````
---
title: "第二次作业"
author: "万丽娟2024281050972"
date: "`r Sys.Date()`"
output:
  html_document:
    df_print: paged
  pdf_document:
    latex_engine: yes
documentclass: ctexart
---

```{r setup, include=FALSE}
knitr::opts_chunk$set(
  out.width = "100%",
  fig.show = TRUE,
  df_print = "tibble",
  paged.print = FALSE,
  split = FALSE
)
```


```{r message=FALSE, warning=FALSE, paged.print=FALSE}
library(readxl)
library(tidyverse)
library(stats)
library(lubridate)
library(infer)
library(kableExtra)
library(scales)
library(showtext)
showtext_auto()
```


```{r}
file_path1<-read_csv("/Users/Air/Desktop/BigBangTheory.csv")
print(file_path1)
```
第一题
```{r}
#Question 1:
file_path1 <- read_csv(str_c
                       ("/Users/Air/Desktop/BigBangTheory.csv")) %>%
  rename(viewers = `Viewers (millions)`, air_date = `Air Date`) %>%
  mutate(air_date = mdy(air_date))
```


```{r}
#a.
min_value <- min(file_path1$viewers)
max_value<-max(file_path1$viewers)
print(min_value)
print(max_value)
```
min=13.3 max=16.5
````

```{r}
#b
Mean_value<-mean(file_path1$viewers)
Median_value<-median(file_path1$viewers)
Mode_value<-names(which.max(table(file_path1$viewers)))
print(Mean_value)
print(Median_value)
print(Mode_value)
```
mean=15.04286, median=15, and mode=13.6.
```{r}
#c.
Q1 <-quantile(file_path1$viewers,probs = 0.25) ;
Q3 <-quantile(file_path1$viewers,probs = 0.75)
print(Q1)
print(Q3)
```
the first quartiles:25%=14.1 ;third quartiles:75%=16
```{r}
#d.
ggplot(file_path1,aes(air_date,viewers)) +
  geom_point() + geom_line(color="blue") +
  scale_x_date(breaks = file_path1$air_date) +
  theme(axis.text.x = element_text(angle = 30))
```

从上图看出2011年-2012年，不能得出增长还是下降的趋势。

第二题
```{r}
#Question 2
file_path2<-read.csv("/Users/Air/Desktop/NBAPlayerPts.csv")
print(file_path2)
```

```{r}
#a.
frequency<-table(cut_width(file_path2$PPG,2,boundary = 10))
print(frequency)
```

```{r}
#b
Relative <-table(cut_width(file_path2$PPG,2,boundary = 10))/50
print(Relative)
```

```{r}
#c
Cumulative <- cumsum(table(cut_width(file_path2$PPG,2,boundary = 10))/50)
print(Cumulative)
```

```{r}
#d
ggplot(file_path2,aes(PPG)) + geom_histogram(binwidth = 5,color =
"black",fill="yellow")
```

```

```{r}
#e：向右偏
```
向右偏
```{r}
#f
num_high_PPG <- sum(file_path2$PPG>= 20)
total_players <- length(file_path2$PPG)
percentage <- (num_high_PPG / total_players) *100
print(percentage)
```


第三题
```{r}
#Question 3
#a:根据公式：SE=σ/√n得出 SE=20，σ=500，所以
n<-(500/20)^2
print(n)
```
该调查使用的样本大小是625

```{r}
#b
#在正态分布下，标准化的z分数公式为：$Z=(X-\mu)/S$
#假设点估计的误差范围为 ±25，那么：
z<-25/20
print(z)
probability <- pnorm(1.25)
print(probability)
```
#因为结果0.89大于0.05，属于正态分布，点估计在总体均值的±25范围内的概率是89.435%

第四题
```{r}
#Q4
data_q4 <- read_csv(str_c("/Users/Air/Desktop/Professional.csv")) %>%
  rename(age = Age, gender = Gender, real_estate = `Real Estate
Purchases?`, investments = `Value of Investments ($)`, num_trans =
`Number of Transactions`, has_broadband = `Broadband Access?`, income =
`Household Income ($)`, have_children = `Have Children?`) %>%
  select(age:have_children) %>%
  mutate_if(is.character, as.factor)
```
```{r}
#a计算描述性统计
skimr::skim(data_q4)
```
#描述性：
#年龄：订阅者的最小年龄为19岁，最大年龄为42岁，平均年龄为30.11岁，中位数年龄为30岁，表明订阅者主要为30岁左右的年轻专业人士。
#投资价值：订阅者家庭的平均投资价值为$28，538，中位数为$24,800，表明投资分布有一定的
```

偏斜。
#交易次数：过去一年中，订阅者平均进行了5.973次股票/债券/共同基金交易。
#家庭收入：平均家庭收入为$74,460，中位数为$66,050，反映出一定的收入分布不均。
#子女情况：41%的订阅者有子女。
```{r}
#b
# 直接使用t.test获取95%置信区间
t.test(data_q4$age)[[4]]
t.test(data_q4$income)[[4]]
```
我们有95%的信心可以得出结论平均年龄置信区间$(29.72,30.50)$家庭收入的置信区间为$(71,079, 77839.77)$
```{r}
#c
# 计算家中有宽带接入的比例的95%置信区间
prop_test(data_q4,response = has_broadband,success = "Yes")

# 计算有子女的比例的95%置信区间
prop_test(data_q4, response = have_children,success = "Yes")
```
我们有95%的信心可以得出结论:家中有宽带接入的比例置信区间$(0.57, 0.67)$有子女的比例置信区间为$(0.484, 0.537)$

#d .在线券商广告投放潜力:考虑到67%的订阅者家中有宽带接入，以及平均家庭收入较高，这表明《年轻专业人士》杂志的订阅者群体具有较高的金融投资潜力。因此，该杂志是在线券商广告的良好投放渠道。

#e.教育软件和儿童电脑游戏广告投放潜力:由于53.7%的订阅者有子女，且家庭收入较高，这表明有一定的市场潜力。然而，考虑到杂志的目标受众是年轻专业人士，可能更多关注职业发展和个人理财，因此教育软件可能比儿童电脑游戏更受欢迎。

#f. 读者感兴趣的文章类型
#职业发展：鉴于订阅者处于职业生涯的早期阶段，有关职业规划、技能提升和行业动态的文章可能受欢迎。
#个人理财：投资策略、资产管理和财务规划相关的文章可能吸引订阅者。
#技术与创新：新兴技术、创新应用和互联网趋势的文章可能符合订阅者的互联网使用习惯。
#生活品质：健康生活、旅游和文化活动等提高生活品质的内容可能引起兴趣。
#综上所述，通过分析订阅者的统计数据，我们可以看到《年轻专业人士》杂志不仅适合金融和投资领域的广告商，也适合教育软件和家庭相关产品的广告商。同时，杂志编辑可以根据这些数据来选择更符合读者兴趣的文章主题。

第五题
```{r}
#Q5
data_q5 <- read.csv("/Users/Air/Desktop/Quality.csv")
colnames(data_q5) <- c("Sample 1", "Sample 2", "Sample 3","Sample 4")
data_q5
```

```{r}
#a
cal_p <- function(vec,miu,sigma,n){
```

```
  a <- mean(vec) - miu
  if(a >=0) {
    return(2*(1-pnorm(a/(sigma/sqrt(n)))))}
  else
    return(2*pnorm(a/(sigma/sqrt(n)))) }
cal_p(data_q5$'Sample 1', miu = 12,sigma = 0.21,n=30)
cal_p(data_q5$'Sample 2', miu = 12,sigma = 0.21,n=30)
cal_p(data_q5$'Sample 3', miu = 12,sigma = 0.21,n=30)
cal_p(data_q5$'Sample 4', miu = 12,sigma = 0.21,n=30)
```
对每个样本执行假设检验,1号、2号和4号样本不拒绝零假设，过程令人满意,3号样本拒绝零假设,需要纠正。
```{r}
#b
sample_sds <- apply(data_q5, 2, sd)
print(sample_sds)
```
计算每个样本的标准差，通过下面的计算结果评估假设总体标准差为0.21的是合理的
```{r}
#c 样本均值的控制限为11.89432 --12.10568
alpha <- 0.01
sigma <- 0.21
n <- 30
mu <- 12
df <- n - 1
t_critical <- qt(1 - alpha/2, df)
p_values <- numeric(4)
control_limits <- c(mu - t_critical * sigma / sqrt(n), mu + t_critical *
sigma / sqrt(n))
print(control_limits)
```
样本均值的控制限为11.89432 --12.10568

d. 讨论改变显著性水平的影响：增加显著性水平会增加第一类错误的风险（错误地拒绝零假设）。

第六题

```{r}
#q6
data_q6 <- read.csv("/Users/Air/Desktop/Occupancy.csv")
```

```{r}
# a.
sum(data_q6$March_2007 %in% c("Yes"))/200
sum(data_q6$March_2008 %in% c("Yes"))/150
```
2007年入住单位的比例为0.35，2008年入住单位的比例为0.47
```{r}
#b
pa <- sum(data_q6$March_2007%in% c("Yes"))/200
pb <- sum(data_q6$March_2008 %in% c("Yes"))/150
e <- qnorm(0.975) * sqrt(pa*(1-pa)/200 + pb*(1-pb)/150)
f<-(pa-pb)+e
d<-(pa-pb)-e
```

```
print(f)
print(d)
```
 求95%的置信区间差异：因为pa-pb=-0.09，在置信区间0.013和0.22之间，是可以接受的。
```{r}
#c.
c(pa-pb-e,pa-pb+e)

```
由于这个置信区间的下限是 -0.23，且上限是 -0.01，且整个区间都小于 0，这意味着在 95% 的置信水平下，我们可以认为 2008 年的出租比例显著低于 2007 年的出租比例。如果出租比例下降，通常可以推测出租金的上涨压力可能较小，甚至可能出现下降的趋势。这是因为出租比例下降可能意味着市场供给过剩，或者需求减弱，从而对租金产生下行压力。

第七题
```{r}
#7
data_q7 <- read.csv("/Users/Air/Desktop/Training.csv")
```
```{r}
# a.
skimr::skim(data_q7)
```
```{r}
# b.
t.test(data_q7$Current,data_q7$Proposed)
```
 在0.05显著水平上，两组间无差异。

```{r}
# c.
map(data_q7,sd)
map(data_q7,var)
t_test <- t.test(data_q7$Current, data_q7$Proposed)
print(t_test)
```
结论:sd或方差是不同的。现行方法方差较大。

```{r}
#d
var.test(data_q7$Current,data_q7$Proposed)
```
结论:sd或方差是不同的。现行方法方差较大。

#e基于现有的数据，建议的方法是首选的。两种方法的平均完成时间非常接近，差异的95%置信区间为-1.55 ~ 0.83小时。然而，所提出的方法具有较低的方差。在提出的方法下，学生更有可能在大约 相同的时间内完成训练。应该减少快的学生等待慢的学生完成训练的机会。

#f在做出最终决定之前，我们建议收集两种方法下学习量的数据。时间数据倾向于采用该方法。然而，在培训项目结束时，两组学生都可以参加考试。对考试成绩的分析将决定这些课程在提供的学习量方面是相似还是不同。这种分析应该在最终决定是否切换到所建议的方法之前进行。

第八题
```{r}
```

```
#Q8
data_q8 <- read_csv("/Users/Air/Desktop/Camry.csv") %>%
  rename(miles = `Miles (1000s)`, price = `Price ($1000s)`)
#a:
ggplot(data = data_q8, mapping = aes(x = miles, y = price)) +
  geom_point() +
  geom_smooth()
```

#b这两个变量之间似乎有一种负的关系，可以用一条直线近似表示。还有一种观点认为，这种关系可能是 曲线的，因为在某一点上，一辆车的行驶里程是如此之多，以至于它的价值变得非常小

```{r}
# c.
lm_camry <- lm(price ~ miles, data = data_q8)
summary(lm_camry)
```

Price = 16.470 - 0.059 * miles 回归方程：价格= 16.470 - 0.059 *英里

d. # p-value = 0.000348 < α = .05 p-value = 0.000348 < α = 0.05
#e r平方= 0.5387;考虑到汽车的状况也是决定价格的一个重要因素，这是一个相当好的匹配。
#f估计的回归方程的斜率为-0.059。因此，x值增加一个单位与y值减少等于0.059一致。因为这些数据 记录了数千次，汽车里程表上每多跑1000英里，预计价格就会下降59.0美元。

g 2007款凯美瑞6万英里的预期价格是= 16.47 -.0588(60)= 12.942或12942美元。由于其他因素，如果条件和卖家是私人或经销商，这可能不是你会为汽车提供的价格。但是，这应该是一个很好的起点， 知道该向卖家提供什么。

第九题
```{r}
#Q9 加载必要的库
library(readxl)
# 读取数据
data_q9 <- read_excel("/Users/Air/Desktop/WE.xlsx")

# 查看数据结构
str(data_q9)

# 转换数据类型
data_q9$流失 <- as.factor(data_q9$流失)

# 客户幸福指数
ggplot(data_q9, aes(x = 流失, y = 当月客户幸福指数)) +
  geom_boxplot() +
  labs(title = "客户幸福指数与流失状态", x = "流失状态", y = "客户幸福指数")

# 客户支持
ggplot(data_q9, aes(x = 流失, y = 当月客户支持)) +
  geom_boxplot() +
  labs(title = "客户支持与流失状态", x = "流失状态", y = "客户支持")

# 服务优先级
```

```
ggplot(data_q9, aes(x = 流失, y = 当月服务优先级)) +
  geom_boxplot() +
  labs(title = "服务优先级与流失状态", x = "流失状态", y = "服务优先级")

# 登录次数
ggplot(data_q9, aes(x = 流失, y = 当月登录次数)) +
  geom_boxplot() +
  labs(title = "登录次数与流失状态", x = "流失状态", y = "登录次数")

# 访问次数
ggplot(data_q9, aes(x = 流失, y = 访问次数相比上月的增加)) +
  geom_boxplot() +
  labs(title = "访问次数与流失状态", x = "流失状态", y = "访问次数增加")
```

```{r}
#b定义一个函数来计算均值差异
# 定义一个函数来执行t检验并返回p值
t_test_mean <- function(data_q9, group, variable) {
  t.test(group1, group2)
}

# 执行t检验
t_test_results <- data.frame(
  Variable = character(),
  P_Value = numeric()
)

t_test_results$P_Value <- sapply(t_test_results$Variable, function(var) {
  t_test_mean(data_q9, "流失", var)$p.value
})

# 查看结果
summary(data_q9$当月客户幸福指数_mean)
summary(data_q9$当月客户支持_mean)
summary(data_q9$当月服务优先级_mean)
summary(data_q9$当月登录次数_mean)
summary(data_q9$访问次数相比上月的增加_mean)
```

```{r}
#c
# 建立逻辑回归模型
model <- glm(流失 ~ 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 + 博客数相比上月的变化 + 访问次数相比上月的增加 + 客户使用期限 + 访问间隔变化,
             data = data_q9, family = binomial)

# 输出模型摘要
summary(model)
```