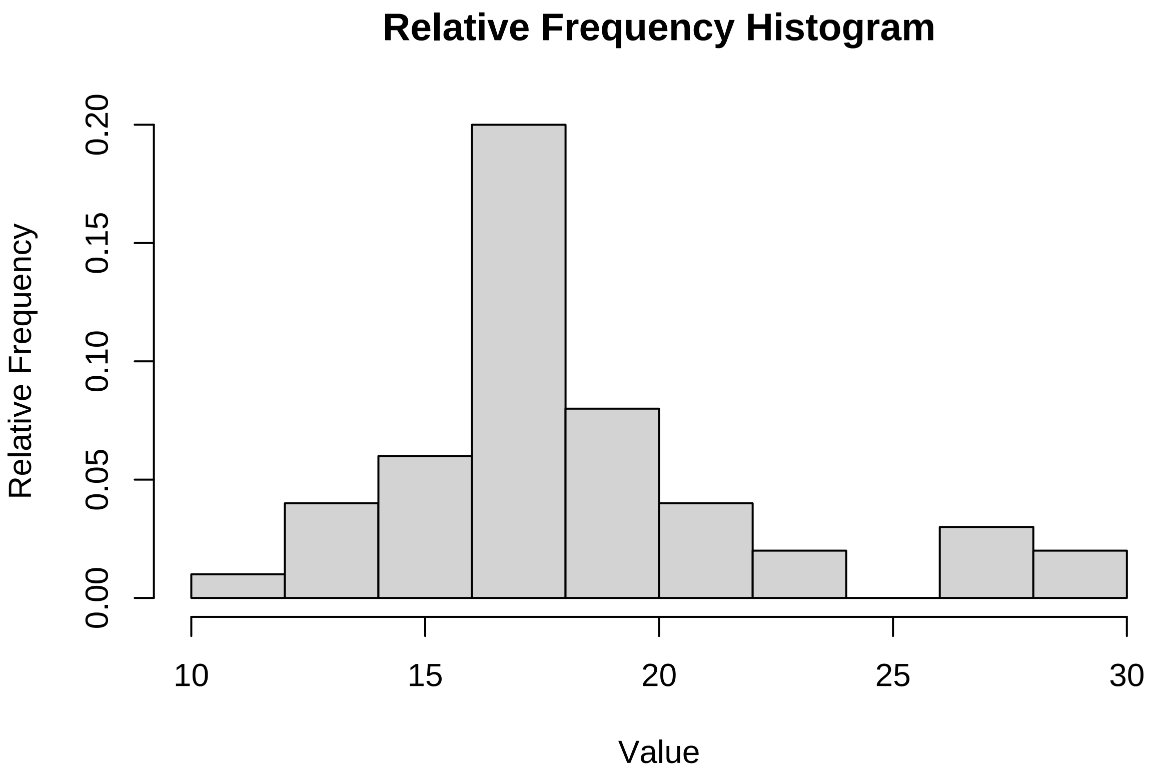


8 Q2 b. Show the relative frequency distribution.

```
freq_table <- table(nba$PPG_category)
# 用频率表中的数字/总长度就是相对频率
relative_freq <- freq_table / length(nba$PPG_category)
relative_freq
```

```
#>
#> (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
```

```
nba_ppg <- c(nba$PPG)
#freq = FALSE 代表相对分布
hist(nba_ppg, freq = FALSE, main = "Relative Frequency Histogram", xlab = "Value", ylab = "Relativ
```



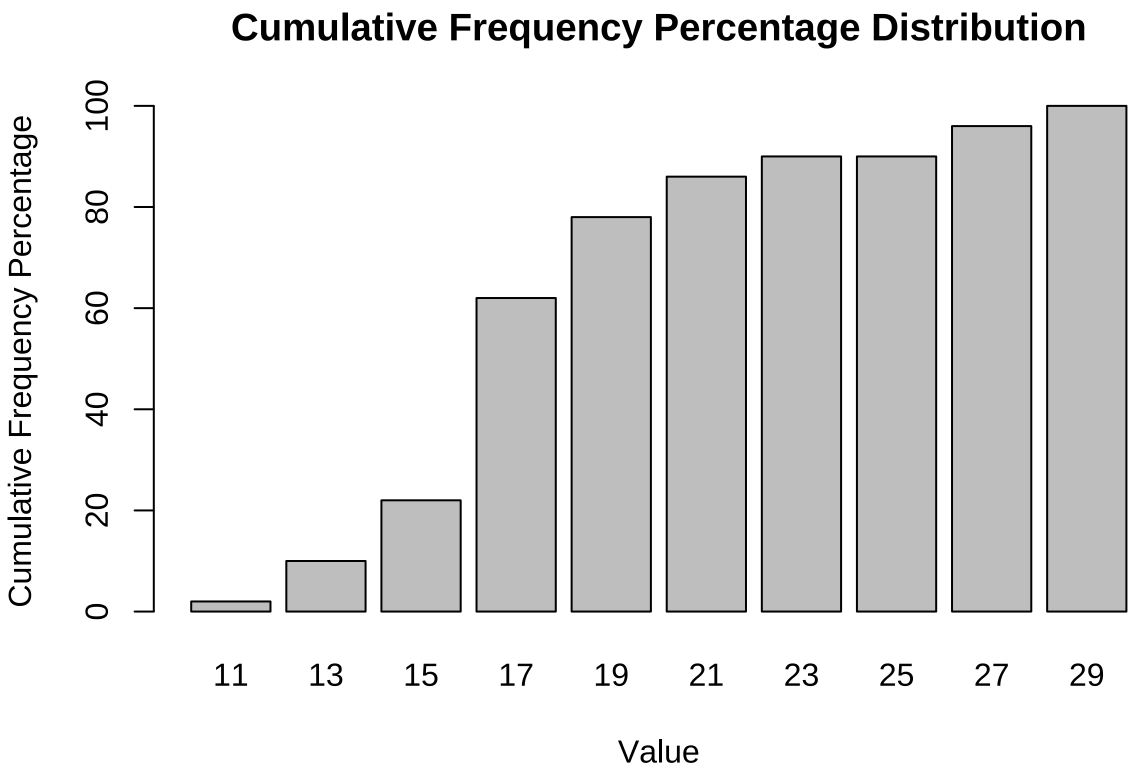
9 Q2 c. Show the cumulative percent frequency distribution.

```
# 先计算频率
freq_table <- table(nba$PPG_category)
# 再用频率表中的数字/总长度就是相对频率
relative_freq <- freq_table / length(nba$PPG_category)
# 再将相对频率相加获得累计频率
cumulative_freq <- cumsum(relative_freq)
cumulative_freq
```

```
#> (10,12] (12,14] (14,16] (16,18] (18,20] (20,22] (22,24] (24,26] (26,28] (28,30]
#>    0.02    0.10    0.22    0.62    0.78    0.86    0.90    0.90    0.96    1.00
```

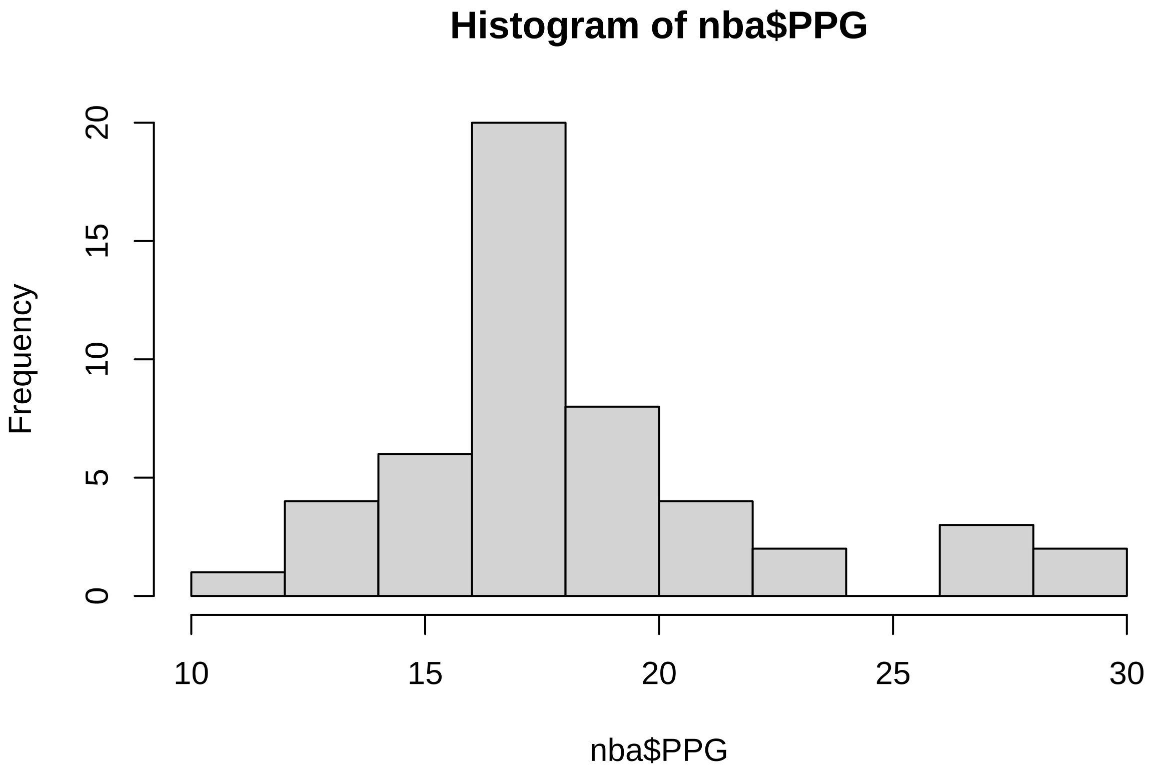
```
# 创建包含数据点的向量
nba_ppg <- c(nba$PPG)
# 使用 hist 函数创建一个直方图对象，但设置 plot = FALSE 以避免立即绘制直方图。这个直方图对象包含了关于数
```

```
nba_hist <- hist(nba_ppg,plot = FALSE)
# 使用 cumsum 函数计算直方图中每个区间频数的累积和。这将得到一个向量，表示小于或等于每个区间上限的数据点
cumulative_frequency <- cumsum(nba_hist$counts)
# 将累积频数转换为累积频率百分比。首先将累积频数除以数据向量的总频数（通过 sum(hist_data$counts) 计算得
cumulative_percentage <- cumulative_frequency / sum(nba_hist$counts) * 100
# 使用 barplot 函数绘制累积频率百分比的柱状图。cumulative_percentage 作为纵坐标，names.arg 设置横坐标
barplot(cumulative_percentage, names.arg = nba_hist$mids, main = "Cumulative Frequency Percentage
```



10 Q2 d. Develop a histogram for the average number of points scored per game.

```
hist(nba$PPG)
```



11 Q2 e. Do the data appear to be skewed? Explain.

偏度是描述数据分布不对称程度的统计量。如果偏度为 0，表示数据分布是对称的；如果偏度大于 0，表示数据分布有正偏态（右侧长尾）；如果偏度小于 0，表示数据分布有负偏态（左侧长尾）。

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

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

```
# 结果> 0，右偏
```

Q2 f. What percentage of the players averaged at least 20 points per game?

```
nba_hight <- sum(nba$PPG>20)/length(nba$PPG)
nba_hight
```

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

12 Question 3: A researcher reports survey results by stating that the standard error of the mean is 20. The population standard deviation is 500.

Q3 a. How large was the sample used in this survey?

```r
# 标准误差（Standard Error, SE）的计算公式为：S^2= ^2/n，其中是  总体标准差，n 是样本量。
#n= ^2/s^2
n <- 500^2/20^2
n
```

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

13 Q3 b. What is the probability that the point estimate was within ±25 of the population mean?

```r
#pnorm 函数是 R 中用于计算标准正态分布的累积分布函数值的函数
#20 是此题的标准误差
pnorm(25/20) - pnorm(-25/20)
```

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

14 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.

15 Q4 a. Develop appropriate descriptive statistics to summarize the data.

```r
pro <- read.csv2("Professional.csv",sep = ",")
column_names <- names(pro)
# 文件中有空值占据列，仅留有效数据列
pro <- pro[,1:8]
# 遍历每一列，并用 na.strings="NA" 指定了将字符 NA（注意这里是字符 NA，不是缺失值 NA 的概念）当作缺失值
for (col_name in column_names) {
  if (is.character(pro[[col_name]])) {
    pro[[col_name]] <- as.factor(pro[[col_name]])
  }
}
summary(pro)
```

```
#>       Age           Gender    Real.Estate.Purchases. Value.of.Investments....
#>  Min.   :19.00   Female:181   No :229                Min.   :     0
#>  1st Qu.:28.00   Male  :229   Yes:181                1st Qu.: 18300
#>  Median :30.00                                       Median : 24800
#>  Mean   :30.11                                       Mean   : 28538
#>  3rd Qu.:33.00                                       3rd Qu.: 34275
#>  Max.   :42.00                                       Max.   :133400
#>  Number.of.Transactions Broadband.Access. Household.Income.... Have.Children.
#>  Min.   : 0.000         No :154           Min.   : 16200       No :191
#>  1st Qu.: 4.000         Yes:256           1st Qu.: 51625       Yes:219
#>  Median : 6.000                           Median : 66050
#>  Mean   : 5.973                             Mean   : 74460
#>  3rd Qu.: 7.000                             3rd Qu.: 88775
#>  Max.   :21.000                             Max.   :322500
```

16 Q4 b. Develop 95% confidence intervals for the mean age and household income of subscribers.

```
# 方式一：公式计算
# 求均值
a <- mean(pro$Age)
# 下限样本均值 +Z（ /2） * 标准差/根号 n，qnorm(" 比例") 比例即面积只能是正
b <- a + qnorm(0.025)*sd(pro$Age)/sqrt(length(pro$Age))
# 上限样本均值-Z（ /2） * 标准差/根号 n
c <- a - qnorm(0.025)*sd(pro$Age)/sqrt(length(pro$Age))
age_an <- c(b,c)
age_an
```

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

```
# 方案二：函数计算
#[[4]] 表示取计算出来的第四项数据
t.test(pro$Household.Income....)[[4]]
```

```
#> [1] 71079.26 77839.77
#> attr(,"conf.level")
#> [1] 0.95
```

17 Q4 c. Develop 95% confidence intervals for the proportion of subscribers who have broadband access at home and the proportion of subscribers who have children.

```
#sum 可以对逻辑值计数求和，length 仅可对数值求和
br_p <- sum(pro$Broadband.Access.== "Yes")/length(pro$Broadband.Access.)
p1 <- br_p + qnorm(0.025)*sqrt(br_p*(1-br_p)/length(pro$Broadband.Access.))
```

```
p2 <- br_p - qnorm(0.025)*sqrt(br_p*(1-br_p)/length(pro$Broadband.Access.))
cp <- c(p1,p2)
cp
```

```
#> [1] 0.5775140 0.6712665
```

```
br_c <- sum(pro$Have.Children.== "Yes")/length(pro$Have.Children.)
c1 <- br_c + qnorm(0.025)*sqrt(br_c*(1-br_c)/length(pro$Have.Children.))
c2 <- br_c - qnorm(0.025)*sqrt(br_c*(1-br_c)/length(pro$Have.Children.))
cc <- c(c1,c2)
cc
```

```
#> [1] 0.4858615 0.5824312
```

18 Q4 d. Would Young Professional be a good advertising outlet for online brokers? Justify your conclusion with statistical data.

回答：适合。原因：样本中 62.4% 的人有网，具备网络交易基本物质条件。样本中 62.7% 的家庭投资金额占家庭收入 30% 以上，样本中一半的人一年交易在 6 次及以上，具备投资交易习惯，属于网络交易高潜用户。

```
# 有网络人数占比
br_p
```

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

```
# 投资交易数据
mean(pro$Number.of.Transactions)
```

```
#> [1] 5.973171
```

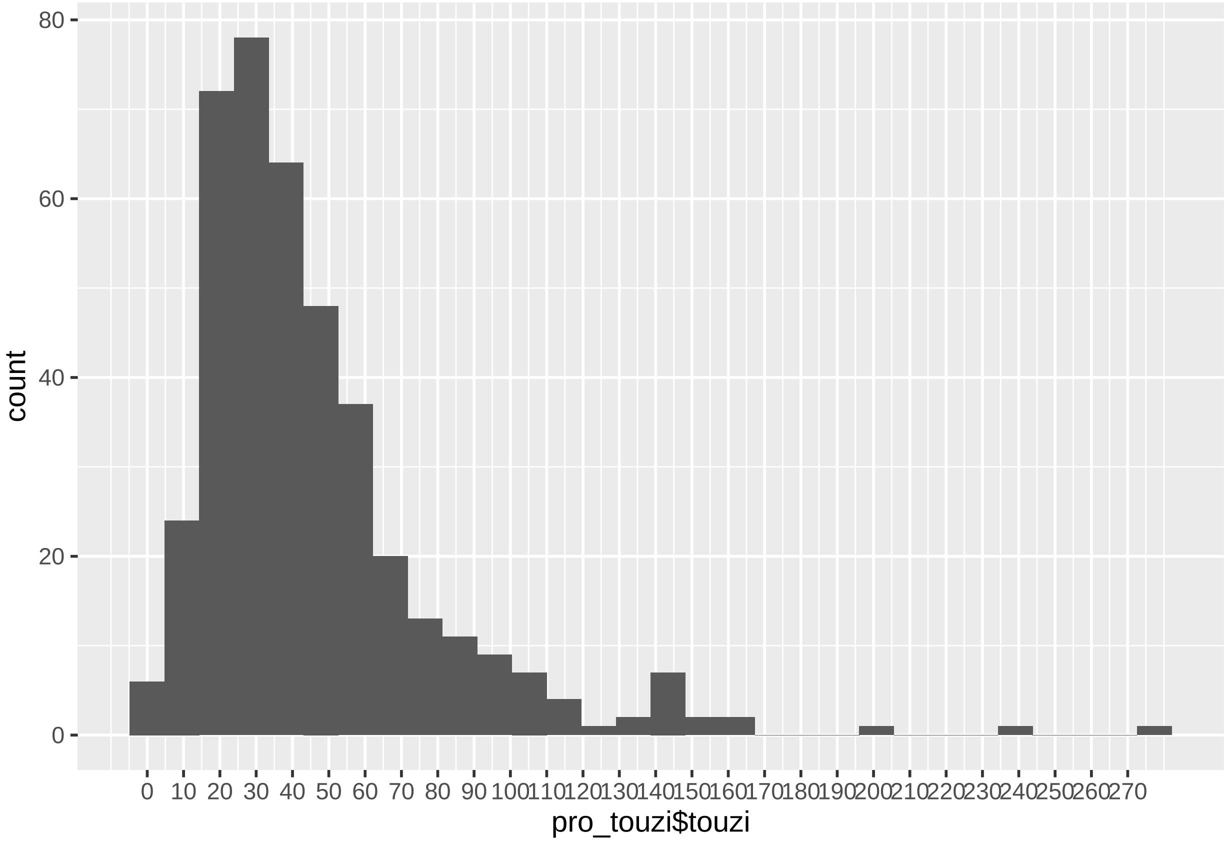
```
median(pro$Number.of.Transactions)
```

```
#> [1] 6
```

```
# 金融投资占收入的占比数据
touzi <- (pro$Value.of.Investments..../pro$Household.Income....)*100
touzi_p <- sum(touzi>30)/length(touzi)
touzi_p
```

```
#> [1] 0.6268293
```

```
pro_touzi <- cbind(pro,touzi=touzi)
ggplot(pro_touzi,aes(x=pro_touzi$touzi))+
  geom_histogram()+
# 调节 x 轴间隔
   scale_x_continuous(breaks = seq(min(pro_touzi$touzi), max(pro_touzi$touzi), by = 10))
```



19 Q4 e. Would this magazine be a good place to advertise for companies selling educational software and computer games for young children?

回答：适合。原因：样本中 53.4% 的家庭是有小孩的，且样本人员年龄平均在 30，半数在 30 岁以下，即大部分是有小孩且父母年轻的家庭，是小孩教育与游戏的受众人群。

```
# 有孩家庭占比
chi_pr <- sum(pro$Have.Children.=="Yes")/length(pro$Have.Children.)
chi_pr
```

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

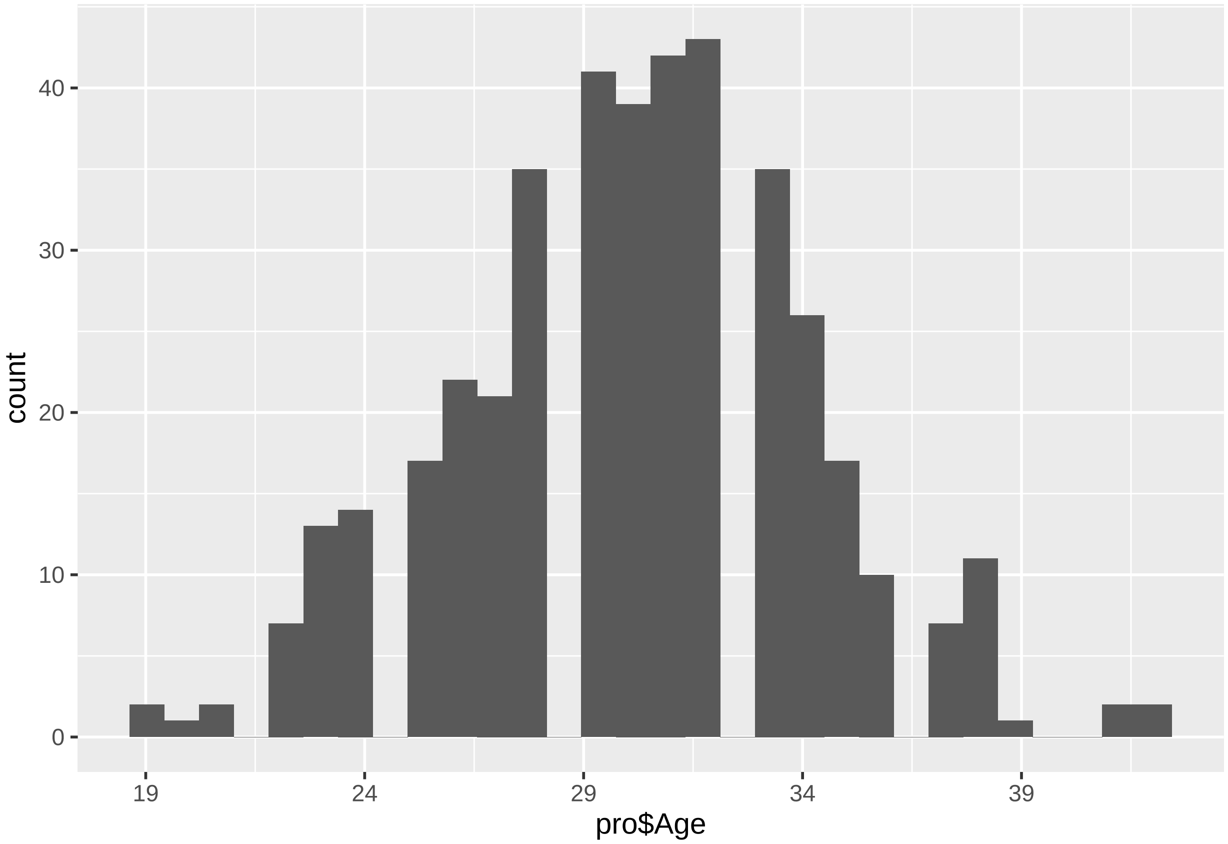
```
# 样本人员年龄分布
mean(pro$Age)
```

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

```
median(pro$Age)
```

```
#> [1] 30
```

```
ggplot(pro,aes(x=pro$Age))+
  geom_histogram()+
   scale_x_continuous(breaks = seq(min(pro$Age), max(pro$Age), by = 5))
```



20 Q4 f. Comment on the types of articles you believe would be of interest to readers of Young Professional.

回答：根据数据样本摘要发现样本具有这样几个特征：1、采样人群年龄偏年轻，集中在 30 岁左右；2、62% 的人有互联网；3、53% 的人家里有小孩；4、半数人投资金额占家庭收入 36% 及以上，且半数人一年内投资交易次数在 6 次及以上；结合 1、2 条猜测一些新奇、猎奇、科技相关的客户尝试推荐；结合 1、2、3 条推测家庭教育相关他们会比较感兴趣；结合 1、2、4 条数据推测投资理财相关可以尝试。

```
summary(pro_touzi)
```

```
#>       Age            Gender    Real.Estate.Purchases. Value.of.Investments....
#>  Min.   :19.00   Female:181   No :229                Min.   :     0
#>  1st Qu.:28.00   Male  :229   Yes:181                1st Qu.: 18300
#>  Median :30.00                                       Median : 24800
#>  Mean   :30.11                                       Mean   : 28538
#>  3rd Qu.:33.00                                       3rd Qu.: 34275
#>  Max.   :42.00                                       Max.   :133400
#>  Number.of.Transactions Broadband.Access. Household.Income.... Have.Children.
#>  Min.   : 0.000         No :154           Min.   : 16200       No :191       
#>  1st Qu.: 4.000         Yes:256           1st Qu.: 51625       Yes:219       
#>  Median : 6.000                           Median : 66050                     
#>  Mean   : 5.973                           Mean   : 74460                     
#>  3rd Qu.: 7.000                           3rd Qu.: 88775                     
#>  Max.   :21.000                           Max.   :322500                     
#>      touzi       
#>  Min.   :  0.00  
#>  1st Qu.: 23.96  
#>  Median : 36.18  
#>  Mean   : 45.62  
#>  3rd Qu.: 55.83  
#>  Max.   :277.34  
```

21 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. H0: =12 H1: 12 Corrective action will be taken any time H0 is rejected. Data are available in the data set Quality.

读取文件

```
qua <- read.csv("Quality.csv")
```

22 Q5 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.

sample.3 拒绝 H0，其他几个都是接受 H0

```
# 双边检测所以 *2
p1 <- 2*pnorm((mean(qua$Sample.1)-12)/(0.21/sqrt(30)))
p2 <- 2*(1-pnorm((mean(qua$Sample.2)-12)/(0.21/sqrt(30))))
p3 <- 2*pnorm((mean(qua$Sample.3)-12)/(0.21/sqrt(30)))
p4 <- 2*(1-pnorm((mean(qua$Sample.4)-12)/(0.21/sqrt(30))))
p1
```

```
#> [1] 0.2810083
```

```
p2
```

```
#> [1] 0.4546503
```

```
p3
```

```
#> [1] 0.003790318
```

```
p4
```

```
#> [1] 0.03389336
```

23 Q5 b. compute the standard deviation for each of the four samples. does the assumption of .21 for the population standard deviation appear reasonable?

回答：我觉得 0.21 的标准差不是很合理，因为按 0.21 的标准差算出来的置信区间是 11.91081-12.08919，但是通过直方图简单对比所有抽样的结果发现，高重合区域在 11.75-12.25，比以 sd=0.21 算出来的区间要大。建议可以增大到 0.59

```
# 赋值向量
vactors <- list(sd_1=qua$Sample.1,
                sd_2=qua$Sample.2,
                sd_3=qua$Sample.3,
                sd_4=qua$Sample.4)
# 对每个向量进行标准差计算，sapply（计算向量名称，计算的步骤）
sds <- sapply(vactors,sd)
sds
```

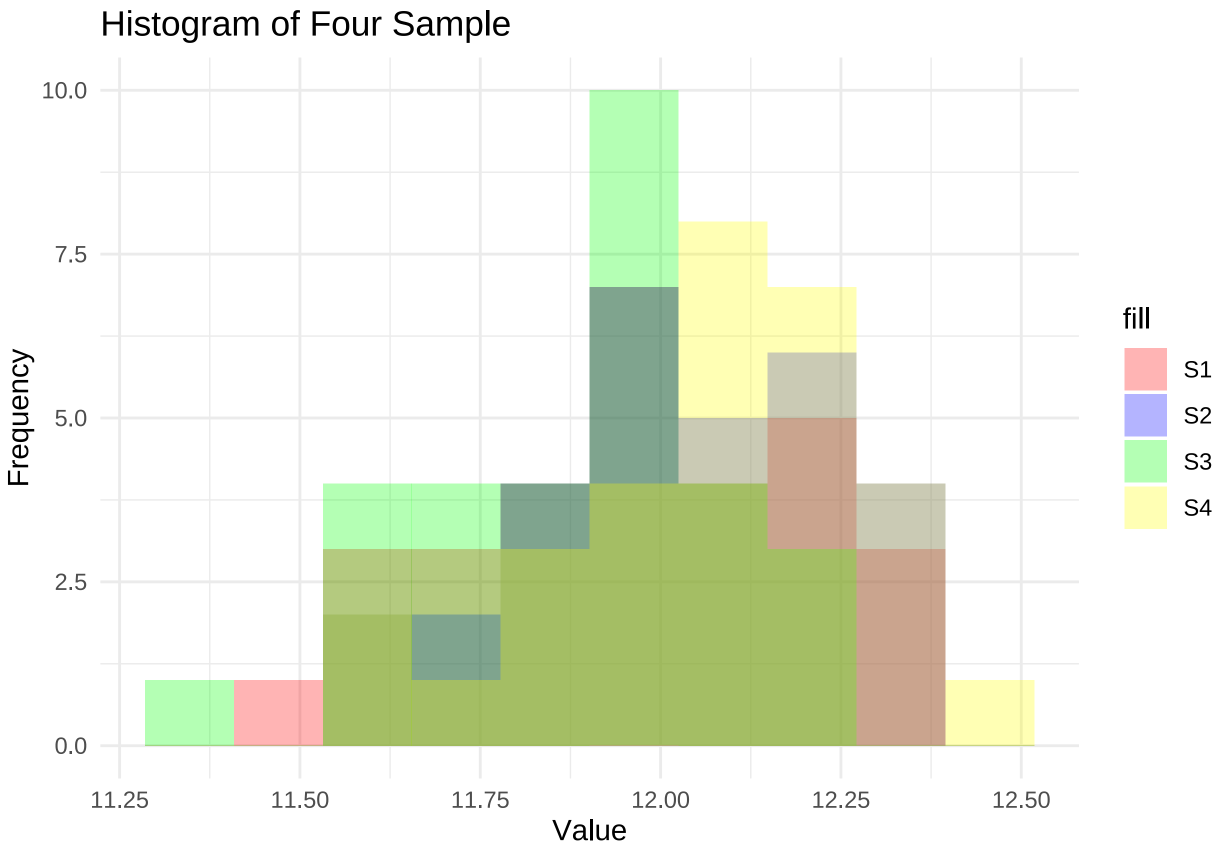
```
#>      sd_1      sd_2      sd_3      sd_4
#> 0.2203560 0.2203560 0.2071706 0.2061090
```

```
#0.01 时的置信区间
c1 <-12+qnorm(0.01)*(0.21/sqrt(30))
c2 <-12-qnorm(0.01)*(0.21/sqrt(30))
z <-c(c1,c2)
z
```

```
#> [1] 11.91081 12.08919
```

```
df <- data.frame(
  s1 = qua$Sample.1,
  s2 = qua$Sample.2,
  s3 = qua$Sample.3,
  s4 = qua$Sample.4
)

ggplot(df, aes(x = s1, fill = "S1")) +
  geom_histogram(bins = 10, alpha = 0.3) +  # 绘制第一组数据的直方图
  geom_histogram(aes(x = s2, fill = "S2"), bins = 10, alpha = 0.3) +  # 第二组数据
  geom_histogram(aes(x = s3, fill = "S3"), bins = 10, alpha = 0.3) +  # 第三组数据
  geom_histogram(aes(x = s4, fill = "S4"), bins = 10, alpha = 0.3) +  # 第四组数据
  scale_fill_manual(values = c("S1" = "red", "S2" = "blue", "S3" = "green", "S4" = "yellow")) +  #
  labs(title = "Histogram of Four Sample", x = "Value", y = "Frequency") +
  theme_minimal()  # 使用简洁的主题
```



```
sd_suggest <- 0.25*sqrt(30)/qnorm(0.01)
sd_suggest
```

```
#> [1] -0.5886078
```

```
c3 <-12+qnorm(0.01)*(0.59/sqrt(30))
c4 <-12-qnorm(0.01)*(0.59/sqrt(30))
z2 <-c(c3,c4)
z2
```

```
#> [1] 11.74941 12.25059
```

24 Q5 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.

```
#0.01 时的置信区间
c1 <-12+qnorm(0.01)*(0.21/sqrt(30))
c2 <-12-qnorm(0.01)*(0.21/sqrt(30))
z <-c(c1,c2)
z
```

```
#> [1] 11.91081 12.08919
```

25 Q5 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?

回答：其他变量不变的情况下，置信水平增加，置信区间变窄，那么第一类错误 错误增加

```
#0.01 时的置信区间
c1 <-12+qnorm(0.01)*(0.21/sqrt(30))
c2 <-12-qnorm(0.01)*(0.21/sqrt(30))
z <-c(c1,c2)
z
```

```
#> [1] 11.91081 12.08919
```

```
#0.05 时的置信区间
c5 <-12+qnorm(0.1)*(0.21/sqrt(30))
c6 <-12-qnorm(0.1)*(0.21/sqrt(30))
z3 <-c(c5,c6)
z3
```

```
#> [1] 11.95086 12.04914
```

26 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.

准备工作读取数据

```
# 读取 CSV 文件,skip=1 表示跳过第一行
occ<- read.csv("Occupancy.csv", skip = 1, header = TRUE, stringsAsFactors = TRUE)
```

27 Q6 a. Estimate the proportion of units rented during the first week of March 2007 and the first week of March 2008.

```
summary(occ)
```

```
#>  March.2007 March.2008
#>  No :130        :50
#>  Yes: 70    No :80
#>             Yes:70
```

```
p_7 <- 70/(70+130)
p_7
```

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

```
p_8 <- 70/(70+80)
p_8
```

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

28 Q6 b. Provide a 95% confidence interval for the difference in proportions.

双样本 t 检验

```
bw <- qnorm(0.025)*sqrt(p_7*(1-p_7)/(130+70)+p_8*(1-p_8)/(80+70))
bw
```

```
#> [1] -0.1036515
```

```
qujian <- c(p_7-p_8+bw,p_7-p_8-bw)
qujian
```

```
#> [1] -0.22031818 -0.01301516
```

29 Q6 c. On the basis of your findings, does it appear March rental rates for 2008 will be up

from those a year earlier? 回答：租金应该不会上涨，一方面是看似入住率增长了，但是入住率增长的原因是因为房屋少了，分母少了，其实分子式没变的，都是 70 个入住；令一方面，假设 07 和 08 年入住率在 0.05 水平下有显著差异，计算对应 p 值为 0 对 07 年和 08 年的入住率做对比发现 p 值 =0.0137 并不显著

```
occ$March.2007 <- ifelse(occ$March.2007 == "Yes", 1, 0)
occ$March.2008<- ifelse(occ$March.2008 == "Yes", 1, 0)
head(occ)
```

```
#>   March.2007 March.2008
#> 1          1          0
#> 2          0          1
#> 3          1          1
#> 4          0          0
#> 5          0          1
#> 6          1          0
```

```
a <- (p_7-p_8)/sqrt((p_7*(1-p_7)/200)+(p_8*(1-p_8)/150))
pnorm(a)
```

```
#> [1] 0.01368956
```

30 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). 准备工作，导入数据

```
tra <- read.csv("Training.csv")
```

31 Q7 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?

回答：相似之处：两者均值相似、p50 相等，即两者分布对称轴相近；不同之处：Proposed 的 sd 小于 Current 的，即 Current 的相对差异性较大，Proposed 的相对分布较为平均。Proposed 的 p0、p25 高于 Current 的，但 p75、p100 略低于 Current 的。

```
skimr::skim(tra)
```

表 1: Data summary

Name	tra
Number of rows	61
Number of columns	2
Column type frequency:	
numeric	2
Group variables	None

Variable type: numeric

skim_variable	n_missing	complete_rate	mean	sd	p0	p25	p50	p75	p100	hist
Current	0	1	75.07	3.94	65	72	76	78	84	
Proposed	0	1	75.43	2.51	69	74	76	77	82	

skim() 函数是 R 语言中 skimr 包提供的一个用于数据摘要和探索性数据分析（EDA）的函数。它的目的是快速查看数据集的结构和内容，提供数据的概览，包括数据类型、缺失值、唯一的值、摘要统计量（如均值、中位数、标准差等）以及分位数等。skim(data：指定要分析的数据集,n_max：设置显示的最大行数,n_min：设置显示的最小行数,n：覆盖 n_max 和 n_min 的值,max_chars：设置显示的最大字符数) n_missing: 每个变量的缺失值数量。complete_rate: 每个变量的完整率，即非缺失值的比例。完整率 =1，表示没有缺失值。

32 Q7 b. Comment on any difference between the population means for the two methods. Discuss your findings.

回答：p 值 =0.548，不显著，两者均值无显著差异。

```
t.test(tra$Current,tra$Proposed,var.equal = TRUE)
```

```
#>
#>  Two Sample t-test
#>
```

```
#> data:  tra$Current and tra$Proposed
#> t = -0.60268, df = 120, p-value = 0.5479
#> alternative hypothesis: true difference in means is not equal to 0
#> 95 percent confidence interval:
#>  -1.5454793  0.8241679
#> sample estimates:
#> mean of x mean of y
#>  75.06557  75.42623
```

33 Q7 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.

回答：F test检验方差齐性，P值=0.00058，显著，Proposed与Current方差差异较大。

```r
list <- list(tra$Current,tra$Proposed)
sapply(list, sd)
```

```
#> [1] 3.944907 2.506385
```

```r
sapply(list, var)
```

```
#> [1] 15.562295  6.281967
```

```r
var.test(tra$Current,tra$Proposed)
```

```
#> 
#>  F test to compare two variances
#> 
#> data:  tra$Current and tra$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 函数用于执行方差齐性检验，F-test 检验方差齐性。sapply(列表, 计算方式) 对每列执行对应计算

34 Q7 d. what conclusion can you reach about any differences between the two methods? what is your recommendation? explain.

回答：虽然两者均值相等，但是 Proposed 的标准差和方差都小于 Current 的，即 Current 的相对差异性较大，Proposed 的相对分布较为平均，且 Proposed 的 p0、p25 高于 Current 的，说明 Proposed 的方法对学生整体节省时间是有帮助的，建议采纳 Proposed 的方法

35 Q7 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?

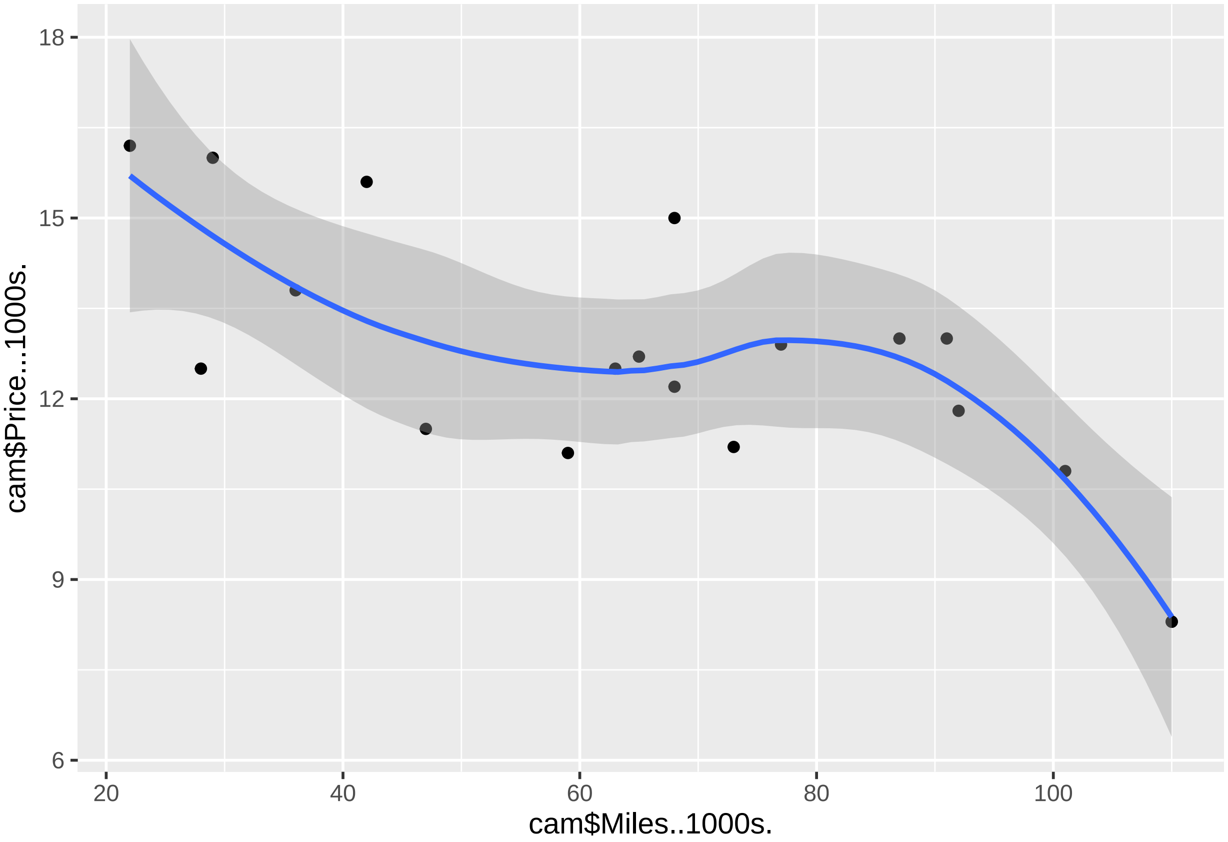
回答：目前的数据对比只能看到学生学习事件上的变化，但是学习质量上的变化无从判断。最终方案确定前，建议采取也做一些学习质量方面的检测（比如考试、完成任务等等），在时间和质量上双平衡后再确定最终方案。

36 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).

```r
cam <- read.csv("Camry.csv")
```

37 Q8 a. Develop a scatter diagram with the car mileage on the horizontal axis and the price on the vertical axis.

```r
library(ggplot2)
ggplot(cam, aes(x = cam$Miles..1000s.,y= cam$Price...1000s.)) +
  geom_point() +
  geom_smooth()
```



38 Q8 b. what does the scatter diagram developed in part (a) indicate about the relationship between the two variables?

回答：根据散点及曲线图发现，随着 mile 的增加，price 下降。

39 Q8 c. Develop the estimated regression equation that could be used to predict the price ($1000s) given the miles (1000s).

lm(y ~ x1 + x2 + … + xn, data = your_data)

```
cam_lm <- lm(cam$Price...1000s.~ cam$Miles..1000s., data = cam)
cam_lm
```

```
#> 
#> Call:
#> lm(formula = cam$Price...1000s. ~ cam$Miles..1000s., data = cam)
#> 
#> Coefficients:
#>       (Intercept)  cam$Miles..1000s.  
#>          16.46976           -0.05877  
```

```
summary(cam_lm)
```

```
#> 
#> Call:
#> lm(formula = cam$Price...1000s. ~ cam$Miles..1000s., data = cam)
#> 
#> 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 ***
#> cam$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
```

残差（Residuals）：实际观测值与模型预测值之间的差异 Min: -2.32408（最小残差）1Q: -1.34194（第一四分位数）Median: 0.05055（中位数）3Q: 1.12898（第三四分位数）Max: 2.52687（最大残差）

回归系数（Coefficients）截距（Intercept）: 16.46976 当行驶里程为 0 时，汽车价格的预期值。斜率（Slope）: -0.05877 行驶里程每增加 1 千英里，汽车价格预计减少 0.05877 千美元。

标准误差（Std. Error）: 截距的标准误差为 0.94876。斜率的标准误差为 0.01319。

t 值（t value）: 截距的 t 值为 17.359。斜率的 t 值为-4.455。

p 值（Pr(>|t|)）: 截距的 p 值为 2.99e-12。斜率的 p 值为 0.000348。这些 p 值远小于 0.05，表明截距和斜率在统计上显著不为 0。

显著性代码（Signif. codes）*** 表示 p 值小于 0.001，** 表示 p 值小于 0.01，*表示 p 值小于 0.05。在这个模型中，截距和斜率的 p 值都小于 0.001，标记为* **。

模型拟合优度（Model Fit）残差标准误差（Residual standard error）: 1.541 表示模型预测值与实际观测值之间的平均差异。R 平方（Multiple R-squared）: 0.5387 表示模型解释了 53.87% 的因变量变异。调整后的 R 平方（Adjusted R-squared）: 0.5115 调整后的 R 平方考虑了模型中变量的数量，提供了一个更为严格的模型拟合度量。F 统计量（F-statistic）: 19.85 用于检验模型中至少有一个系数显著不为 0 的统计量。F 统计量的 p 值: 0.0003475 远小于 0.05，表明模型整体上是统计显著的。

结论：这个线性回归模型表明，汽车的行驶里程和价格之间存在显著的负相关关系。模型的拟合度适中，但需要注意的是，R 平方值相对较低，意味着还有相当一部分变异没有被模型解释。此外，模型的解释变量可能需要进一步扩展，以提高模型的预测能力。

40 Q8 d. Test for a significant relationship at the .05 level of significance.

回答：截距的 p 值为 2.99e-12，斜率的 p 值为 0.000348，F 统计量的 p 值 0.0003475，这些 p 值远小于 0.05，表明截距和斜率在统计上显著。

41 Q8 e. Did the estimated regression equation provide a good fit？ Explain.

回答：残差标准误差 1.541，R 方 0.5115 模型的拟合度适中，R 平方值 0.5387 意味着还有相当一部分变异没有被模型解释。

42 Q8 f. Provide an interpretation for the slope of the estimated regression equation.

回答：-0.05877 行驶里程每增加 1 千英里，汽车价格预计减少 0.05877 千美元。

43 Q8 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.

```
# 由 c 测算可得方程 Price=16.46976-0.05877*Miles
#Miles=60000, 单位换算后 =60
Price <- 16.46976-0.05877*60
Price
```

```
#> [1] 12.94356
```

# 问题 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。

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

44 Q9 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？

回答：对比流失客户与不流失客户各项行为指标均值发现似乎应该” 流失 “指标中 0 表示不流失，”1“表示流失。因为流失为 0 的客户例如登录次数、博客次数、访问次数等行为指标时高于流失为 1 的客户的，同时流失为 0 的客户登录间隔也更短，表明这部分客户又频繁登录的习惯，而且流失为 0 的客户幸福指数也是更高的。

```
# 按是否流失分类求均值
mean_values <- we %>%
  group_by(流失) %>%
  summarise(across(everything(), mean, na.rm = TRUE))
mean_values
```

```
#> # A tibble: 2 x 13
#>    流失 客户ID 当月客户幸福指数 客户幸福指数相比上月变化 当月客户支持
#>    <dbl>  <dbl>            <dbl>                        <dbl>        <dbl>
#> 1      0  3219.             88.6                         5.53        0.724
#> 2      1  2330.             63.3                        -3.74        0.372
#> # i 8 more variables: 客户支持相比上月的变化 <dbl>, 当月服务优先级 <dbl>,
#> #   服务优先级相比上月的变化 <dbl>, 当月登录次数 <dbl>,
#> #   博客数相比上月的变化 <dbl>, 访问次数相比上月的增加 <dbl>,
#> #   客户使用期限 <dbl>, 访问间隔变化 <dbl>
```

45 Q9 b. 通过均值比较的方式验证上述不同是否显著。

回答：当月客户幸福指数、当月客户支持、当月服务优先级、当月登录次数均显著不同

```
t_test_happiness <- t.test(当月客户幸福指数 ~ 流失, data = we)
t_test_support <- t.test(当月客户支持 ~ 流失, data = we)
t_test_serve <- t.test(当月服务优先级 ~ 流失, data = we)
t_test_num <- t.test(当月登录次数 ~ 流失, data = we)
t_test_happiness
```

```
#> 
#>  Welch Two Sample t-test
#> 
#> data:  当月客户幸福指数 by 流失
#> t = 7.6242, df = 369.36, p-value = 2.097e-13
#> alternative hypothesis: true difference in means between group 0 and group 1 is not equal to 0
#> 95 percent confidence interval:
#>  18.79956 31.86737
#> sample estimates:
#> mean in group 0 mean in group 1 
#>        88.60591        63.27245
```

```
t_test_support
```

```
#> 
#>  Welch Two Sample t-test
#> 
#> data:  当月客户支持 by 流失
#> t = 5.5099, df = 419.22, p-value = 6.281e-08
#> alternative hypothesis: true difference in means between group 0 and group 1 is not equal to 0
#> 95 percent confidence interval:
#>  0.2269082 0.4785969
#> sample estimates:
#> mean in group 0 mean in group 1
#>       0.7242696       0.3715170
```

```
t_test_serve
```

```
#> 
#>  Welch Two Sample t-test
#> 
#> data:  当月服务优先级 by 流失
#> t = 5.1428, df = 373.13, p-value = 4.381e-07
#> alternative hypothesis: true difference in means between group 0 and group 1 is not equal to 0
#> 95 percent confidence interval:
#>  0.2038355 0.4562009
#> sample estimates:
#> mean in group 0 mean in group 1
#>       0.8295759       0.4995577
```

```
t_test_num
```

```
#> 
#>  Welch Two Sample t-test
#> 
#> data:  当月登录次数 by 流失
#> t = 3.5709, df = 362.67, p-value = 0.0004037
#> alternative hypothesis: true difference in means between group 0 and group 1 is not equal to 0
#> 95 percent confidence interval:
#>   3.628884 12.525166
#> sample estimates:
#> mean in group 0 mean in group 1
#>        16.13894         8.06192
```

46 Q9 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。

46 *Q9 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

```r
we_lm <- lm(流失~ 当月客户幸福指数+当月客户支持+当月服务优先级, data = we)
we_lm
```

```
#> 
#> Call:
#> lm(formula = 流失 ~ 当月客户幸福指数 + 当月客户支持 + 
#>     当月服务优先级, data = we)
#> 
#> Coefficients:
#>      (Intercept)  当月客户幸福指数      当月客户支持    当月服务优先级  
#>        0.0757987        -0.0002428        -0.0009905        -0.0037046  
```

```r
summary(we_lm )
```

```
#> 
#> Call:
#> lm(formula = 流失 ~ 当月客户幸福指数 + 当月客户支持 + 
#>     当月服务优先级, data = we)
#> 
#> Residuals:
#>      Min       1Q   Median       3Q      Max 
#> -0.07580 -0.06706 -0.04958 -0.03307  0.99100 
#> 
#> Coefficients:
#>                    Estimate Std. Error t value Pr(>|t|)    
#> (Intercept)       7.580e-02  4.568e-03  16.594  < 2e-16 ***
#> 当月客户幸福指数 -2.428e-04  4.496e-05  -5.400 6.91e-08 ***
#> 当月客户支持     -9.905e-04  2.116e-03  -0.468    0.640    
#> 当月服务优先级   -3.705e-03  2.836e-03  -1.306    0.191    
#> ---
#> Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
#> 
#> Residual standard error: 0.219 on 6343 degrees of freedom
#> Multiple R-squared:  0.007722,   Adjusted R-squared:  0.007253 
#> F-statistic: 16.45 on 3 and 6343 DF,  p-value: 1.197e-10
```

46 *Q9 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

```r
# 由于“流失”是二元变量（0 或 1），逻辑回归（logistic regression）会更准确地建模这种二分类问题。
# 建立逻辑回归模型，以流失为因变量，当月客户幸福指数和当月客户支持为自变量
we_glm  <- glm(流失 ~ 当月客户幸福指数 + 当月客户支持+当月服务优先级, data = we, family = binomial)
# 打印逻辑回归模型的摘要
summary(we_glm )
```

```
#> 
#> Call:
#> glm(formula = 流失 ~ 当月客户幸福指数 + 当月客户支持 + 
#>     当月服务优先级, family = binomial, data = we)
#> 
#> Coefficients:
#>                    Estimate Std. Error z value Pr(>|z|)    
#> (Intercept)      -2.4491241  0.0834370 -29.353  < 2e-16 ***
#> 当月客户幸福指数 -0.0052554  0.0009973  -5.270 1.37e-07 ***
#> 当月客户支持     -0.0587277  0.0721328  -0.814    0.416    
#> 当月服务优先级   -0.0747291  0.0747327  -1.000    0.317    
#> ---
#> 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: 2500.9  on 6343  degrees of freedom
#> AIC: 2508.9
#> 
#> Number of Fisher Scoring iterations: 6
```

系数（Coefficients）: (Intercept): 截距项的估计值为-2.4564321，标准误差为 0.0831971，z 值为-29.525，p 值小于 2e-16，表示截距项在统计上非常显著。当月客户幸福指数: 系数为-0.0054455，标准误差为 0.0009824，z 值为-5.543，p 值为 2.98e-08，表示这个变量在统计上非常显著，且对数几率的变化与当月客户幸福指数的增加呈负相关，即当月客户幸福指数每增加一个单位，流失的概率降低。当月客户支持: 系数为-0.1116664，标准误差为 0.0569560，z 值为-1.961，p 值为 0.0499，表示这个变量在统计上显著，且对数几率的变化与当月客户支持的增加呈负相关，即当月客户支持每增加一个单位，流失的概率降低。

模型拟合度（Model Fit）: Null deviance: 空模型（只有截距项）的偏差为 2553.1，自由度为 6346。Residual deviance: 拟合模型后的残差偏差为 2501.9，自由度为 6344。残差偏差的减少表明模型对数据的拟合有所改善。AIC: 赤池信息准则（Akaike Information Criterion）为 2507.9，用于模型选择，值越小表示模型越好。

47 Q9 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。

```r
# 首先，使用 glm() 函数构建逻辑回归模型，预测流失。使用 predict() 函数计算每个客户的流失概率，并将结果存
we$流失概率 <- predict(glm(流失 ~ 当月客户幸福指数 + 当月客户支持 + 当月服务优先级,
                             family = binomial, data = we),
                        type = "response")

# 然后，使用 order() 函数对流失概率进行降序排序，并选择前 100 名客户的 ID 和流失概率。
top_100_customers <- we[order(-we$流失概率), c("客户ID", "流失概率")][1:100, ]

# 输出前 100 名用户 ID 及其流失概率
print(top_100_customers)
```

```
#> # A tibble: 100 x 2
#>    客户ID 流失概率
#>     <dbl>    <dbl>
#>  1      1   0.0795
#>  2      3   0.0795
#>  3     14   0.0795
#>  4     18   0.0795
#>  5     21   0.0795
#>  6     57   0.0795
#>  7     59   0.0795
#>  8     60   0.0795
#>  9     94   0.0795
#> 10    101   0.0795
#> # i 90 more rows
```