第二次作业

赵映辉

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).

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

```
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
```

```
library(readxl)
library(janitor)
```

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

```
##
## 载入程序包: 'janitor'
```

```
##
## The following objects are masked from 'package:stats':
##
##     chisq.test, fisher.test
```

```r
table1 = read.csv("./data/BigBangTheory.csv")

table1 = clean_names(table1)

# 获取最大值
max(table1$viewers_millions)
```

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

```r
# 获取最小值
min(table1$viewers_millions)
```

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

b. Compute the mean, median, and mode.

```r
# 平均数
mean(table1$viewers_millions)
```

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

```r
# 中位数
median(table1$viewers_millions)
```

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

```r
# 众数
result = table(table1$viewers_millions, useNA = "no")

as.numeric(names(result[result == max(result)]))
```

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

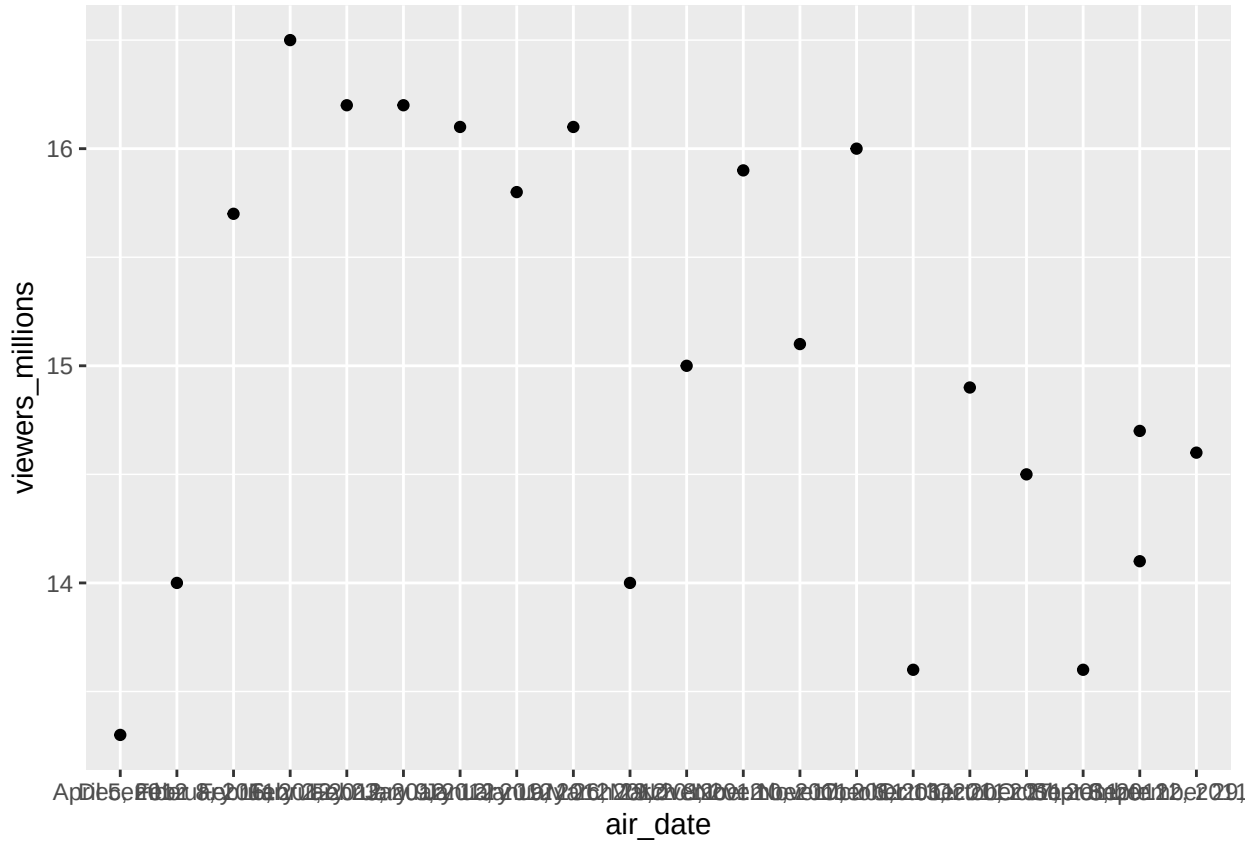
c. Compute the first and third quartiles

```r
# 四分
quantile(table1$viewers_millions, probs = c(0.25, 0.75))
```

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

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

```r
ggplot(table1, aes(x = air_date, y = viewers_millions)) + geom_point()
```



第二季开始的时候很好看，后续应该是质量下滑导致收视率下降

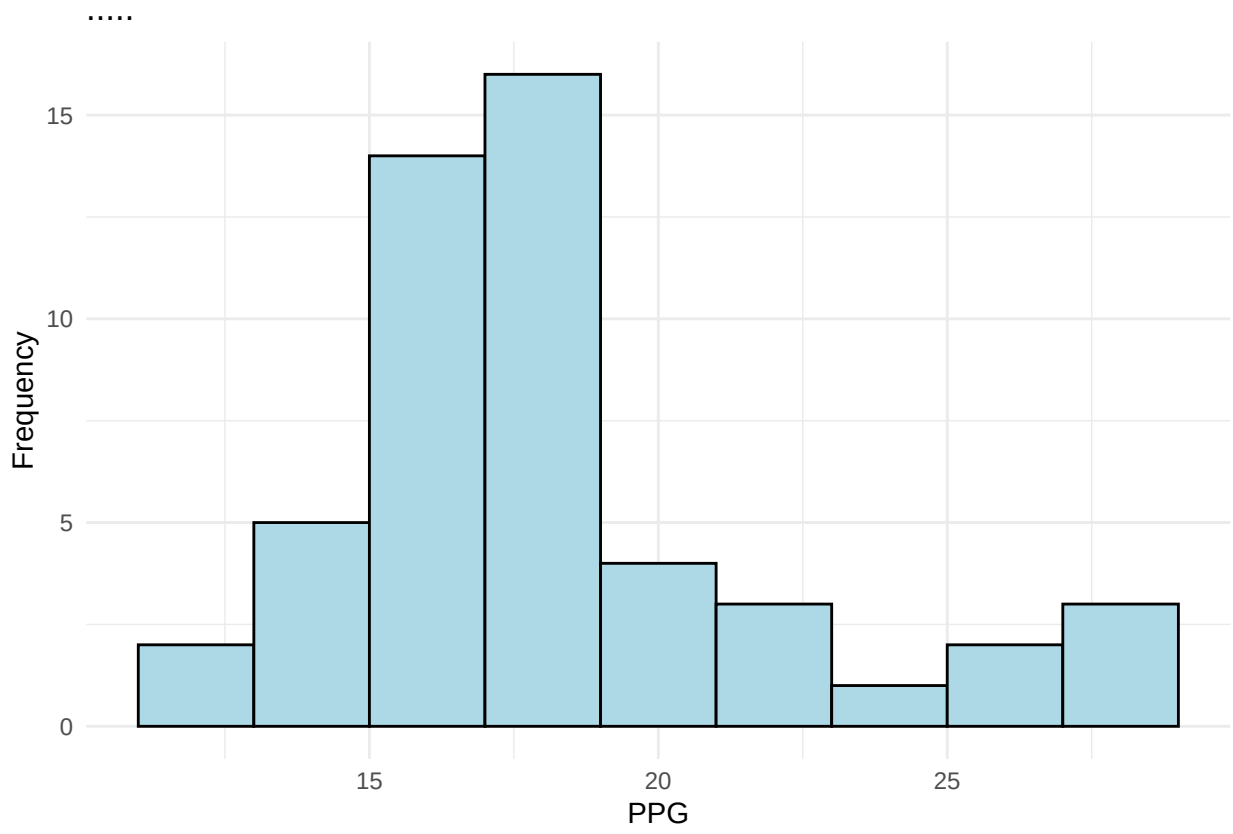
Question #2: NBAPlayerPts. (Attached Data: NBAPlayerPts)

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.

a.Show the frequency distribution.

```
# 实现频率直方图
library(tidyverse)
table2 = read.csv("./data/NBAPlayerPts.csv")

ggplot(table2, aes(x = PPG)) + geom_histogram(binwidth = 2, fill = "lightblue", color = "black") +
```

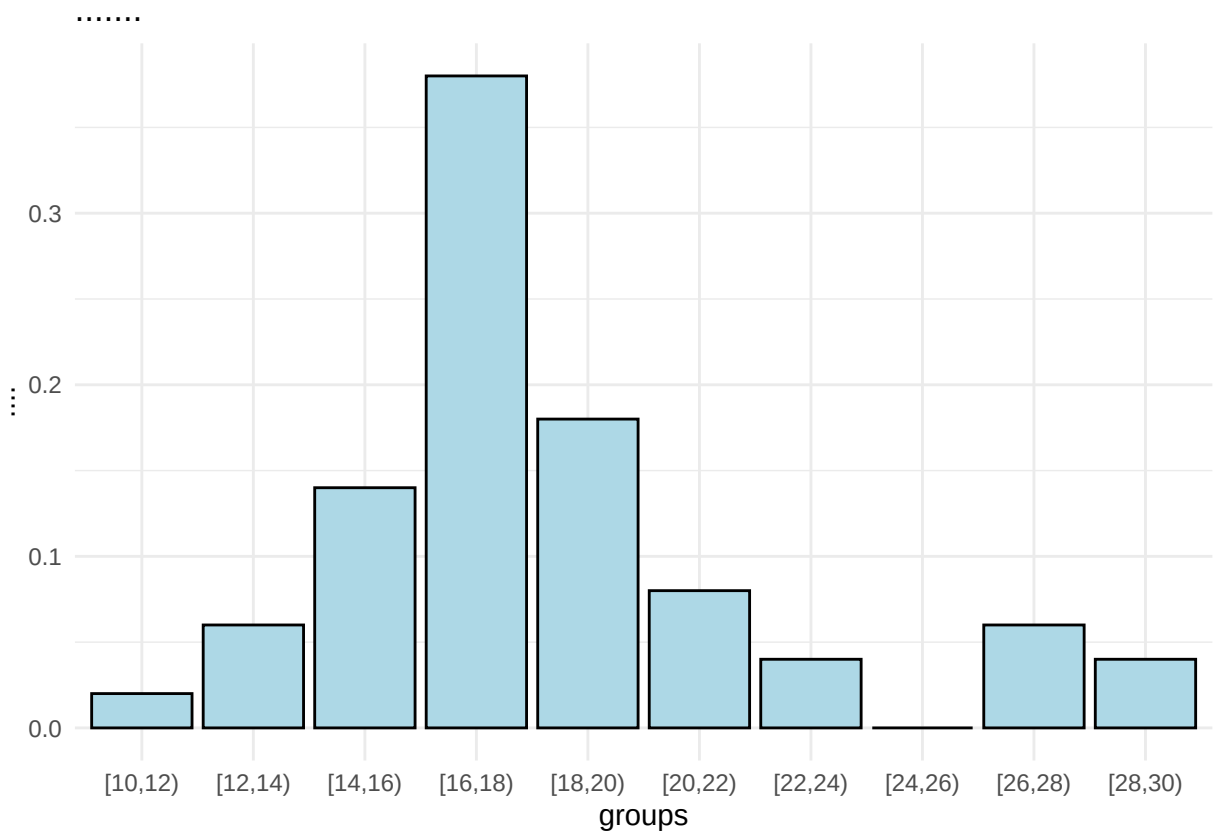
```
.....
```



b. Show the relative frequency distribution.

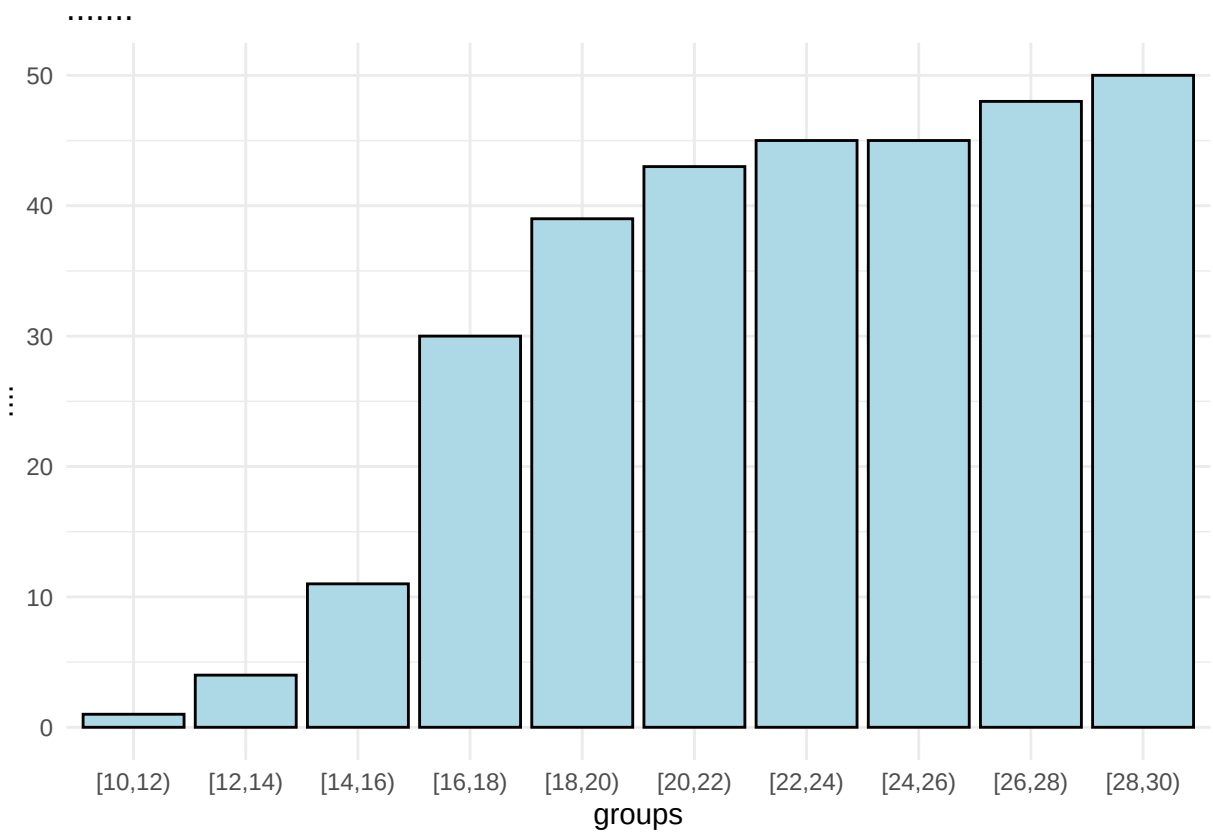
```
# 相对频率分布直方图
library(tidyverse)

breaks = seq(10, 30, by = 2)

groups = cut(table2$PPG, breaks, right = FALSE)

freq_table = as.data.frame(table(groups))
```

```r
freq_table = freq_table %>% mutate(relativeFrequency = Freq / sum(Freq))

ggplot(freq_table, aes(x = groups, y = relativeFrequency)) + geom_bar(stat = "identity", fill = "l
```



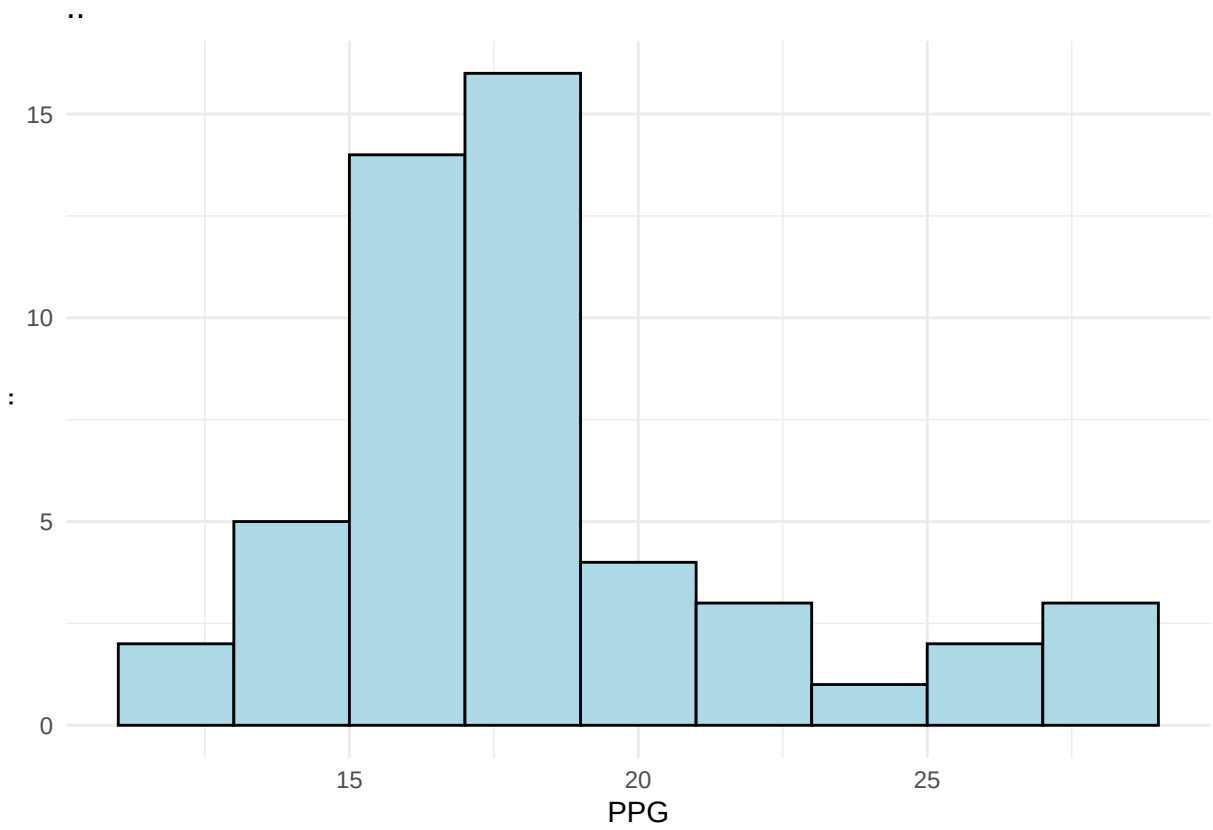
c. Show the cumulative percent frequency distribution.

```r
# 累积频率分布直方图
library(tidyverse)

freq_table = freq_table %>% mutate(cumulativeFrequency = cumsum(Freq))

ggplot(freq_table, aes(x = groups, y = cumulativeFrequency)) + geom_bar(stat = "identity", fill =
```



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

```
# 分数分布直方图
library(tidyverse)

ggplot(table2, aes(x = PPG)) + geom_histogram(binwidth = 2, fill = "lightblue", color = "black") +
```



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

```
library("moments")
skewness(table2$PPG)
```

```
## [1] 1.158609
```

```
median(table2$PPG)
```

```
## [1] 17.4
```

是的，可以发现频率分布直方图右偏，在累计分布的时候可以发现在 [14, 16), (16, 18) 这个区间是快速上升的，数据分布不均匀的

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

```
# 分数分布直方图
library(tidyverse)

nrow(table2 %>% filter(PPG >= 20)) / nrow(table2)
```

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

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

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

```
(500 / 20) ^ 2
```

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

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

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

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

如果用总体均值在 +-25 的时候，78% 的点会落到这个区间，证明这个波动是可靠的

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:

What is your age?

Are you: Male_____ Female_____

Do you plan to make any real estate purchases in the next two years?

Yes_____ No_____

What is the approximate total value of financial investments, exclusive of your

home, owned by you or members of your household?

How many stock/bond/mutual fund transactions have you made in the past year?

Do you have broadband access to the Internet at home? Yes_____ No_____

Please indicate your total household income last year. _____

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.

a. Develop appropriate descriptive statistics to summarize the data.

```
# 分数分布直方图
library(skimr)
library(kableExtra)
```

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

```
##
## 载入程序包: 'kableExtra'
```

```
## The following object is masked from 'package:dplyr':
##
##     group_rows
```

```
table3 = clean_names(read.csv("./data/Professional.csv")) %>% select(age:have_children)

skimr::skim(table3) %>% kable() %>% kable_styling()
```

skim_type	skim_variable	n_missing	complete_rate	character.min	character.max	character.em
character	gender	0	1	4	6	
character	real_estate_purchases	0	1	2	3	
character	broadband_access	0	1	2	3	
character	have_children	0	1	2	3	
numeric	age	0	1	NA	NA	

numeric	value_of_investments	0	1	NA	NA
numeric	number_of_transactions	0	1	NA	NA
numeric	household_income	0	1	NA	NA

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

```
# 计算致信区间
t.test(table3$age, conf.level = 0.95)$conf.int[1]
```

```
## [1] 29.72153
```

```
t.test(table3$age, conf.level = 0.95)$conf.int[2]
```

```
## [1] 30.50286
```

```
t.test(table3$household_income, conf.level = 0.95)$conf.int[1]
```

```
## [1] 71079.26
```

```
t.test(table3$household_income, conf.level = 0.95)$conf.int[2]
```

```
## [1] 77839.77
```

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

```
# 计算致信区间

# 总量
n_broadband = nrow(table3)

# 为 YES 的量
x_broadband = nrow(table3 %>% filter(broadband_access == 'Yes'))

result_broadband <- prop.test(x_broadband, n_broadband, conf.level = 0.95)
result_broadband$conf.int[1]
```

```
## [1] 0.5753252
```

```r
result_broadband$conf.int[2]
```

```
## [1] 0.6710862
```

```r
# 有孩子的量
x_children = nrow(table3 %>% filter(have_children == 'Yes'))
result_children <- prop.test(x_children, n_broadband, conf.level = 0.95)
result_children$conf.int[1]
```

```
## [1] 0.4845521
```

```r
result_children$conf.int[2]
```

```
## [1] 0.5830908
```

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

```r
# 计算致信区间
nrow(table3 %>% filter(broadband_access == 'Yes')) / nrow(table3)
```

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

大概有 62% 的人接入了互联网，所以是一个上网的良好的渠道

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

```r
# 计算百分比
nrow(table3 %>% filter(broadband_access == 'Yes')) / nrow(table3)
```

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

大概有 53% 的人有孩子，所以感觉也是个非常好的渠道

f. Comment on the types of articles you believe would be of interest to readers of Young Professional.
我个人认为应该是对金融投资或者电子游戏类的杂志感兴趣

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.

Corrective action will be taken any time is rejected.

Data are available in the data set Quality.

Managerial Report

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.

```
table4 = clean_names(read.csv("./data/Quality.csv"))

avg1 = mean(table4$sample_1)
avg2 = mean(table4$sample_2)
avg3 = mean(table4$sample_3)
avg4 = mean(table4$sample_4)

u = 12;
s = 0.21
n = nrow(table4)

tValue1 = (avg1 - u) / (s / sqrt(n))
pValue1 = if(tValue1 > 0) 2 * (1 - pnorm(tValue1)) else 2 * pnorm(tValue1)

tValue2 = (avg2 - u) / (s / sqrt(n))
pValue2 = if(tValue2 > 0) 2 * (1 - pnorm(tValue2)) else 2 * pnorm(tValue2)

tValue3 = (avg3 - u) / (s / sqrt(n))
pValue3 = if(tValue3 > 0) 2 * (1 - pnorm(tValue3)) else 2 * pnorm(tValue3)
```

```
tValue4 = (avg4 - u) / (s / sqrt(n))
pValue4 = if(tValue4 > 0) 2 * (1 - pnorm(tValue4)) else 2 * pnorm(tValue4)
```

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

```
sd(table4$sample_1)
```

```
## [1] 0.220356
```

```
sd(table4$sample_2)
```

```
## [1] 0.220356
```

```
sd(table4$sample_3)
```

```
## [1] 0.2071706
```

```
sd(table4$sample_4)
```

```
## [1] 0.206109
```

合理

c. compute limits for the sample mean around such that, as long as a new sample mean is within those limits, the process will be considered to be operating satisfactorily. if exceeds the upper limit or if 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.

```
z_interval <- function(x_bar, sigma, conf_level = 0.95, n) {
  z_star <- qnorm(1 - (1 - conf_level) / 2)  # 计算临界值
  margin_error <- z_star * (sigma / sqrt(n)) # 计算误差范围
  lower <- x_bar - margin_error              # 下限
  upper <- x_bar + margin_error              # 上限
  return(c(lower, upper))                    # 返回置信区间
}

z_interval(12, 0.21, 0.01, 30)
```

```
## [1] 11.99952 12.00048
```

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?

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.

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

```
table5 = clean_names(read.csv("./data/Occupancy.csv", skip = 1))

nrow(table5 %>% filter(march_2007 == 'Yes')) / nrow(table5)
```

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

```
nrow(table5 %>% filter(march_2008 == 'Yes')) / nrow(table5 %>% filter(march_2008 != ''))
```

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

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

```
#march2027 总量
n_2027 = nrow(table5)

#march2027 为 YES 的量
x_2027 = nrow(table5 %>% filter(march_2007 == 'Yes'))

#march2028 总量
n_2028 = nrow(table5 %>% filter(march_2008 != ''))

#march2027 为 YES 的量
x_2028 = nrow(table5 %>% filter(march_2007 == 'Yes'))

prop.test(x_2027, n_2027, conf.level = 0.95)$conf.int[1]
```

```
## [1] 0.2849421
```

```r
prop.test(x_2027, n_2027, conf.level = 0.95)$conf.int[2]
```

```
## [1] 0.4209231
```

```r
prop.test(x_2028, n_2028, conf.level = 0.95)$conf.int[1]
```

```
## [1] 0.3854464
```

```r
prop.test(x_2028, n_2028, conf.level = 0.95)$conf.int[2]
```

```
## [1] 0.5496202
```

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

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

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?

```r
table6 = read.csv("./data/Training.csv")

skimr::skim(table6) %>% kable() %>% kable_styling()
```

skim_type	skim_variable	n_missing	complete_rate	numeric.mean	numeric.sd	numeric.p0	numeric.p25
numeric	Current	0	1	75.06557	3.944907	65	72
numeric	Proposed	0	1	75.42623	2.506385	69	74

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

```r
t.test(table6$Current, table6$Proposed)
```

```
##
##  Welch Two Sample t-test
##
## data:  table6$Current and table6$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
```

pValue 为 0.54，2 组数据没有证据证明其有显著差异

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
var(table6$Current)
```

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

```r
var(table6$Proposed)
```

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

标准差: 3.944907 2.506385 方差: 15.5623 6.281967

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

我认为计算机辅助这种是比较好的，可以看到其方差较小，分布很集中，说明其效率对于大多数同学是比较有效的

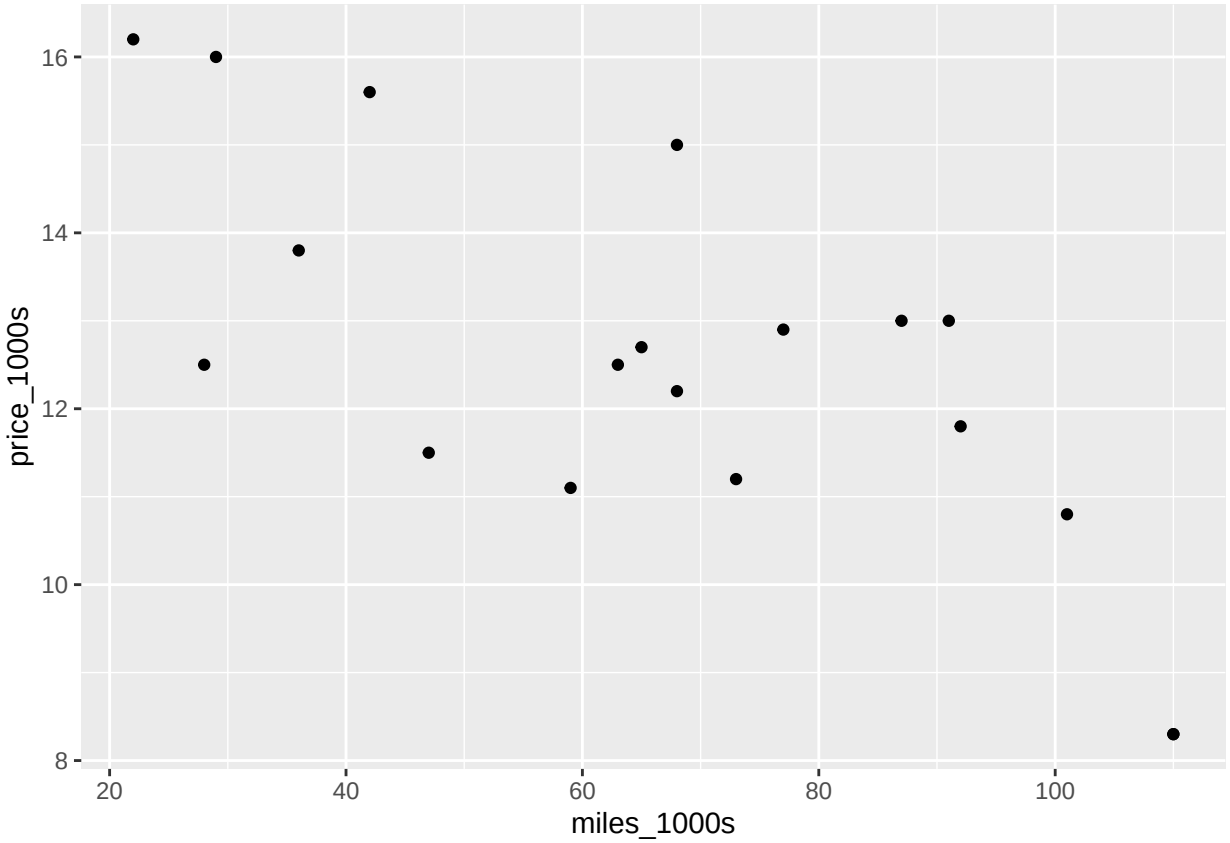
e. can you suggest other data or testing that might be desirable before making a final decision on the training program to be used in the future?

我认为需要再计入学生的年龄，性别，学习时间段等，这些变量也是能够影响最终结果的变量

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).

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

```
table7 = clean_names(read.csv("./data/Camry.csv"))

ggplot(data = table7, aes(x = miles_1000s, y = price_1000s)) + geom_point()
```



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

汽车行驶的里程越多的话，价格就卖的越少

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

```
modelCar = lm(price_1000s ~ miles_1000s, data = table7)

library(car)
```

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

## 载入需要的程序包: carData

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

##
## 载入程序包: 'car'
```

```
## The following object is masked from 'package:dplyr':
##
##     recode

## The following object is masked from 'package:purrr':
##
##     some
```

```
summary(modelCar)
```

```
##
## Call:
## lm(formula = price_1000s ~ miles_1000s, data = table7)
##
## 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
```

```
(16.47 - 1) / 0.059
```

```
## [1] 262.2034
```

结果大概是 262

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

0.0003475 $<$ 0.5, 所以该数据是可信的

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

大概是 0.5387 的拟合，个人感觉是一个算好的拟合

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

斜率为 0.59，每开 1000miles，价格就会下降 59 美刀

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.

(16.47 - 60 * 0.059) * 1000

大概是 12930 美刀的价格

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。

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？

```
table8 = clean_names(read_xlsx("./data/WE.xlsx"))

numeric_columns = names(table8)[sapply(table8, is.numeric) & names(table8) != "liu_shi" & names(ta

t_test_summary = sapply(numeric_columns, function(col) {
  formula <- as.formula(paste(col, "~ liu_shi"))
  test <- t.test(formula, data = table8)
  c(t_value = test$statistic, p_value = test$p.value, mean_group1 = test$estimate[1], mean_group2
})

t_test_summary = as.data.frame(t(t_test_summary))

# 致信区间
# t.test(dang_yue_deng_lu_ci_shu ~ liu_shi, data = table8)$conf.int[1]

# 自由度
# t.test(dang_yue_deng_lu_ci_shu ~ liu_shi, data = table8)$parameter

# p 值
```

```r
# t.test(dang_yue_deng_lu_ci_shu ~ liu_shi, data = table8)$p.value

# 均值
# as.numeric(t.test(dang_yue_deng_lu_ci_shu ~ liu_shi, data = table8)$estimate["mean in group 0"])
```

通过一系列比较发现几乎所有的变量，都存在显著的差异，除了” 服务优先级相比上月的变化”，“客户支持相比上月的变化” 2 个参数

b. 通过均值比较的方式验证上述不同是否显著。

第一问已经验证完毕

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。

```r
modelUse = lm(liu_shi ~ dang_yue_ke_hu_xing_fu_zhi_shu + ke_hu_xing_fu_zhi_shu_xiang_bi_shang_yue_

vif(modelUse)
```

```
##                     dang_yue_ke_hu_xing_fu_zhi_shu
##                                           1.460448
## ke_hu_xing_fu_zhi_shu_xiang_bi_shang_yue_bian_hua
##                                           1.338778
##                              dang_yue_ke_hu_zhi_chi
##                                           1.845977
##                          dang_yue_fu_wu_you_xian_ji
##                                           1.875531
##                             dang_yue_deng_lu_ci_shu
##                                           1.382685
##          bo_ke_shu_xiang_bi_shang_yue_de_bian_hua
##                                           1.063321
##    fang_wen_ci_shu_xiang_bi_shang_yue_de_zeng_jia
##                                           1.003266
##                               ke_hu_shi_yong_qi_xian
##                                           1.205960
##                           fang_wen_jian_ge_bian_hua
##                                           1.022844
```

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。不会