Week1 practical

Vectors

Add “1” to all elements of the created vector.

```
Ad_vector<-my_vector+1
```

Extract the first element of the vector, add 3, and add the resultant number as the sixth element of the vector.

```
My_vector<-c(my_vector, my_vector[1]+3)
```

Take the square root from the resultant vector and assign it as a new vector.

```
Sqrt(my_vector)
```

Convert it to a matrix with 3 columns.

```
Matrix(my_vector, ncol=3)   nrow=…
```

```
rnames <- c('R1', 'R2')    ##行名
cnames <- c('C1', 'C2', 'C3')   ##列名
mat3 <- matrix(1:6, nrow=2, dimnames=list(rnames, cnames))    ##通过设定 dimnames 参数添加行列名
可以直接通过 rownames(),colnames()函数给 matrix 添加行名列名
rownames(mat) <- c("R1", "R2", "R3")
colnames(mat) <- c("C1","C2")
```

Subtract the first four elements of originally created vector from the newly converted matrix and round the result to the second decimal.

```
Round(my_matrix-my_vector[1:4], digits=2)
```

Working with tables in R

1. By the direct command data.frame() and specifying the columns;

```
# 创建姓名向量
names <- c ("Alice", "Bob", "Charlie", "David")
# 创建年龄向量
```

```
ages <- c (25, 30, 22, 28)

# 创建成绩向量

scores <- c (95, 89, 75, 92)

# 使用 data.frame 函数创建数据框

my_data <- data.frame(Name = names, Age = ages, Score = scores)


math_scores <- c(80, 85, 90)

english_scores ≤- c(75, 80, 85)

science_scores <- c(90, 95, 80)

df < - data.frame (math = math_scores, english = english_scores, science =

science_scores, row.names =c（“"张三”，"李四”，"王五"）#可以在data.frame()函数内定义

rowname
```

（1） my_data <- data.frame(Name = names, Age = ages, Score = scores, check.names = FALSE) #通过check.name=FALSE使colname不会在R内自动更新来定义colname

（2） 目前一般使用colnames(my_data) <- c("NewName", "NewAge", "NewScore")

2. By concatenating vectors and combining them into a data frame with several commands

```
c() ->

cbind() -> as.data.frame();

# 创建两个向量

vec1 <- c(1, 2, 3)

vec2 <- c(4, 5, 6)

# 使用 cbind 合并两个向量
```

```
# cbind 和 rbind 也可以用于给 matrix 添加新列和新行（e.g. rbind(matrix/dataframe,vector)）

merged_matrix <- cbind(vec1, vec2)

#得到一个 matrix


Data<-as.data.frame(merged_matrix)

#将 matrix 转化为 dataframe


#添加新的一列

Data$english<-c(...)

Data[, ncol(data)+1]<-c(...)

#添加新的一行

Data[nrow(data)+1,]<-c(...)
```

3. What are the properties of the data frame? Use commands head(), tail(), mode(), class(), and str().

内置数据集 iris 导入：data(iris)

```
> mode(iris)

[1] "list"

> str(iris)

'data.frame':  150 obs. of  5 variables:

 $ Sepal.Length: num  5.1 4.9 4.7 4.6 5 5.4 4.6 5 4.4 4.9 ...

 $ Sepal.Width : num  3.5 3 3.2 3.1 3.6 3.9 3.4 3.4 2.9 3.1 ...
```

```
 $ Petal.Length: num  1.4 1.4 1.3 1.5 1.4 1.7 1.4 1.5 1.4 1.5 ...
 $ Petal.Width : num  0.2 0.2 0.2 0.2 0.2 0.4 0.3 0.2 0.2 0.1 ...
 $ Species     : Factor w/ 3 levels "setosa","versicolor",..: 1 1 1 1 1 1 1 1 1 1 ...
> class(iris)
[1] "data.frame"
> mode(iris$Sepal.Length)
[1] "numeric"
#mode()适用于一串（向量）的数据类型
```

4.保存文件

```
save(file = "filename.RData")


#通过 write.csv()保存创建的
my_data <- data.frame(
  Name = c("Alice", "Bob", "Charlie"),
  Age = c(25, 30, 22),
  Score = c(95, 89, 75)
)
# 指定输出文件路径和名称
output_file <- "my_data.csv"
# 使用 write.csv() 将数据框写入 CSV 文件
write.csv(my_data, file = output_file, row.names = FALSE)
```

#write.csv()默认分隔符为逗号

#write.table()具有”sep”参数可以设置分割为其他（默认为逗号）

output_file <- "my_data.tsv" # 使用制表符分隔的文件

write.table(my_data, file = output_file, sep = "\t", row.names = FALSE)

week1 problem set

Importing and exporting .csv tables

read.csv(file, header = TRUE, sep = ",", row.names = FALSE) #默认格式

 # .txt 文件的读取：

data <- read.table("yourfile.txt", header = TRUE, sep = ",")

“/t”制表符号分割：

data <- read.delim("yourfile.txt", header = TRUE)

data <- **read.csv**("ADS2week1.csv", row.names = NULL)

data <- data[,-1] *# **We do not need the first column***

 # Generally, you should use diverse names for your data. Otherwise, R will be confused.

 # For instance, there is a special command `data()` in R.

data <- **read.csv**("ADS2week1.csv", row.names = NULL)

PS1 <- **c**() *# You create an empty vector where you will store your data*

for(i **in** 1:**ncol**(data)){

linE <- data[**!is.na**(data[,i]), i] 可以用于提取指定列非 NA 的行

 # So, in each round, you choose one column from your data: `data[, i]`

 # and you choose only those rows that do not include NAs: `data[!is.na(data[, i]), ]`

```
# As a result, you get a character vector.

#e.g.

> data[!is.na(data[,1]), 1]

[1] "It"         "is"         "a"          "strange" "data"     "frame!"

> !is.na(data[,1])

[1]  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE FALSE FALSE FALSE

linE <- paste(linE, collapse = " ")

# After this, you just combine the vector in a text string

# You are welcome to try to beautify it further.

#add “ “ between the words

linE <- trimws(linE)

# This line will remove white spaces on the around the string. 去除 vector 中每一句话前后的空白

# For instance, the `paste(data, collapse = " ")` command will add " " on the right side of your string.

# trimws(text, which=”left”)/trimws(text, which=”right”)

PS1[i] <- linE # So, you add your string into the PS1 vector you created earlier

#PS1 是一个拥有四个元素的向量（每个元素都是由 “ ” 拼接单词组成的语句）

}

writeLines(PS1, "ADS2_PS1.txt")

# writeLines() 将文本写入文件而不添加行尾符（例如换行符 "\n"）

# 如果需要换行：

lines <- c("This is the first line.", "And this is the second line.", "Final line.")
```

lines_with_newline <- paste(lines, collapse = "\n")

writeLines(lines_with_newline, "example.txt")

Week2 practical

Task1

data <- read.csv("Chicago2013.csv", row.names = NULL)

head(data)

Can you tell what countries the athletes are from and how many there are from each country?

data.frame(data$Country)

country<-data.frame(data$Country)

table(country)

#table()可以用于计算频数

Can you draw a histogram of finishing times?

hist(data$Time,col='pink')

Select 10 people at random from the dataset and draw a histogram of their finishing time

random_vect<-sample(nrow(data),10)

行数

sorted_vect<-sort(random_vect)

selected_rows<-head(sorted_vect,10)

#获得随机的十个行号并进行排序

selected_data<-data[selected_rows,]

hist(selected_data$Time,col='pink')

Task2

```
data_man=rnorm(45,172,7)

data_woman=rnorm(55,158.5,6)

print(mean(data_man))

print(sd(data_man))

boxplot(data_man)

boxplot(data_woman)

max(data_man)

min(data_woman)
```

How many students are taller than you are?

```
print(length(data_man[data_man>165])+length(data_woman[data_woman>165]))
```

Give a short summary about the mean, median, and quantile distribution of the values in both fictional samples.

```
summary(data_man)

summary(data_woman)
```

Task3

Write a command that creates a group of 26 students, and assigns a day of the year to each of them as their birthday.

```
birth<-sample(365,26)
```

Are there shared birthdays in this group?

```
length(birth)>length(unique(birth))

#Judge whether it is true
```

How would you go about computing the overall probability of a shared birthday for n=26?

```
1-prod((365:(365-26+1))/365)
```

#prod()*是连积乘“：”可以用于指代从 365 一直到（365-26+1）*

How about computing and plotting the probabilities of shared birthdays from n=1 to n=50?

```
data_birth=c()
for (n in 1:50){
  data_birth=c(data_birth,1-prod((365:(365-n+1))/365))
}
plot(data_birth,main="Birthday problem",xlab="class size",ylab="Probability of a shared birthday")
```

Bonus

Let's try to work with probability distributions. Recently, your colleagues measured the activity of serum alanine aminotransferase (ALT) in young healthy mice. Here are the values: 33.45, 24.67, 24.16, 21.27, 26.86, 27.38, 27.91, 26.15, 31.63, 28.12 IU/L.

Considering that this sample represents the whole population, calculate its mean value and standard deviation (SD) and identify the probability of getting the following readings derived from the same population randomly:

1. 40.2 IU/L;

2. higher than 33 IU/L;

3. 22–25 IU/L;

4. 27–31 IU/L;

5. which values are within 40–65% highest values;

6. which value is less than 99.995% of all the other values derived from this population.

```
data<-c(33.45,24.67,24.16,21.27,26.86,27.38,27.91,26.15,31.63,28.12)
print(mean(data))
print(sd(data))
print(dnorm(40.2,mean=mean(data),sd=sd(data)))
print(pnorm(33,mean=mean(data),sd=sd(data),lower.tail=FALSE))
print(pnorm(25,mean=mean(data),sd=sd(data),lower.tail=TRUE)-pnorm(22,mean=mean(data),sd=sd(data),lower.tail=TRUE))
print(pnorm(31,mean=mean(data),sd=sd(data),lower.tail=TRUE)-pnorm(27,mean=mean(data),sd=sd(data),lower.tail=TRUE))
print(qnorm(0.99995,mean=mean(data),sd=sd(data)))
```

Week2 problem set

In order to count students with grades lower or higher than two numbers, make use of the

“or” operator: In R, you can write A | B for “A or B”

grade_81_91 <- sum(class<81 | class>91)

grade_81_91

rep(x,n)

#获得 n 个重复的 x

Round(n) #将 n 四舍五入

Round(n,1) #保留一位小数

Scenario 1: Random correct answers

The instructors have randomly chosen A, B, C, or D to be the right answer for each

question.

If we want to run a quick simulation, we can generate a list of correct answers by drawing

20 random numbers between 1 and 4 (1 standing for A, 2 for B etc.)

correct <- sample(1:4,20,replace=TRUE)

Similarly, we can generate a list of a student’s guesses:

guesses <- sample(1:4,20,replace=TRUE)

how many correct answers that student can get just by guessing

score <- sum(correct==guesses)

And is that score equal to or bigger than 10?

```
score >= 10
```

then we can repeat it for many times

the probability of getting any 10 questions right (and the other 10 wrong) is

In R:

```
0.25^10*0.75^10*choose(20,10)
```

You don't only pass with 10 correct answers, you also pass with 11, 12, 13, 14, etc.

And then we just have to sum over all those to get the passing probability.

```
p_passing = 0

for (s in 10:20) {

p_s = 0.25^s*0.75^(20-s)*choose(20,s)

p_passing = p_passing+p_s

}

p_passing
```

Week5 practical

Width of the sampling distribution and sample size

You were asked to take 1000 samples of size 5 and record the mean – i.e. create a sampling distribution for samples of size 5. You were then asked to repeat this for samples of size 100.

This time, we want to be systematic about it. We would like to know what's the **standard deviation of the sampling distribution** as a function of sample size (for sample sizes between 5 and 100).

```
population<-rnorm(1e6,100,5)
sd=c()
for (n in 5:100){
  #sample size between 5 and 100
   mean=c()
  for(i in 1:1000){
     #重复 1000 次得到 sampling distribution
     sample_1<-sample(population,n)
     sample_mean<-mean(sample_1)
     mean=c(mean,sample_mean)
  }
  sd_sample=sd(mean)
  sd=c(sd,sd_sample)
}
#注意{}包括哪些 code（尤其是当循环套循环的时候）
plot(sd,ylab="sd of sampling distribution",xlab="sample size")
```

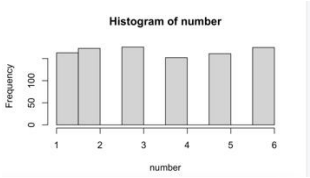
Rolling dice （验证 the central limit theorem in general）

• Roll a six-faced dice in R

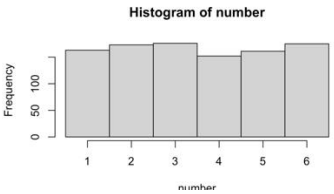
• This should be a uniform distribution – every number from 1 to 6 is equally likely to appear. Roll your dice 1000 times and visualize the outcome to convince yourself that this is true

• If you drew a histogram and it looks a bit weird, this is because of the way that R decides where to break between each of the bins. Try adding breaks=0.5:6.5 to your hist command to manually set break points.

```
number=c()
for(i in 1:1000){
  dice<-sample(1:6,1)
  number=c(number,dice)
}
hist(number)
```




```
hist(number, breaks=0.5:6.5) #break 默认的 size=1 如果需要调整则 break=c(0.5,6.5)*0.5
```



• Everything changes though when you roll 2 dice and add up their numbers. You may know from your own experience what number is the most frequent. Why?

• Let's do this 1000 times and see what we get

• Does this look normal to you? Why or why not?

```
sum=c()
for(i in 1:100){
  dice1<-sample(1:6,1)
  dice2<-sample(1:6,1)
  sumnumber=dice1+dice2
  sum=c(sum,sumnumber)
}
hist(sum)#the sum is normally distributed
```

Dragon wingspans (check the central limit theorem)

• The file dragons.csv contains wing spans for a random sample of 500 dragons. Import the file and visualize the data to convince yourself that this is indeed a non–normal distribution.

• Create a sampling distribution for a sample size of your choice. Here is one we did with n=10

```
dragon<-read.csv("dragons.csv")
mean=c()
for (i in 1:1000){
  sample_wing<-sample(dragon$wingspan,10)
  mean_sample<-mean(sample_wing)
  mean=c(mean,mean_sample)
}
hist(mean)
```

Week5 problem set

Rolling Dice (some mathematical problems on it)

You can go through all the numbers and convince yourself that for i where $2 \leqslant i \leqslant 7$, the probability $P(i)$ of throwing i with two dice is:

$P(i) = P(14 - i) = (i - 1)/36$

the diagram looks like a pyramid

Check the normality– remember the 68–95–99.7 rule!

```
minus1sd = mean(dicerolls)–sd(dicerolls)

plus1sd = mean(dicerolls)+sd(dicerolls)

sum(dicerolls>minus1sd & dicerolls<plus1sd)/10

#实验重复了 1000 次“/10”可以变为 100 来算百分比
```

[1] 65.7

#按照 mean+/–2sd, mean+/–3sd 逐个验证是否与 68，95，99.7 相近

Bean machines

Let's make our own bean machine! Let's assume there are 8 layers of pegs. We don't need to simulate the beads physically going through the machine. All we need to know is that each bead has to make 8 left/right decisions and that we want to record the position of the bead at the end of those decisions.

My code:

```
pos=c()

sum=0

for(x in 1:1000){

  for(i in 1:8){
```

```
    lr<-sample(1:2,1)
    sum<-sum+lr
  }
  pos=c(pos,sum)
  sum=0
}
hist(pos)
```

Code in note:

```
ends <- c()
for (i in 1:1000) {
end <- sum(sample(c(0,1),8,replace=T))
ends <- c(ends, end)
}
hist(ends, main="Bean machine", col="thistle", xlab="", axes=F, ylab="")
#col=”thistle” set the color, axes=F hide the axis
```

What would happen if the probability to go left or right at each peg was not 50:50? For instance, what if there was an 80% probability to go left and only a 20% probability to go right at each peg?

My code:

```
pos=c()
sum=0
for(x in 1:1000){
```

```
  for(i in 1:8){
    probability<-sample(1:5,1)
    if(probability==5){
      lr=2
    }
    if(probability<5){
      lr=1
    }
    sum<-sum+lr
  }
  pos=c(pos,sum)
  sum=0
}
hist(pos)
```

Code in note:

```
sample(c(0,1),8,replace=T,p=c(0.8,0.2))
```

sample() 函数还可以设置选择的概率

```
# we can just use the parameter "p" in sample() to determine the probability
```

Class grades

In a previous problem set, we modelled the grades that students received on a multiple-choice test as a normal distribution. We added as a caveat that this is an assumption, and that the distribution could have looked different.

• From what you have learned this week, do you think test scores can be modelled as normally distributed? Why or why not? Think about how a test score is computed, and work from there.

For simplicity, let's just look at multiple-choice quizzes with four answer choices per question, where exactly one choice is correct and where all questions are weighted equally.

Of course, a test score is just a sum of individual question scores, and we can model each question score as a binary random variable (the two possible outcomes being "correct" and "incorrect").

Note that in order to do this, we need not assume that the underlying process is truly "random". If it is (i.e. the student is just guessing), then we would have $P(incorrect) = 0.75$ and $P(correct) = 0.25$. But even if the student is not guessing, we can still model this as a random variable, just (hopefully) with a bigger $P(correct)$.

Let's assume that this $P(correct)$ is the same for all questions and all students. Since we are computing the overall quiz score by summing the number of correct answers, we can again use the Central Limit Theorem - as long as there are enough questions, overall quiz scores will be approximately normally distributed.

But what if $P(correct)$ is not the same for all qusetions and all students?

Then it depends. If questions are just of different levels of difficulty, then you can still model the quiz as a sum of binary random processes and just use the average $P(correct)$. You are only interested in the sum (overall score) anyways. (Basically if a test contains one very easy question and one very hard question, the overall score will be the same as if they were just two medium questions instead.)

If different students have different probabilities of answering a question correctly, then this averaging of $P(correct)$ will not work anymore, and we may indeed end up with a distribution that is not normal.

Week7 practical

```
install.packages("tidyverse")
```

```
library(tidyverse)
```

Practice 1. West Nile Virus (WNV) Mosquito Test Background

```
wnv <- read.csv("Week_7_WNV_mosquito_test_results.csv" )
```

```
head(wnv, 10)
```

```
summary(wnv)
```

```
#给出每一列的 min max...
```

```
str(wnv)
```

```
class(wnv)
```

```
## [1] "data.frame"
```

```
class(wnv$TEST.DATE)
```

[1] "character"

table(mosquito$LOCATION)

#得到 value 的 frequency

Screen and diagnosis.

• Is this a long or wide form?

• Are the variable names informative and precise? (**Hint**: The name of the first variable SEASON.YEAR is not accurate. Please change it to YEAR.)

• Are there missing values? anyNA()?

• Are there duplicated values?

• Are there any strange patterns? You will need to make different diagnostic plots to spot any strange patterns. Try to be creative!

Our plan is:

• Rename the SEASON.YEAR column.

• Reformat the data set so all the columns have appropriate types of data. For instance, TEST.DATE

must be treated as a **date**.

• Separate LOCATION to LONGITUDE and LATITUDE.

• Check for the presence of duplicated rows and NA values.

Codes in note:

anyNA(wnv)

duplicated(wnv) %>% sum()

my codes:

is.null(mos)

sum(is.null(mos))

sum(duplicated(mos))

- Is.null()没效果
- is.na() 检查对象中的元素是否包含缺失值 NA，用于检查缺失值的存在。

How to delete these duplicated rows and rows containing “na”?

Anyduplicated→unique()

anyNA>na.omit()

```
# Let's check the `TEST.DATE` column

class(wnv$TEST.DATE)

## [1] "character"

# The format is wrong

wnv <- wnv %>%

rename(YEAR = SEASON.YEAR) %>%

relocate(1, 6) %>%

#将第一列移动到第 6 列的位置上

relocate()函数的一些常见参数包括：

    ...：列的名称，可以指定多个列（用逗号隔开）。
    .before, .after：指定列的新位置，可以是列名或位置索引。

relocate(Salary, .after = Age) 将 Salary 列移动到 Age 列的后面。

mutate(TEST.DATE = as.POSIXct(wnv$TEST.DATE, format = "%m/%d/%Y %H:%M:%S",

tz="America/Chicago"))
```

```
attributes(wnv$TEST.DATE)

## $class

## [1] "POSIXct" "POSIXt"

## $tzone

## [1] "America/Chicago"

dat1 <- wnv$TEST.DATE[1] // dat1<-mos[1,'TEST.DATE']

#得到 dat1

attributes(dat1)$tzone <- "America/Los_Angeles"

# Let's check all the columns for anything strange:

for(i in c(1, 2, 4:ncol(wnv))){

writeLines(colnames(wnv)[i])

print(table(wnv[, i]))

writeLines("\n")

}

#输出 1，2，4 到所有列的相关 value 的一个频数表


wnv_out <- which(wnv$NUMBER.OF.MOSQUITOES> 70)

#获得符合该条件的 index

wnv_out_block <- wnv %>%

slice(wnv_out) %>%

select(BLOCK) %>% unlist()

#unlist()可以将值转化为一维向量“…”
```

```
# Let's see values around this block

wnv %>%

filter(BLOCK == wnv_out_block)

# Let's see values around this row

wnv %>%

slice((wnv_out-4):(wnv_out+4))


#获得频数表前三个频数的名称

table(wnv$LOCATION) %>% head(3) %>% dimnames()

## [[1]]

## [1] "" #总的  不是 cell 中数值

## [2] "(41.644720066326094, -87.60185152802353)"

## [3] "(41.64831068933974, -87.55963204714429)"


将空值转化为 NA

wnv$LOCATION[wnv$LOCATION == ""] <- NA

# As you can see, each value is in brackets. We do not need them.

# “perl” 使用了正则表达式 默认为 False

wnv$LOCATION <- wnv$LOCATION %>%

gsub(pattern= "[()]", replacement = "", wnv$LOCATION, perl = T)

# Now, we can separate our `LOCATION` column into two more columns

#gsub()函数可以用于替换
```

```
wnv <- separate(data = wnv, col = LOCATION,

into = c("LATITUDE", "LONGITUDE"),

sep = ",", remove = F, fill = "left", convert = T)
```

- remove = F: 不删除原始列。
- fill = "left": 如果某些行的值不足两个，用左侧填充。
- convert = T: 尝试将拆分的结果转换为数值型。

My codes:

```
mos_new<-mos_1 %>% mutate(LOCATION=gsub("\\(","",LOCATION))

mos_new<-mos_new %>% mutate(LOCATION=gsub("\\)","",LOCATION))

head(mos_new)

mos_new1<-mos_new %>% separate(LOCATION,c("LATITUDE","LONGITUDE"),sep=", ")
```

summarize the data:

```
# ggplot2----

wnv_plot <- ggplot(data = drop_na(wnv)) +

theme_classic()

# By Year

wnv_plot +

geom_boxplot(mapping = aes(x = as.factor(YEAR),
```

In ggplot2, the as.factor() function is often used when you want to treat a numeric variable as a categorical (factor) variable for plotting purposes.

```
y = NUMBER.OF.MOSQUITOES))

# By Latitude
```

```
wnv_plot +
geom_boxplot(mapping = aes(x = round(LATITUDE, 3),
y = NUMBER.OF.MOSQUITOES,
group = round(LATITUDE, 3)),
position = "identity",
outlier.shape = NA) +
scale_x_continuous()
#调整刻度范围
# By Longitude
wnv_plot +
geom_boxplot(mapping = aes(x = round(LONGITUDE, 3),
y = NUMBER.OF.MOSQUITOES,
group = round(LONGITUDE, 3)),
position = "identity",
outlier.shape = NA) +
scale_x_continuous()
# By Trap type
wnv_plot +
geom_boxplot(mapping = aes(x = TRAP_TYPE,
y = NUMBER.OF.MOSQUITOES,
group = TRAP_TYPE),
#确保每一个箱子对应一个 trap_type
```

```
position = "identity")

#让箱线位置与 x 轴位置相对应
```

Practice 2. Tests for antibodies to trachoma PGP3 antigen

```
pgp3 <- read.table(file = "Week_7_Tests_PGP3.txt", sep = "\t")

pgp3[which(duplicated(pgp3)),]

pgp3[which(pgp3$SampleID %in% c(2,7,46,201)),]

##%in%可以代表包含这些数值

# Remove duplicated rows

pgp3 <- pgp3 %>% .[-which(duplicated(.)),] %>%

# And convert to the wide format

spread(key = measured, value = value)


#列举频数以及 NA 个数

for(i in 1:ncol(pgp3)){

if(i == 1){

writeLines(colnames(pgp3)[i])

writeLines("NAs in the column:")

sum(is.na(pgp3[, i])) %>% print()

writeLines("\n\n")

}else{

writeLines(colnames(pgp3)[i])

table(pgp3[, i]) %>% print()
```

```
writeLines("NAs in the column:")

sum(is.na(pgp3[, i])) %>% print()

writeLines("\n\n")

} }


# Change codes for `sex`

pgp3$sex[which(pgp3$sex == 1)] <- "Male"

pgp3$sex[which(pgp3$sex == 2)] <- "Female"

# Substitute "" with NAs

pgp3$age.f[pgp3$age.f == ""] <- NA

# Reformat and reshape the dataset

pgp3 <- pgp3 %>%

mutate_at(.vars = c(1,2,5), .funs = as.factor) %>%

mutate_at(.vars = c(3,4), .funs = as.numeric) %>%

# mutate_at()分别转化为 factor 类型和 numeric 类型

# di<-diamonds%>%mutate_at(vars(5),funs(sqrt))

relocate(1,5,2,4,3) %>%

#relocate 这里表示重新排列为 15243 的顺序

# relocate(.data, ..., .before, .after)

例如，将数值列移到 name 列的后面：

df %>%

  relocate(where(is.numeric), .after = name)
```

```
gather(key = "Time.point", value = "OD", c(4,5)) %>%

drop_na(c(2,3)) %>%

# It is used to remove rows with missing values in columns 2 and 3.

mutate(Time.point = factor(Time.point, levels = c("elisa.pre.od", "elisa.od"), ordered =

T)) %>%

#转化为 factor 类型  根据 level 参数内相关变量的值进行排序

droplevels()



ggplot(data = pgp3, aes(x = age.f, y = OD, fill = Time.point)) + geom_boxplot(color =

"black") +

theme_classic() +

scale_fill_manual(labels = c("elisa.pre.od", "elisa.od"),

name = "Measurement",

values = c("#ADD8E6", "#FF6347"),

breaks = c("elisa.pre.od", "elisa.od")) +

theme(legend.position = "right") +

guides(color = guide_legend(override.aes = list(size = 3) ) )

#设置颜色图例的位置大小
```

My codes:

```
testx<-gather(test_nna,key="time.point",value="ELISA.od", elisa.od, elisa.pre.od,

factor_key = TRUE)
```

#factor_key=True 如果原数据类型是 factor 则会保留

testx$age.f<-as.factor(testx$age.f)

testx$ELISA.od<-as.numeric(testx$ELISA.od)

e<-ggplot(testx,aes(x=age.f,y=ELISA.od,fill=time.point))+geom_boxplot()+theme_classic()

Week7 problem set

apply(is.na(diamond_sample), 2, which)

#返回各列中 NA 值所在的 index

#”2”指的是维度

Treat “NA”

anyNA(diamond_sample)

dim(diamond_sample)

data.noNA = diamond_sample[**complete.cases**(diamond_sample),]

dim(data.noNA)

#complete.cases 返回有关于有无缺失值的逻辑值

#diamond_sample[**complete.cases**(diamond_sample),]用于筛选不含 NA 的行

Treat duplicated value

frw.idx = which(duplicated(data.noNA))

#duplciated() will only give you the duplicated rows,

 # but not the original rows, so we need the next line to get the originals

rvs.idx = which(duplicated(data.noNA, fromLast = TRUE))

data.noNA[c(frw.idx, rvs.idx),]

#通过合并从后向前和从前向后找的行 可以得到所有的重复行以及其原始行

```
data.noNA.noDup = data.noNA[!duplicated(data.noNA),]
```

Treat strange pattern

```
data.noNA.noDup$volume = data.noNA.noDup$x * data.noNA.noDup$y *

data.noNA.noDup$z %>% round(2)


plot(x = data.noNA.noDup$carat, y = data.noNA.noDup$volume,

pch = 20, col = "darkgoldenrod4",

las = 1, xlab = "carat", ylab = "volume",

main = "diamond carat ~ volume", bty = "l")

text(data.noNA.noDup$carat, data.noNA.noDup$volume,

labels = data.noNA.noDup$X, col = "dimgray",

cex = 0.7, pos = 4)
```

- `y = data.noNA.noDup$volume` ：指定散点图的 y 轴数据来自 `data.noNA.noDup` 数据框的 `volume` 列。
- `pch = 20` ：指定散点的形状为实心圆。
- `col = "darkgoldenrod4"` ：指定散点的颜色为深金黄色。
- `las = 1` ：调整坐标轴标签的方向。
- `xlab = "carat"` ：x 轴标签为 "carat"。
- `ylab = "volume"` ：y 轴标签为 "volume"。
- `main = "diamond carat ~ volume"` ：设置图表标题为 "diamond carat ~ volume"。
- `bty = "l"` ：指定绘制图表的边框类型为线条。

2. `text(data.noNA.noDup$carat, data.noNA.noDup$volume, labels = data.noNA.noDup$X, col = "dimgray", cex = 0.7, pos = 4)` ：

- `text()` 函数用于在散点图上添加文本标签。
- `data.noNA.noDup$carat, data.noNA.noDup$volume` ：指定文本标签的位置，x 轴坐标为 `carat` 列，y 轴坐标为 `volume` 列。
- `labels = data.noNA.noDup$X` ：设置文本标签内容为 `data.noNA.noDup` 数据框中的 `X` 列。
- `col = "dimgray"` ：设置文本标签的颜色为暗灰色。
- `cex = 0.7` ：设置文本标签的缩放比例为 0.7，使其相对于默认大小稍微小一些。
- `pos = 4` ：设置文本标签相对于位置的位置代码，这里是右侧。

```
data.noNA.noDup[data.noNA.noDup$X == 9, c(9:11)]

data.noNA.noDup.noStrg = data.noNA.noDup[-which(data.noNA.noDup$X == 9), ]
```

Correct typos in the dataset

Screen

```
table(data.noNA.noDup.noStrg$cut)
```

Diagnosis

```
data.noNA.noDup.noStrg %>%

filter(cut == "Idea")

## X carat cut color clarity depth table price x y z volume

## 1 95 0.33 Idea G IF 61.7 58 968 4.42 4.46 2.74 54.01417
```

Treat

```
data.noNA.noDup.noStrg.noTypo <- data.noNA.noDup.noStrg

data.noNA.noDup.noStrg.noTypo[data.noNA.noDup.noStrg.noTypo$X == 95,

"cut"] <- "Ideal"
```

Find outliers in the dataset.

Screen

```
plot(x = data.noNA.noDup.noStrg.noTypo$carat, y =

data.noNA.noDup.noStrg.noTypo$price,

pch = 20, col = "darkslateblue", las = 1, xlab = "carat", ylab = "price", main = "diamond

carat ~ price")

text(data.noNA.noDup.noStrg.noTypo$carat, data.noNA.noDup.noStrg.noTypo$price,

labels = data.noNA.noDup.noStrg.noTypo$ID, col = "dimgray", cex = 0.7, pos = 4)
```

Diagnose

```
data.noNA.noDup.noStrg.noTypo %>% filter(price > 40000)

#Even those two outliers look very suspicious, but we didn't have evidence to show it's

wrong.So we decide to let's view some IDs surrondings to those two strange IDs.

data.noNA.noDup.noStrg.noTypo[which(data.noNA.noDup.noStrg.noTypo$X ==
```

27−2):which(data.noNA.noDup.noStrg.noTypo$X == 27+2),]

Treat

outlier.idx = rep(0, nrow(data.noNA.noDup.noStrg.noTypo))

#创建了一个名为 outlier.idx 的向量，其长度为 data.noNA.noDup.noStrg.noTypo 数据框的行数，并将其中的所有元素初始化为 0。

outlier.idx[which(data.noNA.noDup.noStrg.noTypo$X == 27)] = 1

outlier.idx[which(data.noNA.noDup.noStrg.noTypo$X == 96)] = 1

data.noNA.noDup.noStrg.notypo.mkOtlr = data.frame(data.noNA.noDup.noStrg.noTypo,

otlr = outlier.idx)

#创建了异常值列 异常值被标为 1 其他均为 0

Week8 practical(only my codes)

Background information:
Imagine that a country runs an exam system using normative scoring, where each student first gets a raw score based on performance in the exam, and later student's normative score is computed as corresponding to the percentile of all students in the country, hence it can be effectively approximated using a uniform distribution between 0 and 100 (for the sake of simplicity let's assume all scores are not rounded). So, for example, if raw scores of 10 students were 17, 17.5, 16, 16.4, 18.9, 18.3, 18.6, 20, 15.5, 18.1, their normative scores would be 35, 45, 15, 25, 85, 65, 75, 95, 5, 55 respectively.

Question 1

Let's assume that a class has 26 students whose score distribution follows the same one as in the whole country. We want to know the probability that the mean of their normative scores is lower than 40. How can we find this out using simulation?

Normative scores->百分制 直接转化为 runif(26,0,100)

count=0

loopnum=1000

```
meanstu<-c()

for (i in 1:loopnum){

  norm_stu<-runif(26,0,100)

  meanstu<-c(meanstu,mean(norm_stu))

  if (mean(norm_stu)<40){

    count=count+1

  }

}

prob<-count/loopnum

hist(meanstu)
```

Question 2

Now imagine that one class of 26 students had a careless administrator who didn't notice the 5th exam question and only printed the first four. Let's

assume that as a result of this, the normative scores of students from this

class followed a uniform distribution between 0 and 80 (with the mean of 40). We now want to know what is the probability that this class did better than another class of 26 that didn't have such bad luck. How can we do this using simulation?

```
library(tidyr)

library(ggplot2)

uni_80<-replicate(10000,mean(runif(26,0,80)))

uni_100<-replicate(10000,mean(runif(26,0,100)))

mean(uni_80>uni_100)
```

```
data.frame("score"=c(uni_80,uni_100),"student"=rep(c("badluck","goodluck"),each=10000)
)%>%ggplot()+theme_classic()+geom_histogram(aes(x=score,y=..density..,fill=student))
```

#..density..可以更好展示概率密度分布情况 而不仅仅是频率（这里指的是 score 的分布）

Question 3

As this exam had a clear scoring system and students could take a copy of their papers, a student Leonie found that she got a 64. Based on statistics from previous years of students taking a similar exam, she has found out that their raw scores follow a normal distribution with a mean of 50 and standard deviation of 10 (for which you can use function *rnorm*).

What would be the expected normative score Leonie likely got? Why? Leonie and her three friends got the following raw scores: 64, 63, 62, 59. Her friend Sheldon is a very bright student from another class. He and his friends got the following raw marks: 70, 63, 61, 56. As a result, Sheldon is boastful that their average is higher. However, having a good understanding of data science Leonie thinks that her team will have the last laugh once the normative scores are out. Is she right? Why or why not? Use simulations and plots to support your argument.

```
pnorm(64,50,10)*100

#Leonie likely got 91.9 as her normative score

lteam<-mean(pnorm(c(64,63,62,59),50,10)*100)

steam<-mean(pnorm(c(70,63,61,56),50,10)*100)

lteam>steam

#Leonie's team wins
```

Question 4

Now remember the unlucky class for which one exam question was not printed. What is wrong with the assumption in bold in question 2? Let’s find and plot a distribution of normative (percentile-based) scores for this class if 1 out of 5 answers are missing, hence their raw scores are on average 20%

lower but follow the same distribution (normal with mean = 40 and standard deviation = 8).
How does this distribution look? Why is it such? What is its mean value? What is the probability that this class got a higher mean normative score than another class of 26 that didn't have bad luck of missing one problem?

```
stu1<-rnorm(26,mean=40,sd=8)

stu2<-rnorm(26,mean=50,sd=10)

norm_stu1<-c()

norm_stu2<-c()

for (i in stu1){

  norm_stu1<-c(norm_stu1,pnorm(i,40,8))

}

mean(norm_stu1)

for (i in stu2){

  norm_stu2<-c(norm_stu2,pnorm(i,50,10))

}

mean(norm_stu2)

mean(norm_stu2>norm_stu1)

#the probability of this class got a higher mean normative score is 34.6%
```

Does the school's principal have a valid reason to worry that the unlucky class would be the worst in the country, assuming it has about 10000 classes of the same size?

```
mean_country<-c()

for (i in 1:10000){

  stucountry<-rnorm(26,50,10)

  stuclass<-rnorm(26,40,8)
```

```
  norm_country<-c()
  for (i in stucountry){
    norm_country<-c(norm_country,pnorm(i,50,10))
  }
  norm_class<-c()
  for (i in stuclass){
    norm_class<-c(norm_class,pnorm(i,40,8))
  }
  mean_country<-c(mean_country,mean(norm_class>norm_country))
}
hist(mean_country)
```

Now assume the number of students per class actually varies, following a uniform
distribution between 5 and 40 (with 10000 classes in total). Would that alleviate the
principal's worries of having a class with the worst normative score average in the country? Why or why not?

```
class_size<-sample(5:40,10000,replace = TRUE)
mean_classc<-c()
for (i in 1:10000){
  classc<-rnorm(class_size[i],50,10)
  norm_classc<-pnorm(classc,50,10)
  mean_classc<-c(mean_classc,mean(norm_classc))
}
worst_class<-min(mean_classc)
```

worst_class

Practical 9: Visualizing data

1. Overplotting

library(ggplot2)

g<-ggplot(data=diamonds,mapping=aes(x=carat,y=price,group=cut))

#stat='identity' means that it will use the real data in dimonds

#alpha means the transparency

#shape=18 makes the shape of the points like dimonds

#如果不设置 shape 的话 shape 默认为 round point

g1<-g+geom_point(stat='identity',aes(color=cut),size=1,alpha=0.2,shape=5)

g1

#ggplot 中的 group=cut 和 geom_point 中的 colour=cut 代表不同的含义：（1）group=cut 代表 ggplot 按照 cut 列中的值将数据分组，指定了绘图元素（连接线或分组的点）的组成方式

（2）colour=cut 是给不同的 cut 值着色

2. Rewrite the code

In the second of lecture slides (choosing different geom_XXXX and stat_XXXX functions), we use the stat_bin function to generate two plots. Please try to rewrite the code and use geom_XXX and stat_XXX to generate the same plots.

gg1 <-

ggplot(data = diamonds,

mapping = aes(x = carat, y = price, group = cut))

g2 <- gg1 + stat_identity(

```
mapping = aes(color = cut),

size = 0.6,

geom = "point",

shape = 3,

position = "jitter"

#“jitter”表示抖动，可以防止 point 的 overlapping

)

g2
```

“stat”和“geom”函数可以相互转化

```
d <- ggplot(diamonds, aes(carat)) + xlim(0, 3)

d + stat_bin(aes(ymax = ..count..), binwidth = 0.1, geom = "area")

d + stat_bin(aes(size = ..density..), binwidth = 0.1, geom = "point", position="identity" )

d<-ggplot(data=diamonds,aes(carat))+xlim(0,3)

a<-d+geom_area(stat="bin",aes(y=..count..),binwidth=0.1)

b<-d+geom_point(stat="bin",aes(size = ..density..), binwidth = 0.1,position="identity")
```

3. Build plots layer by layer

1. Plot the following boxplot from the dataset diamonds.

2. generate another layer of linear fitting using geom_smooth, use method lm. Save the geom_smooth to a new object sm.

3. Apply faceting to the plot by cut, color and cut~color.

4. Add another layer to add a title to the plot using labs

5. Change the color key using scale_color_brewer.

```
png(file="my_plot.png", width=500, height=500, units="px")

……(codes)

dev.off()#创建图片并储存

g <- ggplot(data = diamonds, mapping = aes(x = clarity, y = carat))

g3.1 <- g + geom_boxplot(outlier.size = 0.8)

g3.1



sm <- geom_smooth(

mapping = aes(x = as.integer(clarity)),

# Notice how the x-value was changed (x is not numeric)

method = "lm", se = F, size = 0.7

# 使用线性回归进行拟合

# 不显示拟合直线的置信区间

# 设置拟合直线的线宽

)

g3.2 <- g3.1 + sm

g3.2



g <- ggplot(data = diamonds, mapping = aes(x = clarity, y = carat, color = cut))

# As we need to split our data according to the `cut` variable, I decided to

# change the very original `ggplot` object

g3.3 <- g + geom_boxplot(outlier.size = 0.8) + sm
```

```
g3.3



g3.4 <-

g3.3 + ggtitle("Carats vs clarity") +

# I added the general title by `ggtitle`

theme( plot.title = element_text(hjust = 0.5),

# I decided to make this label(general title) in the middle

axis.text.x = element_text(angle = 45),

# x axis title rotates 45

legend.position = "bottom"

# I decided to change the location of the graph legend(图例)

)

g3.5 <- g3.4 + scale_color_brewer(palette = 3)

#颜色主题色

g3.5 + facet_wrap(~ color)

#按照 color 分成不同的区域块
```

4. Scale the y axis

1. Start from the plot in 3.2, transform the y-axis scale to log10 using scale_y_continuous. Pay attention to the change of y-axis. What should be the unit? Please change the y-axis label to include the unit using ylab.

```
g4.1 <- g3.2 + scale_y_continuous(trans = "log10") + ylab("Carats, log10")
```

g4.1

my codes:

```
g<-ggplot(data=diamonds,mapping=aes(x=clarity,y=carat))

sm<-geom_smooth(mapping=aes(x=as.integer(clarity),color=cut),method ="lm")

g+geom_boxplot(aes(color=cut))+sm+scale_y_log10()+labs(y="carat(log10)")
```

2. redo the boxplot in 3.1 by changing the y aesthetics to log10(carat). Compare the y-axis here with the one in 3.1. Change the y-axis label to include the unit.

```
library(cowplot)

# This package allows you locating several `ggplot` objects one by another. “plot_grid”有用到

g4.2.1 <- g + geom_boxplot(mapping = aes(y = carat)) + theme(legend.position = "none")

+

ggtitle("Original plot")

g4.2.2 <- g + geom_boxplot(mapping = aes(y = log10(carat))) + ggtitle("Change the aesthetic")

#改变美学属性

g4.2.3 <- g4.2.1 + scale_y_continuous(trans = "log10") + ggtitle("Change the Y-scale")

#改变 y 轴的比例尺

plot_grid(g4.2.1, g4.2.2, g4.2.3, NULL, rel_widths = c(0.75, 1), nrow = 2, byrow = T)
```

- **rel_widths = c(0.75, 1)**：这个参数指定了每个图形的相对宽度，其中第一个图形的宽度是第二个图形的 0.75 倍。这样的设置使得第一个图形相对较窄，而后两个图形相对较宽。

- **nrow = 2**：这个参数指定了网格的行数为 2。
- **byrow = TRUE**：这个参数指定了按行填充，即先填充第一行，然后填充第二行。
- NULL 参数表示没有额外的图形要添加

3. Add a layer of linear fitting to the plot in 4.2 by + sm from point 3.2. Do you see a problem? Please fix the problem by change the aesthetics in sm.

```
g4.3.1 <- g4.2.2 + sm + theme(legend.position = "none")

g4.3.2 <- g4.2.2 + geom_smooth( mapping = aes(x = as.integer(clarity), y = log10(carat)),

# Note how I changed aesthetics in this layer

method = "lm", se = F, size = 0.7 )

plot_grid(g4.3.1, g4.3.2, rel_widths = c(0.75, 1), labels = c("Just apply `sm`", "Fix aesthetics in `sm`"))

##一个是采用了原始数值 一个取了 carat 的对数
```

4. change the range of y-axis in the plot 4.1. set the limits to 0.3 to 1 using scale_y_continuous.

```
plot_grid(g4.1, g4.1 + ylim(c(0.3, 1)))

#对 y 轴范围进行限制
```

My codes:

```
g<-ggplot(data=diamonds,mapping=aes(x=clarity,y=carat))

sm<-geom_smooth(mapping=aes(x=as.integer(clarity),color=cut),method ="lm")
```

```
g+geom_boxplot(aes(color=cut))+sm+scale_y_continuous(trans="log10",limits =
c(0.3,3.0))
```

5. Jitter plot and scales.

Suppose, you want to show the relationship of the cut quality, clarity, mass, and price of the diamond on the same plot.

1. Sample out 100 cases from diamonds dataset. Generate a plot with a layer of boxplot and a layer of jitter plot like this using geom_jitter.

```
samp <- sample(x = 1:nrow(diamonds), size = 200)
gg5 <- ggplot(data = diamonds[samp,], mapping = aes(x = clarity, y = carat))
g5.1 <- gg5 + geom_boxplot(outlier.size = 0.8) + geom_jitter(width = 0.1, mapping =
aes(size = carat, color = price))
g5.1
```

2. Reset the color scale to only color the diamonds in the price range(10000, 15000). Try scale_color_gradient.

3. Reset the size scale to only show the diamonds in the carat range(1.2, 2). Try scale_size_continous.

```
g5.2 <- g5.1 + scale_color_gradient(limits = c(10000, 15000))
g5.3 <- g5.1 + scale_size_continuous(limits = c(1.2, 2))
plot_grid( g5.2, g5.3, ncol = 2, nrow = 1, labels = c("scale_color_gradient",
"scale_size_continous"), hjust = -0.26 )
```

6. Position.

In the lecture, we discussed different position adjustments. Generate the plots that can describe the number of different categories of diamonds according to their cut and clarity similar to the one described in the fourth part of the lecture (the slide about the position adjustment). Use the position argument in geom_bar.

```
gg6 <- ggplot(data = diamonds, mapping = aes(x = clarity, group = cut))

g6.1 <- gg6 + geom_bar(mapping = aes(fill = cut), position = "stack") + ggtitle("Stacking")

+ theme( plot.title = element_text(hjust = 0.5), axis.text.y = element_text(size = 10),

axis.text.x = element_text(size = 7), legend.position = "n", aspect.ratio = 7 / 7 )

#“aspect.ratio”设置了长宽比

#“legend.position=n”隐藏了图例

g6.2 <- gg6 + geom_bar(mapping = aes(fill = cut), position = "fill") + ggtitle("Filling") +

theme(

plot.title = element_text(hjust = 0.5), axis.text = element_text(size = 9), axis.text.y =

element_text(size = 10), axis.text.x = element_text(size = 7), legend.position = "right",

aspect.ratio = 7 / 5 )


g6.3 <- gg6 + geom_bar(mapping = aes(fill = cut), position = "dodge") +

ggtitle("Positioning near") + theme( plot.title = element_text(hjust = 0.5), axis.text.x =

element_text(size = 7), axis.text.y = element_text(size = 10), legend.position = "n",

aspect.ratio = 7 / 7 )
```

```
plot_grid( plotlist = list(g6.1, g6.2, g6.3, NULL), rel_widths = c(0.75, 1.25), rel_ heights = 1 )
```

Practical 9 problem set(W8.R)

Practical 10 : t-test

We will use the t.test R function. To perform a one sample t-test we need to specify the parameter mu, which represents the mean we are testing against. We can also specify the parameter alternative to test whether the test is one- or two-sided. In this case, since we are not testing for a change in a specific direction we use a two-sided test. If we wanted to test whether the temperature was higher (or lower) than 37◦C then we would use a one-sided test (alternative = "greater" or alternative = "lower")

```
temp<-read.csv("OrionTemp.csv")

head(temp)

summary(temp)

mysample<-sample(temp$Temperature,10)

#这个地方不需要设置 replace=TRUE

t.test(mysample,mu=37,alternative="two.sided")

#t.test()会显示 sample mean/p-value/t value...
```

Generate and plot the null distribution for your test. The t-test uses a t-distribution as a null, that you can recreate in R using the dt function, to which you need to pass the number of degrees of freedom (parameter df). Do you remember how to calculate the degrees of freedom? *Hint*: the null is defined between $-\infty$ and $+\infty$, but you can simply calculate it between, for example -10 and 10.

```
x<-seq(-10,10,sep=0.01)

y<-dt(x,9)
```

- seq(−10, 10, by = 0.01) creates a sequence of values from −10 to 10 with a step size of 0.01 and assigns it to the variable 'x'.
- dt(x, 9) calculates the probability density function of the t-distribution with 9 degrees of freedom for each value in 'x' and assigns it to the variable 'y'.

```
plot(x,y)
```

You can store the p-values for each point into a variable that you initialise (outside the for loop) to NULL. You can then use c to add new values to it. For example

```
res <- t.test(...)

res$p.value # This is the p value!

p.values <- NULL

for (...)

{

<calculate the test here>

p.values <- c(p.values, test$p.value)

}
```

Week 10 : t-test problem set

Guinness Quality Control

1. Is the brewery adding enough barley?

```
barley<-scan("barley.txt")

t.test(barley, mu = 50)
```

2. Is the *t*–distribution an appropriate test to answer this question? Do these data meet the assumptions required?

a. The first assumption is that the volumes recorded are continuous and randomly selected. A quick look at the data will reveal that the values are continuous.

```
head(barley)
```

b. The second assumption is that the sample is normally distributed. This can be done by plotting a histogram to determine whether the data looks normal enough to me – note here you have to make your own judgement.

```
hist(barley, xlab = "Barley weights (g)")

#judge whether the data is normally distributed

#the null hypothesis is normal distribution

shapiro.test(barley)
```

c. The third assumption is that the mean and standard error are independent. I test this by performing random sampling, where I record the standard error as well as the mean. A plot of these two values against each other suggests that there is little relationship between them, and this can be confirmed by performing a linear regression. The estimated value of the coefficient describing the relationship between the sampled standard errors and sample means is small (exactly how small will change during each round of sampling) and its associated p–value is large (often > 0.05) suggesting that there is little statistical support for this relationship

```
sampling_errors<-vector()

sampling_means<-vector()
```

```
for (replicate in 1:100){
barley_sample<-sample(barley, size = length(barley), replace = TRUE)
standard_error<-sd(barley_sample)/sqrt(length(barley_sample))
sampling_errors<-c(sampling_errors, standard_error)
sampling_means<-c(sampling_means, mean(barley_sample))
}
plot(sampling_means, sampling_errors, xlab = "Sample mean", ylab = "Standard error")
lmfit<-lm(sampling_errors~sampling_means)
abline(lmfit, col = 'red')
summary(lmfit)
```

3.The *t*-distribution is most useful when there are small sample sizes, but how small is small? Can you run a simulation to determine the power of the *t*-distribution as the number of pints sampled decreases? What is the minimum number of pints we need to sample to have 95% confidence that we could detect any real difference from the required 50 g?

As above, the power of the *t*-distribution can be explored through sampling. I run 100 replicates for sample sizes between 5 and 50. For each individual sample size, I have recorded the number of replicates for which I received a significant result (as simply judged by $p <= 0.05$) and plotted this against the sample size.

```
sig_results<-vector()
for (sample_size in 5:50){
found_sigs<-0
```

```
for (replicate in 1:100){

barley_sample<-sample(barley, size = sample_size, replace = FALSE)

p_value<-t.test(barley_sample, mu = 50)$p.value

if (p_value <= 0.05){

found_sigs = found_sigs + 1

}

}

sig_results<-c(sig_results, found_sigs)

}

plot(sig_results, type = 'l', ylab = "No. significant results/100", xlab = "Sample size")

abline(h = 95, col = 'red')


min(which(sig_results >= 95))
```

Practical 11: Choosing t-tests for different conditions

1) The "ToothGrowth" dataset comes with R. Please read the instructions for the dataset. and import the data using the function data

a. Let us assume that the average tooth length of normal guinea pigs is 8.5, then do you think vitamin C has a significant effect on tooth growth? Which t-test do you perform?

b. We want to know if the delivery methods could lead to significant differences. Please perform t-tests to support your conclusions. Does the dosage level matter?

```
library(datasets)

data("ToothGrowth")
```

```
t.test(ToothGrowth$len,mu=8.5,alternative="greater")

#get specific rows

dose_0.5<-subset(ToothGrowth,dose==0.5)

dose_1.0<-subset(ToothGrowth,dose==1.0)

dose_2.0<-subset(ToothGrowth,dose==2.0)

t.test(dose_0.5$len,dose_2.0$len)
```

2) Import another dataset “iris”. Apply a t-test to see if the average sepal length of setosa and versicolor is significantly different. How do you set the argument var.equal? Why?

```
data("iris")

head(iris)

pick_setosa<-subset(iris,Species=="setosa")

pick_versicolor<-subset(iris,Species="versicolor")

var.test(pick_setosa$Sepal.Length,pick_versicolor$Sepal.Length)

t.test(pick_setosa$Sepal.Length,pick_versicolor$Sepal.Length,var.equal = FALSE)
```

3) The blood pressure data record the blood pressure of patients before and after a treatment (https: //github.com/Opensourcefordatascience). You can import the dataset from the CSV file. Now you can apply t-tests to see if the treatment is useful. Which t-test do you use?

```
#read.table() for ".txt" file

#"header=TRUE" makea the first row of the file to be the header

bp_data<-read.table("blood_pressure.txt",header = TRUE)

#"paired=TRUE"->before and after: compare the effect
```

```
t.test(bp_data$bp_before,bp_data$bp_after,alternative = "greater",paired=TRUE)

#test by the rank

wilcox.test(bp_data$bp_before,bp_data$bp_after,alternative = "greater",paired=TRUE)
```

week 11 problem set

Exercise 1

Field A: 10.2 10.7 15.5 10.4 9.9 10.0 16.6 15.1 15.2 13.8 14.1 11.4 11.5 11.0

Field B: 8.1 8.5 8.4 7.3 8.0 7.1 13.9 12.2 13.4 11.3 12.6 12.6 12.7 12.4 11.3 12.5

Use a 5% level of significance to test the claim that field A has, on average higher soil water content than field B.

looking at the distribution of the data

```
Field_A <- c(10.2, 10.7, 15.5, 10.4 , 9.9, 10.0, 16.6, 15.1, 15.2, 13.8, 14.1, 11.4 , 11.5 , 11.0)

Field_B <- c( 8.1, 8.5, 8.4 , 7.3, 8.0, 7.1, 13.9 , 12.2, 13.4, 11.3, 12.6, 12.6, 12.7 , 12.4 , 11.3, 12.5)

hist(Field_A)

hist(Field_B)
```

Which parameters shall we use?

1) H0: Field A does NOT have higher soil water content than Field B.

2) HA: Field A does have higher soil water content than Field B.

3) Therefore, we should perform a 1–tail test with the ‘alternative’ parameter set as ‘greater’.

4) The two fields are not paired, therefore we perform an unpaired test.

```
wilcox.test(Field_A, Field_B, alternative = 'greater', paired = F)
```

Exercise 2

You measured the weights of 10 mice before and after you fed them with high–fat high–sugar diet for 3 weeks. You want to know if the treatment (high–fat, high–sugar diet) has effects on the weight. Import the “mice_weights.txt” data and perform an appropriate statistical test to make a decision whether your treatment has an effect.

looking at the distribution of the data

```
mice_data = read.csv('mice_weights.txt')
```

Option 1, judge by eyes based on the density plot

```
library(ggplot2)

ggplot(data=mice_data) +

geom_line(aes(x=before, col='before'), stat='density') +

geom_line(aes(x=after,col='after'), stat='density') + theme_classic()
```

Option 2, judge by eyes based on the Q–Q plot This is another neat visualization methods to examine whether

the distribution is different from the normal distribution.

```
library(ggpubr)

## Warning: package 'ggpubr' was built under R version 4.3.2

ggqqplot(mice_data$before)

ggqqplot(mice_data$after)
```

Option 3, use Shapiro test

```
shapiro.test(mice_data$before)

shapiro.test(mice_data$after)

#p-value<0.05 是非正态分布
```

2.2 Performing the t-test.

1) Ho: The treatment does not have an effect on the mice weights.

2) Ha: The treatment has an effect on the mice weights

3) Therefore, we should perform a 2-tail t-test.

```
# Are the variance same or different?

var.test(mice_data$before, mice_data$after)

t.test(mice_data$before, mice_data$after, paired = TRUE, var.equal=TRUE)
```

Practical 12: R Markdown

Installation instructions

```
install.packages("rmarkdown")

tinytex::install_tinytex()

library(tinytex)

library(rmarkdown)
```

Text editing in R markdown

```
## R Markdown
```

较大的字体

```
I am the *best* student
```

斜体

```
- Prepare tutorial

- Go to class
```

– Revise lecture

– Finish ICA report

圆点

1. Eat breakfast

2. Eat lunch

4. Eat dinner

2. Sleep

自动按照数字顺序

eval=FALSE and echo=FALSE

eval=FALSE 看不见 output

echo=FALSE 看不见 input

Tables, images, and comments

```
library(knitr)
```
只需要输入一次

```
kable(head(Dragons, 5))
```

```
mice <- read.csv("mouse_report.csv")
```

```
kable(mice[1:6,c(2,4,5)])
```

```
kable(summary(mice))
```

Images

```
![Figure legend](path_to_file.png){Figure options (width or heigth; this part is optional)}
```

不需要放在 r chunk 中

```{r chunk-name, out.width='100%', fig.align='center',

fig.caption='The description of the figure.'}

include_graphics(path = 'image_name.type')

```

```{r chunk-name, out.width='50%', fig.align='center',fig.caption='Figure1: ZJU

International Campus bridge in the snow. By Melanie Stefan, 2018 (CC-BY-SA 4.0)'}

include_graphics(path ="./Bridge(1).png")

```

Comments

注意一行不要太多字

Changing Image Sizes and Alignments

There are a bunch of options for adjusting the placement of the figures which R produces. fig.align controls the horizontal **alignment** (left, right, or center).

When producing PDF, the options out.height and out.width let you specify the desired height or width of the figure, in inches, centimeters, or multiples of pre-defined lengths (from LaTeX). So for instance ```{r, out.height="3in"} forces the image to be 3 inches high, while ```{r, out.width="0.48\\textwidth"} forces the image's width to be a bit less than half of the total width of the text on the page (so that two such images will fit side by side).

```{r, fig.width=6, fig.height=4}

# Your R code for generating a histogram or ggplot goes here

## Week 12: problem set
```

Prof. Xu wants to take a look at the data every month, but is very busy, and therefore just wants a short overview. This overview should

• be in pdf format

• be no longer than 2 pages

• contain a nice title/heading including the current date

• contain summary statistics about the overall weight and age of mice

• show boxplots of the mouse weight and age by genotype

• contain as much text as is needed to understand the report (this could either be in the form of section headings or short sentences or figure legends)

My codes:

Overall weight and age of mice

```{r,echo=FALSE}

mice <- read.csv("mouse_report(1).csv")

library(knitr)

mice_select<-mice[,4:5]

kable(summary(mice_select))

```

Box plot about the mouse age and weight

- Age

```{r,echo=FALSE}
```

library(ggplot2)

mice$genotype<-as.factor(mice$genotype)

ggplot(mice,aes(x=genotype,y=age))+geom_boxplot()

```

- Weight

```{r,echo=FALSE}

ggplot(mice,aes(x=genotype,y=weight))+geom_boxplot()

```

Codes in note:

```{r setup, include=FALSE}

#r setup 用于设置全局

# `include=FALSE` means that this chunk will be executed, the results of this code will affect this R session, but the output will not be stored.

knitr::opts_chunk$set(echo = F) # You can set the default option for the code chunks by changing `opts_chunk`. Here, we set the `echo` attribute for all the following chunks to FALSE by default.

# Keep in mind that normally we are interested in the code you use to prepare your final document!

library(tidyverse)

library(knitr)

# Also, remember: do not use comments within the code chunks too much.

# And do not make them too long. This code chunk is an exception.

# The results of this chunk will not be exported further.

```

Introduction

This is the monthly mouse report, based on data collected by the lab assistant, Dr. Lu.

```{r}

mouse_report <- read.csv("mouse_report(2).csv") %>%

  mutate(genotype = as.factor(genotype))

```

Overall statistics

Total number of mice: `r nrow(mouse_report)` animals. Their distribution according to the genotype is:

```{r}

table(mouse_report$genotype) %>% kable(col.names = c("Genotype", "Number"))

```

The distribution of age and weight of mice is shown in figure 1.

```{r, fig.ncol=2, out.width="50%"}

- fig.ncol=2 specifies that the figures should be arranged in two columns.
- out.width="50%" sets the width of the output figures to 50% of the text width.

geno <- ggplot(data = mouse_report, aes(x = genotype, y = weight, col = genotype)) +
  geom_boxplot(outlier.shape = NA) +
  geom_jitter(width = 0.1) +
#数据点在 x 轴上的位置可以在 −0.1 到 0.1 之间进行小范围的随机抖动。
  theme_classic()
age <- ggplot(data = mouse_report, aes(x = genotype, y = age, col = genotype)) +
  geom_boxplot(outlier.shape = NA) +
  geom_jitter(width = 0.1) +
  theme_classic()
print(geno)
print(age)
```

Fun with maths

\(E = mc^2\)
```

$K_d = \frac{k_b}{k_f}$

$v = \frac{v_{max}[S]}{K_M + [S]}$

$\sigma = \sqrt{\text{var}}$

$\mu = \frac{1}{n} \sum_{k=1}^{n} x_k$

Mock Coding Challenge

1. Stroop test

Import the data and plot them in a useful way.

```{r}
data<-read.csv("stroop_test.csv")
library(ggplot2)
ggplot(data,aes(x=Time,y=Score,fill=Time))+geom_boxplot()+labs(
  title="Morning vs Afternoon",x="Time of Day",y="Score")+theme_classic()
```

```
stroop <- read.csv("stroop_test.csv", header=TRUE)
head(stroop)
library(ggplot2)
p <- ggplot(stroop,aes(x=Time,y=Score)) + geom_boxplot()
```

p

Is there a difference in performance on the Stroop task between the morning and afternoon group?

```{r}

data_morning<-data[data$Time=="Morning",]

data_afternoon<-data[data$Time=="Afternoon",]

shapiro.test(data_morning$Score)

shapiro.test(data_afternoon$Score)

# We firstly use shapiro.test to test the normality of both samples

# It is found that the samples are all not normally distributed

# with p-value<0.05

# Then we can use wilcox.test to judge whether there is

# significant difference

wilcox.test(data_morning$Score,data_afternoon$Score)

# From wilcox test, we can get the result that "p-value=0.3774"

# so there is no significant difference in performance

# on the Stroop task

# between the morning and afternoon group

```

```
hist(stroop[stroop$Time=="Morning","Score"],main="Morning", xlab="Score")

hist(stroop[stroop$Time=="Afternoon","Score"],main="Afternoon", xlab="Score")
```

```
wilcox.test(Score~Time,stroop, alternative="two.sided")
```

Name one way in which the study could be improved or followed up on.

- Has a suggestion been made for improvement or follow-up?
- Is the suggestion concrete? For instance, you can't just say "Do better experiments!" Also, you can't just say "Gather more data points in both groups". Because why do you think there aren't enough data points? How many data points would you need? Unless you can give an explanation for that, "collect more data" is just a lazy suggestion.

The same group of participants should do the test both in the morning and afternoon instead of some participants do the test in the morning and some in the afternoon.

2. Marathon finishing times

Import the dataset, and plot it in a way that addresses the question we are interested in.

```{r}
runner<-read.csv("Chicago2013_random_finishers.csv")
ggplot(data=runner,mapping=aes(x=Age,y=Time,colour=Gender))+geom_point()+geom_smooth()+labs(
  title="finishing time related to age and gender",x="Age of finishers",y="Finishing time")
```

```
marathon <- read.csv("Chicago2013_random_finishers.csv")
```

```
head(marathon)

q <- ggplot(marathon, aes(x=Age,y=Time, col=Gender))

q <- q+xlab("Age") + ylab("Time [hours]")

q <- q + geom_point()

q
```

What are the average finishing times and standard deviation for each gender? What are the average finishing times and standard deviation for each age quartile?

```{r}

runner_man<-runner[runner$Gender=="M",]

runner_woman<-runner[runner$Gender=="F",]



# the average finishing time for man



round(mean(runner_man$Time),2)



# the average finishing time for woman



round(mean(runner_woman$Time),2)



# the standard deviation for man
```

```
round(sd(runner_man$Time),2)

# the standard deviation for woman

round(sd(runner_woman$Time),2)

quantile_boundary<-quantile(runner$Age,probs=c(0,0.25,0.5,0.75,1))
runner$Age_quartile<-cut(runner$Age,breaks = quantile_boundary,
                         labels=c("Q1","Q2","Q3","Q4"))

# mean and standard deviation for age Q1

mean(na.omit(runner[runner$Age_quartile=="Q1",]$Time))
sd(na.omit(runner[runner$Age_quartile=="Q1",]$Time))

# mean and standard deviation for age Q2

mean(na.omit(runner[runner$Age_quartile=="Q2",]$Time))
sd(na.omit(runner[runner$Age_quartile=="Q2",]$Time))

# mean and standard deviation for age Q3
```

```
mean(na.omit(runner[runner$Age_quartile=="Q3",]$Time))

sd(na.omit(runner[runner$Age_quartile=="Q3",]$Time))


# mean and standard deviation for age Q4


mean(na.omit(runner[runner$Age_quartile=="Q4",]$Time))

sd(na.omit(runner[runner$Age_quartile=="Q4",]$Time))
```

对于取四分位数部分：

```
q25 <- quantile(marathon$Age,0.25)

q50 <- mean(marathon$Age)

q75 <- quantile(marathon$Age,0.75)

youngest <- marathon[marathon$Age <= q25, "Time"]

young_middle <- marathon[(marathon$Age > q25 & marathon$Age <= q50), "Time"]

old_middle <- marathon[(marathon$Age > q50 & marathon$Age <= q75), "Time"]

oldest <- marathon[marathon$Age > q75, "Time"]
```

If you had to suggest a statistical test to determine the effect of age quartile and gender on marathon finishig time, what test would you suggest, and why?

I would choose to use Two-Way ANOVA. Because age and gender are two different factors acting on marathon finishing time and AVOVA is useful for such comparison.

Two-Way ANOVA（双因素方差分析）是一种统计方法，用于分析两个因素对一个连续变量的影响，并检验这些因素是否显著地影响了变量的均值。这种方法常用于实验设计和研究中，以确定两个不同的因素（通常称为因子）对于观察到的变化是否有显著影响。

以下是 Two-Way ANOVA 的主要用途：

1. **识别因素之间的相互作用：** Two-Way ANOVA 允许分析两个因素是否相互作用，即它们的组合是否对响应变量产生不同的效应。这对于了解两个因素之间是否存在复杂的关系很重要。
2. **分解总方差：** Two-Way ANOVA 通过分解总方差为来自两个因素的方差、因子之间的交互作用方差以及未解释的方差，帮助我们理解观察到的变化来源。这有助于确定哪些因素对于解释变量的变异程度贡献较大。
3. **检验因素的主效应：** Two-Way ANOVA 可用于检验每个因素的主效应，即每个因素独立地是否对响应变量产生显著的影响。这有助于确定哪个因素对变量的均值产生了显著影响。
4. **比较组之间的均值：** Two-Way ANOVA 通过比较不同组之间的均值，帮助确定哪些组之间存在显著差异。如果 Two-Way ANOVA 表明因素或因素的交互作用显著，进一步的事后检验可能需要进行，以确定具体哪些组之间存在差异。

```
data <- data.frame(  Factor1 = rep(c("A", "B", "C", "D"), each = 25),  Factor2 =
rep(c("X", "Y"), each = 50),  Response = rnorm(100))
# 使用 aov 函数进行 Two-Way ANOVA
model <- aov(Response ~ Factor1 * Factor2, data = data)
# 查看 ANOVA 表
summary(model)
```

An appropriate test is suggested, **and** the selection of test is explained. For instance, here, because we are looking at two factors (gender and age group), a two-way ANOVA would be a good choice (provided assumptions are met!) If you suggest an ANOVA, you should also explain whether it should be with or without interactions, and why. Suggesting the

test and explaining the suggestion is enough. Note that we don't ask you to run the test. You will not get additional points for running the test! All it makes you do is lose time!

3. Antiviral drug

What are the Null Hypothesis and the Alternative Hypothesis in Prof. Liu's trial?

Null hypothesis: The rats from two groups take the same time to fully recover.

Alternative hypothesis: The rats that were given the new drug take shorter time than the control group to fully recover.

H0 and HA have been formulated, for instance:

- H0: Recovery time is the same independent of treatment
- HA: Reocvery time differs according to what treatment has been administered.

H0 and HA have to be formulated so that exactly one of them has to be true. A commonly–made error is this:

- H0: There is no difference between treatments.
- HA: The new treatment works better than the current treatment.

What is the problem here? Well, what if the new treatment works *worse* than the current treatment? Then H0 is false **and** HA is also false. That can't be happening. The universe will explode! Please take care of the universe! (Actually, the universe is OK. But hypothesis testing won't work, because it relies on the fact that either H0 or HA is true.)

For her statistical analysis, Prof. Liu uses the commonly used significance level α of 0.05. What type of error does this relate to, and how?

This related to type 1 error. A significance level α of 0.05 means that Prof. Liu has 5% chance of making type 1 error which means when null hypothesis is true and we incorrectly reject it in favor of alternative hypothesis.

There is a precise explanation, along these lines: If H0 is true (there is no effect), then the p-value is the probability of seeing data as or more extreme than the one we saw in our experiment. We reject H0 if p is smaller than α. That means that we accept a probability of up to α of mistakenly rejecting H0, even if it's true. This is exactly the probability of making a Type 1 error.

After ensuring that the sample size is big enough and that the assumptions of the statistical tests are met, Prof. Liu runs a statistical test and gets a p-value of 0.059. What is a p-value? Based on this result, what should Prof. Liu report as the outcome of this study?

p-value means the probability of observing a value as or more extreme as the one you observe if the null hypothesis is true. p-value=0.059>0.05 so we can't reject the null hypothesis that the rat from two groups might take the same time to recover. This antiviral drug is not pretty effective.

A p–value is correctly defined: It is the probability of seeing data as or more extreme as the data seen in the experiment *if* H0 is true. Go through your words very carefull!! A p–value is **NOT** the probability that H0 is true. It is **NOT** the probability of the data.

It is correctly identified that 0.059 > 0.05 and therefore H0 cannot be rejected.

An interpretation is given in terms of the original biomedical problem, in this case: We cannot conclude that the new antiviral treatment is any different from the already available treatment. There is no trying to turn this into a positive result, for instance by saying that 0.059 is quite close to 0.05 so it still kind of counts. (It doesn't. That's what α is for, it's a decision cutoff.) Similarly you can't say that the only reason it's not significant is that there isn't enough data, because the instructions clearly say that Prof. Liu has already ensured that the sample is big enough. Finally, you can't say that the result is "trending towards significance", "approaching significance", "very nearly almost significant" or anything similar. Significance testing is a yes/no thing.