(本頁中文版如下一頁)

National Chengchi University, 113-1 Academic Year
Midterm Exam of Statistics (I), Bonus Test, R Programming

Department/Grade: _____ ID: _____ Name: _____
Subject: Statistics (I)
Date: 2024/11/14
This test consists of 5 major questions. (20% each, total score: 100%)
Time period: 15:00~16:00 (total 60 minutes)

Notes:

1. Download the R exam sheet (`113-1-Stat-R-Midterm.zip`) from the course website and unzip in your laptop. The zip file contain the question sheet, the answer sheet, and the datasets.
2. Answers for this exam should be provided using the R programming language (either Rgui or RStudio). Other programming languages are not permitted.
3. During the exam, you may refer to textbooks, lecture notes (including videos, Please bring your own headphones), or browse the internet. However, the use of communication software/APP such as Messenger, IG, Line, etc., is strictly prohibited.
4. Any form of cheating or suspicious behavior is not allowed.
5. On this answer sheet, please ensure you copy the "executed code and its results (including graphics)" from the **R Console** and paste it here (in `Courier New font`, size 10, black text on a white background). This should include both the code and the output, not just one or the other. Finally, the answers for each sub-question should be highlight by yellow color (not just printing the report; the TA shouldn't have to search for the answers)
6. Please label your answers in sequence, e.g., (1)a, (1)b, (2)a, etc.
7. After completing your answers, save this Word document with the filename "`StudentID-FamilyName-Midterm.docx`" (replace with your actual "`Student ID` and `FamilyName`") and upload it to http://hmwu.nccu.edu.tw/login.html .
8. Username: `stat113`, Password: (classroom number) `26xxxx`, Folder: "`20241114-MidtermExam`".
9. If the upload site displays a "blank page", move your cursor to the "address bar" and press "Enter". If that doesn't work, try using a different browser (IE/Edge/Firefox/Chrome).
10. Uploaded files cannot be deleted. If you need to upload a revised file, please add "-2" to the main filename, e.g., " `StudentID-FamilyName-Midterm-2.docx` ".

_____Wishing you a successful exam_____

(English version on the previous page)

國立政治大學 113 學年度第一學期

統計學(一) 期中 R 程式加分考

系級:_____ 學號:_____ 姓名:_____

考試科目: 統計學(一)

考試日期:2024/11/14

本試題共 5 大題 (各 20%)

考試時間:15:00~16:00 (共 60 分鐘)

注意事項:

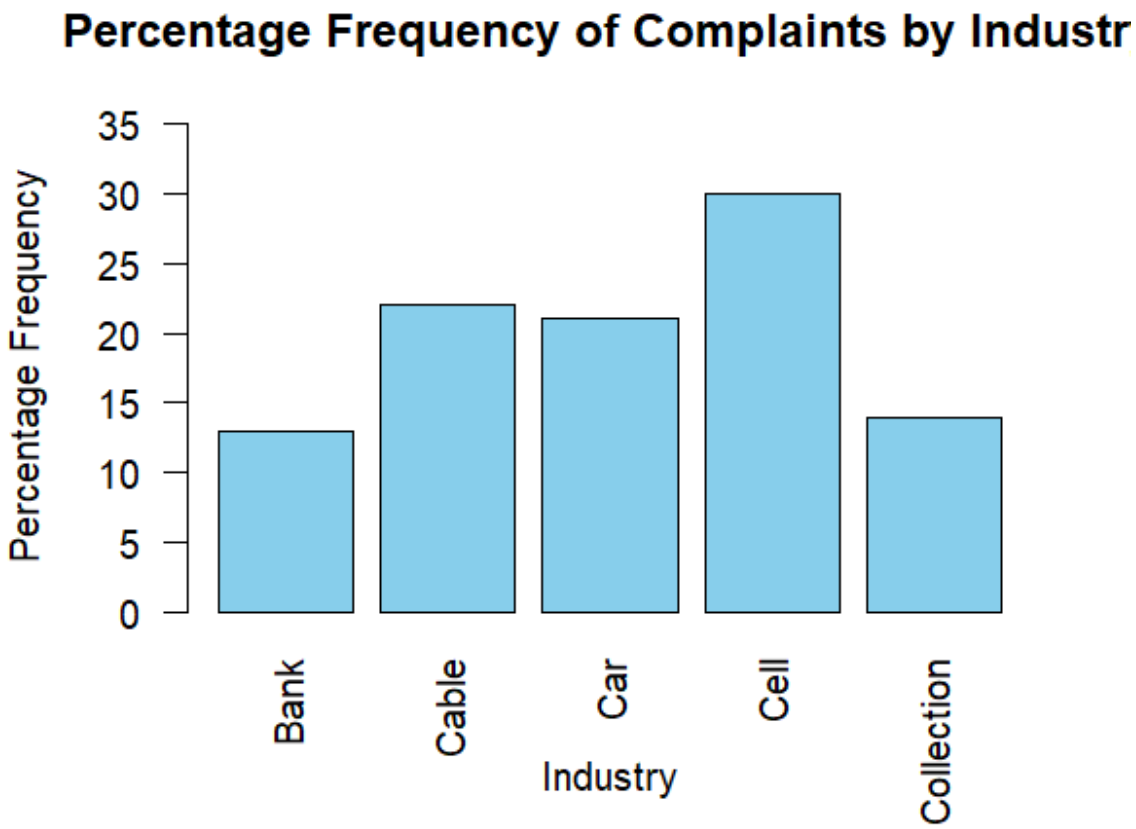
1. 從教學網站下載電子考卷 (`113-1-Stat-R-Midterm.zip`)，並於自己的筆電解壓縮。壓縮檔包含題目卷、答案卷和資料集。
2. 本次考題以 R 程式(Rgui 或 RStudio)方式作答，其他程式不允許。
3. 考試過程中可查詢書本、教學講義或上網(含上課影片，請自備耳機)，禁止利用 messenger, IG, Line 等等通訊軟體。
4. 禁止疑似作弊行為。
5. 本答案卷上請務必於 R Console 內**複制「執行後的程式碼及結果(含圖形)」**，於本答案卷貼上(`Courier New, 10 點字，白底黑字`)，不是只有程式碼，不是只有報表。最後，將每小題之答案以黃色底高亮起來(不能只印出報表，要助教去找答案)。
6. 請依序註明題號: (1)a, (1)b, (2)a 等等。
7. 作答完請將此 word 檔存檔，檔名為「`StudentID-FamilyName-Midterm.docx`」(更改成自己「學號」、「姓」)並上傳至 http://hmwu.nccu.edu.tw/login.html
8. 帳號: `stat113`，密碼: (上課教室號碼) `26xxxx`，資料夾:「`20241111-MidtermExam`」
9. 如果上傳網站出現「空白頁」，請將滑鼠移至「網址列」後，按「Enter」即可。若再不行，請換其它瀏覽器(IE/Edge/Firefox/Chrome)
10. 上傳檔案無法刪除，若要上傳更新檔，請於主檔名後加「-2」，例如:「`StudentID-FamilyName-Midterm-2.docx`」。

_____祝考試順利_____

(1)	**Data file: BBB** **Complaints Reported to BBB.** Consumer complaints are frequently reported to the Better Business Bureau (BBB). Some industries against whom the most complaints are reported to the BBB are banks; cable and satellite television companies; collection agencies; cellular phone providers; and new car dealerships (*USA Today*). The results for a sample of 200 complaints are contained in the file *BBB*. a. Show the frequency and percent frequency of complaints by industry. b. Construct a bar chart of the percent frequency distribution. c. Which industry had the highest number of complaints? d. Comment on the percentage frequency distribution for complaints.

```
# a)
> print(result)
      Company Frequency Percentage
1        Bank        26         13
2       Cable        44         22
3         Car        42         21
4        Cell        60         30
5 Collection        28         14

> # b). Constructing the bar chart for percent frequency
> # Calculate the frequency and percentage
> freq_table <- table(industries)
> percent_table <- prop.table(freq_table) * 100
> # Plot the bar chart
> barplot(
+   percent_table,
+   main = "Percentage Frequency of Complaints by Industry",
+   xlab = "Industry",
+   ylab = "Percentage Frequency",
+   col = "skyblue",
+   las = 2,
+   ylim = c(0, max(percent_table) + 5)
+ )
```




```
> # c). Determining the industry with the highest number of complaints
> highest_complaints <- names(which.max(freq_table))
> cat("The industry with the highest number of complaints is:", highest_complaints,
 "\n")
The industry with the highest number of complaints is: Cell

> # d). Comment on the percentage frequency distribution for complaints
> summary_stats <- round(percent_table, 2)
> cat("Percentage Frequency Distribution of Complaints:\n")
Percentage Frequency Distribution of Complaints:
> print(summary_stats)
industries
      Bank      Cable        Car       Cell Collection
        13         22         21         30         14
> cat("\nComments:\n")

Comments:
> cat("The majority of complaints come from", highest_complaints,
+     "with a notable percentage share, indicating a potential service or satisfact
ion issue in this industry.")
The majority of complaints come from Cell with a notable percentage share, indicati
ng a potential service or satisfaction issue in this industry.
> cat("The distribution shows that some industries have significantly lower complai
nt percentages, suggesting variability in complaint volume across industries.\n")
The distribution shows that some industries have significantly lower complaint perc
entages, suggesting variability in complaint volume across industries.
```

(2)	**Household Incomes.** The following data represent a sample of 14 household incomes ($1000s). Answer the following questions based on this sample. 49.4 52.4 53.4 51.3 52.1 48.7 52.1 52.2 64.5 51.6 46.5 52.9 52.5 51.2 a. What is the median household income for these sample data? b. According to a previous survey, the median annual household income five years ago was $55,000. Based on the sample data above, estimate the percentage change in the median household income from five years ago to today. c. Compute the first and third quartiles. d. Provide a five-number summary. e. Using the *z*-score approach, do the data contain any outliers? Does the approach that uses the values of the first and third quartiles and the interquartile range to detect outliers provide the same results?

```
# a. Median household income
> median_income <- median (household_incomes)
> cat ("Median household income:", median_income, "\n")
Median household income: 52.1
>
> # b. Percentage change in median income
> previous_median <- 55.0 # in $1000s
> percentage_change <- ((median_income - previous_median) / previous_median) * 100
> cat("Percentage change in median income:", percentage_change, "%\n")
Percentage change in median income: -5.272727 %
>
> # c. First and third quartiles
> first_quartile <- quantile(household_incomes, 0.25)
> third_quartile <- quantile(household_incomes, 0.75)
> cat("First quartile (Q1):", first_quartile, "\n")
First quartile (Q1): 51.225
> cat("Third quartile (Q3):", third_quartile, "\n")
Third quartile (Q3): 52.475
>
> # d. Five-number summary
> five_number_summary <- summary(household_incomes)
> cat("Five-number summary:\n")
Five-number summary:
> print(five_number_summary)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  46.50   51.23   52.10   52.20   52.48   64.50
>
> # e. Outlier detection
> # Using Z-score approach
> mean_income <- mean(household_incomes)
> sd_income <- sd(household_incomes)
> z_scores <- (household_incomes - mean_income) / sd_income
> outliers_z <- household_incomes[abs(z_scores) > 3]
> cat("Outliers based on Z-score approach:", outliers_z, "\n")
Outliers based on Z-score approach: 64.5
>
> # Using IQR approach
> iqr <- IQR(household_incomes)
> lower_bound <- first_quartile - 1.5 * iqr
```

```
> upper_bound <- third_quartile + 1.5 * iqr
> outliers_iqr <- household_incomes[household_incomes < lower_bound | household_inc
omes > upper_bound]
> cat("Outliers based on IQR approach:", outliers_iqr, "\n")
Outliers based on IQR approach: 48.7 64.5 46.5
```

(3) **Data file: NFLTeamValue**

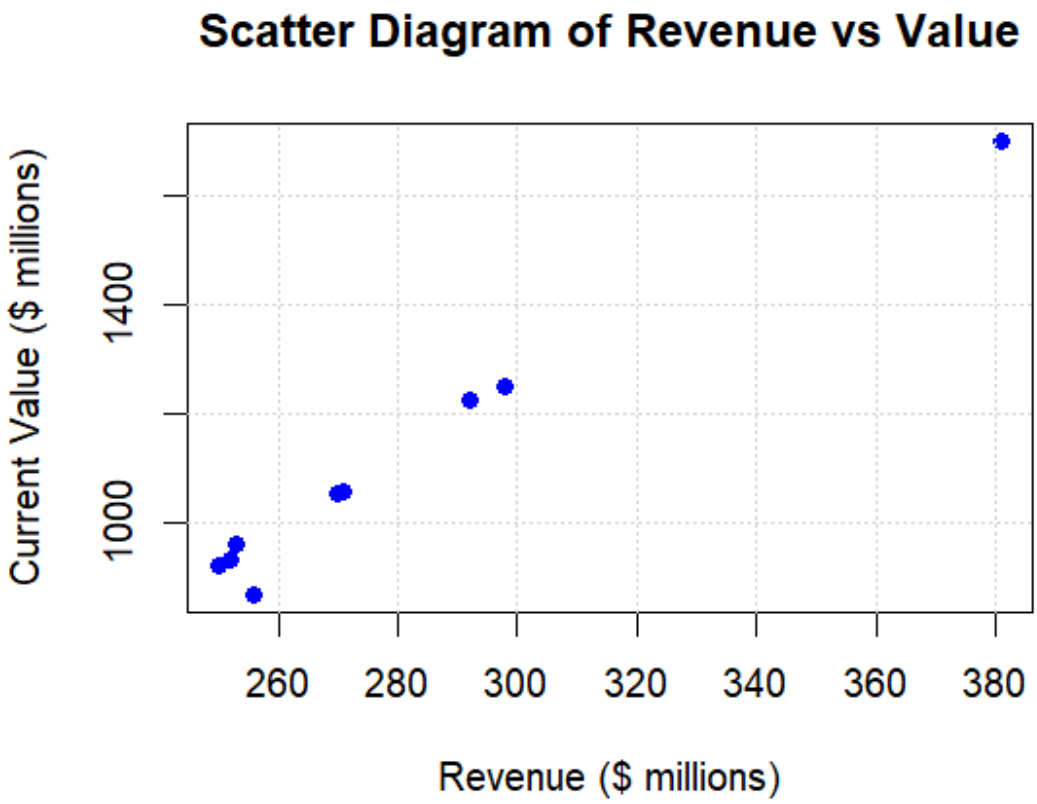
NFL Teams Worth. In 2014, the 32 teams in the National Football League (NFL) were worth, on average, $1.17 billion, 5% more than in 2013. The following data show the annual revenue ($ millions) and the estimated team value ($ millions) for the 32 NFL teams in 2014 (*Forbes* website).

Team	Revenue ($ millions)	Current Value ($ millions)
Arizona Cardinals	253	961
Atlanta Falcons	252	933
Baltimore Ravens	292	1227
Buffalo Bills	256	870
Carolina Panthers	271	1057
Chicago Bears	298	1252
Cincinnati Bengals	250	924
Tennessee Titans	270	1055
Washington Redskins	381	1700

a. Develop a scatter diagram with Revenue on the horizontal axis and Value on the vertical axis. Does there appear that there is any relationship between the two variables?

b. What is the sample correlation coefficient? What can you say about the strength of the relationship between Revenue and Value?

```
# Q.3
> # Data for the NFL teams
> teams <- c("Arizona Cardinals", "Atlanta Falcons", "Baltimore Ravens", "Buffalo B
ills",
+            "Carolina Panthers", "Chicago Bears", "Cincinnati Bengals", "Tennessee
 Titans",
+            "Washington Redskins")
> revenue <- c(253, 252, 292, 256, 271, 298, 250, 270, 381)
> value <- c(961, 933, 1227, 870, 1057, 1252, 924, 1055, 1700)
> # a) Scatter diagram
> plot(revenue, value,
+      main = "Scatter Diagram of Revenue vs Value",
+      xlab = "Revenue ($ millions)",
+      ylab = "Current Value ($ millions)",
+      pch = 16, col = "blue")
> grid()
```




```
>
> # b) Correlation coefficient
> correlation <- cor(revenue, value)
> cat("The correlation coefficient between Revenue and Value is:", correlation, "\n
")
The correlation coefficient between Revenue and Value is: 0.9881032
>
> # Interpretation
> if (abs(correlation) > 0.8) {
+   cat("There is a strong linear relationship between Revenue and Value.\n")
+ } else if (abs(correlation) > 0.5) {
+   cat("There is a moderate linear relationship between Revenue and Value.\n")
+ } else {
+   cat("There is a weak or no linear relationship between Revenue and Value.\n")
+ }
There is a strong linear relationship between Revenue and Value.
```

(4) **Data file: CodeChurn**

Code Churn. Code Churn is a common metric used to measure the efficiency and productivity of software engineers and computer programmers. It's usually measured as the percentage of a programmer's code that must be edited over a short period of time. Programmers with higher rates of code churn must rewrite code more often because of errors and inefficient programming techniques. The following table displays sample information for 10 computer programmers.

Programmer	Total Lines of Code Written	Number of Lines of Code Requiring Edits
Liwei	23,789	4,589
Andrew	17,962	2,780
Jaime	31,025	12,080
Sherae	26,050	3,780
Binny	19,586	1,890
Roger	24,786	4,005
Dong-Gil	24,030	5,785
Alex	14,780	1,052
Jay	30,875	3,872
Vivek	21,546	4,125

a. Use the data in the table above and the relative frequency method to determine probabilities that a randomly selected line of code will need to be edited for each programmer.
b. If you randomly select a line of code from Liwei, what is the probability that the line of code will require editing?
c. If you randomly select a line of code from Sherae, what is the probability that the line of code will *not* require editing?
d. Which programmer has the lowest probability of a randomly selected line of code requiring editing? Which programmer has the highest probability of a randomly selected line of code requiring editing?

```
> # Q.4.
> # Data for the programmers
> programmers <- c("Liwei", "Andrew", "Jaime", "Sherae", "Binny", "Roger", "Dong-Gi
l", "Alex", "Jay", "Vivek")
> total_lines <- c(23789, 17962, 31025, 26050, 19586, 24786, 24030, 14780, 30875, 2
1546)
> lines_requiring_edits <- c(4589, 2780, 12080, 3780, 1890, 4005, 5785, 1052, 3872,
 4125)
>
> # a) Relative frequency of lines requiring edits for each programmer
```

```
> prob_requiring_edit <- lines_requiring_edits / total_lines
> names(prob_requiring_edit) <- programmers
> cat("Probabilities of a randomly selected line requiring edits for each programme
r:\n")
Probabilities of a randomly selected line requiring edits for each programmer:
> print(prob_requiring_edit)
     Liwei     Andrew      Jaime     Sherae      Binny      Roger   Dong-Gil
Alex        Jay      Vivek
0.19290428 0.15477118 0.38936342 0.14510557 0.09649750 0.16158315 0.24074074 0.0711
7727 0.12540891 0.19145085
>
> # b) Probability a randomly selected line of code from Liwei requires editing
> prob_liwei_edit <- prob_requiring_edit["Liwei"]
> cat("Probability that a randomly selected line of code from Liwei requires editin
g:", prob_liwei_edit, "\n")
Probability that a randomly selected line of code from Liwei requires editing: 0.19
29043
>
> # c) Probability a randomly selected line of code from Sherae does not require ed
iting
> prob_sherae_no_edit <- 1 - prob_requiring_edit["Sherae"]
> cat("Probability that a randomly selected line of code from Sherae does not requi
re editing:", prob_sherae_no_edit, "\n")
Probability that a randomly selected line of code from Sherae does not require edit
ing: 0.8548944
>
> # d) Programmers with the lowest and highest probabilities of requiring editing
> min_prob <- min(prob_requiring_edit)
> max_prob <- max(prob_requiring_edit)
> programmer_min_prob <- names(prob_requiring_edit[prob_requiring_edit == min_prob]
)
> programmer_max_prob <- names(prob_requiring_edit[prob_requiring_edit == max_prob]
)
>
> cat("Programmer with the lowest probability of a randomly selected line requiring
 editing:", programmer_min_prob,
+     "with probability:", min_prob, "\n")
Programmer with the lowest probability of a randomly selected line requiring editin
g: Alex with probability: 0.07117727
> cat("Programmer with the highest probability of a randomly selected line requirin
g editing:", programmer_max_prob,
+     "with probability:", max_prob, "\n")
Programmer with the highest probability of a randomly selected line requiring editi
ng: Jaime with probability: 0.3893634
```

(5)	**Americans Without Health Insurance.** The National Center for Health Statistics, housed within the U.S. Centers for Disease Control and Prevention (CDC), tracks the number of adults in the United States who have health insurance. According to this agency, the uninsured rates for Americans in 2018 are as follows: 5.1% of those under the age of 18, 12.4% of those ages 18–64, and 1.1% of those 65 and older do not have health insurance (CDC website). Approximately 22.8% of Americans are under age 18, and 61.4% of Americans are ages 18–64. a. What is the probability that a randomly selected person in the United States is 65 or older? b. Given that the person is an uninsured American, what is the probability that the person is 65 or older?

```
# Q.5.
> # Population proportions for each age group
> prop_under_18 <- 0.228
> prop_18_to_64 <- 0.614
> prop_65_or_older <- 1 - (prop_under_18 + prop_18_to_64)
>
> # Uninsured rates for each age group
> uninsured_under_18 <- 0.051
> uninsured_18_to_64 <- 0.124
> uninsured_65_or_older <- 0.011
>
> # a) Probability that a randomly selected person is 65 or older
> prob_65_or_older <- prop_65_or_older
> cat("Probability that a randomly selected person is 65 or older:", prob_65_or_old
er, "\n")
Probability that a randomly selected person is 65 or older: 0.158
>
> # b) Conditional probability: Given that the person is uninsured, what is the pro
bability they are 65 or older?
> # Numerator: Proportion of people who are uninsured and 65 or older
> num_uninsured_65_or_older <- prop_65_or_older * uninsured_65_or_older
>
> # Denominator: Total proportion of uninsured people
> total_uninsured <- (prop_under_18 * uninsured_under_18) +
+ (prop_18_to_64 * uninsured_18_to_64) +
+ (prop_65_or_older * uninsured_65_or_older)
>
> # Conditional probability
> prob_65_or_older_given_uninsured <- num_uninsured_65_or_older / total_uninsured
> cat("Probability that a person is 65 or older given they are uninsured:", prob_65
_or_older_given_uninsured, "\n")
Probability that a person is 65 or older given they are uninsured: 0.01941856
```