

109-1 資料科學應用 - R語言篇

期末考

學號:A106260093 姓名:王緯華

15 一月 2021

- 注意事項
 - 下載題目卷
 - 考試期間
 - 答題檔案原則
 - 上傳答題檔案
- 1 抽球算機率
 - 1.1 直接算機率
 - 1.2 模擬抽球
 - 1.3 重覆實驗10次
 - 1.4 重覆實驗1000次
- 2 字串處理
 - 2.1 讀取資料
 - 2.2 屏蔽字元
 - 2.3 排序資料
- 3 屬質變異指數(IQV)

注意事項

下載題目卷

- 於課程網站(<http://www.hmwu.idv.tw/web/SHU/> (<http://www.hmwu.idv.tw/web/SHU/>))下載題目卷。

考試期間

- 請按照平時上課之座位入座。
- 可參考課本、上課講義(包含電子檔)及其它資料，但不能與別人討論。
- 可使用計算機、自己的筆記型電腦、平板電腦及手機。
- 全程可上網查詢，但不能用通訊軟體(例如: FB/LINE/IG)討論，也不可抄襲網路上之程式碼。
- 不按照規定作答者，酌量扣分。
- 有問題者，請舉手發問，勿與同學交談。
- 不可使用它人之隨身碟。
- 「作弊」或「疑似作弊」，本學期總成績不予評分。
- 程式請隨時存檔，避免突然意外發生，程式檔不見。

答題檔案原則

- 若程式碼直接複製(或照抄)講義上的以不給分為原則。
- 程式碼請直接寫在本 Rmd 檔。經由 Knit 編譯出 .html 檔。

Loading [MathJax]/jax/output/HTML-CSS/fonts/TeX/fontdata.js
• 作答檔案，請隨時存檔並備份，勿直接存於公用電腦(例如: 桌面)。

上傳答題檔案

- 上傳方式同平時作業方式。
- 請上傳三個檔案: 「學號-姓名-SHU-R-FinalExam.Rmd」、「學號-姓名-SHU-R-FinalExam.html」及「學號-姓名-SHU-R-FinalExam.pdf」。其中 pdf 檔為使用瀏覽器(IE, Edge, Firefox, Chrome)開啟 .html 印出成PDF檔。
- 若上傳檔案格式錯誤, 內容亂碼, 空檔等等問題。請自行負責。

1 抽球算機率

一袋中有5顆紅球及3顆白球, 小明由袋中隨機抽球, 每次取一球, 共取4次, 令 A 為抽出2次白球的事件, 計算此事件 分別在放回(replacement)、不放回(without replacement)兩種情況下之機率 $P(A_r), P(A_w)$ 。

$$\text{放回: } P(A_r) = C_2^4 \left(\frac{5}{8}\right)^2 \left(\frac{3}{8}\right)^2$$

$$\text{不放回: } P(A_w) = \frac{C_2^5 C_2^3}{C_4^8}$$

1.1 直接算機率

請利用上式(C 為組合數), 使用 R 指令直接計算上述之機率 (分別命名為 `Prob.Ar`, `Prob.Aw`)並印出。

```
# your source code here
# my anser
Prob.Ar <- choose(4,2)*(5/8)^2*(3/8)^2
Prob.Aw <- (choose(5,2)*choose(3,2)) / choose(8,2)

cat("Prob.Ar:", Prob.Ar, "\n")
```

```
## Prob.Ar: 0.3295898
```

```
cat("Prob.Aw:", Prob.Aw, "\n")
```

```
## Prob.Aw: 1.071429
```

1.2 模擬抽球

小明今天想要以寫 R 程式的方式來模擬此隨機實驗, 計算抽球的機率, 若設定`{set.seed(123456)}`, 列出「一袋中有5顆紅球及3顆白球, 小明由袋中隨機抽球, 分別在放回(replacement)、不放回(without replacement)兩種情況下, 每次取一球, 共取4次」實驗一次的結果, 並計數印出白球出現之個數。(不需寫成 R 函式) (提示: `sample`, `table`)

Loading [MathJax]/jax/output/HTML-CSS/fonts/TeX/fontdata.js

```
# your source code here
# my anser
set.seed(123456)
ball <- c("白球", "紅球")
bag <- rep(ball, c(5, 3))
s <- sample(bag, 4)
cat("抽出不放回")
```

```
## 抽出不放回
```

```
table(s)
```

```
## s
## 白球 紅球
##      3      1
```

```
set.seed(123456)
ball <- c("白球", "紅球")
bag <- rep(ball, c(5, 3))
b <- sample(bag, 4, T)
cat("抽出放回")
```

```
## 抽出放回
```

```
table(b)
```

```
## b
## 白球
##      4
```

1.3 重覆實驗10次

同上小題，寫一 R 函式(命名為 `Draw_Ball`)，沒有輸入，輸出為白球 分別在放回、不放回兩種情況下的個數。重覆上述實驗10次，印出分別在放回、不放回兩種情況下白球出現的個數。(提示：`as.data.frame`，`replicate`)

```
# your source code here
# my anser

Draw_Ball <- function(){
  ball <- c("白球", "紅球")
  bag <- rep(ball, c(5, 3))
  ans <- sample(bag, 4)
  table(factor(ans, levels=ball))
}

set.seed(123456)
DrawResult <- as.data.frame(t(replicate(10, Draw_Ball())))
cat("抽出不放回")
```

```
## 抽出不放回
```

```
DrawResult
```

```
##      白球 紅球
## 1         3     1
## 2         2     2
## 3         2     2
## 4         2     2
## 5         2     2
## 6         3     1
## 7         2     2
## 8         2     2
## 9         2     2
## 10        1     3
```

```
Draw_Ball_1 <- function(){
  ball.t <- c("白球", "紅球")
  bag.t <- rep(ball, c(5, 3))
  ans.t <- sample(bag, 4, T)
  table(factor(ans.t, levels=ball.t))
}

set.seed(123456)
DrawResult_1 <- as.data.frame(t(replicate(10, Draw_Ball_1())))
cat("抽出放回")
```

```
## 抽出放回
```

```
DrawResult_1
```

```
##      白球 紅球
## 1      4      0
## 2      1      3
## 3      3      1
## 4      2      2
## 5      1      3
## 6      2      2
## 7      2      2
## 8      1      3
## 9      1      3
## 10     2      2
```

1.4 重覆實驗1000次

同上小題，重覆上述實驗1000次，計算在放回、不放回兩種情況下，抽到2顆白球的機率。(提示: `as.data.frame, replicate, sum, ==`)

```
# your source code here
# my anser

n <- 1000
set.seed(123456)
DrawResult <- as.data.frame(t(replicate(n, Draw_Ball())))
#sum((DrawResult$白球==2))/n
cat("抽出不放回:", sum((DrawResult$白球==2))/n, "\n")
```

```
## 抽出不放回: 0.432
```

```
n <- 1000
set.seed(123456)
DrawResult_1 <- as.data.frame(t(replicate(n, Draw_Ball_1())))
#sum((DrawResult_1$白球==2))/n
cat("抽出放回:", sum((DrawResult_1$白球==2))/n, "\n")
```

```
## 抽出放回: 0.339
```

2 字串處理

某商業公司舉行抽獎活動，中獎名單紀錄於 `award-list.xlsx` 檔中，包含會員姓名、會員卡號及得獎金額。

2.1 讀取資料

請讀取此檔案，並印出全部中獎名單。

```
# your source code here
#my answer
library(readxl)
award.list <- read_excel("D:/R專案2020/SHU-R/1091-SHU-R-FinalExam/award-list.xlsx", sheet =
  "工作表1", na = "NA")
award.list
```

```
## # A tibble: 10 x 3
##   會員姓名  會員卡號 得獎金額
##   <chr>      <dbl>    <dbl>
## 1 沈俞予    7113235607    500
## 2 簡惠榕    8010785376   1000
## 3 徐一良    9010344896   2000
## 4 賴淨茹    3010789872   1500
## 5 林金玲    5011213845   4500
## 6 吳彩鳳    2592903839   1000
## 7 江德翰    3714483694   3000
## 8 葉建鴻    4012123657   2500
## 9 阮通全    3053398421   5000
## 10 黃玉鈴    3317005422   3500
```

2.2 屏蔽字元

因考量個資法，公告名單不能將全名公開，請你幫此名單，每一中獎者的姓名及會員卡號，部份字元打上*，例如第一筆紀錄為「沈俞予 7113235607」，請改為「沈*予 7113 *** 607」，印出修改後可公告之名單。(提示: substr)

```
# your source code here
#my answer
str(award.list)
```

```
## tibble [10 x 3] (S3: tbl_df/tbl/data.frame)
## $ 會員姓名: chr [1:10] "沈俞予" "簡惠榕" "徐一良" "賴淨茹" ...
## $ 會員卡號: num [1:10] 7.11e+09 8.01e+09 9.01e+09 3.01e+09 5.01e+09 ...
## $ 得獎金額: num [1:10] 500 1000 2000 1500 4500 1000 3000 2500 5000 3500
```

```
name <- substr(award.list$會員姓名,2,2)
name
```

```
## [1] "俞" "惠" "一" "淨" "金" "彩" "德" "建" "通" "玉"
```

```
gsub(name, "*", award.list)
```

```
## Warning in gsub(name, "*", award.list): 引數 'pattern' 長度 > 1 且只有第一個元素
## 將被使用
```

```
## [1] "c(\\沈*予\\", \\簡惠榕\\", \\徐一良\\", \\賴淨茹\\", \\林金玲\\", \\吳彩鳳\\", \\江德翰\\", \\葉建鴻\\", \\阮通全\\", \\黃玉鈴\\")"
## [2] "c(7113235607, 8010785376, 9010344896, 3010789872, 5011213845, 2592903839, 3714483694, 4012123657, 3053398421, 3317005422)"
## [3] "c(500, 1000, 2000, 1500, 4500, 1000, 3000, 2500, 5000, 3500)"
```

```
number <- substr(award.list$會員卡號,5,7)
number
```

Loading [MathJax]/jax/output/HTML-CSS/fonts/TeX/fontdata.js

```
## [1] "235" "785" "344" "789" "213" "903" "483" "123" "398" "005"
```

```
gsub(number, "****", award.list$會員卡號)
```

```
## Warning in gsub(number, "****", award.list$會員卡號): 引數 'pattern' 長度 > 1 且
## 只有第一個元素將被使用
```

```
## [1] "7113***607" "8010785376" "9010344896" "3010789872" "5011213845"
## [6] "2592903839" "3714483694" "4012123657" "3053398421" "3317005422"
```

2.3 排序資料

承上小題，請將修改後之名單，依照「得獎金額」由多至少的順序，全部印出。

```
# your source code here
#my answer
sort(award.list$得獎金額, decreasing=TRUE)
```

```
## [1] 5000 4500 3500 3000 2500 2000 1500 1000 1000 500
```

3 屬質變異指數(IQV)

計算名目變數(nominal variable)的變異分散程度，其中Index of Qualitative Variation (IQV)是一個指標(其數值是介於0與1中間)。公式如下：

$$IQV = \frac{k(n^2 - \sum f^2)}{n^2(k-1)},$$

其中 k 是類別數或組數， n 是樣本數， $\sum f^2$ 是將各類別次數之平方加起來之總和。假設有一名目變數資料(nv)如下，試寫一 R 函式，計算IQV。(提示: table)

```
set.seed(12345)
no <- sample(20:100, 1)
nv <- LETTERS[sample(1:26, 5)][sample(1:5, no, replace=T)]
```

```
# your source code here
set.seed(12345)
no <- sample(20:100, 1)
nv <- LETTERS[sample(1:26, 5)][sample(1:5, no, replace=T)]
IQV <- function(){
}
}
```