國立政治大學 114 學年度第一學期
統計學(一) 期中 R 程式加分考-解答卷
(助教：陳宣岑)

(1) **(5 分)**	**用 R 印出下列字句(姓名改為自己的姓名):** "本人(學號)(姓名)恪遵各項考試規則，若如違反，願受校方最嚴厲處罰，謹誓。"

```
> # ex1
> cat("本人(12345)(吳漢銘)恪遵各項考試規則，若如違反，願受校方最嚴厲處罰，謹誓。")
本人(12345)(吳漢銘)恪遵各項考試規則，若如違反，願受校方
```

(2)
(20 分)

Data file: Hypertension.xlsx

Hypertension and Heart Disease. People often wait until middle age to worry about having a healthy heart. However, many studies have shown that earlier monitoring of risk factors such as blood pressure can be very beneficial (*The Wall Street Journal*). Having higher than normal blood pressure, a condition known as hypertension, is a major risk factor for heart disease. Suppose a large sample of individuals of various ages and gender was selected and that each individual's blood pressure was measured to determine if they have hypertension. For the sample data, the following table shows the percentage of individuals with hypertension.

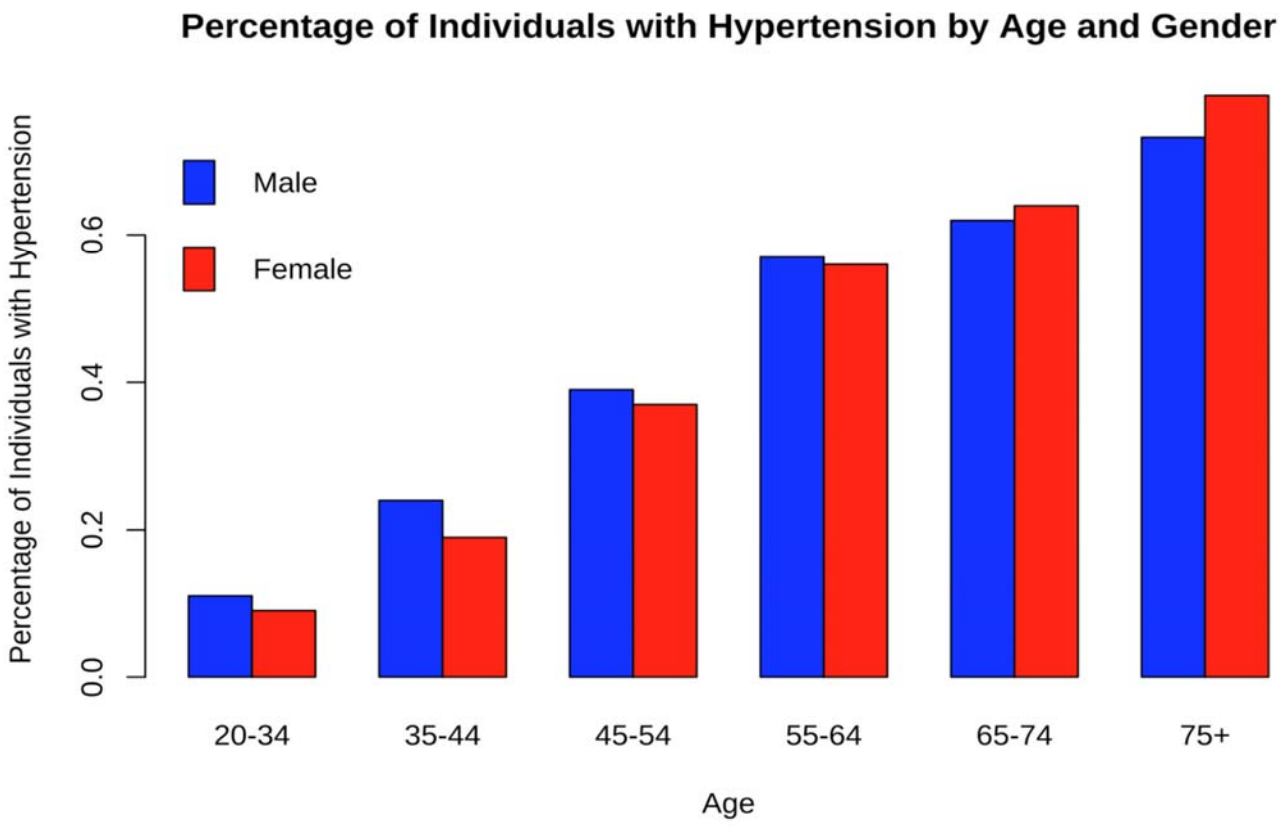
Age	Male	Female
20–34	11.00%	9.00%
35–44	24.00%	19.00%
45–54	39.00%	37.00%
55–64	57.00%	56.00%
65–74	62.00%	64.00%
75+	73.30%	79.00%

a. Develop a side-by-side bar chart with age on the horizontal axis, the percentage of individuals with hypertension on the vertical axis, and side-by-side bars based on gender.

```
> #Ex2
```

```
> #a.
> Hypertension_data <- read_excel("/Users/chenhsuantsen/Downloads/114-1-Stat-R-
Midterm/data/Hypertension.xlsx")

> Hypertension_matrix <- t(Hypertension_data[, 2:3])
> colnames(Hypertension_matrix) <- Hypertension_data[[1]]
> barplot(Hypertension_matrix,
+         beside = TRUE,
+         col = c("blue", "red"),
+         main = "Percentage of Individuals with Hypertension by Age and Gender",
+         xlab = "Age",
+         ylab = "Percentage of Individuals with Hypertension",
+         legend.text = rownames(Hypertension_matrix),
+         args.legend = list(x = "topleft", bty = "n"))
```



(3)
(25 分)

Data file: Colleges

Colleges' Year Founded and Percent Graduated. Refer to the data set in Table 2.18.

TABLE 2.18 Data for a Sample of Private Colleges and Universities

School	Year Founded	Tuition & Fees	% Graduate
American University	1893	$36,697	79.00
Baylor University	1845	$29,754	70.00
Belmont University	1951	$23,680	68.00
.	.	.	.
.	.	.	.
.	.	.	.
Wofford College	1854	$31,710	82.00
Xavier University	1831	$29,970	79.00
Yale University	1701	$38,300	98.00

a. Construct a crosstabulation with Year Founded as the row variable and % Graduate as the column variable. Use classes starting with 1600 and ending with 2000 in increments of 50 for Year Founded. For % Graduate, use classes starting with 35% and ending with 100% in increments of 5%.
b. Compute the row percentages for your crosstabulation in part (a).
c. Comment on any relationship between the variables.
d. Construct a histogram.
e. Construct a scatter diagram to show the relationship between Year Founded and Tuition & Fees.

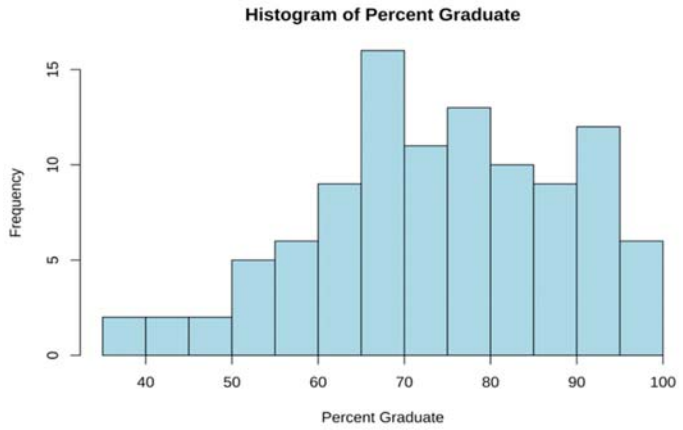
```
> #Ex3
> Colleges_data <- read_excel("/Users/chenhsuantsen/Downloads/114-1-Stat-R-
Midterm/data/Colleges.xlsx")
> YearFounded_col <- as.numeric(Colleges_data[['Year Founded']])
> PercentGraduate_col <- as.numeric(Colleges_data[['% Graduate']])
> TuitionFees_col <- as.numeric(Colleges_data[['Tuition & Fees']])

> #a.
> breaks_year <- seq(1600, 2000, by = 50)
> breaks_grad <- seq(35, 100, by = 5)
>
> YearFounded_cut <- cut(YearFounded_col, breaks = breaks_year, right = FALSE,
include.lowest = TRUE)
> PercentGraduate_cut <- cut(PercentGraduate_col, breaks = breaks_grad, right =
FALSE, include.lowest = TRUE)
>
> crosstabulation <- table(YearFounded_cut, PercentGraduate_cut)
> print(crosstabulation)
```

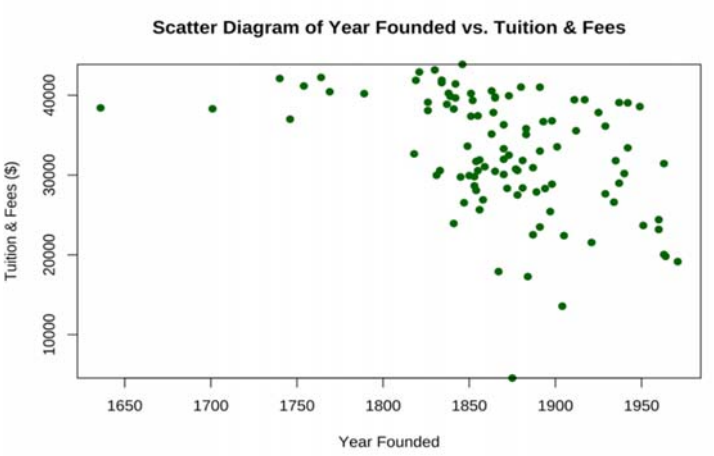
```
                  PercentGraduate_cut
YearFounded_cut [35,40) [40,45) [45,50) [50,55) [55,60) [60,65) [65,70) [70,75) [75,80) [80,85) [85,90) [90,95) [95,100]
   [1600, 1650)       0       0       0       0       0       0       0       0       0       0       0       0        1
   [1650, 1700)       0       0       0       0       0       0       0       0       0       0       0       0        0
   [1700, 1750)       0       0       0       0       0       0       0       0       0       0       0       0        3
   [1750, 1800)       0       0       0       0       0       0       0       0       0       0       0       1        3
   [1800, 1850)       0       0       0       0       0       1       2       4       2       3       4       3        2
   [1850, 1900)       0       0       1       2       4       3      11       5       9       6       3       4        1
   [1900, 1950)       1       1       1       0       1       3       0       3       2       4       1       1        0
   [1950, 2000]       1       0       1       3       0       0       2       0       0       0       0       0        0
> #b.
> row_percentages <- prop.table(crosstabulation, margin = 1) * 100
> print(row_percentages)
                  PercentGraduate_cut
YearFounded_cut    [35,40)    [40,45)    [45,50)    [50,55)    [55,60)    [60,65)    [65,70)    [70,75)    [75,80)    [80,85)    [85,90)    [90,95)   [95,100]
   [1600, 1650)   0.000000   0.000000   0.000000   0.000000   0.000000   0.000000   0.000000   0.000000   0.000000   0.000000   0.000000   0.000000 100.000000
   [1650, 1700)
   [1700, 1750)   0.000000   0.000000   0.000000   0.000000   0.000000   0.000000   0.000000   0.000000   0.000000   0.000000   0.000000   0.000000 100.000000
   [1750, 1800)   0.000000   0.000000   0.000000   0.000000   0.000000   0.000000   0.000000   0.000000   0.000000   0.000000   0.000000  25.000000  75.000000
   [1800, 1850)   0.000000   0.000000   0.000000   0.000000   0.000000   4.761905   9.523810  19.047619   9.523810  14.285714  19.047619  14.285714   9.523810
   [1850, 1900)   0.000000   0.000000   2.040816   4.081633   8.163265   6.122449  22.448980  10.204082  18.367347  12.244898   6.122449   8.163265   2.040816
   [1900, 1950)   5.555556   5.555556   5.555556   0.000000   5.555556  16.666667   0.000000  16.666667  11.111111  22.222222   5.555556   5.555556   0.000000
   [1950, 2000]  14.285714   0.000000  14.285714  42.857143   0.000000   0.000000  28.571429   0.000000   0.000000   0.000000   0.000000   0.000000   0.000000
> #c.
```

This suggests a slight negative correlation; that is, institutions founded earlier tend to have a higher percentage of graduates.

```
> #d.
> hist(PercentGraduate_col,
+      main = "Histogram of Percent Graduate",
+      xlab = "Percent Graduate",
+      ylab = "Frequency",
+      breaks = breaks_grad,
+      col = "lightblue",
+      border = "black")
```




```
> #e.
> plot(YearFounded_col, TuitionFees_col,
+      main = "Scatter Diagram of Year Founded vs. Tuition & Fees",
+      xlab = "Year Founded",
+      ylab = "Tuition & Fees ($)",
+      pch = 19,
+      col = "darkgreen")
```



(4)
(25 分)

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?

```
> #Ex4
> Household_Incomes <- c(53.4, 49.4, 51.3, 52.1, 48.7, 52.1, 52.9, 52.2, 64.5,
51.6, 46.5, 52.5, 51.2, 52.4)
> #a.
> median_income <- median(Household_Incomes)
> cat("a. Median Household Income (in $1000s): ", median_income, "\n")
a. Median Household Income (in $1000s):  52.1

> #b.
> median_five_years_ago <- 55
> percentage_change <- ((median_income - median_five_years_ago) /
median_five_years_ago) * 100
> cat("b. Percentage Change in Median Household Income: ", percentage_change,
"%\n")
b. Percentage Change in Median Household Income:  -5.272727 %
```

```
> #c.
> quartiles <- quantile(Household_Incomes, probs = c(0.25, 0.75), type = 6)
> Q1 <- quartiles["25%"]
> Q3 <- quartiles["75%"]
> cat("c. First Quartile (Q1) (in $1000s): ", Q1, "\n")
c. First Quartile (Q1) (in $1000s):  50.75
> cat("c. Third Quartile (Q3) (in $1000s): ", Q3, "\n")
c. Third Quartile (Q3) (in $1000s):  52.6

> #d.
> five_number_summary <- summary(Household_Incomes)
> 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.
> Detect_Outlier <- function(x){
+   Q1 <- quantile(x, type = 6, probs = 0.25)
+   Q3 <- quantile(x, type = 6, probs = 0.75)
+   IQR.x <- IQR(x, type = 6)
+   Lower.Limit <- Q1 - 1.5 * IQR.x
+   Upper.Limit <- Q3 + 1.5 * IQR.x
+   outlier_idx <- which(x < Lower.Limit | x > Upper.Limit)
+   if(length(outlier_idx) > 0){
+     return(data.frame(Method = "IQR", Outliers = x[outlier_idx], Indices =
outlier_idx))
+   } else {
+     return(data.frame(Method = "IQR", Outliers = "None", Indices = "None"))
+   }
+ }
> #e.Z-Score
> mean_income <- mean(Household_Incomes)
> sd_income <- sd(Household_Incomes)
> z_scores <- (Household_Incomes - mean_income) / sd_income
> z_outliers <- Household_Incomes[abs(z_scores) > 3]
> cat("e. Z-Score :\n")
e. Z-Score:
> if (length(z_outliers) > 0) {
+   cat("Outliers (in $1000s): ", z_outliers, "\n")
+ } else {
+   cat("No outliers detected using the Z-Score approach.\n")
+ }
Outliers (in $1000s):  64.5
> #e.IQR
> iqr_outliers <- Detect_Outlier(Household_Incomes)
> cat("e. IQR:\n")

e. IQR:
> print(iqr_outliers)
  Method Outliers Indices
1    IQR     64.5       9
2    IQR     46.5      11

They do not have same results.
```

(5) **(25 分)**	**Data file: Coldstream12** **Golf Scores.** During the summer of 2018, Coldstream Country Club in Cincinnati, Ohio, collected data on 443 rounds of golf played from its white tees. The data for each golfer's score on the twelfth hole are contained in the DATAfile *Coldstream12*. a. Construct an empirical discrete probability distribution for the player scores on the twelfth hole. b. A *par* is the score that a good golfer is expected to get for the hole. For hole number 12, par is four. What is the probability of a player scoring less than or equal to par on hole number 12? c. What is the expected score for hole number 12? d. What is the variance for hole number 12? e. What is the standard deviation for hole number 12?

```
> #Ex5
> Coldstream12.tmp <- read_excel("/Users/chenhsuantsen/Downloads/114-1-Stat-R-
Midterm/data/Coldstream12.xlsx")
> Coldstream12 <- Coldstream12.tmp[[1]]
>
> #a.
> score_table <- table(Coldstream12)
> score_prob_dist <- data.frame(
+   Score_x = as.numeric(names(score_table)),
+   Frequency_f = as.numeric(score_table),
+   Prob_f_x = as.numeric(score_table) / total_rounds
+ )
> print(score_prob_dist)
  Score_x Frequency_f    Prob_f_x
1       3           4 0.009029345
2       4          57 0.128668172
3       5         212 0.478555305
4       6         139 0.313769752
5       7          27 0.060948081
6       8           4 0.009029345

> #b.
> par_score <- 4
> prob_le_par <- sum(score_prob_dist$Prob_f_x[score_prob_dist$Score_x <=
par_score])
> cat("b. Probability of scoring <= Par (4): ", prob_le_par, "\n")
b. Probability of scoring <= Par (4):  0.1376975

> #c.
> expected_score <- sum(score_prob_dist$Score_x * score_prob_dist$Prob_f_x)
> cat("c. Expected Score (E(X)) for hole 12: ", expected_score, "\n")
c. Expected Score (E(X)) for hole 12:  5.316027

> #d.
> variance <- sum((score_prob_dist$Score_x - expected_score)^2 *
score_prob_dist$Prob_f_x)
> cat("(5) d. Variance (Var(X)) for hole 12: ", variance, "\n")
(5) d. Variance (Var(X)) for hole 12:  0.7037386

> #e.
```

```
> standard_deviation <- sqrt(variance)
> cat("e. Standard Deviation (SD(X)) for hole 12: ", standard_deviation, "\n")
e. Standard Deviation (SD(X)) for hole 12:  0.8388913
```

(6) (20 分)	**Web Browser Market Share.** Market-share-analysis company Net Applications monitors and reports on Internet browser usage. According to Net Applications, in the summer of 2014, Google's Chrome browser exceeded a 20% market share for the first time, with a 20.37% share of the browser market (*Forbes* website). For a randomly selected group of 20 Internet browser users, answer the following questions. a. Compute the probability that exactly 8 of the 20 Internet browser users use Chrome as their Internet browser. b. Compute the probability that at least 3 of the 20 Internet browser users use Chrome as their Internet browser. c. For the sample of 20 Internet browser users, compute the expected number of Chrome users. d. For the sample of 20 Internet browser users, compute the variance and standard deviation for the number of Chrome users. **((a)(b)禁止用 R 直接加減乘除做計算，請利用 R 套件或指令運算)**

```
> #Ex.6
> n <- 20
> p <- 0.2037
>
> #a.
> prob_exactly_8 <- dbinom(x = 8, size = n, prob = p)
> cat("dbinom(x = 8, size = 20, prob = 0.2037) = ", prob_exactly_8, "\n")
dbinom(x = 8, size = 20, prob = 0.2037) =  0.02427258

> # b.
> prob_at_least_3 <- 1 - pbinom(q = 2, size = n, prob = p)
> cat("1 - pbinom(q = 2, size = 20, prob = 0.2037) = ", prob_at_least_3, "\n")
1 - pbinom(q = 2, size = 20, prob = 0.2037) =  0.8050771

> #c.
> expected_number <- n * p
> cat("20 * 0.2037 = ", expected_number, "\n")
20 * 0.2037 =  4.074

> #d.
> variance <- n * p * (1 - p)
> std_dev <- sqrt(variance)
> cat("Variance (Var(X) = n*p*(1-p)): ", variance, "\n")
Variance (Var(X) = n*p*(1-p)):  3.244126
> cat("Standard Deviation (SD(X) = sqrt(Var(X))): ", std_dev, "\n")
Standard Deviation (SD(X) = sqrt(Var(X))):  1.801146
```