Bài tập tuần 8

Câu 1: Theo dõi ngẫu nhiên 35 hộ ở Hà Nội được bảng số liệu về tiền tiêu thụ điện trong một tháng (đơn vị nghìn đồng) như sau:

Tiền tiêu thụ điện (x_i)	1700	1800	1900	2000	2100	2200	2300
Số hộ (r_i)	3	4	5	7	8	5	3

a. Nhập từ bàn phím dữ liệu của 35 hộ vào R.

b. Tính các giá trị trung bình và độ lệch tiêu chuẩn mẫu (s), biết trung bình và độ lệch tiêu chuẩn mẫu được tính theo công thức

$$\bar{x} = \frac{1}{n}\sum_{i}^{k} x_i r_i \quad \text{và} \quad s^2 = \frac{1}{n-1}\sum r_i \left(x_i - \bar{x}\right)^2$$

c. Có ý kiến cho rằng: "Số điện trung bình của một hộ ở Hà Nội là $\mu_0 = 190$ số điện". Để kiểm tra ý kiến trên có đúng không, ta đặt T được tính bằng công thức dưới đây là Test thống kê.

$$T = \frac{\bar{x} - \mu_0}{s}.\sqrt{n}$$ với n là cỡ mẫu (length)

Nếu $T > 1.757$ thì ta có thể đưa ra kết luận ý kiến trên là sai.

Dựa vào bộ dữ liệu của 35 hộ dân trên và sử dụng lệnh "print" để đưa ra kết luận cho bài toán

d. Hãy viết một hàm với biến đầu vào là dữ liệu x và hằng số μ_0 để kiểm tra tính chính xác của các khẳng định như trên.

e. Hãy sinh ra một véc tơ mới, gồm 150 giá trị, là tiền điện tiêu thụ của 150 hộ dân. Biết rằng, véc tơ này có phân phối chuẩn với cùng trung bình và độ lệch tiêu chuẩn của dữ liệu ban đầu.

f. Hãy kiểm tra khẳng định : "Số điện trung bình của một hộ ở Hà Nội là $\mu_0 = 190$ số điện" có chính xác trên dữ liệu vừa sinh ra không ?

Câu 2: Đo chiều dài của 100 ống tuýp do một xí nghiệp sản xuất được kết quả:

Chiều dài (cm)	178	179	180	181	182
Số ống	12	18	35	20	15

a. Nhập từ bàn phím dữ liệu của 100 ống tuýp vào R và đặt tên là "dulieu".

b. Tính các giá trị trung bình và độ lệch tiêu chuẩn mẫu (s), biết trung bình và độ lệch tiêu chuẩn mẫu được tính theo công thức

$$\bar{x} = \frac{1}{n}\sum_{i}^{k} x_i r_i \quad \text{và} \quad s^2 = \frac{1}{n-1}\sum r_i \left(x_i - \bar{x}\right)^2$$

c. Viết một hàm để xác định xem một số μ_0 có thể đại diện cho chiều dài của các ống tuýp do xí nghiệp x sản xuất hay không, biết μ_0 có thể đại diện cho chiều dài của các ống tuýp do xí nghiệp trên sản xuất nếu μ_0 nằm trong khoảng

$$\left(\overline{x} - 1.96\frac{s}{\sqrt{n}}; \overline{x} + 1.96\frac{s}{\sqrt{n}}\right)$$

Trong đó, $\overline{x}$ là trung bình của dữ liệu x, s là độ lệch tiêu chuẩn của dữ liệu x.

d. Số $\mu_0 = 185$, có đại diện cho dữ liệu ban đầu hay không?

e. Hãy sinh ra một véc tơ mới (vecto y), gồm 1500 giá trị, là chiều dài của 1500 bóng tuýp. Biết rằng, véc tơ này có phân phối chuẩn với cùng trung bình và độ lệch tiêu chuẩn của dữ liệu ban đầu.

f. Số $\mu_0 = median(y)$ có đại diện cho dữ liệu y vừa sinh ra không?

Câu 3: Một công ty làm thống kê thời gian đi từ nhà đến công ty của 350 nhân viên, thu được kết quả sau:

Thời gian (Phút)	40	42	44	46	48	50
Số nhân viên	25	50	60	110	90	15

a. Nhập từ bàn phím dữ liệu của 350 nhân viên vào R.

b. Tính các giá trị trung bình và độ lệch tiêu chuẩn mẫu (s), biết trung bình và độ lệch tiêu chuẩn mẫu được tính theo công thức

$$\overline{x} = \frac{1}{n}\sum_{i}^{k} x_i r_i \quad \text{và} \quad s^2 = \frac{1}{n-1}\sum r_i \left(x_i - \overline{x}\right)^2$$

c. Viết một hàm để xác định xem một số có thể đại diện cho "Tỉ lệ nhân viên đi làm hết nhiều hơn 46 phút" hay không, biết nếu có thể đại diện cho "Tỉ lệ nhân viên đi làm hết nhiều hơn 46 phút" thì sẽ nằm trong khoảng

$$\left(f - 1.96 * \frac{\sqrt{f*(1-f)}}{\sqrt{n}}; f + 1.96 * \frac{\sqrt{f*(1-f)}}{\sqrt{n}}\right)$$

với n là cỡ mẫu (length), m là số nhân viên đi làm hết nhiều hơn 46 phút và .

d. Dựa vào số liệu trên hãy trả lời câu hỏi cho đó cho: và

Câu 4: Dữ liệu "winequality.red.5.csv" cho biết các thông tin liên quan đến chất lượng rượu vang đỏ. Thực hiện các thao tác sau với phần mềm R.

a. Nhập dữ liệu từ tệp đã cho vào R và đặt tên là "red". Giải thích ý nghĩa của lệnh dim.

b. Trong dữ liệu về total.sulfur.dioxide ta thấy các giá trị 29 và 102 bị nhập sai, nó phải là 92. Tìm và thay thế nó.

c. Trong dữ liệu về citric.acid, để khắc phục lỗi kĩ thuật khi đo đạc ta sẽ cộng thêm một lượng là 0,01 với những giá trị **Không** lớn hơn 0,02. Viết một hàm để thực hiện thao tác đó.

d. Giả sử rằng: rượu được đánh giá là tốt nếu hằng số k > 8,3, trong đó $k = \frac{alcohol.quality}{pH}$. Viết một hàm để phân loại chất lượng rượu từ dữ liệu đã cho ("Tốt", "Xấu").

e. Từ việc phân loại trên, hãy tính trung bình, độ lệch tiêu chuẩn của độ *pH* và mật độ rượu *density* cho mỗi nhóm.

Câu 5: Cho file dữ liệu Sampledatasafety ("Sampledatasafety.csv") cho biết các thông tin về số ca tai nạn của một công ty trong hai năm 2020 và 2021 như sau: Ngày nhập viện (Date), Vị trí chấn thương (Injury Location), Giới tính (Gender), Độ tuổi (Age Group), Nguyên nhân chấn thương (Incident Type), Số ngày công mất đi do tai nạn (Days Lost), Địa chỉ (Plant), Phân loại tổn thương (Report Type), Thời điểm xảy ra chấn thương (Shift), Bộ phận làm việc (Department), Chi phí điều trị (Incident Cost), Thời điểm thanh toán hóa đơn điều trị (WKDay-Month-Year)

Thực hiện các thao tác sau với phần mềm Rstudio

a) Nhập dữ liệu ("Sampledatasafety.csv") và đặt tên là data vào R. Loại bỏ dữ liệu trống nếu có.

b) Trích ra một bộ dữ liệu con về nguyên nhân chấn thương do bỏng (Burn) và tính tổng chi phí điều trị (Incident Cost).

c) Trích ra một dữ liệu về số ca tai nạn trong năm 2020, đặt tên là data1, và một dữ liệu về số ca tai nạn trong năm 2021, đặt tên là data 2. So sánh tổng số ngày công mất đi do tai nạn (Days Lost) trong hai năm đó.

d) Hãy cho biết nguyên nhân gây chấn thương (Incident Type) chủ yếu trong năm 2020, 2021.

e) Hãy cho biết vị trí chấn thương thường gặp nhất của các công nhân.

f) Viết một hàm tính tổng chi phí điều trị cho mỗi loại nguyên nhân gây chấn thương.

g) Công ty báo cáo rằng tổng chi phí cần điều trị (Incident Cost) của các công nhân ở bộ phận vận chuyển (Shipping) là lớn nhất. Kết luận này đúng hay sai?