授課教師：曾智義　　課程名稱：進階 R 資料分析與應用　學期：112 第二學期

一、第三週課堂練習

單樣本 t 檢定、雙樣本 t 檢定、成對雙樣本 t 檢定、ANOVA 檢定

二、個人/成員：

A1093325 黃紹瑜 資訊管理學系

三、議題規劃

(一) 單樣本 t 檢定：檢視此資料中平均 BMI 值為多少。
(二) 雙樣本 t 檢定：檢視男女間 BMI 值是否有顯著差異
(三) 成對雙樣本 t 檢定：檢視學生在 technology 和 entertainment 支出間是否有顯著差異
(四) ANOVA 檢定：比較金額與購買的類別是否有差

四、問題定義

(一) 單樣本 t 檢定：由《馬偕護理雜誌》第 8 卷 1 期中得知慢性精神病人因疾病症狀及藥物服用，常導致體重過重，其平均 BMI 為 30kg/m^2。故想驗證此資料是否為慢性精神病患者的資料。
(二) 雙樣本 t 檢定：想了解男女之間的肥胖程度有沒有差異。
(三) 成對雙樣本 t 檢定：想了解在和學生最相關的娛樂與科技兩項支出會不會有明顯的支出差異。
(四) ANOVA 檢定：想了解商場中的各個類別的收入是否有差異。

五、程式碼設計 和 執行結果

(一) 資料集介紹（以下只列出部分表頭說明）：

1. data_insurance：由 kaggle 中找到的醫療保險費用
   (1). age: The insured person's age.
   (2). sex: Gender (male or female) of the insured.
   (3). bmi: A measure of body fat based on height and weight.
2. data_student：由 kaggle 中找到的學生支出習慣表
   (1). age: Age of the student (in years)
   (2). gender: Gender of the student (Male, Female, Non-binary)
   (3). year_in_school: Year of study (Freshman, Sophomore, Junior, Senior)
   (4). major: Field of study or major
   (5). technology: Expenses for technology (in dollars)
   (6). entertainment: Expenses for entertainment (in dollars)
3. data_shopping：由 kaggle 中找到的顧客消費傾向
   (1). Id - Unique identifier for each customer
   (2). Age - Age of the customer
   (3). Category - Category of the item purchased
   (4). Purchase - The amount of the purchase in USD

(二) 程式碼：

```
library(readr) # 載入讀取資料套件
library(ggplot2) # 載入繪圖套件
library(plyr) # 載入資料處理套件

```

```r
## 資料建立
# 讀取 mock.csv 檔案
data_insurance <- read_csv("/Users/shaoyu/Desktop/🏬/2 進階 R/dataset/insurance.csv")

# 取前面 10 筆資料查看
head(data_insurance)

# 顯示出不同的申請條件
unique(data_insurance$sex)

# ------------------------------------------------------- #

## 單樣本 t 檢定
# 以雙邊建立 t 檢定
# 結果：30.5 非真正平均值
t.test(data_insurance$bmi, alternative = "two.sided", mu = 30.5)

# 建立一個 t 分佈
randT <- rt(30000, df=NROW(data_insurance)-1)

#進行 t 檢定
chargeTTest <- t.test(data_insurance$bmi, alternative="two.sided", mu=30.50)

# 繪製密度圖
# 繪圖得：實線於信賴區間內，表示平均與 30.5 並無顯著差異
ggplot(data.frame(x=randT)) +
  geom_density(aes(x=x), fill="grey", color="grey") +
  geom_vline(xintercept=chargeTTest$statistic) +
  geom_vline(xintercept=mean(randT) +
               c(-2, 2)*sd(randT), linetype=2)

# 以單邊建立 t 檢定
# 結果：真正平均值大於 30.5
t.test(data_insurance$bmi, alternative = "greater", mu = 30.5)

# ------------------------------------------------------- #

## 雙樣本 t 檢定
# 計算每個性別 bmi 的變異數
aggregate(bmi ~ sex, data=data_insurance, var)

# 進行 Shapiro-Wilk normality test
shapiro.test(data_insurance$bmi) # 檢定所有樣本的 BMI
shapiro.test(data_insurance$bmi[data_insurance$sex == "male"]) # 檢定男性樣本的 BMI
shapiro.test(data_insurance$bmi[data_insurance$sex == "female"]) # 檢定女性樣本的 BMI

```

```r
# 繪製直方圖
ggplot(data_insurance, aes(x=bmi, fill=sex))  +
  geom_histogram(binwidth=.5, alpha=1/2) +
  theme(text=element_text(family="Helvetica", size=14))

# 使用 ansari.test 執行方差齊性檢定
# 結果： p-value > 0.05
ansari.test(bmi ~ sex, data_insurance)

# 使用 t.test 執行等方差的獨立樣本 t 檢定
# 結果： t < 1.96 不顯著
t.test(bmi ~ sex, data = data_insurance, var.equal = TRUE)

# 根據 sex 進行分組
# 計算每個組的 BMI 平均值和標準差
# 計算 95%的信心區間的上下界
resultSummary <- ddply(data_insurance, "sex", summarize,
                       money.mean=mean(bmi), money.sd=sd(bmi),
                       Lower=money.mean - 2*money.sd/sqrt(NROW(bmi)),
                       Upper=money.mean + 2*money.sd/sqrt(NROW(bmi)))
resultSummary

# 畫圖顯示信賴區間
# 結果： 信賴區間重疊表不顯著
ggplot(resultSummary, aes(x=money.mean, y=sex)) + geom_point() +
  geom_errorbarh(aes(xmin=Lower, xmax=Upper), height=.2)

# ------------------------------------------------------------ #
## 成對雙樣本 t 檢定
data_student <- read_csv("/Users/shaoyu/Desktop/🏠/2 進階 R/dataset/student_spending.csv")
head(data_student)

# 比較學生在 technology 支出和 entertainment 支出之間的平均差異
# 結果： t-value > 1.96 有顯著差異
t.test(data_student$technology, data_student$entertainment, paired = TRUE)

# 計算 technology 支出和 entertainment 支出的差異
entSub <- data_student$technology - data_student$entertainment

# 繪製密度圖
# 結果： 平均數不為零，因此兩者的支出是不相等的
ggplot(data_student, aes(x=technology - entertainment)) +
  geom_density() + # 添加密度曲線
  geom_vline(xintercept=mean(entSub)) + # 添加平均值的垂直線
  geom_vline(xintercept=mean(entSub) + # 添加 95%信賴區間的垂直線， 以標示差異的統計顯著性
               2*c(-1, 1)*sd(entSub)/sqrt(nrow(data_student)),
             linetype=2)
```

```r

# ---------------------------------------------------------- #
## 變異數分析 (比較各商品的售出金額)
data_shopping <- read_csv("/Users/shaoyu/Desktop/🏬/2 進階 R/dataset/shopping_trends.csv")
names(data_shopping) <- c("Id", "Age", "Gender", "Item",
                          "Category", "Purchase", "Location", "Size",
                          "Color", "Season", "Rating",
                          "Subscribe", "Payment", "Ship_Type", "Discount",
                          "PromoCode", "Previous_Pay", "Prefer_Payment",
                          "Frequency")
head(data_shopping)

# 無截距的 ANOVA 分析，分析了 Category 對 Purchase 的影響
shop_Anova <- aov(Purchase ~ Category - 1, data_shopping)
# 有截距的 ANOVA 分析，分析了 Category 對 Purchase 的影響
shop_Intercept <- aov(Purchase ~ Category, data_shopping)

# 檢視係數
shop_Anova$coefficients
shop_Intercept$coefficients

# 對無截距的 ANOVA 分析進行摘要統計
# Category 有潛著顯著差異
summary(shop_Anova)

# 對 data_shopping 資料依據 Category 進行分組
# 計算每個分組的 Purchase 平均值、標準差、樣本數以及 95%的信賴區間
shop_DDply <- ddply(data_shopping, "Category", summarize,
                    Purchase.mean=mean(Purchase), Purchase.sd=sd(Purchase),
                    Length=NROW(Purchase),
                    tfrac=qt(p=.90, df=Length-1),
                    Lower=Purchase.mean - tfrac*Purchase.sd/sqrt(Length),
                    Upper=Purchase.mean + tfrac*Purchase.sd/sqrt(Length)
)

# 畫圖顯示信賴區間
# 結果：outerwear 與其餘三種不重疊，有顯著不同
ggplot(shop_DDply, aes(x=Purchase.mean, y=Category)) + geom_point() +
  geom_errorbarh(aes(xmin=Lower, xmax=Upper), height=.3)
```

(三) 執行結果：

1. head(data_insurance)

```
# A tibble: 6 × 7
    age sex      bmi children smoker region    charges
  <dbl> <chr>  <dbl>    <dbl> <chr>  <chr>       <dbl>
1    19 female  27.9        0 yes    southwest  16885.
2    18 male    33.8        1 no     southeast   1726.
3    28 male    33          3 no     southeast   4449.
4    33 male    22.7        0 no     northwest  21984.
5    32 male    28.9        0 no     northwest   3867.
6    31 female  25.7        0 no     southeast   3757.
```

圖一、data_insurance 資料集展示

2. unique(data_insurance$sex)

```
[1] "female" "male"
```

圖 二、顯示出不同的申請條件

3. t.test(data_insurance$bmi, alternative = "two.sided", mu = 30.5)

```
        One Sample t-test

data:  data_insurance$bmi
t = 0.9801, df = 1337, p-value = 0.3272
alternative hypothesis: true mean is not equal to 30.5
95 percent confidence interval:
 30.33635 30.99045
sample estimates:
mean of x
  30.6634
```

圖 三、雙邊建立單樣本 t 檢定

4. ggplot(data.frame(x=randT)) +
   geom_density(aes(x=x), fill="grey", color="grey") +
   geom_vline(xintercept=chargeTTest$statistic) +
   geom_vline(xintercept=mean(randT) +
   c(-2, 2)*sd(randT), linetype=2)

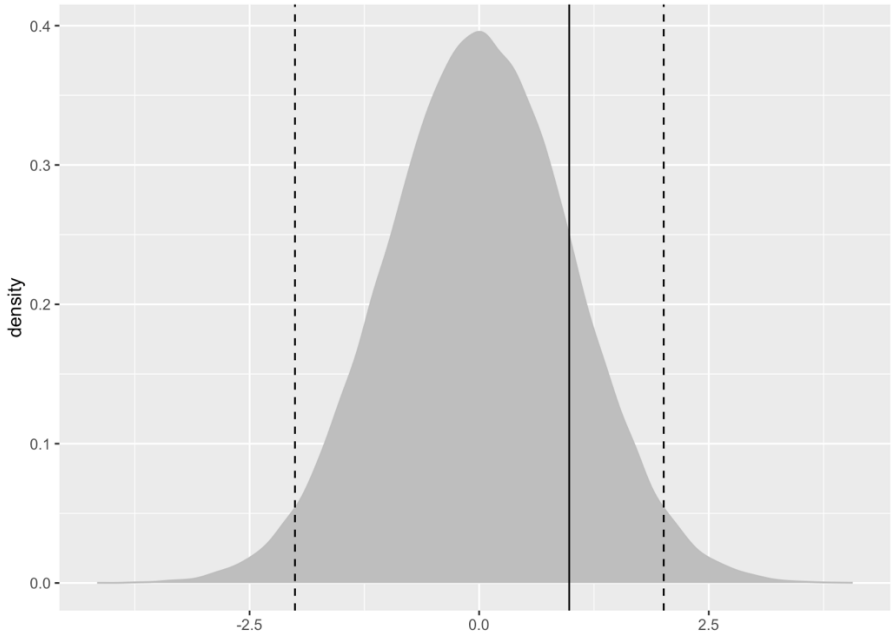

圖 四、密度圖

5. t.test(data_insurance$bmi, alternative = "greater", mu = 30.5)

```
        One Sample t-test

data:  data_insurance$bmi
t = 0.9801, df = 1337, p-value = 0.1636
alternative hypothesis: true mean is greater than 30.5
95 percent confidence interval:
 30.38899      Inf
sample estimates:
mean of x
  30.6634
```

圖 五、以單邊建立 t 檢定

6. aggregate(bmi ~ sex, data=data_insurance, var)

```
     sex      bmi
1 female 36.55440
2   male 37.70494
```

圖 六、每個性別 bmi 的變異數

7. shapiro.test(data_insurance$bmi) I

```
        Shapiro-Wilk normality test

data:  data_insurance$bmi
W = 0.99389, p-value = 2.605e-05
```

圖 七、檢定所有樣本的 BMI

8. shapiro.test(data_insurance$bmi[data_insurance$sex == "male"])

```
        Shapiro-Wilk normality test

data:  data_insurance$bmi[data_insurance$sex == "male"]
W = 0.99305, p-value = 0.003133
```

圖 八、檢定男性樣本的 BMI

9. shapiro.test(data_insurance$bmi[data_insurance$sex == "female"])

```
        Shapiro-Wilk normality test

data:  data_insurance$bmi[data_insurance$sex == "female"]
W = 0.99303, p-value = 0.003543
```

圖 九、檢定女性樣本的 BMI

10. ggplot(data_insurance, aes(x=bmi, fill=sex)) +
    geom_histogram(binwidth=.5, alpha=1/2) +
    theme(text=element_text(family="Helvetica", size=14))

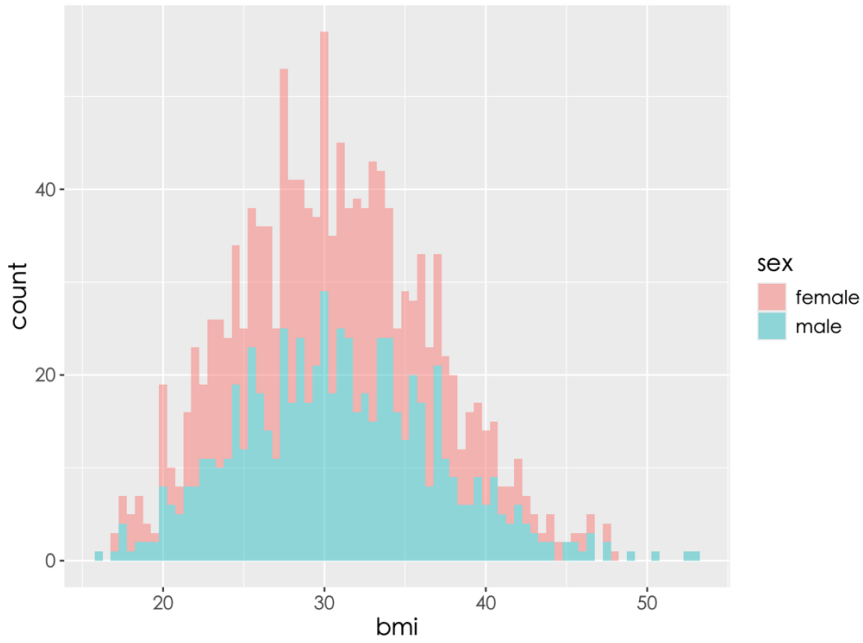

圖 十、以 BMI 根據性別繪出直方圖

11. ansari.test(bmi ~ sex, data_insurance)

```
        Ansari-Bradley test

data:  bmi by sex
AB = 220629, p-value = 0.7467
alternative hypothesis: true ratio of scales is not equal to 1
```

圖 十一、方差齊性檢定結果

12. t.test(bmi ~ sex, data = data_insurance, var.equal = TRUE)

```
        Two Sample t-test

data:  bmi by sex
t = -1.6968, df = 1336, p-value = 0.08998
alternative hypothesis: true difference in means between group female and group male is not equal
to 0
95 percent confidence interval:
 -1.21905646  0.08829755
sample estimates:
mean in group female   mean in group male
            30.37775             30.94313
```

圖 十二、等方差的獨立樣本 t 檢定結果

13. resultSummary

```
     sex money.mean money.sd    Lower    Upper
1 female   30.37775 6.046023 29.90778 30.84772
2   male   30.94313 6.140435 30.47079 31.41547
```

圖 十三、顯示個性別的 BMI 平均值和標準差與 95%的信心區間的上下界

14. ggplot(resultSummary, aes(x=money.mean, y=sex)) + geom_point() +
    geom_errorbarh(aes(xmin=Lower, xmax=Upper), height=.2)

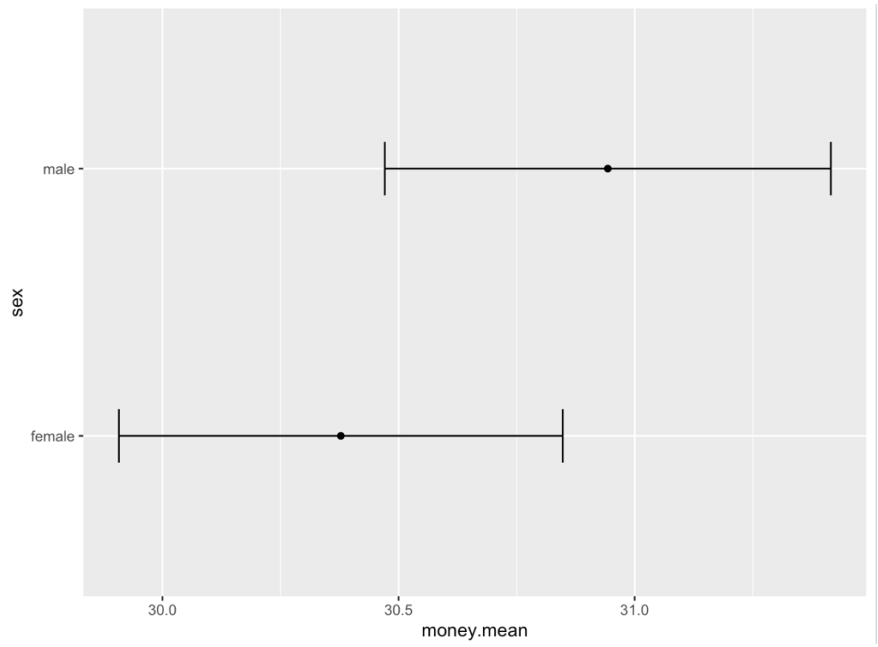

圖 十四、信賴區間

15. head(data_student)

```
# A tibble: 6 × 18
   ...1   age gender     year_…¹ major month…² finan…³ tuition housing  food trans…⁴ books…⁵
  <dbl> <dbl> <chr>      <chr>   <chr>   <dbl>   <dbl>   <dbl>   <dbl> <dbl>   <dbl>   <dbl>
1     0    19 Non-binary Freshm… Psyc…     958     270    5939     709   296     123     188
2     1    24 Female     Junior  Econ…    1006     875    4908     557   365      85     252
3     2    24 Non-binary Junior  Econ…     734     928    3051     666   220     137      99
4     3    23 Female     Senior  Comp…     617     265    4935     652   289     114     223
5     4    20 Female     Senior  Comp…     810     522    3887     825   372     168     194
6     5    25 Non-binary Sophom… Comp…     523     790    3151     413   386     122     131
```

圖 十五、data_student 資料

16. t.test(data_student$technology, data_student$entertainment, paired = TRUE)

```
        Paired t-test

data:  data_student$technology and data_student$entertainment
t = 36.606, df = 999, p-value < 2.2e-16
alternative hypothesis: true mean difference is not equal to 0
95 percent confidence interval:
 88.47822 98.50178
sample estimates:
mean difference
          93.49
```

圖 十六、學生在 technology 支出和 entertainment 支出之間的平均差異

17. ggplot(data_student, aes(x=technology - entertainment)) +
    geom_density() +
    geom_vline(xintercept=mean(entSub)) +
    geom_vline(xintercept=mean(entSub) +
    2*c(-1, 1)*sd(entSub)/sqrt(nrow(data_student)),
    linetype=2)

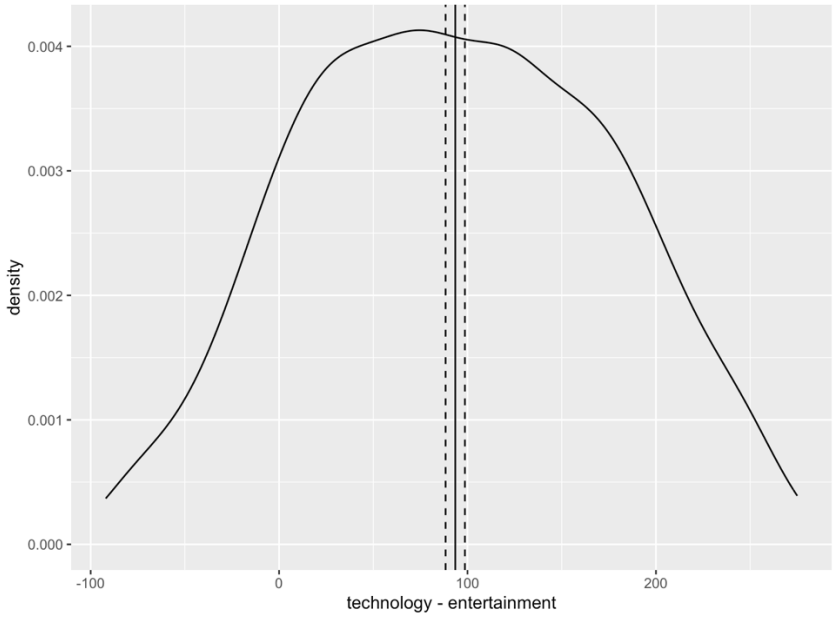

圖 十七、密度圖

18. head(data_shopping)

```
# A tibble: 6 × 19
     Id   Age Gender Item  Categ…¹ Purch…² Locat…³ Size  Color Season Rating Subsc…⁴ Payment
  <dbl> <dbl> <chr>  <chr> <chr>     <dbl> <chr>   <chr> <chr> <chr>   <dbl> <chr>   <chr>
1     1    55 Male   Blou… Clothi…      53 Kentuc… L     Gray  Winter    3.1 Yes     Credit…
2     2    19 Male   Swea… Clothi…      64 Maine   L     Maro… Winter    3.1 Yes     Bank T…
3     3    50 Male   Jeans Clothi…      73 Massac… S     Maro… Spring    3.1 Yes     Cash
4     4    21 Male   Sand… Footwe…      90 Rhode … M     Maro… Spring    3.5 Yes     PayPal
5     5    45 Male   Blou… Clothi…      49 Oregon  M     Turq… Spring    2.7 Yes     Cash
6     6    46 Male   Snea… Footwe…      20 Wyoming M     White Summer    2.9 Yes     Venmo
```

圖 十八、data_shopping 資料

19. shop_Anova$coefficients

```
      CategoryAccessories     CategoryClothing     CategoryFootwear     CategoryOuterwear
                 59.83871             60.02533             60.25543             57.17284
```

圖 十九、無截距的 ANOVA 分析（Category 對 Purchase 的影響）

20. shop_Intercept$coefficients

```
     (Intercept)  CategoryClothing  CategoryFootwear CategoryOuterwear
      59.8387097         0.1866214         0.4167160        -2.6658702
```

圖 二十、有截距的 ANOVA 分析（Category 對 Purchase 的影響）

21. summary(shop_Anova)

```
                Df   Sum Sq Mean Sq F value Pr(>F)
Category     4 13932382 3483096    6211 <2e-16 ***
Residuals 3896  2184885     561
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

圖 二十一、無截距的 ANOVA 分析進行摘要統計

22. ggplot(shop_DDply, aes(x=Purchase.mean, y=Category)) + geom_point() +
    geom_errorbarh(aes(xmin=Lower, xmax=Upper), height=.3)

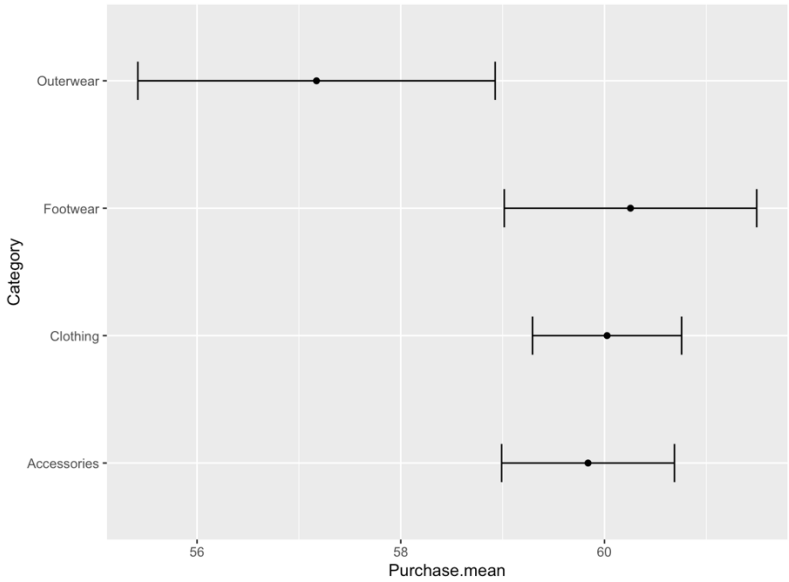

圖 二十二、信賴區間

六、意涵詮釋

(一) 學習如何讀取 CSV 檔案並整理表頭。
(二) 如何繪製簡易視覺化圖形
(三) 第 17~38 行（檢視 BMI 平均值是否為 30.5）：藉由 t 檢定的單一樣本找出此資料集內的平均 BMI 值與 30.5 不存在顯著差異。
(四) 第 46~76 行（檢視男女間 BMI 值是否有顯著差異）：先由繪圖顯示出並不是常態分佈，再由雙樣本 t 檢定看出 t 值未超過 1.96，因此男女的 BMI 值差不多。再由建立信賴區間圖看出有重疊部分，再次驗證了男女的 BMI 值差不多。
(五) 第 79~91 行（檢視學生在 technology 和 entertainment 支出間是否有顯著差異）：由成對 t 檢定看出兩者支出有顯著差異，再由繪製密度圖得知兩者的支出是不相等的。
(六) 第 100~136 行（比較金額與購買的類別是否有差）：由 ANOVA 檢定看出有顯著差異，並由信賴區間圖中可以明顯看出 outerwear 與 footwear、clothing、accessories 三者有顯著不同。

七、參考說明

無