商业数据分析第 2 次作业

胡雄雁

2024-11-30

准备工作：导包

```
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(dplyr)
library(readxl)
library(knitr)
library(tidyr)
library(tibble)
```

第 1 题：BigBang Theory

数据载入

```
filepath <- r"(C:/Users/huxio/Desktop/武汉大学 MEM/2. 商业数据分析/作业 2/BigBangTheory.csv)"
df_bigbang <- read.csv(filepath)
```

```r
df_bigbang$Air.Date <- mdy(df_bigbang$Air.Date)
head(df_bigbang,5)
```

```
##     Air.Date Viewers..millions.
## 1 2011-09-22                14.1
## 2 2011-09-22                14.7
## 3 2011-09-29                14.6
## 4 2011-10-06                13.6
## 5 2011-10-13                13.6
```

数据分析

1. 最小和最大观众人数

- 最小值为 13.3，最大值为 16.5

2. 均值、中位数和众数

- 均值为 15.04，中位数为 15.0，众数有 4 个，分别为：13.6,14,16.1,16.2

3. 第一和第三四分位数

- 第一分位为 14.1，第三分位为 16

```r
## 统计概览
summary(df_bigbang)
```

```
##     Air.Date          Viewers..millions.
##  Min.   :2011-09-22   Min.   :13.30
##  1st Qu.:2011-10-20   1st Qu.:14.10
##  Median :2011-12-08   Median :15.00
##  Mean   :2011-12-17   Mean   :15.04
##  3rd Qu.:2012-02-09   3rd Qu.:16.00
##  Max.   :2012-04-05   Max.   :16.50
```

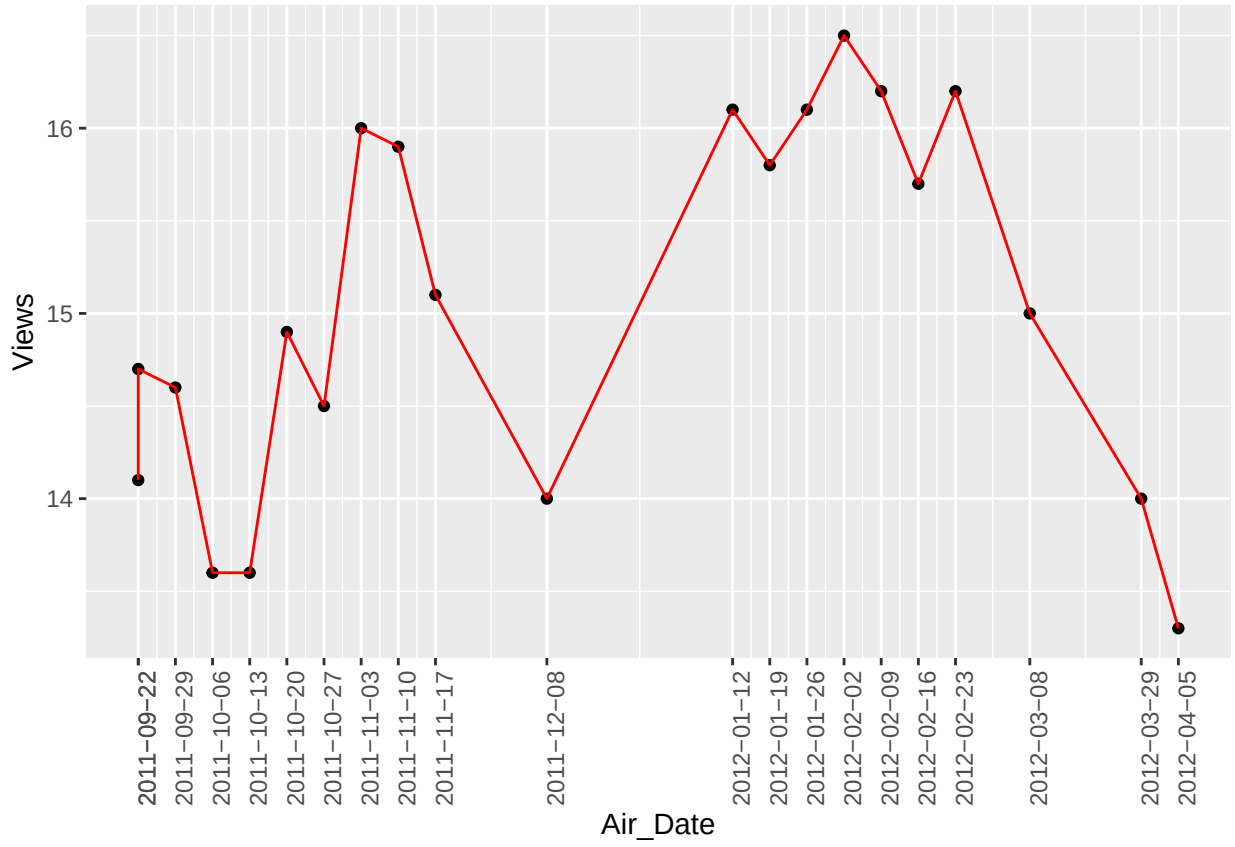
```r
## 计算众数
t_mode <- df_bigbang$Viewers..millions. %>%
  as_factor() %>%
  table()
t_mode[t_mode==max(t_mode)]
```

```
## .
## 13.6   14 16.1 16.2
##    2    2    2    2
```

4. 增长趋势判断

- 不具备明显的增长趋势。21 年和 22 年观看人数有升有降，虽 22 年的最高观看人数大于 21 年的最高观看人数，但无法看出明显的增长趋势。

```
ggplot(data=df_bigbang,mapping=aes(x=Air.Date,y=Viewers..millions.))+
  geom_point()+
  geom_line(color='red')+
  scale_x_date(date_labels='%Y-%m-%d',breaks=df_bigbang$Air.Date)+ # 格式化日期且只显示有值的 X 轴标
  theme(axis.text.x=element_text(angle=90,hjust=1))+ # 竖着显示日期
  labs(x='Air_Date',y='Views')
```



第 2 题：NBA Players

数据载入

```
filepath <- r"(C:/Users/huxio/Desktop/武汉大学 MEM/2. 商业数据分析/作业 2/NBAPlayerPts.csv)"
df_nba <- read.csv(filepath)
head(df_nba,5)
```

```
##   Rank                 Player  PPG
## 1    1      LeBron James, MIA 27.0
## 2    2      Kevin Durant, OKC 28.8
## 3    3      James Harden, HOU 26.4
## 4    4       Kobe Bryant, LAL 27.1
## 5    5 Russell Westbrook, OKC 22.9
```

数据分析

```
breaks <- seq(10,30,by=2)
data_cut <- cut(df_nba$PPG,breaks=breaks,right=FALSE)  # 左闭右开区间
# levels(data_cut) # 显示分组因子
group_cut <- table(data_cut)
group_cut
```

1. 显示频率分布

```
## data_cut
## [10,12) [12,14) [14,16) [16,18) [18,20) [20,22) [22,24) [24,26) [26,28) [28,30)
##       1       3       7      19       9       4       2       0       3       2
```

```
group_cut_rate <- group_cut/length(df_nba$PPG)
group_cut_rate
```

2. 显示相对频率分布

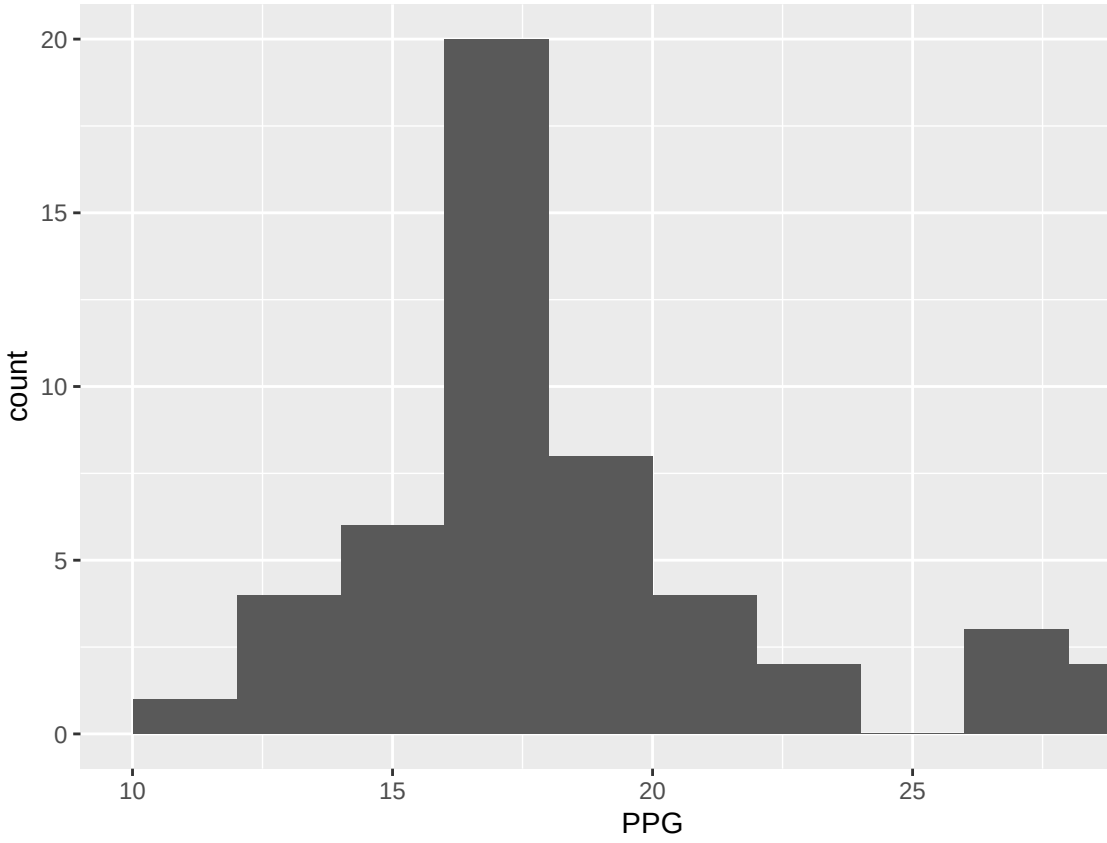
```
## data_cut
## [10,12) [12,14) [14,16) [16,18) [18,20) [20,22) [22,24) [24,26) [26,28) [28,30)
##    0.02    0.06    0.14    0.38    0.18    0.08    0.04    0.00    0.06    0.04
```

```
group_cut_cumsum <-cumsum(group_cut_rate)
group_cut_cumsum
```

3. 显示累积百分比频率分布

```
## [10,12) [12,14) [14,16) [16,18) [18,20) [20,22) [22,24) [24,26) [26,28) [28,30)
##    0.02    0.08    0.22    0.60    0.78    0.86    0.90    0.90    0.96    1.00
```

```
ggplot(data=df_nba,mapping=aes(x=PPG))+
  geom_histogram(binwidth=2,breaks=breaks)  # 以 2 为增量分组
```



4. 制作场均得分的直方图

5. 解释数据是否呈现偏斜

- 从直方图的分布来看，呈现出右偏特征。

6. 有多少百分比的球员场均得分至少为 20 分

- 从累计百分比频率分布知为：1-0.78=0.22

第 3 题：样本量估计

1. 计算调查样本总数

- 调查样本总数为为 625。

```
real_std = 500
mean_std = 20
n = (real_std/mean_std)^2
n
```

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

2. 计算点估计的概率

- 点估计的概率为：0.7887005

```
p = pnorm(25/mean_std)-pnorm(-25/mean_std)
p
```

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

第 4 题：Young Professional Magazine

数据载入

```
filepath <- r"(C:/Users/huxio/Desktop/武汉大学 MEM/2. 商业数据分析/作业 2/Professional.csv)"
df_profession <- read.csv(filepath)
head(df_profession,5)
```

```
##   Age Gender Real.Estate.Purchases. Value.of.Investments....
## 1  38 Female                    No                    12200
## 2  30   Male                    No                    12400
## 3  41 Female                    No                    26800
## 4  28 Female                   Yes                    19600
```

```
## 5  31 Female                    Yes                   15100
##   Number.of.Transactions Broadband.Access. Household.Income.... Have.Children.
## 1                      4               Yes                75200            Yes
## 2                      4               Yes                70300            Yes
## 3                      5               Yes                48200             No
## 4                      6                No                95300             No
## 5                      5                No                73300            Yes
##    X X.1 X.2 X.3 X.4 X.5
## 1 NA      NA  NA  NA  NA
## 2 NA      NA  NA  NA  NA
## 3 NA      NA  NA  NA  NA
## 4 NA      NA  NA  NA  NA
## 5 NA   .  NA  NA  NA  NA
```

```
summary(df_profession)
```

```
##       Age           Gender          Real.Estate.Purchases.
##  Min.   :19.00   Length:410         Length:410
##  1st Qu.:28.00   Class :character   Class :character
##  Median :30.00   Mode  :character   Mode  :character
##  Mean   :30.11
##  3rd Qu.:33.00
##  Max.   :42.00
##  Value.of.Investments.... Number.of.Transactions Broadband.Access.
##  Min.   :     0           Min.   : 0.000         Length:410
##  1st Qu.: 18300           1st Qu.: 4.000         Class :character
##  Median : 24800           Median : 6.000         Mode  :character
##  Mean   : 28538           Mean   : 5.973
##  3rd Qu.: 34275           3rd Qu.: 7.000
##  Max.   :133400           Max.   :21.000
##  Household.Income.... Have.Children.        X               X.1
##  Min.   : 16200       Length:410         Mode:logical   Length:410
##  1st Qu.: 51625       Class :character   NA's:410       Class :character
##  Median : 66050       Mode  :character                  Mode  :character
##  Mean   : 74460
##  3rd Qu.: 88775
##  Max.   :322500
##    X.2             X.3             X.4             X.5
##  Mode:logical   Mode:logical   Mode:logical   Mode:logical
##  NA's:410       NA's:410       NA's:410       NA's:410
```

##
##
##
##

数据分析

```
# 对变量按照文本型和数值型进行分类
c_character <- c()
c_logical <- c()
for (col in names(df_profession)){
  col_class = class(df_profession[[col]])
  if (col_class=='character'){
    c_character <- c(c_character,col)}
  else{c_logical <- c(c_logical,col)}
}
```

```
# 数值型变量描述性统计
df_profession_logical <-  select(df_profession,all_of(c_logical))
summary(df_profession_logical)
```

1. 描述性统计

```
##       Age        Value.of.Investments.... Number.of.Transactions
##  Min.   :19.00   Min.   :     0           Min.   : 0.000
##  1st Qu.:28.00   1st Qu.: 18300           1st Qu.: 4.000
##  Median :30.00   Median : 24800           Median : 6.000
##  Mean   :30.11   Mean   : 28538           Mean   : 5.973
##  3rd Qu.:33.00   3rd Qu.: 34275           3rd Qu.: 7.000
##  Max.   :42.00   Max.   :133400           Max.   :21.000
##  Household.Income....    X             X.2            X.3
##  Min.   : 16200       Mode:logical   Mode:logical   Mode:logical
##  1st Qu.: 51625       NA's:410       NA's:410       NA's:410
##  Median : 66050
##  Mean   : 74460
##  3rd Qu.: 88775
##  Max.   :322500
```

```
##    X.4             X.5
##  Mode:logical   Mode:logical
##  NA's:410       NA's:410
##
##
##
##
```

```
# 文本型变量描述性统计
df_profession_character <-  select(df_profession,all_of(c_character))
lapply(df_profession_character,table)
```

```
## $Gender
##
## Female   Male
##    181    229
##
## $Real.Estate.Purchases.
##
##  No Yes
## 229 181
##
## $Broadband.Access.
##
##  No Yes
## 154 256
##
## $Have.Children.
##
##  No Yes
## 191 219
##
## $X.1
##
##       .
## 409   1
```

2. 为订阅者的平均年龄和家庭收入制定 95% 的置信区间

- 使用 t 分布进行总体均值的区间估计：平均年龄置信区间为 (29.72153,30.50286), 平均家庭收入置信

区间为 (71079.26,77839.77)

```
t_evaluate <- function(df_x,r){
  c_r <- c((1-r)/2,1-(1-r)/2)  # 置信水平上下限
  n <- length(df_x)
  df_n <- n-1
  mean_x <- mean(df_x)
  sd_x <- sd(df_x)/sqrt(n)
  t_ci <- mean_x+qt(c_r,df_n)*sd_x
  t_ci
}
cat(t_evaluate(df_profession$Age,0.95)
    ,';'
    ,t_evaluate(df_profession$Household.Income,0.95))
```

```
## 29.72153 30.50286 ; 71079.26 77839.77
```

3. 为家中拥有宽带接入的订阅者比例和有孩子的订阅者比例制定 95% 的置信区间

- 有宽带接入的订阅者比例 95% 的置信区间为：c(0.5775140,0.6712665)
- 有孩子的订阅者比例 95% 的置信区间为：c(0.4858615,0.5824312)

```
norm_p_evaluate <- function(df_x,r){
  c_r <- c((1-r)/2,1-(1-r)/2)  # 置信水平上下限
  n <- length(df_x)
  mean_p_x <- sum(df_x=='Yes')/n
  sd_p_x <- sqrt(mean_p_x*(1-mean_p_x)/n)  # 均值标准差
  p_ci <- mean_p_x+qnorm(c_r)*sd_p_x
  p_ci
}
cat(norm_p_evaluate(df_profession$Broadband.Access.,0.95)
    ,';'
    ,norm_p_evaluate(df_profession$Have.Children.,0.95))
```

```
## 0.577514 0.6712665 ; 0.4858615 0.5824312
```

4. 对在线代理商而言，《年轻专业人士》是否是一个好的广告渠道？用统计数据来支持您的结论

- 是一个好的广告渠道，分析如下：

- 有超 62% 的杂志订阅用户接入了宽带，即具备线上观看广告的条件；
- 该杂志订阅用户超 98% 都有投资活动，其中超 35% 的用户投资额超 30000，甚至存在有近 0.5% 的用户投资额超 90000
- 该杂志订阅用户基本都有投资活动且投资相对频繁，其中超 65% 的用户交易次数不少于 5 次，超 11% 的用户交易次数不少于 10 次

```
## 杂志订阅用户接入宽带分析
ttl_count <- dim(df_profession)[1]
p_broad_access <- sum(df_profession$Broadband.Access.=='Yes')/ttl_count
p_broad_access
```

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

```
## 该杂志订阅用户投资活动分析
p_have_investment <- sum(df_profession$Value.of.Investments....>0)/ttl_count
p_have_investment
```

```
## [1] 0.9853659
```

```
breaks <- seq(0,150000,by=30000)
investment_cut <- cut(df_profession$Value.of.Investments....,breaks=breaks,right=FALSE)  # 左闭右开
investment_cut_rate <- table(investment_cut)/ttl_count
investment_cut_rate
```

```
## investment_cut
##          [0,3e+04)      [3e+04,6e+04)      [6e+04,9e+04)    [9e+04,1.2e+05)
##        0.648780488        0.300000000        0.046341463        0.002439024
## [1.2e+05,1.5e+05)
##        0.002439024
```

```
## 该杂志订阅用户交易频次分析
p_have_transaction <- sum(df_profession$Number.of.Transactions>0)/ttl_count
p_have_transaction
```

```
## [1] 0.995122
```

```
breaks <- seq(0,25,by=5)
transaction_cut <- cut(df_profession$Number.of.Transactions,breaks=breaks,right=FALSE)  # 左闭右开
transaction_cut_rate <- table(transaction_cut)/ttl_count
transaction_cut_rate
```

```
## transaction_cut
##       [0,5)      [5,10)     [10,15)     [15,20)     [20,25)
## 0.339024390 0.546341463 0.097560976 0.014634146 0.002439024
```

5. 这本杂志是否适合为销售幼儿教育软件和电脑游戏的公司做广告

- 适合，原因如下：
  - 由上分析知，有超 62% 的杂志订阅用户接入了宽带，即具备接入软件或者游戏的条件
  - 平均年龄为 30.1 岁且 53% 的用户有小孩，用户群体整体年轻，一反面依然处于对电脑游戏有兴趣的年龄层，一方面家里有幼儿需要教育

```
mean_age <- mean(df_profession$Age)
mean_age
```

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

```
p_have_children <- sum(df_profession$Have.Children.=='Yes')/ttl_count
p_have_children
```

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

6. 就您认为《年轻专业人士》的读者会感兴趣的文章类型发表评论

- 房地产、游戏、育儿、投资风向判断等都可能是读者感兴趣的文章，分析如下：
  - 订阅用户有 56% 的用户待购置房产，因此推断房产会是一个感兴趣的话题；
  - 从上分析可知该用户群体较为年轻、超半数有小孩，因此推断游戏和育儿也是不错的话题；
  - 从上分析可知该用户群体基本都有投资活动且投资相对频繁，因此推断投资风向判断也是不错的话题。

```
p_have_estate <- sum(df_profession$Real.Estate.Purchases.=='No')/ttl_count
p_have_estate
```

```
## [1] 0.5585366
```

第 5 题：假设检验-质量问题

数据载入

```r
filepath <- r"(C:/Users/huxio/Desktop/武汉大学 MEM/2. 商业数据分析/作业 2/Quality.csv)"
df_quality <- read.csv(filepath)
head(df_quality,5)
```

```
##   Sample.1 Sample.2 Sample.3 Sample.4
## 1    11.55    11.62    11.91    12.02
## 2    11.62    11.69    11.36    12.02
## 3    11.52    11.59    11.75    12.05
## 4    11.75    11.82    11.95    12.18
## 5    11.90    11.97    12.14    12.11
```

数据分析

1. 提供每个样本的检验 p 值

- Sample.1:p 值 0.2810083 ; 样本均值: 11.95867 临界区间:z 11.85991 或 z 12.05743
- Sample.2:p 值 0.4546503 ; 样本均值: 12.02867 临界区间:z 11.92991 或 z 12.12743
- Sample.3:p 值 0.003790318 ; 样本均值: 11.889 临界区间:z 11.79024 或 z 11.98776
- Sample.4:p 值 0.03389336 ; 样本均值: 12.08133 临界区间:z 11.98257 或 z 12.18009

```r
## 计算每个样本的 p 值和临界区间 (双侧检验, 拒绝 H0 时)
s_p_value <- function(df_s,alpha,n,sigma,miu_0){
  mean_s = mean(df_s)
  sd = sd(df_s)
  sem = sigma/sqrt(n)
  z = abs(mean_s-miu_0)/sem
  s_p_value = (1-pnorm(z))*2
  ubound = qnorm(1-alpha/2)*sem+mean_s
  lbound = -qnorm(1-alpha/2)*sem+mean_s
  cat('样本标准差',sd,'\n')
  cat('p 值',s_p_value,';','样本均值:',mean_s,'临界区间:z ',lbound,'z ',ubound,'\n')
}

cat(s_p_value(df_quality$Sample.1,0.01,30,0.21,12)
,s_p_value(df_quality$Sample.2,0.01,30,0.21,12)
,s_p_value(df_quality$Sample.3,0.01,30,0.21,12)
,s_p_value(df_quality$Sample.4,0.01,30,0.21,12)
)
```

```
## 样本标准差 0.220356
## p值 0.2810083 ; 样本均值: 11.95867 临界区间:z 11.85991 z 12.05743
## 样本标准差 0.220356
## p值 0.4546503 ; 样本均值: 12.02867 临界区间:z 11.92991 z 12.12743
## 样本标准差 0.2071706
## p值 0.003790318 ; 样本均值: 11.889 临界区间:z 11.79024 z 11.98776
## 样本标准差 0.206109
## p值 0.03389336 ; 样本均值: 12.08133 临界区间:z 11.98257 z 12.18009
```

2. 标准差制定是否合理

- 结果为 0.2134979，故定标准差为 0.21 合理

```
mean(c(0.220356,0.220356,0.2071706,0.206109))
```

```
## [1] 0.2134979
```

3. 计算样本均值的上下限

- 答案如题 1

4. 提高显著性水平的意义

- 将显著性水平（ ）提高到一个更大的值时，这意味着我们降低了拒绝原假设 (H0) 的门槛，即更容易拒绝原假设，将会导致观察到的数据与原假设之间的差异可能并不足以表明存在真正的效应或差异

第 6 题：Vacation occupancy rates

数据载入

```
filepath <- r"(C:/Users/huxio/Desktop/武汉大学 MEM/2. 商业数据分析/作业 2/Occupancy.csv)"
df_vacation <- read.csv(filepath)
df_vacation_cleaned <- df_vacation[-1,]  # 数据清洗
names(df_vacation_cleaned) <- c('March_2007','March_2008')
head(df_vacation_cleaned,5)
```

```
##   March_2007 March_2008
## 2        Yes         No
## 3         No        Yes
```

```
## 4         Yes         Yes
## 5          No          No
## 6          No         Yes
```

数据分析

1. 估算 2007 年 3 月第一周和 2008 年 3 月第一周的出租单位比例

- 2007 年 3 月第 1 周：0.35；2008 年 3 月第 1 周：0.4666667

```
p1 = sum(df_vacation_cleaned$March_2007=='Yes')/sum(df_vacation_cleaned$March_2007=='Yes' | df_vac
p2 = sum(df_vacation_cleaned$March_2008=='Yes')/sum(df_vacation_cleaned$March_2008=='Yes' | df_vac
print(p1)
```

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

```
cat(p1,p2)
```

```
## 0.35 0.4666667
```

2. 为这两个比例之差提供一个 95% 的置信区间

- 置信区间为：(-0.22031818,-0.01301516)

```
## 由于 n1*(1-p1) 和 n2*(1-p2) 都大于 5，可使用正态分布开展区间估计
n1 = sum(df_vacation_cleaned$March_2007=='Yes' | df_vacation_cleaned$March_2007=='No')
n2 = sum(df_vacation_cleaned$March_2008=='Yes' | df_vacation_cleaned$March_2008=='No')
x = p1 - p2
sem = sqrt(p1*(1-p1)/n1 + p2*(1-p2)/n2)
c(x+qnorm(0.025)*sem,x+qnorm(1-0.025)*sem)
```

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

3. 租赁费率判断

- 设立下侧检验为：h0:P1-P2>=0,h1:P1-P2<0, 在置信度为 0.05 为前提下，拒绝域为样本均值需小于等于-0.08698709, 因为 p1-p2 为-0.117, 小于拒绝域上限，故拒绝原假设，即 2008 年 3 月的租赁费率会高于 1 年前。

```r
ubound = qnorm(0.05)*sem
ubound
```

```
## [1] -0.08698709
```

第 7 题：Air Force Training Program

数据导入

```r
filepath <- r"(C:/Users/huxio/Desktop/武汉大学 MEM/2. 商业数据分析/作业 2/Training.csv)"
df_train <- read.csv(filepath)
head(df_train,5)
```

```
##   Current Proposed
## 1      76       74
## 2      76       75
## 3      77       77
## 4      74       78
## 5      76       74
```

数据分析

1. 使用适当的描述性统计量来总结每种方法的培训时间数据

- 提议的方法与当前的方法比较结果为：最短学习时长增加，从原本的 65 增加到 69；中位数一致，都位 76；平均学习时长稍有增加，从 75.07 增加到 75.43；最大学习时长有所改善，从 84 降到了 82。

```r
summary(df_train)
```

```
##     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
```

2. 评论两种方法下总体均值的差异

- 从均值结果知两者方法的总体均值一致，但无法确定这是否为偶然现象, 进一步通过 t 检验发现，在置信度为 0.05 的基础下，两种学习方式对降低平均学习时长无显著差异。
  - 构建右侧假设 t 检验

```
# 基于提议的方法有所改善学习时长构建右侧假设检验
# H0: _d=0 H1: _d 0 置信度为 0.05 其中 p_d 为当前学习方式与提议学习方式的总体均值差
alpha = 0.05
df_train$d <- df_train$Current - df_train$Proposed # 构建差值统计量
mean_d = mean(df_train$d)
sd_d = sd(df_train$d)
mean_d
```

```
## [1] -0.3606557
```

```
sd_d
```

```
## [1] 4.423546
```

```
# 由于总体方差未知，此处使用 t 检验
n = nrow(df_train)-1
sem = sd_d/sqrt(n)
sem
```

```
## [1] 0.5710773
```

- 接受域为 (-1.519227,1.519227), 因为样本均值-0.3606557 落在了接受域内，故接受原假设，两种学习方式对降低平均学习时长无显著差异

```
t_critical <- qt(alpha/2, n)
t_critical
```

```
## [1] -2.000298
```

```
c(t_critical*sem,abs(t_critical)*sem)
```

```
## [1] -1.142325  1.142325
```

- p 值为 0.5300891 大于置信度 0.05，故接受原假设，两种学习方式对降低平均学习时长无显著差异，与上述接受域检验的结果一致

```
pt((mean_d-0)/sem,n)*2
```

```
## [1] 0.5300891
```

3. 计算每种培训方法的标准差和方差及每种培训方法的总体方差是否相等假设检验

- 两种培训方式的标准差和方差：当前培训方式和提议方式方差分别为：15.5623 6.281967 当前培训方式和提议方式标准差分别为：3.944907 2.506385

```
sd_current = sd(df_train$Current)
sd_proposed = sd(df_train$Proposed)
var_current = sd_current^2
var_proposed = sd_proposed^2
var_d = var_current-var_proposed
cat('当前培训方式和提议方式方差: ',var_current,var_proposed
    ,'\n'
    ,'当前培训方式和提议方式标准差: ',sd_current,sd_proposed)
```

```
## 当前培训方式和提议方式方差:  15.5623 6.281967
##  当前培训方式和提议方式标准差:  3.944907 2.506385
```

- 使用 F 检验法开展假设检验假设检验：H0: 方差相等 H1: 方差不等，置信度为 0.05，在这样的条件下：因为样本构建 F 统计量的值为 2.477296 不在接受域范围内，故拒绝原假设，即两种培训方法的总体方差不相等；p 值为 0.0005780315, 明显小于 0.05 的显著性水平，故拒绝原假设，即两种培训方法的总体方差不相等；

```
# 求接受域
c(qf(alpha/2,df1=n,df2=n),qf(1-alpha/2,df1=n,df2=n))
```

```
## [1] 0.5999553 1.6667908
```

```
f_sem = var_current/var_proposed
f_sem
```

```
## [1] 2.477296
```

```r
# 求 P 值
p_value = (1-pf(f_sem,df1 = n,df2=n))*2
p_value
```

```
## [1] 0.0005780315
```

4. 您能得出关于这两种方法之间差异的什么结论？您有什么建议？解释原因。

- 基于 0.05 的置信度水平，这 2 种方法测试出的学生平均学习时长一致，即可认为这 2 种方式在降低学生平均学习时长方面无显著差异；
- 但从方差来看，提议方式的方差小于当前学习方式的方差，可认为提议方式可减少快速学习的学生需要等待慢速学习的学生的时间，可降低大家学习时长的差距。

5. 在就未来要使用的培训计划做出最终决定之前，您能否建议其他数据或测试？

- 建议扩大样本容量，降低第 I 类错误发生的概率；
- 建议增加学习成果检验测试。学习时长不能代表学习效果，建议增加学习成果检验的考试，以考试成绩来说明这 2 种方式的差异。

第 8 题：Camry 二手车价格预测

数据导入

```r
filepath <- r"(C:/Users/huxio/Desktop/武汉大学 MEM/2. 商业数据分析/作业 2/Camry.csv)"
df_camry <- read.csv(filepath)
df_camry <- rename(df_camry,miles=Miles..1000s.,price=Price...1000s.,) # 重命名字段
head(df_camry,5)
```

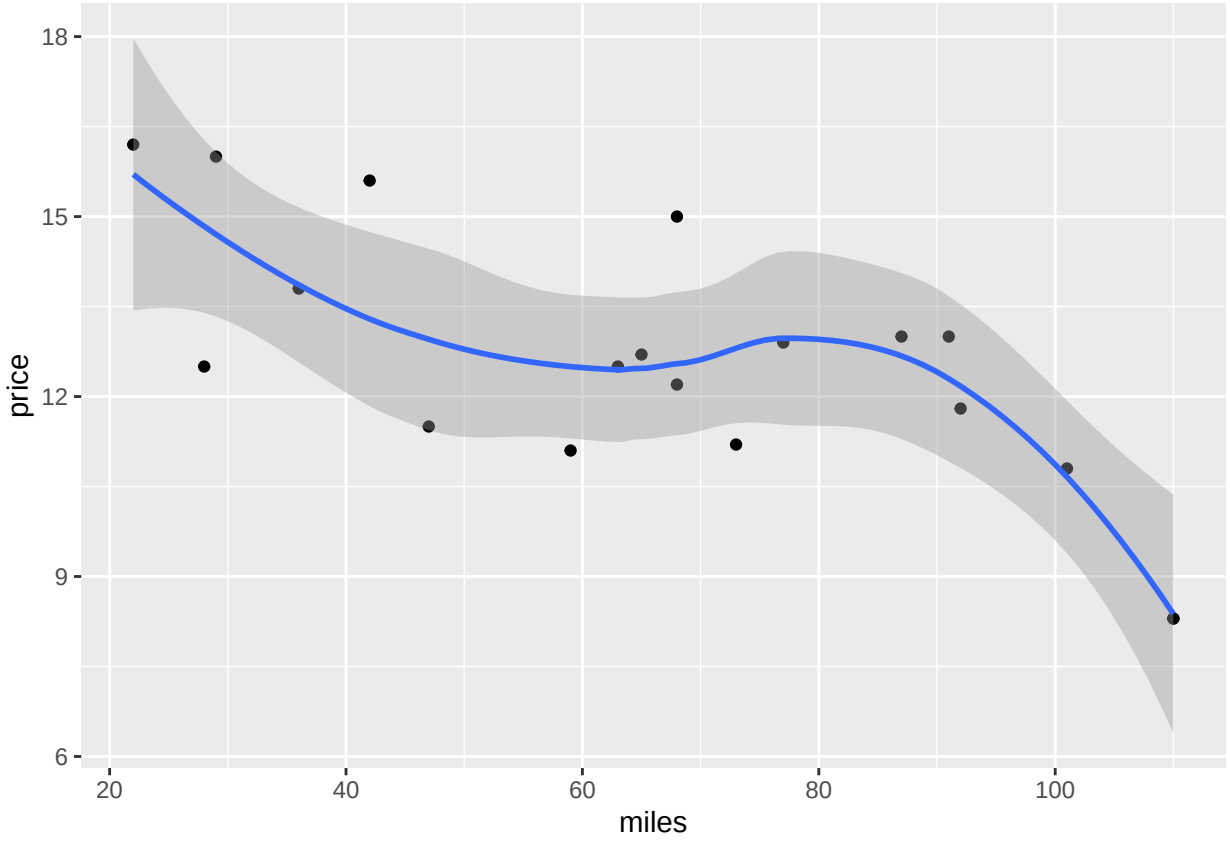
```
##   miles price
## 1    22  16.2
## 2    29  16.0
## 3    36  13.8
## 4    47  11.5
## 5    63  12.5
```

数据分析

```
ggplot(data=df_camry,mapping=aes(x=miles,y=price))+
  geom_point()+
  geom_smooth()
```

1. 制作一个散点图，横轴表示汽车里程数，纵轴表示价格

```
## `geom_smooth()` using method = 'loess' and formula = 'y ~ x'
```



2. 在部分（a）中制作的散点图表明了两个变量之间的什么关系

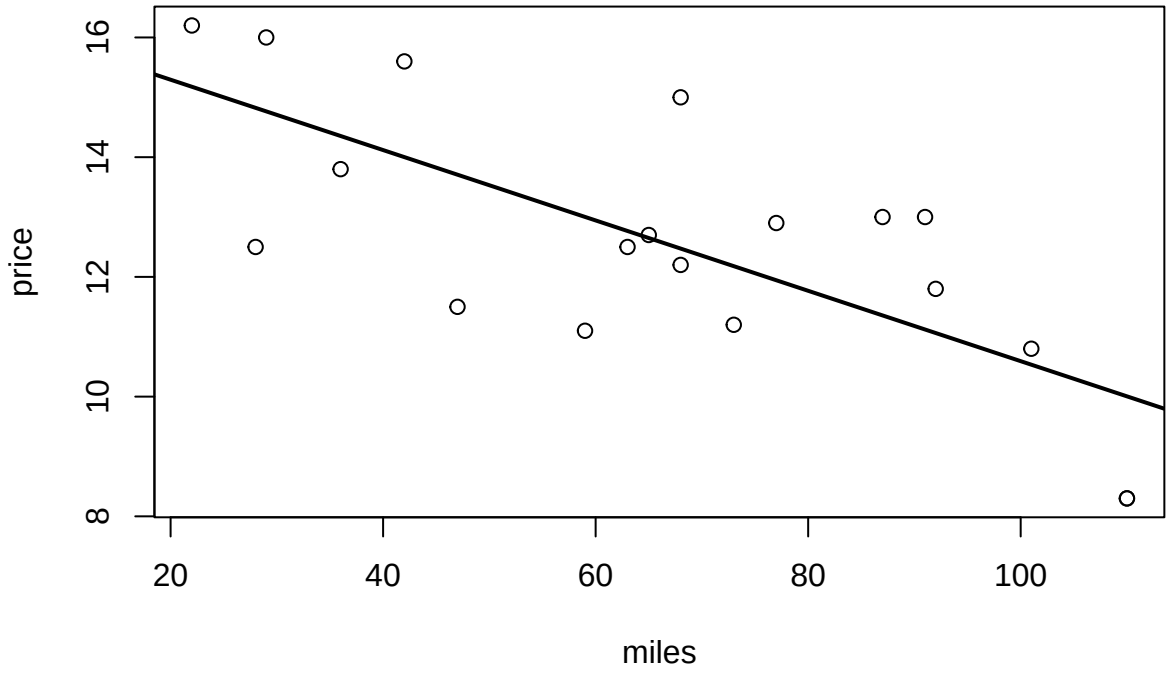
- 整体呈现出随着里程数的增加，销售价格下降的趋势。

3. 开发一个估计回归方程，用于在给定里程数（千英里）的情况下预测价格（千美元）

- 估计回归方程为:price = -0.05877*miles+16.46976

```r
plot(df_camry)
abline(lm(df_camry$price ~ df_camry$miles), color="red", lwd=2)
```

```
## Warning in int_abline(a = a, b = b, h = h, v = v, untf = untf, ...):
## "color"不是图形参数
```




```r
model <- lm(df_camry$price ~ df_camry$miles)
summary(model)
```

```
##
## Call:
## lm(formula = df_camry$price ~ df_camry$miles)
##
## 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 ***
## df_camry$miles -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
```

4. 在 0.05 的显著性水平下检验是否存在显著关系

- 里程数和价格之间存在显著性关系。通过构建建设检验：H0:k=0,H1:k 0，其中 k 为 miles 的系数, 当 k 为 0 说明 price 与 miles 不相关，假设检验结果为：
  - 从 p-value 看：当前 p 值为 0.0003475 小于显著性水平 0.05，即拒绝了原假设，故在显著性水平 0.05 的假设下，即里程数和价格之间存在显著性关系；
  - 从拒绝域来看: 拒绝域为 (-∞,0.001011395) 和 (6.042013,+∞), 样本的 F 值为 19.85 落在右侧拒绝域内，故拒绝原假设，即里程数和价格之间存在显著性关系。

```
ubound = qf(1-0.025,1,17)
lbound = qf(0.025,1,17)
cat(lbound,ubound)
```

```
## 0.001011395 6.042013
```

5. 估计的回归方程是否拟合良好？

- 从 summary 的结果值, 拟合优度为 0.5387, 调整后的拟合优度为 0.5115, 说明里程数是影响销售价格的一个重要的影响因素。

6. 对估计回归方程的斜率进行解释

- 从 summary 的结果值, 斜率为-0.05877, 含义为里程数每增加 1000 英里，价格预计下降 58.77 美金。

7. 价格预测: 里程为 60000 公里, 预测价格，且这是您会向卖家提供的价格吗？

- 根据模型预测出价格为 12943.56 美元，这个价格可以作为参考价格，但最后出价还需结合剩余保修期、品牌售后服务水平和卖方身份如个人还是代理商等综合考虑。

```
x_miles =  60 # 单位为千美元
predict_price = -0.05877*x_miles+16.46976
predict_price
```

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

第 9 题：网站客户流失分析

数据导入

```
filepath <- r"(C:/Users/huxio/Desktop/武汉大学 MEM/2. 商业数据分析/作业 2/WE.xlsx)"
df_we <- read_excel(filepath)
df_we <- rename(df_we
                ,customer_id=客户 ID,is_lost=流失,happy_index=当月客户幸福指数
                ,happy_index_change=客户幸福指数相比上月变化,customer_support=当月客户支持
                ,support_change=客户支持相比上月的变化,service_priority=当月服务优先级
                ,service_priority_change=服务优先级相比上月的变化,log_times=当月登录次数
                ,blog_change=博客数相比上月的变化,log_times_change=访问次数相比上月的增加
                ,used_duration=客户使用期限,log_lag_change=访问间隔变化)
head(df_we,5)
```

```
## # A tibble: 5 x 13
##   customer_id is_lost happy_index happy_index_change customer_support
##         <dbl>   <dbl>       <dbl>              <dbl>            <dbl>
## 1           1       0           0                  0                0
## 2           2       0          62                  4                0
## 3           3       0           0                  0                0
## 4           4       0         231                  1                1
## 5           5       0          43                 -1                0
## # i 8 more variables: support_change <dbl>, service_priority <dbl>,
## #   service_priority_change <dbl>, log_times <dbl>, blog_change <dbl>,
## #   log_times_change <dbl>, used_duration <dbl>, log_lag_change <dbl>
```

数据分析

```
df_we_mean <- select(df_we,2:length(df_we)) %>%
  group_by(is_lost) %>%
```

```
  summarize_all(mean)
kable(select(df_we_mean,1:6),align='l')
```

1. 获取所有变量按照流失和未流失分类的均值

is_lost	happy_index	happy_index_change	customer_support	support_change	service_priority
0	88.60591	5.530213	0.7242696	-0.0092961	0.8295759
1	63.27245	-3.736842	0.3715170	0.0371517	0.4995577

```
kable(select(df_we_mean,1,7:12),align='l')
```

is_lost	service_priority_change	log_times	blog_change	log_times_change	used_duration	log_lag_change
0	0.0326818	16.13894	0.1711487	106.6096	18.81873	3.511454
1	-0.0166962	8.06192	-0.1021672	-95.7678	20.35294	8.486068

```
df_result <- tibble(
  variable = character()
  ,estimate = numeric()
  ,estimate_lost = numeric()
  ,estimate_not_lost = numeric()
  ,statistic = numeric()
  ,p_value = numeric()
  ,parameter = numeric()
  ,conf_low = numeric()
  ,conf_up = numeric()
  ,method = character()
)

simple_variable_t_test <- function(df_result,col){
  x <- df_we[[col]][df_we$is_lost==0]
  y <- df_we[[col]][df_we$is_lost==1]
  mean_x = mean(x)
  mean_y = mean(y)
  mean_df = mean_x - mean_y
  # 假设检验: H0:x_d_miu=0,H1:x_d_miu 0
```

```
  t_result <- t.test(x,y)
  lbound = t_result$conf.int[1] # 置信区间下限
  ubound = t_result$conf.int[2] # 置信区间上限
  p_value = t_result$p.value # p 值
  t_statistic = t_result$statistic # 样本统计量
  t_method = t_result$method # 检验
  parameter = t_result$parameter # 自由度
  # 结果输入
  df_result <- add_row(df_result,variable=col,estimate=mean_df,estimate_lost=mean_x,estimate_not_l
          ,statistic=t_statistic,p_value=p_value,parameter=parameter,conf_low=lbound,conf_up=uboun
  return(df_result)
}

for (col in names(df_we)[3:length(df_we)]){
  df_result <- simple_variable_t_test(df_result,col)
}
df_result
```

2. 计算每个变量流失均值和未流失均值的差值是否显著 (流失为 **0**, 未流失为 **1**)

```
## # A tibble: 11 x 10
##    variable          estimate estimate_lost estimate_not_lost statistic  p_value
##    <chr>                <dbl>         <dbl>             <dbl>     <dbl>    <dbl>
##  1 happy_index        25.3          88.6               63.3        7.62  2.10e-13
##  2 happy_index_chan~   9.27          5.53              -3.74       5.78  1.57e- 8
##  3 customer_support    0.353         0.724              0.372      5.51  6.28e- 8
##  4 support_change     -0.0464       -0.00930            0.0372    -0.632 5.28e- 1
##  5 service_priority    0.330         0.830              0.500      5.14  4.38e- 7
##  6 service_priority~   0.0494        0.0327            -0.0167     0.641 5.22e- 1
##  7 log_times           8.08         16.1                8.06       3.57  4.04e- 4
##  8 blog_change         0.273         0.171             -0.102      2.53  1.16e- 2
##  9 log_times_change  202.          107.                -95.8        1.91  5.63e- 2
## 10 used_duration      -1.53         18.8               20.4       -2.98  3.06e- 3
## 11 log_lag_change     -4.97          3.51               8.49      -4.10  5.22e- 5
## # i 4 more variables: parameter <dbl>, conf_low <dbl>, conf_up <dbl>,
## #   method <chr>
```

3. 以” 流失 “为因变量，其他你认为重要的变量为□变量（提示：**a**、**b** 两步的发现），建□回归□程对是否流失进□预测

- 从上述 P 值知，客户支持相比上月的变化和服务优先级相比上月的变化这 2 个变量不存在显著性差异，故回归方程建立仅考虑除这 2 个变量之外的所有变量。

```
we_model <- glm(formula = is_lost ~ happy_index + happy_index_change + customer_support
      + service_priority + log_times + blog_change + log_times_change +used_duration+log_lag_chang
    ,family = binomial(link = "logit"), data = df_we)
```

```
## Warning: glm.fit:拟合概率算出来是数值零或一
```

```
summary(we_model)
```

```
##
## Call:
## glm(formula = is_lost ~ happy_index + happy_index_change + customer_support +
##     service_priority + log_times + blog_change + log_times_change +
##     used_duration + log_lag_change, family = binomial(link = "logit"),
##     data = df_we)
##
## Coefficients:
##                       Estimate Std. Error z value Pr(>|z|)
## (Intercept)        -2.874e+00  1.215e-01 -23.661  < 2e-16 ***
## happy_index        -5.225e-03  1.161e-03  -4.500 6.78e-06 ***
## happy_index_change -9.501e-03  2.424e-03  -3.920 8.87e-05 ***
## customer_support   -3.522e-02  7.438e-02  -0.474  0.63581
## service_priority   -3.727e-02  7.514e-02  -0.496  0.61985
## log_times           9.104e-04  1.952e-03   0.466  0.64098
## blog_change        -2.357e-05  2.080e-02  -0.001  0.99910
## log_times_change   -1.170e-04  4.069e-05  -2.877  0.00401 **
## used_duration       1.418e-02  5.260e-03   2.696  0.00701 **
## log_lag_change      1.700e-02  4.277e-03   3.975 7.03e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 2553.1  on 6346  degrees of freedom
## Residual deviance: 2445.9  on 6337  degrees of freedom
## AIC: 2465.9
##
## Number of Fisher Scoring iterations: 6
```

```r
df_lost_predict <- df_we[df_we$is_lost==1,]
df_lost_predict <- df_lost_predict %>%  # 流失即为 1
  mutate(prediction = predict(we_model,newdata=df_lost_predict,type='response')) %>%
  arrange(prediction) %>%
  select(customer_id, is_lost,prediction,everything()) # 将预测列移动到流失值后面
head(df_lost_predict,100)
```

4. 根据上□步预测的结果，对尚未流失（流失 =0）的客户进□流失可能性排序，并给出流失可能性最□的前 100 名□户 ID 列表

```
## # A tibble: 100 x 14
##    customer_id is_lost prediction happy_index happy_index_change
##          <dbl>   <dbl>      <dbl>       <dbl>               <dbl>
##  1        4455       1     0.0137         159                   0
##  2        2704       1     0.0147         231                  51
##  3        1006       1     0.0161          37                 -15
##  4         764       1     0.0164         104                  71
##  5        3984       1     0.0165         173                  62
##  6         978       1     0.0167         161                  53
##  7         886       1     0.0176         181                  27
##  8        5560       1     0.0179          73                  73
##  9        1067       1     0.0187         130                  57
## 10        1803       1     0.0187           0                 -12
## # i 90 more rows
## # i 9 more variables: customer_support <dbl>, support_change <dbl>,
## #   service_priority <dbl>, service_priority_change <dbl>, log_times <dbl>,
## #   blog_change <dbl>, log_times_change <dbl>, used_duration <dbl>,
## #   log_lag_change <dbl>
```