R语言笔记

0 通用代码

- Clear your screen清空

`rm=(list=ls())`

- 加载包

`library(mosaic)`

`library(fBasics)`

`library(readxl)`

- 输入数据

  `library(readxl)`
  `L2E1 <- read_excel("D:/IM课程/Moodle 1.3/Exercises Excel files-20201118/Lecture 2/L2E1.xlsx")`
  `View(L2E1)`

- 使用数据

```
attach(L1E1) #使用表格L1E1
LEVELeducation <- LEVELeducation #创建变量
Gender <- Gender
Extroversion<- Extroversion
Age <- Age
```

1 描述性统计（List 1 + 2)

- 绘图
  - 茎叶图

    `stem(variable_name,scale=2)`
  - 直方图

    `gf_histogram(~variable,data=DATA)`

    `hist(variable)`
- 数值计算
  - 数据预处理
    - 删除缺失值 `na.omit(variable)`
    - 数值排序 `sort(variable)`
  - 创建vector

    `DATA1 <- c(4,6,6,5,6,7)`
  - 求中位数

    `median(variable)`
  - 求均值

    `mean(variable)`

> e.g. Interpretation: The mean(average) driver's head injury serverity in head-on collisions is 603.7.

- 求众数

```
getmode <- function(v) {
  uniqv <- unique(v)
  uniqv[which.max(tabulate(match(v, uniqv)))]
}
mode <- getmode(Dioxide)
print(mode)
```

- 离散趋势
  - 标准差

    `sd(variable)`
  - 方差

    `var(variable)`
  - 标准差和方差区别

    > Variance tells us about the total distribution of the data, but standard deviation is even better because it is the average of your variance, meaning it is how far a value on average is from the mean. The standard deviation is best to describe the distribution of the 992 support levels as it also takes into account the number of observations that contribute to the total variance.
  - 极差

    `range(variable)`
  - 计算分位数

    ```
    arranged_L2E7 <- arrange(L2E7)  #对整个表格排序
    quantile(arranged_L2E7$NUMSITES, c(.10))    #10th分位数，means at least 10% of the observations are less than 0
    #或者用sort()函数
    sorted_num <- sort(NUMSITES)
    quantile(sorted_num, c(.10))
    ```
  - 用极差估计标准差

    > Using Chebyshev's Rule, at least 8/9 of the measurements ill fall within 3 standard deviations of the mean. Thus, the range of data would be around 6 standard deviations. Using the empirical Rule approximately 95% of the observations are within 2 standard deviations of the mean. Thus, the range of the data would be around four standard deviations. We would expect the standard deviation to be somewhere between Range 6 and Range 4. 标准差介于R/6~R/4之间
- 描述性统计函数

  `favstats(Variable)`
- 过滤分类后求中位数

  `median(filter(L1E5, Oil == "No")$Dioxide)`

  `median(filter(L1E5, Oil == "Yes")$Dioxide)`

- Summary statistics

```
L2E18C <- cbind(x1=REQUEST, x2=MARKET, x3=ENGINEER, x4=ACCOUNT,
x5=Total)
#组合这几列形成新数据
summary(L2E18C)
```

- 删除数组第n个元素

`PLANTS <- PLANTS[-n]`

- 删除极端值

```
PLANTS_ <- sort(PLANTS)
PLANTS_1 <- PLANTS_[-length(PLANTS)]
PLANTS_1
var(PLANTS_1)
sd(PLANTS_1)
```

2.7 异常值：箱线图和z分数

- 过滤后画箱线图

```
JOINT <- dplyr::filter(L2E9, PLAN == "JOINT")
NONE <- dplyr::filter(L2E9, PLAN == "NONE")
PREPACK <- dplyr::filter(L2E9, PLAN == "PREPACK")
boxplot(JOINT$TIME, ylab='Time', main='Boxplot Joint')
```

- 箱线图

```
boxplot(Score, ylab='Score', main='Boxplot Score')
boxplot(TIME~PLAN, main="Boxplot of three different types of Plans", col=
rainbow(3)) #一个变量有三类时，画在同一个图上
```

- 离群值定位、箱线图参数

`boxplot.stats(TIME[PLAN=="JOINT"])`

- for循环遍历整列求z分数

```
#for循环遍历整列
for (i in 1:length(Score))
{Z_score[i] <- (Score[i]-mean(Score))/sd(Score)
if (abs(Z_score[i])>3)
  print(Score[i])
}
# Intepret the z-score of -1.06: The value (-1.06) is fairly low and is not
an unusual value to observe.
```

2.8 画图表示二元变量关系

- 绘制散点图

```
xyplot(X ~ Y, data=L2E11)
xyplot(X ~ Y,type=c("p","smooth"),data=L2E11) #拟合散点图
```

- 画条形图barplot

```
GROUNDING <- filter(L2E17, Cause =="Grounding")
HULLFAIL <- filter(L2E17, Cause =="HullFail")
UNKNOWN <- filter(L2E17, Cause =="Unknown")
NEW_DATA <-
c(length(COLLISION$Spillage),length(FIRE$Spillage),length(GROUNDING$Spillage
),length(HULLFAIL$Spillage),length(UNKNOWN$Spillage))
#List2-17涉及到了分类计数，然后画条形频数统计图,发现fire和grounding是最多的两个原因
barplot(NEW_DATA,xlab="Causes", ylab="Spillage",col="grey", main="Barplot",
        border="black", names.arg = c("Collision", "Fire", "Grounding",
"Hullfail",
                                      "Unknown"))
```

- 条形图的“误导”

> One way the bar graph can mislead the viewer is that the vertical axis has been cut off. Instead of starting at 0, the vertical axis starts at 12. Another way the bar graph can mislead the viewer is that as the bars get taller, the widths of the bars also increase.

2 概率分布（List 3)

- 二项分布

```
dbinom(x, size, prob, log = FALSE) # 二项分布概率密度
sum(dbinom(3:10,10,.09)) #给出概率密度分布，所以要把大于等于3的离散概率累加起来。x是数
字的向量。n是观察的数量。size是试验的数量。prob是每个试验成功的概率。或者可以用1-
pbinom(2,10,0.09)
```

- 正态分布

```
xpnorm(-3, mean=0, sd=3)     # 正态分布，给出X~N(0,3)时的P(x<=3)的概率
```

- 已知二项分布的均值和标准差，确认某个值是不是异常值

  用二项分布近似正态分布list3-5 （教材P165）

$$\mu \pm 3\sigma = np \pm 3\sqrt{npq}$$

  当这个区间落在（0,n）之间时，正态分布对二项分布才有一个比较好的估计，z分数要加0.5

$$z \approx \frac{(a+.5)-n\cdot p}{\sqrt{n\cdot p(1-p)}}$$

- 验证是否符合正态分布

```
#方法1. 图形感受法：建立直方图或者枝干图，看图像的形状是否类似正态曲线，既土墩形或者钟形，
并且两端对称。
hist(FailTime, col="grey")
histogram(~Desire.res, xlab="residuals", fit="normal", data=L8E9,
col="grey", border="black")

#方法2. 计算区间，看落在区间的百分百是否近似于68%，95%，100%。
#经验法则，验证一个标准差范围
mean(FailTime)+sd(FailTime)
mean(FailTime)-sd(FailTime)
INT1 <- dplyr::filter(L3E6, FailTime> 1.006 & FailTime <2.86)
```

```
INT1#可以发现，经过筛选，有33个值落在一个标准差范围内（66%）
mean(FailTime)+2*sd(FailTime)
mean(FailTime)-2*sd(FailTime)
INT2 <- dplyr::filter(L3E6, FailTime> 0.07769 & FailTime <3.792302)
INT2#49个（98%）
mean(FailTime)+3*sd(FailTime)
mean(FailTime)-3*sd(FailTime)
INT3 <- dplyr::filter(L3E6, FailTime> -0.8509533 & FailTime <4.7209533)
INT3#50个（100%）

#方法3. 求IQR和标准差s，计算IQR/s，如若是正态分布，则IQR/s≈1.3。
#再根据IQR检验是否正态分布（大于1.3）
IQR(FailTime)/sd(FailTime)
This is much smaller than the 1.3 we would expect if the data were normal.
This method indicates the data are not normal.

#方法4. 建立正态概率QQ图，如果近似正态分布，点会落在一条直线上。
qqnorm(Desire.res)
qqline(Desire.res, col="red")

#方法5. Jargue-Bera test计算峰度（3）和偏度（0），p值大，说明和正态分布没有显著差异，若p值很小，则拒绝H0: 符合正态分布
library(fBasics)
jarqueberaTest(FailTime)

#方法6：峰度==3,偏度==0
library(fBasics)
kurtosis(Score)
skewness(Score)

#方法7：比较均值==众数==中位数,注意mode众数不能直接得到，需要用函数c
mean(Score)
median(Score)
getmode <- function(v) {
  uniqv <- unique(v)
  uniqv[which.max(tabulate(match(v, uniqv)))]
}
mode <- getmode(Score)
mode
```

3 抽样分布（List 4）

- 将输入的数据合并成表格

```
x_bar <- c(1,1.5,2,2.5,3,3.5,4,4.5,5)
px_bar <- c(.04,.12,.17,.20,.20,.14,.08,.04,.01)
DATA1A <- cbind(x_bar,px_bar)
```

- 已知总体，利用R实现随机抽样

```r
#rep(x,n)令x的值重复n次
sample_means <- rep(NA, 100)
sample_means

for(i in 1:100){
 samp <- sample(L4E5$INTTIME, 40)#从数据INTTIME里进行随机抽样，样本size=40
 sample_means[i] <- mean(samp)
}

hist(sample_means)
```

- 用样本数据估计总体的置信区间

  `t.test(raw,conf.level = 0.95) #list4-6`

- 样本比例的置信区间

  涉及到大样本小样本问题时，e.g. list4-10，计算一下np\nq\n，说明满足大样本条件。尤其对于二项分布和样本比例的分布，样本容量是计算的前提条件。

4 假设检验 (List 5)

- 确定原假设和备择假设
  - Ha是需要努力找出证据证明之后才会被接受的，例如list5里证明药物是安全的

    问题：某个数据是否超过6.500——超过6.500这个结论是需要找出证据来支撑的，所以是Ha
  - H0是现状或者研究者希望证明的论断
  - 等号一定包含在H0中
  - 弃真和取伪都是相对于H0而言的
- 总体比例的假设检验

$$z = \frac{\hat{p} - p_0}{\sigma_{\hat{p}}} = \frac{\hat{p} - p_0}{\sqrt{\frac{p_0 q_0}{n}}}$$

  注意：分母上用的p和q是假设的量，不是样本统计量。（总体方差的假设检验也用到了假设的方差）

5 两个总体的假设 (List 6)

- Q: The sample mean recall scores for the three groups were Xv = 2.08 and xs = 3.17. Explain why one should not draw an inference about differences in the population mean recall scores on the basis of only these summary statistics.为什么我们不可以直接用计算出的这些均值代表整体水平?

  A: The means given are only sample means. If new samples were selected and sample means computed the values and order of the sample means could change. In addition, the variances are not taken into account.

- R语言进行配对样本t检验，可以给出置信区间，t-score以及p值（也可以做方差检验）

```
t.test(u1,u2, paired = TRUE, alternative = "greater")   #右侧检验，Ha: u1-u2>0
t.test(STANDARD, HUFFMAN, paired = TRUE, alternative = "two.sided",
conf.level = 0.95)
```

- 独立抽样的两样本方差检验

```
var.test(DM, HONEY, alternative = "two.sided", conf.level = 0.90)
```

- What assumptions, if any, are required for the inference from the test to be valid? 各个方法的使用条件：大多数情况的大样本总体不需要满足正态的条件，因为可以使用中心极限定理；小样本的时候常要求总体必须满足正态分布。

6 方差分析 (List 7)

- 方差分析的用途：
  - 两个或多个样本均数间的比较；
  - 分析两个或多个因素间的交互作用；
  - 回归方程的线性假设检验；
  - 多元线性回归分析中偏回归系数的假设检验；
  - 两样本的方差齐性检验等。
- Interpret 方差分析结果

  e.g. Since the p-value is less than α (p=0.000<0.01), H0 is rejected. There is sufficient evidence to indicate differences in the mean recall scores among the three viewing groups at α=0.01. The researchers can conclude that the content of the TV show affects the recall of imbedded commercials.
- Check that the ANOVA assumptions are reasonably satisfied. 用R检验数据是否满足进行方差分析的条件：P316F检验所需条件
  - 从k个组（处理）中独立随机抽取样本
  - 所有K个组抽样总体都具有***近似正态***分布（画直方图观察，对每组数据hist(VIOLENT,col = "grey", border = "black")）
  - k个组总体方差是**相等**的，即"齐方差"（通过计算sd(variable)比较大小，注意要剔除空值，e.g. SEX1 <- na.omit(SEX) ）
- 用R进行方差分析（第一步必须检测异常值，list7-4发现了异常值但是选择忽略）

```
MODEL1 <- aov(IMPROVE~ASSIST)
summary(MODEL1)
```

- 箱线图（教材P68)

```
boxplot(IMPROVE~ASSIST, main="Fig.-1: Boxplot of the three different types
of Assistance", col= rainbow(3))
# IMPROVE:因变量；ASSIST：分组的变量
boxplot.stats(IMPROVE[ASSIST=="NO"]) #其中$stats五个值为，箱线图下虚线，Q1，中位数，Q3，上虚线。$n返回样本量，$conf返回置信区间，默认是95的置信区间。$out返回离群值。
```

- 多重均值比较

```
#多重均值比较：两个组之间进行对比，mean difference反应两个组之间的差异，置信区间是否包含0以及是否显著可以对比两个组的高低以及可靠性
Model1 <- aov(RECALL~RATING)

TukeyHSD(Model1, conf.level = 0.95)#输出差异
plot(TukeyHSD(Model1, conf.level = 0.99),las=1, col="red")#可视化difference
install.packages(gplots)
library(gplots)
plotmeans(RECALL~RATING, main="Fig.-2: Mean Plot with 95% Confidence Interval") #可视化means
```

- 析因实验

```
model <- aov(OVERRUN~HOUSING+WTCLASS+HOUSING*WTCLASS)
summary(model)
```

双因素之间的交互效应通过HOUSING*WTCLASS的p值判断是否显著

各个因素各自的主效应通过该因素对应的p值判断是否显著

主效应

```
> summary(aov(OVERRUN~HOUSING+WTCLASS+(HOUSING*WTCLASS), data=L9E6))
                Df Sum Sq Mean Sq F value   Pr(>F)
HOUSING          3  10788    3596  33.061 6.03e-08 ***
WTCLASS          1    329     329   3.026   0.0973 .
HOUSING:WTCLASS  3    248      83   0.759   0.5303
Residuals       20   2175     109
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

交互效应

7 一元回归（List 8）

- 拟合模型

```
# 使用lm()函数拟合数据
Reg<-lm(formula=yi ~ xi, data = L7E3)
summary (Reg)
```

- 画出因变量和自变量的散点图

`plot(xi, yi, main="Scatterplot", col="red")`

- 在回归图上画拟合直线

```
#在已有图上加出拟合直线，命令为lines(x,y),功能相当于plot(x,y,type="1")但plot创建新图
lines(xi, fitted(Reg), col="red")
#添加一条特定直线（纵截距，斜率）
abline(14, -2.5, col="blue")
```

- 回归结果的截距和斜率的解释

  list8-4 Practically interpret the estimated intercept and estimated slope

  纵截距：Since 0 is not in the observed range of the average annual FEMA relief, Beta(0)_hat has no meaning.

  斜率：For each additional dollar in average annual FEMA relief per capita, the mean average annual number of public corruption convictions per 100,000 residents is estimated to increase by 0.00542.

- 方差分析估计σ2

```
anova(lm(AACC ~ AAFEMA, L7E6))      #y~x
```

```
> #方差分析anova()输出方差分析表
> anova(lm(AACC ~ AAFEMA, L7E6))
Analysis of Variance Table

Response: AACC
           Df Sum Sq  Mean Sq F value Pr(>F)
AAFEMA      1 0.0620 0.061997  2.7859 0.1016
Residuals 48 1.0682 0.022253
```

- 标准差和方差哪个可以进行实际解释（标准差的unit单位）
  The standard deviation in part b can be interpreted practically. The standard deviation is measured in the same units as the data. The variance is measured in square units and is very difficult to interpret.
- 回归中计算给定置信度下的置信区间

```
#先回归lm()，再用confint()对回归结果求置信区间
confint(Reg,level=0.95)
#截距置信区间最重要的信息之一就是根据正负性判断线性关系
Reg<-lm(formula = DESIRE~GENDER+SELFESTM+BODYSAT+IMPREAL, data = L8E2)
confint(Reg,"GENDER",level = 0.95)#引号内指定了求置信区间的变量
```

```
> confint(RegA,level=0.95)
                  2.5 %   97.5 %
(Intercept) -36.68443724 8.409384 截距
USBIRTHS     -0.04844045 1.303834 斜率
```

- 落入拒绝域怎么表述： There is sufficient evidence to indicate gender and impression of reality TV interact to affect desire to have cosmetic surgery at alpha = 0.10.

  没落入拒绝域怎么表述： H0 is not rejected. There is insufficient evidence to indicate that repellent type is a useful predictor of costper-use at alpha = .10.

8 多元回归 （List 9）

- 求多元线性回归最小二乘方程

```
Reg<-lm(formula = DESIRE~GENDER+SELFESTM+BODYSAT+IMPREAL, data = L8E2)
summary(Reg)
```

- 对交互效应的拟合回归：

```
reg <- lm(formula = Income~Agree+Gender+Agree*Gender,data = L8E7)
summary(reg)
```

- 将定性变量转换为定量数值再进行回归拟合

```
Type1 <- ifelse(Type == ' Cream', 1, 0)
RegB <-lm(formula = Cost ~ Type1, data = L8E5)
summary (RegB)
```

- List9-1 解释R^2和adjusted R^2的区别

R2 adjusted = R-Squared(adj)=.396. 39.6% of the total sample variation of the y values is explained by the model containing x1 and x2, adjusted for the sample size and the number of parameters in the model.Adjusted R2 is more reliable than R^2 alone. This is because it takes more elements into consideration. Therefore also, it is usually smaller than R^2.

- List9-7 H0:b3=0 H1:b3<0

  a) Conduct the test, part e. Use α = 5%. Is the researchers' theory supported?

  对参数进行假设检验，直接从回归结果查t检验的p值

  错误做法：P= 0.0293 <5%

  重要：R 语言进行的是**双侧检验**，进行单侧检验时要用p/2去和显著性水平α比较

- 如何判断多重共线性

  计算模型中每对自变量之间的相关系数correlation relating number,

  - 绝对值>0.8:严重的多重共线性
  - 0.2<|r|<0.8:一般moderate 多重共线性
  - |r|<0.2:轻微多重共线性

    Cor(DATA)直接出相关系数矩阵

- 残差分析

```
Desire.lm <- lm(DESIRE ~ SELFESTM+BODYSAT+IMPREAL+GENDER, data=L8E9)
Desire.res <- resid(Desire.lm)
Desire.res
#画图判断残差分布情况
plot(L8E9$SELFESTM+BODYSAT+IMPREAL+GENDER,Desire.res,ylab="Residuals",
xlab="SELFESTM+BODYSAT+IMPREAL+GENDER", main="Residual plot for Desire")
plot(Desire.res,ylab="Residuals", xlab="yinbianliang", main="Residual plot
for Desire")
abline(0,0)
#直方图判断正态性
histogram(~Desire.res, xlab="residuals", fit="normal", data=L8E9,
col="grey", border="black")
#正态分布qq图Normal probability plot:
qqnorm(Desire.res)
qqline(Desire.res, col="red")
```

9 复习补充

- R语言可以直接进行T检验，计算置信区间

```
t.test(RAW, conf.level = 0.95)#总体均值的置信区间
t.test(STANDARD, HUFFMAN, paired = TRUE, alternative = "two.sided", con
f.level = 0.95)  #两总体均值差的置信区间
t.test(POSITIVE, NEUTRAL1, alternative="greater")#两样本单边检验，前者是u1，左侧
检验的备择假设是u1-u2大于0
t.test(Before, After, paired = TRUE, alternative = "greater")  #配对样本检验
var.test(DM, HONEY, alternative = "two.sided", conf.level = 0.90)#方差检验
```

- 分析之前考虑数据预处理，可以考虑通过箱线图检查异常值：

  Like other linear model, in ANOVA you should check the presence of outliers. This can be detected with a boxplot. As there are three populations to study, you should use separate boxplot for each of the population. In R, it is done as:

```
boxplot(IMPROVE~ASSIST, main="Fig.-1: Boxplot of the three different types
of Assistance", col= rainbow(3))
boxplot.stats(IMPROVE[ASSIST=="NO"])
```

- 常用词汇
  - observed significance level :p值
  - significance level显著性水平: α
  - confidence level置信度：1-α
  - test statistic 检验统计量
  - analysis of variance 方差分析