yinxuemei_second_assignment

2024-11-19

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)
```

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

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

```
library(dplyr)
bigbang <- read.csv("./data/BigBangTheory.csv")
library(stats)
library(lubridate)
library(readxl)
```

```
#最大值
max(bigbang$Viewers..millions.,na.rm = TRUE)
```

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

```
#最小值
min(bigbang$Viewers..millions.,na.rm = TRUE)
```

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

b. Compute the mean, median, and mode.

```
 # mean
  mean(bigbang$Viewers..millions.,na.rm = TRUE)
```

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

```
 # median
  median(bigbang$Viewers..millions.,na.rm = TRUE)
```

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

```
# mode
  #计算每个date的viewer数的频数
  date_freq <- table(bigbang$Viewers..millions.)
  #获得频数最高的值的索引
  mode_index <- which.max(date_freq)
  #众数
  value  <- names(date_freq)[mode_index]

  value
```

```
## [1] "13.6"
```

c. Compute the first and third quartiles.

```
# 第一四分位数
quantile(bigbang$Viewers..millions.,0.25)
```

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

```
#第三四分位数
quantile(bigbang$Viewers..millions.,0.75)
```

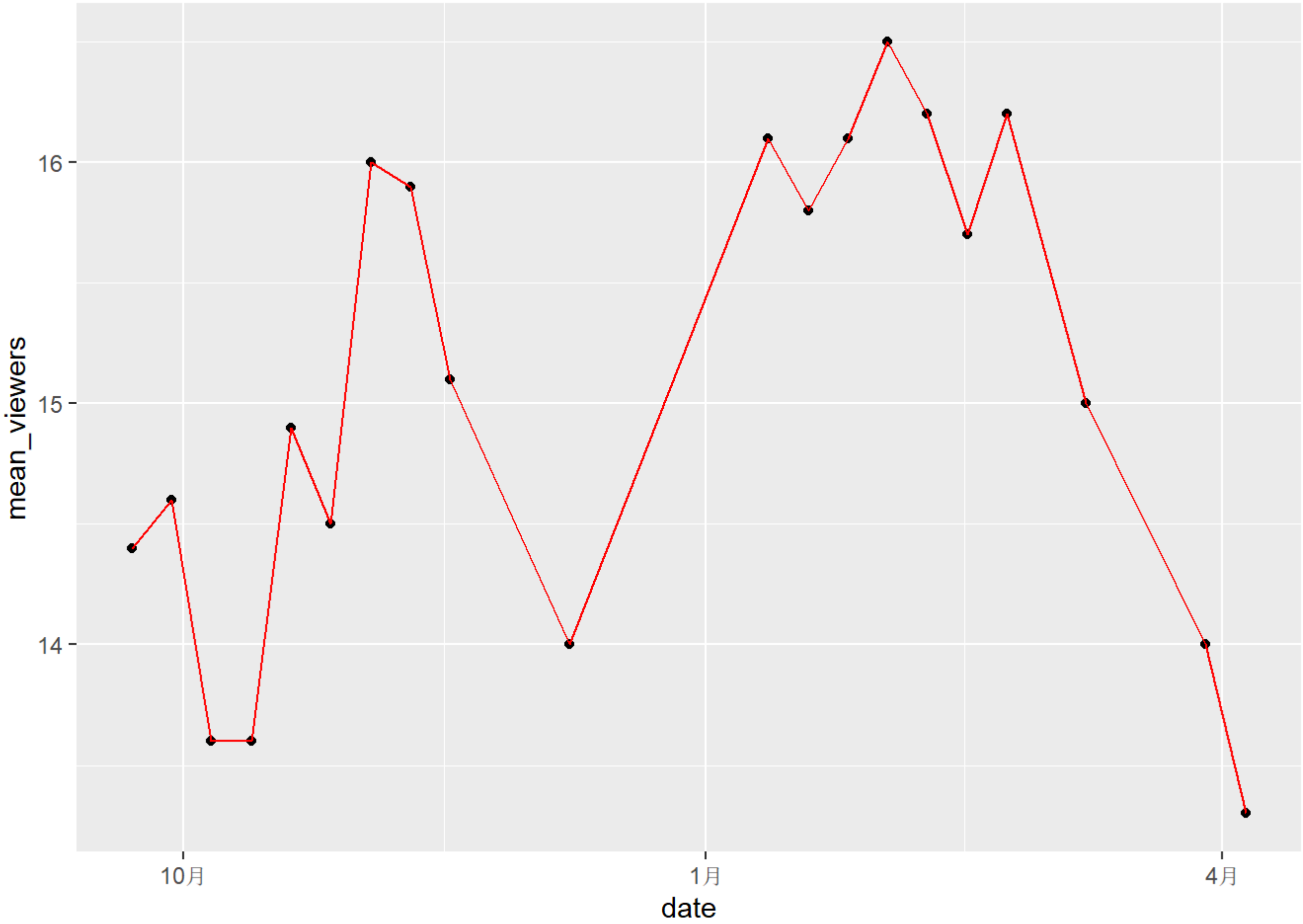
```
## 75%
##  16
```

```
#summary(bigbang)
```

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

没有明显的上升或者下降趋势。

```
library(lubridate)  #转换日期类型
bigbang %>%
  mutate(
    date  = mdy(bigbang$Air.Date),
    viewers = bigbang$Viewers..millions.
    ) %>%
    group_by(date)  %>%
    summarize(
     mean_viewers = mean(viewers)
    ) %>%
  ggplot(mapping = aes(x= date,y=mean_viewers)) +
  geom_point()+
  geom_line( colour = ”red”)
```



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.

```
nba_player <- read.csv(”./data/NBAPlayerPts.csv”)
 #最大值和最小值
max(nba_player$PPG,na.rm = TRUE)
```

```
## [1] 28.8
```

```
min(nba_player$PPG,na.rm = TRUE)
```

```
## [1] 11.7
```

```
#确定组数
#K <- 1 + log(50, base = 2)  #以2为底
#K 约等于 7

groups <- cut(nba_player$PPG, breaks = seq(10, 30, length.out = 7), include.lowest = TRUE)

freq_table <-table(groups)
print(freq_table)
```

```
## groups
##   [10,13.3] (13.3,16.7]   (16.7,20]   (20,23.3] (23.3,26.7]   (26.7,30]
##           2          14          23           6           1           4
```

b. Show the relative frequency distribution.

```
reltiv_freq_table <- freq_table /sum(freq_table)
print("relative frequency distribution:")
```

```
## [1] "relative frequency distribution:"
```

```
print(reltiv_freq_table)
```

```
## groups
##   [10,13.3] (13.3,16.7]   (16.7,20]   (20,23.3] (23.3,26.7]   (26.7,30]
##        0.04        0.28        0.46        0.12        0.02        0.08
```

c. Show the cumulative percent frequency distribution.

```
print("cumulative percent frequency distribution： ")
```

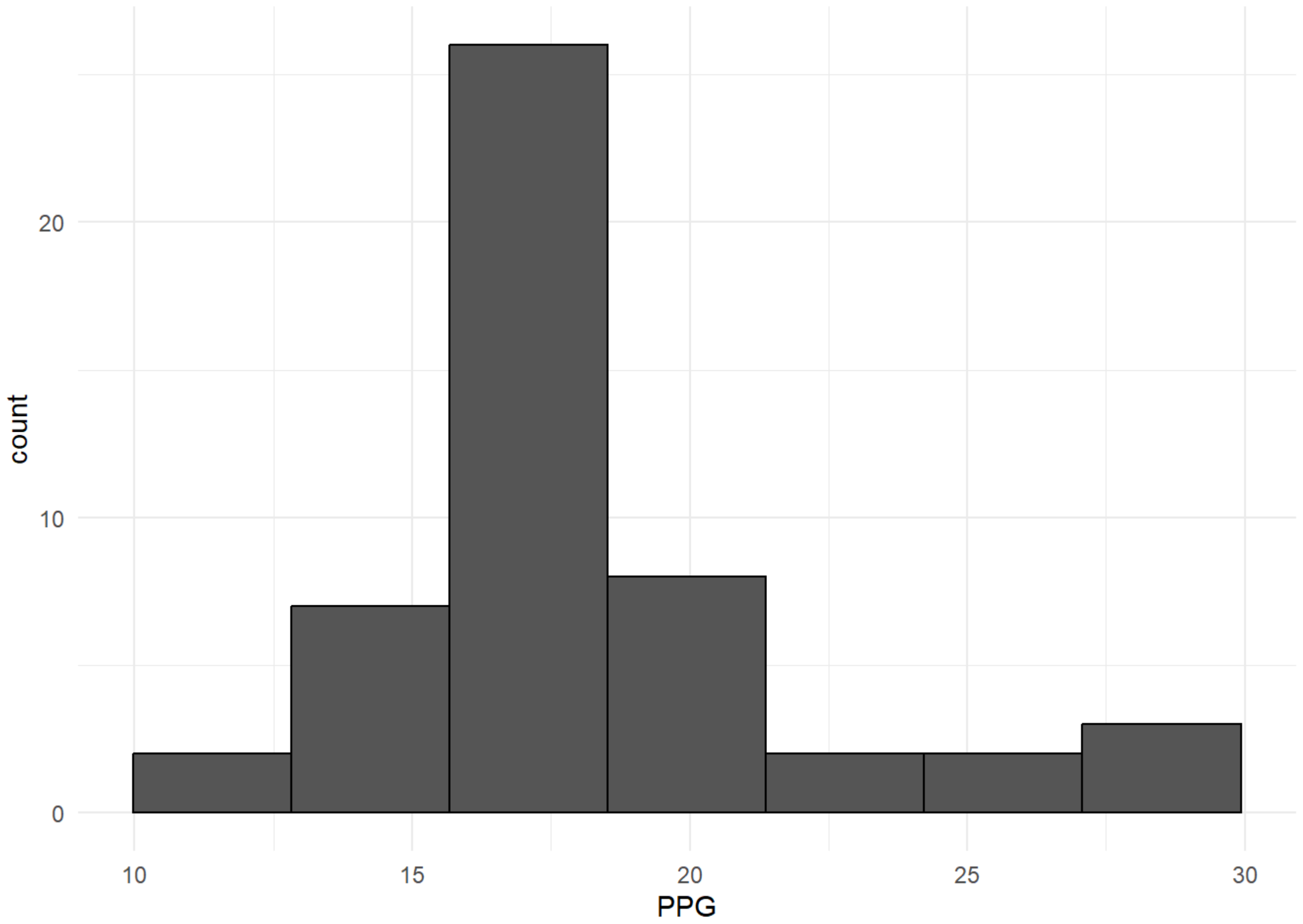
```
## [1] "cumulative percent frequency distribution： "
```

```
cumsum(reltiv_freq_table)
```

```
##   [10,13.3] (13.3,16.7]   (16.7,20]   (20,23.3] (23.3,26.7]   (26.7,30]
##        0.04        0.32        0.78        0.90        0.92        1.00
```

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

```
nba_player %>%
  ggplot(mapping = aes(x=PPG)) +
  geom_histogram(bins = 7, position = "dodge", color = "black")+
  theme_minimal()
```



e.

Do the data appear to be skewed? Explain.

有倾斜。分布向右略微倾斜。

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

```
# PPG大于等于20分的球员人数
count_leq_20 <- sum(nba_player$PPG>=20)

# PPG大于等于20的球员比例
 count_leq_20/length(nba_player$PPG)
```

```
## [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?

```
print("样本量：")
```

```
## [1] "样本量："
```

```
500/20 * 500/20
```

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

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

根据中心极限定理，样本均值服从正态分布

z_score =点估计值与总体均值的差异/平均数的标准误差

z_score = ±25/20 =±1.25 # 题目求 (-1.25 <=z_score<= 1.25)的概率 即P(-1.25 <=z_score<= 1.25) P(-1.25 <=z_score<= 1.25) =P(z_score<= 1.25) -P(z_score<-1.25)

P(z_score<-1.25) = 1- P(z_score<= 1.25) P(-1.25 <=z_score<= 1.25) = 2*P(z_score<= 1.25) -1 #查询标准正态分布表可知，当 P(z_score<= 1.25) =0.8944 P(-1.25 <=z_score<= 1.25)* = 20.8944-1 =0.7888

#所以点估计值在总体均值的±25 的概率约等于 0.79

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.

a. Develop appropriate descriptive statistics to summarize the data.

```
professional <- read.csv("./data/Professional.csv")
#install.packages("tibble")
 library(tibble)
#install.packages("skimr")
library(skimr)

# skim()

skim(professional) %>%
  tibble::as_tibble()
```

```
## # A tibble: 14 × 19
##    skim_type skim_variable   n_missing complete_rate character.min character.max
##    <chr>     <chr>               <int>         <dbl>         <int>         <int>
##  1 character Gender                  0             1             4             6
##  2 character Real.Estate.Pu…         0             1             2             3
##  3 character Broadband.Acce…         0             1             2             3
##  4 character Have.Children.          0             1             2             3
##  5 character X.1                     0             1             0             1
##  6 logical   X                     410             0            NA            NA
##  7 logical   X.2                   410             0            NA            NA
##  8 logical   X.3                   410             0            NA            NA
##  9 logical   X.4                   410             0            NA            NA
## 10 logical   X.5                   410             0            NA            NA
## 11 numeric   Age                     0             1            NA            NA
## 12 numeric   Value.of.Inves…         0             1            NA            NA
## 13 numeric   Number.of.Tran…         0             1            NA            NA
## 14 numeric   Household.Inco…         0             1            NA            NA
## # ℹ 13 more variables: character.empty <int>, character.n_unique <int>,
## #   character.whitespace <int>, logical.mean <dbl>, logical.count <chr>,
## #   numeric.mean <dbl>, numeric.sd <dbl>, numeric.p0 <dbl>, numeric.p25 <dbl>,
## #   numeric.p50 <dbl>, numeric.p75 <dbl>, numeric.p100 <dbl>,
## #   numeric.hist <chr>
```

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

```
#"样本量"
sample_n <-nrow(professional)
#"age样本的均值"
age_mean <- mean(professional$Age)
#"age样本的标准差"
age_sd <-sd(professional$Age)

lower_bond <- age_mean - 1.96*age_sd /20
upper_bond <- age_mean + 1.96*age_sd/20

age_interval <- c(lower_bond,upper_bond)
age_interval
```

```
## [1] 29.71784 30.50655
```

```
print("所以在95%的置信水平下，杂志的订阅者年龄在 30到31岁之间。")
```

```
## [1] "所以在95%的置信水平下，杂志的订阅者年龄在 30到31岁之间。"
```

```
income_lower <- mean(professional$Household.Income....) -1.96*sd(professional$Household.Incom
e....)/20 ;
income_upper <- mean(professional$Household.Income....) +1.96*sd(professional$Household.Incom
e....)/20

income_interval <- c(income_lower,income_upper)
income_interval
```

```
## [1] 71047.33 77871.70
```

```
print("所以在95%的置信水平下，杂志的订阅者的家庭收入在 71047.33 到77871.70 之间。")
```

```
## [1] "所以在95%的置信水平下，杂志的订阅者的家庭收入在 71047.33 到77871.70 之间。"
```

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

have broadband access at home

```
total_amt <- nrow(professional)
##样本比例
has_broadband_prop <- count(filter(professional,Broadband.Access.=='Yes'))/nrow(professional)

#p + 1.96 *(p*(1-p)/n的平方根)
upper_value <-has_broadband_prop + 1.96*sqrt(has_broadband_prop*(1-has_broadband_prop)/total_am
t)
lower_value <-has_broadband_prop - 1.96*sqrt(has_broadband_prop*(1-has_broadband_prop)/total_am
t)
print("95%的置信水平下，订阅者有家庭宽带的比例在")
```

```
## [1] "95%的置信水平下，订阅者有家庭宽带的比例在"
```

```
has_broad_interval <- c(lower_value,upper_value)

has_broad_interval
```

```
## $n
## [1] 0.5775132
##
## $n
## [1] 0.6712673
```

```
has_children_prop <- count(filter(professional,Have.Children.=='Yes'))/nrow(professional)
upper_c <- has_children_prop +1.96*sqrt(has_children_prop*(1-has_children_prop)/total_amt)
lower_c <- has_children_prop -1.96*sqrt(has_children_prop*(1-has_children_prop)/total_amt)
 has_children_interval <-c(lower_c,upper_c)
 print("95%的置信水平下，订阅者有孩子的比例在")
```

```
## [1] "95%的置信水平下，订阅者有孩子的比例在"
```

```
has_children_interval
```

```
## $n
## [1] 0.4858606
##
## $n
## [1] 0.5824321
```

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

```
#professional

#交易次数均值 和标准差
mean(professional$Number.of.Transactions)
```

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

```
sd(professional$Number.of.Transactions)
```

```
## [1] 3.100873
```

```
#95%置信水平下，订阅者买卖股票债券基金的次数
upper_d <-mean(professional$Number.of.Transactions) +1.96*sd(professional$Number.of.Transactions)/20
lower_d <-mean(professional$Number.of.Transactions) -1.96*sd(professional$Number.of.Transactions)/20
transct_times_interval <- c(lower_d,upper_d)
transct_times_interval
```

```
## [1] 5.669285 6.277056
```

```
#家庭除房产的金融投资额
upper_f <- mean(professional$Value.of.Investments....) +1.96*sd(professional$Value.of.Investments....)/20
lower_f <- mean(professional$Value.of.Investments....) -1.96*sd(professional$Value.of.Investments....)/20
invertment_interval <- c(lower_f,upper_f)
invertment_interval
```

```
## [1] 26988.83 30087.75
```

Professional 可以作为online brokers的广告渠道。从数据上显示，Professional的订阅者在95%的置信水平，家庭可投资的金额达到2.6万-3万美金。 在过去一年内交易股票/债券/基金的次数为6次,并且有家庭宽带的占60%左右。

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

professional杂志可以投放教育软件和幼儿游戏广告。因为professional的订阅者中有孩子的比例48.6%-58.2%.

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

```
#计划购买房产的样本比例
estate_purchase_prop <- count(filter(professional,Real.Estate.Purchases.==’Yes’))/nrow(professi
onal)

upper_k <-estate_purchase_prop +1.96 *sqrt(estate_purchase_prop*(1-estate_purchase_prop)/total_
amt)
lower_k <-estate_purchase_prop -1.96 *sqrt(estate_purchase_prop*(1-estate_purchase_prop)/total_
amt)

estate_purchase <-c(lower_k,upper_k)
estate_purchase
```

```
## $n
## [1] 0.3933975
##
## $n
## [1] 0.4895293
```

Young professional可以增加育儿经验或者房产买卖相关的文章。 因为它的订阅者年龄集中在30岁左右，有39%以上的人未来2年打算买房，有孩子的订阅者比例48.6%-58.2%.

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 μ =12 H : 1 μ =□ 12 Corrective action will be taken any time H0 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.

```
quality <- read.csv("./data/Quality.csv")
#总体sd =0.21  样本量 30   （显著性水平0.01  自由度：29  的T检验临界值：2.7564

#平均数的标准误差：SE=0.21/sqrt(30)
t.test(quality$Sample.1,mu=12,conf.level=0.01)$p.value
```

```
## [1] 0.3127296
```

```
t.test(quality$Sample.2,mu=12,conf.level=0.01)$p.value
```

```
## [1] 0.4818209
```

```
t.test(quality$Sample.3,mu=12,conf.level=0.01)$p.value
```

```
## [1] 0.006468822
```

```
t.test(quality$Sample.4,mu=12,conf.level=0.01)$p.value
```

```
## [1] 0.03905895
```

结论：四个样本，其中一个样本的结果拒绝原假设，需要采取行动确认流程是否操作正确。

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

```
sd1<-sd(quality$Sample.1)
sd2<-sd(quality$Sample.2)

sd3<-sd(quality$Sample.3)
sd4<-sd(quality$Sample.4)
```

假设的0.21的总体标准差合理。

c. compute limits for the sample mean x-ba around μ =12 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 x-ba 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.

总体标准差是0.21，使用3σ原则设定上下控制限。 上控制限 12+3*0.12 =12.36 下控制限 12-3*0.12 =11.64

d. discuss the implications of changing the level of significance to a larger value. what mistake or error could increase if the level of significance is increased?

显著性水平提高意味着拒绝域扩大。当显著性水平提高时，意味着犯第一类错误的概率增加，即更容易拒绝原假设

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.

```
occupancy <- read.csv("./data/Occupancy.csv")
#2007
total_2007 <- count(filter(occupancy,March2007=='Yes'|March2007=='No') )
rate_2007 <- count(filter(occupancy,March2007=='Yes'))/total_2007
rate_2007
```

```
##      n
## 1 0.35
```

```
#2008
total_2008 <- count(filter(occupancy,March2008=='Yes'|March2008=='No'))

rate_2008 <- count(filter(occupancy,March2008=='Yes'))/total_2008
rate_2008
```

```
##           n
## 1 0.4666667
```

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

```
diff_prop <- rate_2008 -rate_2007  #比例差的估计值
se <- sqrt(rate_2008*(1-rate_2008)/total_2008+ rate_2007*(1-rate_2007)/total_2007)  #比例差的标准误差
Z_score <- 1.96  ## 95% 置信水平对应的 Z 分数

c(diff_prop-Z_score*se,diff_prop+Z_score*se)
```

```
## $n
## [1] 0.01301325
## 
## $n
## [1] 0.2203201
```

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

从上题08年的出租率和07年的出租率之差95%的可能在0.01到0.22之间，都大于0，所以2008年的出租率是高于2007年的。

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?

```
  training <- read.csv("./data/Training.csv")
  summary(training)
```

```
##     Current         Proposed    
##  Min.   :65.00   Min.   :69.00  
##  1st Qu.:72.00   1st Qu.:74.00  
##  Median :76.00   Median :76.00  
##  Mean   :75.07   Mean   :75.43  
##  3rd Qu.:78.00   3rd Qu.:77.00  
##  Max.   :84.00   Max.   :82.00
```

```
 current_sd<- sd(training$Current)
 proposed_sd <- sd(training$Proposed)

 current_sd
```

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

```
 proposed_sd
```

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

目前的教学方案 和提议的教学方案 结果的均值基本一致，但是目前的教学方案的教学成果波动性大于提议的教学方案。

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

```
mean_current <- mean(training$Current)
mean_proposed <- mean(training$Proposed)

var_current <-var(training$Current)
var_proposed <- var(training$Proposed)
var_current
```

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

```
var_proposed
```

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

```
 # 已知
 n1 <- 61
 n2 <- 61

time_se <- sqrt((var_current/61)+(var_proposed/61))


t_value <- (mean_current-mean_proposed)/time_se
t_value
```

```
## [1] -0.6026832
```

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.

```
#标准差
current_sd
```

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

```
proposed_sd
```

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

```
#方差
var_current
```

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

```
var_proposed
```

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

```
var.test(training$Current,training$Proposed)
```

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

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

目前的教学方法和提议的教学方法标准差和方差不一样，目前的教学方法方差更大。

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?

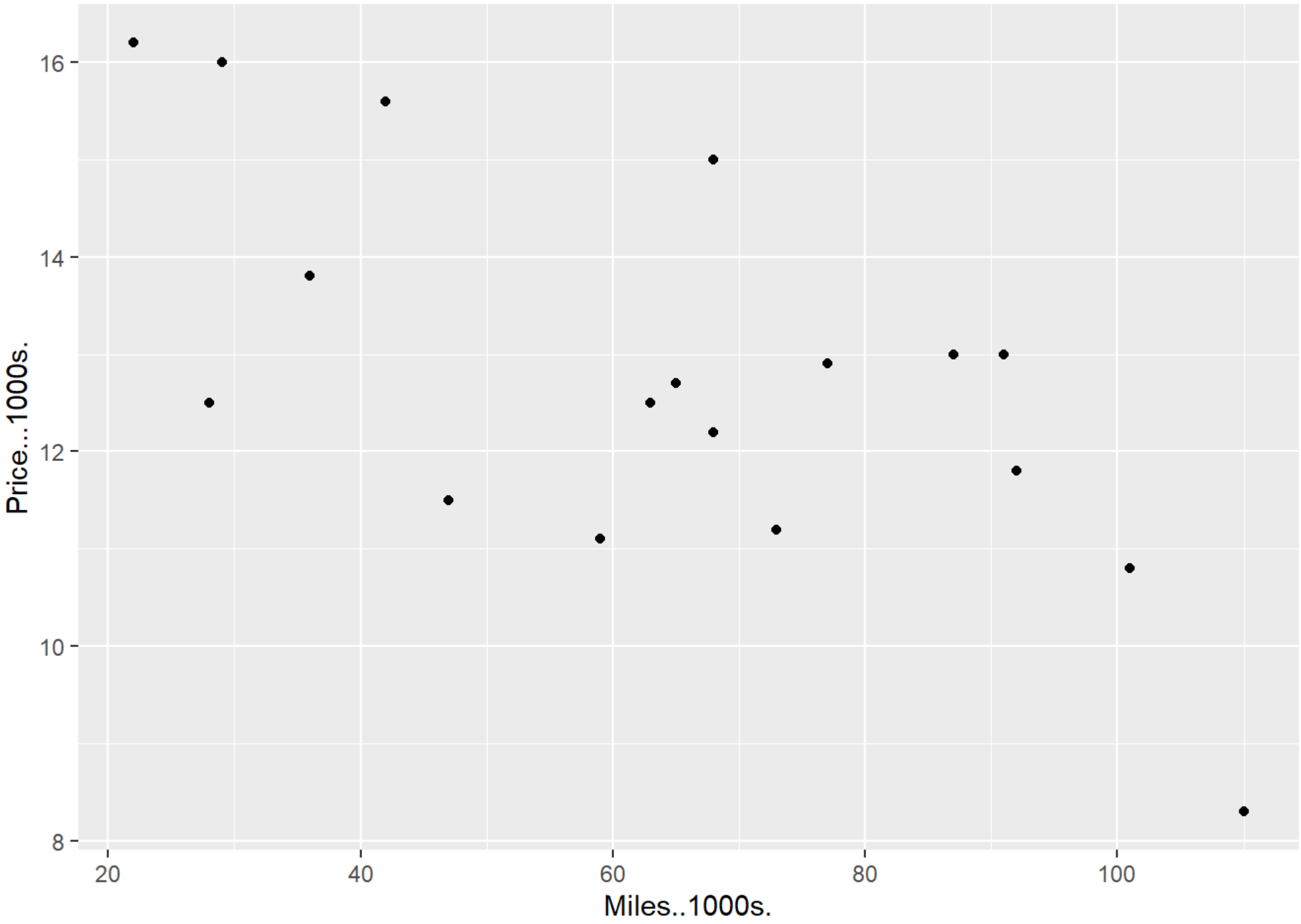
两种教学方法下，学生完成训练的时间均值基本相同，但是两则的方差不同，选择哪一种教学方法还要取决于理想的教学方法的评判标准。 另外样本调查的是2种教学方法下学生完成训练的时间，评估教学方法是不是还要考察其他方面的结果，例如教学结束后一段时间学生掌握程度的区别。

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.

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

ggplot(data=camry,mapping = 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?

从上面的散点图可以看出随里程数的增加，售价呈下降趋势。

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

```
Camry_model <- lm(camry$Price...1000s.~camry$Miles..1000s.,camry)
summary(Camry_model)
```

```
##
## Call:
## lm(formula = camry$Price...1000s. ~ camry$Miles..1000s., data = camry)
##
## 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 ***
## camry$Miles..1000s. -0.05877    0.01319  -4.455 0.000348 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.541 on 17 degrees of freedom
## Multiple R-squared:  0.5387, Adjusted R-squared:  0.5115 
## F-statistic: 19.85 on 1 and 17 DF,  p-value: 0.0003475
```

回归方程式： Price = 16.46976 -0.05877 * Miles

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

```
summary(Camry_model)
```

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

P值小于0.05的显著性水平，说明自变量和因变量之间的关系显著，两者之间存在线性关系。

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

残差（Residuals）： 残差范围从-2.32408到2.52687，这个范围相对合理，没有异常大的残差。 残差的第一四分位数（1Q）和第三四分位数（3Q）之间的范围（-1.34194到1.12898）表明残差的分布相对集中。 系数（Coefficients）： 截距（Intercept）的估计值为16.46976，标准误差为0.94876，t值为17.359，P值非常小（2.99e-12），表明截距在统计上高度显著。 自变量camry$Miles..1000s.（假设这是表示车辆行驶里程的变量，单位为千英里）的系数为-0.05877，标准误差为0.01319，t值为-4.455，P值也非常小（0.000348），表明该自变量在统计上显著。 残差标准误差（Residual standard error）： 残差标准误差为1.541，这个值相对较小，表明模型预测值与实际观测值之间的差异不大。 决定系数（R-squared 和 Adjusted R-squared）： Rsquared为0.5387，表明模型解释了因变量（假设是车辆价格，单位为千美元）53.87%的变异。 Adjusted R-squared为0.5115，考虑了模型中自变量的数量，对R-squared进行了调整。这个值仍然相对较高，表明模型在解释因变量变异方面表现良好。 F统计量（F-statistic）： F统计量为19.85，对应的P值非常小（0.0003475），表明模型整体在统计上显著。 综上所述，该回归方程提供了良好的拟合。模型的系数在统计上显著，残差分布合理，R-squared和Adjusted R-squared值相对较高，F统计量也表明模型整体显著。因此，我们可以认为该回归方程能够较好地解释自变量（车辆行驶里程）与因变量（车辆价格）之间的关系。

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

在回归方程中 Price = 16.46976 -0.05877 * Miles 斜率为-0.05877。

意味着每当车辆的行驶里程（Miles）增加1000英里时，车辆的价格（Price）平均预期会减少0.05877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.

```
new_miles <- 60 # 60千英里
predicted_price <- predict(Camry_model, test_date = data.frame(Miles = new_miles))
predicted_price
```

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

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?

```
#client <- read_excel("data/WE.xlsx")
```

b. 通过均值比较的方式验证上述不同是否显著。

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。

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。