

BÀI TẬP NHÓM

HỌC PHẦN: THÔNG KÊ MÁY TÍNH

Giảng viên: Nguyễn Đức Thuận

Lớp học phần: 61. CNPM-1

NHÓM THỰC HIỆN: 2

Danh sách thành viên:

1. Nguyễn Trọng Hiếu - 61133639
2. Trần Khải Hoàn – 59130790
3. Phạm Minh Hoàng – 61131788
4. Đặng Nguyễn Nhật Hùng – 61133079
5. Huỳnh Ngọc Hưng – 61133707
6. Nguyễn Việt Hưng – 61133712
7. Trần Văn Huy – 61131815
8. Nguyễn Trung Huy – 61133748 – Nhóm trưởng
9. Võ Gia Huy – 61133761
10. Nguyễn Trương Ngọc Huy – 61136483

- Nội dung thực hiện: ex 3.14 – ex 3.25 (P130 - P132)

Bài 3.14:

Compute the mean, median, and mode for the following data:

155	25	30	52	142	35	51	26	2	23
270	74	29	29	29	29	51	83	9	69

Dữ liệu (DL):

155	25	30	52	142	35	51	26	2	23
270	74	29	29	29	29	51	83	9	69

Yêu cầu: Tính mean, median, và mode cho dữ liệu trên.

Giải:

*

- Tính mean (\bar{y}):

$$\bar{y} = \frac{\text{tổng tất cả các phần tử}}{\text{số lượng các phần tử}} = \frac{\sum_{i=1}^n y_i}{n}$$

Với dữ liệu trên và số lượng phần tử $n = 20$, ta có:

$$\bar{y} = \frac{155 + 25 + 30 + 52 + 142 + 35 + 51 + 26 + 2 + 23 + 270 + 74 + 29 + 29 + 29 + 29 + 51 + 83 + 9 + 69}{20} = \frac{1213}{20} = 60,65$$

- Tính median (\tilde{y}):

+ Sắp xếp lại dữ liệu trên theo thứ tự tăng dần như sau:

2	9	23	25	26	29	29	29	29	30
35	51	51	52	69	74	83	142	155	270

+ Với số lượng phần tử $n=20$, ta có được giá trị median như sau:

$$\tilde{y} = \frac{y\left(\frac{n}{2}\right) + y\left(\frac{n}{2} + 1\right)}{2} = \frac{y(10) + y(11)}{2} = \frac{30 + 35}{2} = 32,5$$

- Tìm mode:

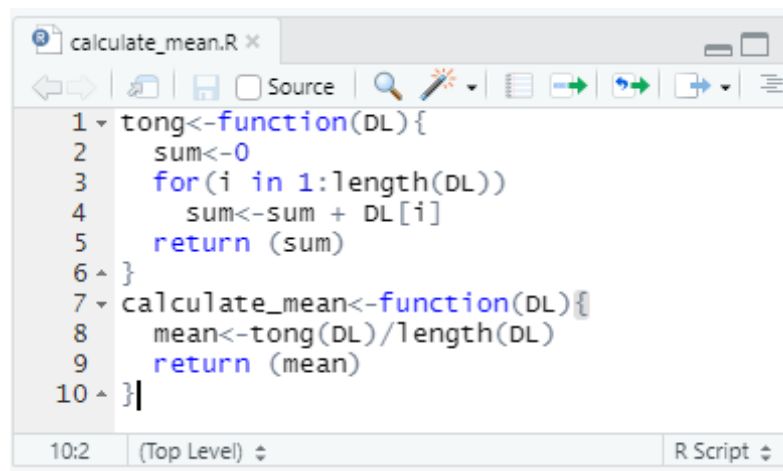
Quan sát dữ liệu trên, ta thấy được giá trị 29 là giá trị được lặp lại nhiều nhất với số lần lặp cao nhất là 4.

Vậy nên, ta có: $mode = 29$

**** : Viết hàm thực hiện trong R**

- Hàm thực hiện tính giá trị cho mean \bar{y}

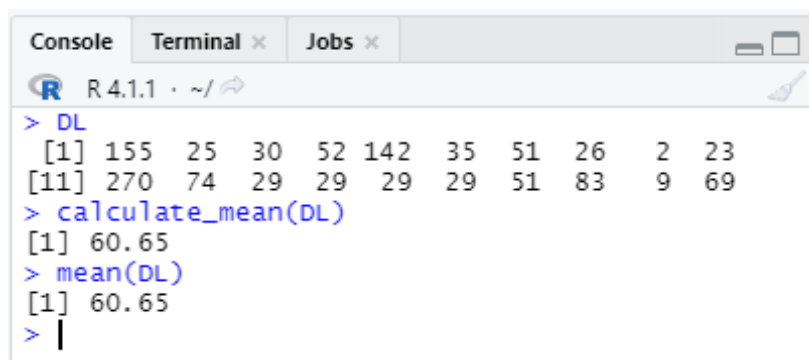
+ Xây dựng hàm:



```
1 tong<-function(DL){
2   sum<-0
3   for(i in 1:length(DL))
4     sum<-sum + DL[i]
5   return (sum)
6 }
7 calculate_mean<-function(DL){
8   mean<-tong(DL)/length(DL)
9   return (mean)
10 }
```

Hình 1: Hàm tính Mean in R

+ Thực thi và kiểm tra kết quả hàm:

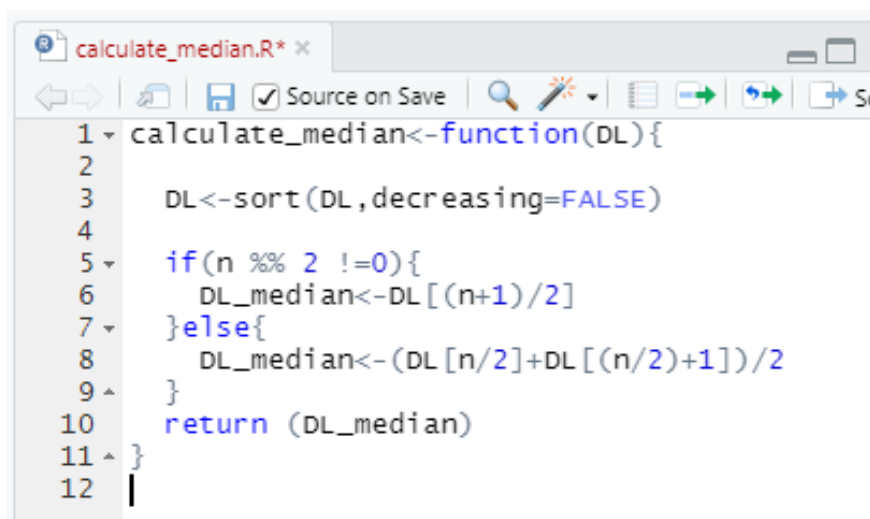


```
> DL
[1] 155  25  30  52 142  35  51  26  2  23
[11] 270  74  29  29  29  29  51  83  9  69
> calculate_mean(DL)
[1] 60.65
> mean(DL)
[1] 60.65
> |
```

Hình 2: Kết quả của Hàm tính Mean in R

- Hàm thực hiện tính giá trị median

+ Xây dựng hàm:



```
1 calculate_median<-function(DL){
2
3   DL<-sort(DL,decreasing=FALSE)
4
5   if(n %% 2 !=0){
6     DL_median<-DL[(n+1)/2]
7   }else{
8     DL_median<-(DL[n/2]+DL[(n/2)+1])/2
9   }
10  return (DL_median)
11 }
12 |
```

Hình 3: Hàm tính Median in R

+ Thực thi và kiểm tra kết quả hàm:

```
Console Terminal x Jobs x
R 4.1.1 · ~/
> DL
[1] 2 270 155 142 83 74 69 52 51 51
[11] 35 30 29 29 29 29 26 25 23 9
> calculate_median(DL)
[1] 32.5
> median(DL)
[1] 32.5
> |
```

Hình 4: Kết quả của Hàm tính Median in R

- Hàm tìm MODE:

+ Xây dựng hàm:

```
Mode.R x
Source on Save
1 demts<-function(x,y){
2   dem<-0
3   for(i in 1:length(y))
4     if(y[i]==x) dem<-dem+1
5   return (dem)
6 }
7 Mode<-function(DL){
8   tam<-unique(DL)
9   kq<-rep(0,length(tam))
10  for(i in 1:length(tam))
11    kq[i]<-demts(tam[i],DL)
12  if(min(kq)!=max(kq))
13    for(i in 1:length(tam))
14      if(kq[i]==max(kq)) print(tam[i])
15 }
16 |
```

Hình 5: Hàm tìm Mode in R

+ Thực thi và kiểm tra kết quả hàm:

```
Console Terminal x Jobs x
R 4.1.1 · ~/
> DL
[1] 2 270 155 142 83 74 69 52 51 51
[11] 35 30 29 29 29 29 26 25 23 9
> Mode(DL)
[1] 29
> |
```

Hình 6: Kết quả Hàm tìm Mode in R

Bài 3.15: Compute the mean, median, and mode for the following data:

35	81	96	45	109	126	71	15	8	79	56
73	58	17	82	29	58	68	24	5	24	

Dữ liệu (DL):

25	81	96	45	109	126	71	15	8	79	56
73	58	17	82	29	58	68	24	5	24	

Yêu cầu: Tính mean, median, và mode cho dữ liệu trên.

Giải:

*

- Tính mean (\bar{y}):

$$\bar{y} = \frac{\text{tổng tất cả các phần tử}}{\text{số lượng các phần tử}} = \frac{\sum_{i=1}^n y_i}{n}$$

Với dữ liệu trên và số lượng phần tử $n = 21$, ta có:

$$\begin{aligned}\bar{y} &= \frac{25 + 81 + 96 + 45 + 109 + 126 + 71 + 15 + 8 + 79 + 56 + 73 + 58 + 17 + 82 + 29 + 58 + 68 + 24 + 5 + 24}{21} = \frac{1149}{21} \\ &= \frac{383}{7} = 54, (714285)\end{aligned}$$

- Tính median (\tilde{y}):

+ Sắp xếp lại dữ liệu trên theo thứ tự tăng dần như sau:

5	8	15	17	24	24	25	29	45	56	58
58	68	71	73	79	81	82	96	109	126	

+ Với số lượng phần tử $n=21$, ta có được giá trị median như sau:

$$\tilde{y} = y\left(\frac{n+1}{2}\right) = y(11) = 58$$

- Tìm mode:

Quan sát dữ liệu trên, ta thấy được giá trị 24 và 58 là hai giá trị được lặp lại nhiều nhất với số lần lặp là 2.

Vậy nên, ta có: $\begin{cases} mode = 24 \\ mode = 58 \end{cases}$

**** : Viết hàm thực hiện trong R**

- Hàm thực hiện tính giá trị cho mean \bar{y}

+ Sử dụng hàm đã xây dựng ở Bài 3.14:

```

1 tong<-function(DL){
2   sum<-0
3   for(i in 1:length(DL))
4     sum<-sum + DL[i]
5   return (sum)
6 }
7 calculate_mean<-function(DL){
8   mean<-tong(DL)/length(DL)
9   return (mean)
10 }

```

Hình 7: Hàm tính Mean in R

+ Thực thi và kiểm tra kết quả hàm:

```

> DL
[1] 25 81 96 45 109 126 71 15 8
[10] 79 56 73 58 17 82 29 58 68
> calculate_mean(DL)
[1] 54.71429
> mean(DL)
[1] 54.71429
>

```

Hình 8: Kết quả của Hàm tính Mean in R

- Hàm thực hiện tính giá trị median

+ Sử dụng hàm đã xây dựng ở Bài 3.14:

```

1 calculate_median<-function(DL){
2
3   DL<-sort(DL,decreasing=FALSE)
4
5   if(n %% 2 !=0){
6     DL_median<-DL[(n+1)/2]
7   }else{
8     DL_median<-(DL[n/2]+DL[(n/2)+1])/2
9   }
10  return (DL_median)
11 }
12

```

Hình 9: Hàm tính Median in R

+ Thực thi và kiểm tra kết quả hàm:

```
Console Terminal x Jobs x
R 4.1.1 · ~/
> DL
[1] 25 81 96 45 109 126 71 15 8 79 56 73
[13] 58 17 82 29 58 68 24 5 24
> calculate_median(DL)
[1] 58
> median(DL)
[1] 58
> |
```

Hình 10: Kết quả của Hàm tính Median in R

- Hàm tìm MODE:

+ Sử dụng hàm đã xây dựng ở Bài 3.14:

```
Mode.R x
Source on Save
1 demts<-function(x,y){
2   dem<-0
3   for(i in 1:length(y))
4     if(y[i]==x) dem<-dem+1
5   return (dem)
6 }
7 Mode<-function(DL){
8   tam<-unique(DL)
9   kq<-rep(0,length(tam))
10  for(i in 1:length(tam))
11    kq[i]<-demts(tam[i],DL)
12  if(min(kq)!=max(kq))
13    for(i in 1:length(tam))
14      if(kq[i]==max(kq)) print(tam[i])
15 }
16 |
```

Hình 11: Hàm tìm Mode in R

+ Thực thi và kiểm tra kết quả hàm:

```
Console Terminal x Jobs x
R 4.1.1 · ~/
> DL
[1] 25 81 96 45 109 126 71 15 8 79 56 73
[13] 58 17 82 29 58 68 24 5 24
> Mode(DL)
[1] 58
[1] 24
> |
```

Hình 12: Kết quả Hàm tìm Mode in R

Bài 3.16:

Refer to the data in Exercise 3.15 with the measurements 109 and 126 replaced by 378 and 517. Recompute the mean, median, and mode. Discuss the impact of these extreme measurements on the three measures of central tendency.

Dữ liệu (DL): Từ DL bài 3.15 thay đổi giá trị 109 và 126 thành 378 và 517

25	81	96	45	378	517	71	15	8	79	56
73	58	17	82	29	58	68	24	5	24	

Yêu cầu:

- Tính mean, median, và mode cho dữ liệu trên.
- Thảo luận về tác động của những giá trị ngoại lệ này (378, 517) đối với 3 phép đo.

Giải:

*

- Tính mean (\bar{y}):

$$\bar{y} = \frac{\text{tổng tất cả các phần tử}}{\text{số lượng các phần tử}} = \frac{\sum_{i=1}^n y_i}{n}$$

Với dữ liệu trên và số lượng phần tử $n = 21$, ta có:

$$\begin{aligned}\bar{y} &= \frac{25 + 81 + 96 + 45 + 378 + 517 + 71 + 15 + 8 + 79 + 56 + 73 + 58 + 17 + 82 + 29 + 58 + 68 + 24 + 5 + 24}{21} = \frac{1809}{21} \\ &= \frac{603}{7} = 86, (142857)\end{aligned}$$

- Tính median (\tilde{y}):

+ Sắp xếp lại dữ liệu trên theo thứ tự tăng dần như sau:

5	8	15	17	24	24	25	29	45	56	58
58	68	71	73	79	81	82	96	378	517	

+ Với số lượng phần tử $n=21$, ta có được giá trị median như sau:

$$\tilde{y} = y\left(\frac{n+1}{2}\right) = y(11) = 58$$

- Tìm mode:

Quan sát dữ liệu trên, ta thấy được giá trị 24 và 58 là hai giá trị được lặp lại nhiều nhất với số lần lặp là 2.

Vậy nên, ta có: $\begin{cases} mode = 24 \\ mode = 58 \end{cases}$

**** : Viết hàm thực hiện trong R**

- Hàm thực hiện tính giá trị cho mean \bar{y}
- + Sử dụng hàm đã xây dựng ở Bài 3.14:


```

1 tong<-function(DL){
2   sum<-0
3   for(i in 1:length(DL))
4     sum<-sum + DL[i]
5   return (sum)
6 }
7 calculate_mean<-function(DL){
8   mean<-tong(DL)/length(DL)
9   return (mean)
10 }

```

Hình 13: Hàm tính Mean in R – Bài 3.15

+ Thực thi và kiểm tra kết quả hàm:

```

> DL
[1] 25 81 96 45 378 517 71 15 8 79 56 73
[13] 58 17 82 29 58 68 24 5 24
> calculate_mean(DL)
[1] 86.14286
> mean(DL)
[1] 86.14286
>

```

Hình 14: Kết quả của Hàm tính Mean in R

- Hàm thực hiện tính giá trị median y

+ Sử dụng hàm đã xây dựng ở Bài 3.14:

```

1 calculate_median<-function(DL){
2
3   DL<-sort(DL,decreasing=FALSE)
4
5   if(n %% 2 !=0){
6     DL_median<-DL [(n+1)/2]
7   }else{
8     DL_median<-(DL [n/2]+DL [(n/2)+1])/2
9   }
10  return (DL_median)
11 }
12

```

Hình 15: Hàm tính Median in R – Bài 3.15

+ Thực thi và kiểm tra kết quả hàm:

```

R 4.1.1 · ~/
> DL
[1] 25 81 96 45 109 126 71 15 8 79 56 73
[13] 58 17 82 29 58 68 24 5 24
> calculate_median(DL)
[1] 58
> median(DL)
[1] 58
> |

```

Hình 16: Kết quả của Hàm tính Median in R

- Hàm tìm MODE:

+ Sử dụng hàm đã xây dựng ở Bài 3.14:

```

Mode.R x
1 demts<-function(x,y){
2   dem<-0
3   for(i in 1:length(y))
4     if(y[i]==x) dem<-dem+1
5   return (dem)
6 }
7 Mode<-function(DL){
8   tam<-unique(DL)
9   kq<-rep(0,length(tam))
10  for(i in 1:length(tam))
11    kq[i]<-demts(tam[i],DL)
12  if(min(kq)!=max(kq))
13    for(i in 1:length(tam))
14      if(kq[i]==max(kq)) print(tam[i])
15 }
16 |

```

Hình 17: Hàm tìm Mode in R – Bài 3.15

+ Thực thi và kiểm tra kết quả hàm:

```

R 4.1.1 · ~/
> DL
[1] 25 81 96 45 378 517 71 15 8 79 56 73
[13] 58 17 82 29 58 68 24 5 24
> Mode(DL)
[1] 58
[1] 24
> |

```

Hình 18: Kết quả Hàm tìm Mode in R

*****: Thảo luận sự tác động của những phần tử ngoại lệ**

Khi thêm các phần tử ngoại lệ vào dữ liệu:

- Đối với mean: Có sự thay đổi

+ Lí do: Do mean là giá trị trung bình của một dãy số, dựa vào các giá trị để tính, mà khi tính cũng bao gồm cả các phần tử ngoại lệ này làm cho tổng chung của dãy có sự chênh lệch rõ rệt.

- Đối với median: Không có sự thay đổi.

+ Lí do: Do median là giá trị trung tâm của một dãy số, mà các phần tử ngoại lệ là các giá trị biên (ngoài tầm).

- Đối với mode: Không có sự thay đổi.

+ Lí do: Do mode là giá trị thường xuyên xuất hiện nhiều nhất trong dãy số, mà các phần tử ngoại lệ chỉ là thiểu số và ít/không có sự trùng lặp hoặc lặp lại các giá trị.

Bài 3.17:

Compute a 10% trimmed mean for the data sets in Exercises 3.15 and 3.16. Do the extreme values in Exercise 3.16 affect the 10% trimmed mean? Would a 5% trimmed mean be as affected by the two extreme values as the 10% trimmed mean?

Yêu cầu:

Tính giá trị trung bình sau khi cắt 10% mẫu cho các tập dữ liệu trong 3.15 và 3.16. Có phải các giá trị ngoại lệ làm ảnh hưởng đến giá trị trung bình sau khi cắt 10% mẫu không? Liệu giá trị trung bình sau khi cắt 5% mẫu có bị ảnh hưởng bởi 2 giá trị ngoại lệ như là giá trị trung bình sau khi cắt 10% mẫu?

Giải:

Tập dữ liệu :

- Dữ liệu 3.15:

35	81	96	45	109	126	71	15	8	79	56	73
58	17	82	29	58	68	24	5	24			

Ban đầu có 21 giá trị quan sát.

Sắp xếp theo thứ tự từ bé đến lớn:

5	8	15	17	24	24	29	35	45	56	58
58	68	71	73	79	81	82	96	109	126	

Cắt giảm 10%, số lượng quan sát cần loại bỏ là $(21 * 0,1) = 2,1 = 2$ giá trị.

Tiến hành cắt 2 giá trị quan sát từ mỗi đầu, ta được :

		15	17	24	24	29	35	45	56	58	58
68	71	73	79	81	82	96					

Giá trị trung bình : $\bar{X} = 53.58824$

```
> d<-c(5,8,15,17,24,24,29,35,45,56,58,58,68,71,73,79,81,82,96,109,126)
> mean(d,trim=0.1)
[1] 53.58824
```

- Dữ liệu 3.16:

35	81	96	45	378	517	71	15	8	79	6	73
58	17	82	29	58	68	24	5	24			

Sắp xếp theo thứ tự từ bé đến lớn:

5	8	15	17	24	24	29	35	45	56	58	58
68	71	73	79	81	82	96	378	517			

Cắt giảm 10%, số lượng quan sát cần loại bỏ là $(21 * 0,1) = 2,1 = 2$ giá trị.

Tiến hành cắt 2 giá trị quan sát từ mỗi đầu, ta được:

		15	17	24	24	29	35	45	56	58	58
68	71	73	79	81	82	96					

Giá trị trung bình : $\bar{X} = 53.58824$

```
> dl<-c(5,8,15,17,24,24,29,35,45,56,58,58,68,71,73,79,81,82,96,378,517)
> mean(dl,trim=0.1)
[1] 53.58824
```

Ta thấy rằng, cả 2 giá trị trung bình sau khi cắt 10% mẫu là bằng nhau, có nghĩa là 2 giá trị ngoại lệ 378 và 517 không ảnh hưởng đến giá trị trung bình sau khi cắt 10% mẫu, bởi vì 2 giá trị đó nằm trong 10% mẫu được cắt từ mỗi đầu của tập dữ liệu.

Xét giá trị trung bình sau khi cắt 5% mẫu.

Số quan sát cần loại bỏ là $(21 * 0.05) = 1.05 = 1$

- Dữ liệu 3.15:

5	8	15	17	24	24	29	35	45	56	58
58	68	71	73	79	81	82	96	109	126	

Tiến hành cắt 2 giá trị quan sát từ mỗi đầu, ta được:

	8	15	17	24	24	29	35	45	56	58	58
68	71	73	79	81	82	96	109				

Giá trị trung bình: $\bar{X} = 54.10526$.

```
> d<-c(5,8,15,17,24,24,29,35,45,56,58,58,68,71,73,79,81,82,96,109,126)
> mean(d,trim=0.05)
[1] 54.10526
```

- Dữ liệu 3.16:

5	8	15	17	24	24	29	35	45	56	58
58	68	71	73	79	81	82	96	378	517	

Tiến hành cắt 2 giá trị quan sát từ mỗi đầu, ta được :

	8	15	17	24	24	29	35	45	56	58	58
68	71	73	79	81	82	96	378				

Giá trị trung bình: $\bar{X} = 68.26316$.

```
> dl<-c(5,8,15,17,24,24,29,35,45,56,58,58,68,71,73,79,81,82,96,378,517)
> mean(dl,trim=0.05)
[1] 68.26316
```

Ta thấy 2 giá trị trung bình sau khi cắt 5% mẫu đã có sự khác nhau, cho nên có thể nói rằng 2 giá trị ngoại lệ có ảnh hưởng đến giá trị trung bình sau khi cắt 5% mẫu.

Bài 3.18: A data set of 75 values is summarized in the following frequency table. Estimate the mean, median, and mode for the 75 data values using the summarized data.

Class Interval	Frequency
2.0–4.9	9
5.0–7.9	19
8.0–10.9	27
11.0–13.9	10
14.0–16.9	5
17.0–19.9	3
20.0–22.9	2

Yêu cầu:

Một tập dữ liệu của 75 giá trị được tóm tắt trong bảng tần suất dưới đây. Hãy ước lượng giá trị trung bình, trung vị, yếu vị của 75 giá trị này.

Giải:

Class Interval	X_i (Mid-group)	Frequency(f)	f_x ($X_i \cdot f$)
2.0-4.9	3.45	9	31.05
5.0-7.9	6.45	19	122.55
8.0-10.9	9.45	27	255.15
11.0-13.9	12.45	10	124.5
14.0-16.9	15.45	5	77.25
17.0-19.9	18.45	3	55.35
20.0-22.9	21.45	2	42.9
Total		$\Sigma f=75$	$\Sigma f_x=708.75$

- Trung bình(Mean):

$$\bar{x} = \frac{\Sigma f_x}{\Sigma f} = \frac{708.75}{75} = 9.45$$

Code R:

```
.... ....
> #Mean (gia tri trung binh)
> Meankhoang<-function(x) {
+   mean<-0
+   y<-x$F
+   mean<- sum(midpoint(x)*y)/sum(y)
+   return (mean)
+ }
```

Kết quả:

```
.. -- --
> Meankhoang(dl)
[1] 9.45
```

- Trung vị(Median):

Class Interval	X_i (Midpoint)	Frequency(f_i)	cf_b (Tần suất tích lũy)
2.0-4.9	3.45	9	9
5.0-7.9	6.45	19	28
8.0-10.9	9.45	27	55
11.0-13.9	12.45	10	65
14.0-16.9	15.45	5	70
17.0-19.9	18.45	3	73
20.0-22.9	21.45	2	75
		$\Sigma f=75$	

- Ta lấy khoảng giá trị khi $cf_b \geq \Sigma f/2$ (37.5)

- Trung vị khi mà giá trị của nó $\geq 50\%$ giá trị của dãy.

$w = 3$ (độ rộng trung bình)

$L = 8$ (giới hạn dưới của nhóm chứa trung vị)

$$\text{Median} = L + \frac{\Sigma f/2 - cf_b}{f(i)} w = 8 + \frac{37.5 - 28}{27} 3 = 8 + 1.05 = \mathbf{9.055556}$$

Code R:

```
> #Median
> Mediankhoang<-function(x) {
+   L<-0
+   z<-length(x$tichluy)
+   k<-sum(x$F)/2
+   #lay ra gia tri meadian
+   for(i in 1:z){
+     if(x$tichluy[i] >= k)
+     {
+       L<-x$GhDuoi[i]
+       f<-x$F[i]
+       cf<-x$tichluy[i-1]
+       w<-dorongtb(x)
+       m<- L+(w/f*(k-cf))
+       return(m)
+       break
+     }
+   }
+ }
```

Kết quả:

```
> Mediankhoang(dl)
[1] 9.055556
```

- Yếu vị(Mode):

$$D_1 = f_i - f_{i-1} = 27 - 19 = 8$$

$$D_2 = f_i - f_{i+1} = 27 - 10 = 17$$

$$\text{Mode} = L + \frac{D_1}{D_1 + D_2} w = 8 + \frac{8}{8 + 17} 3 = 8 + 0.96 = \mathbf{8.96}$$

Code R:

```
> #Tim khoang co tan so cao nhat lay ra vi tri
> Modekhoang<-function(x){
+   m<-x$F
+   gh<-x$GhDuoi
+   n<-length(x$F)
+   L<-0
+   yeuvi<-0
+   #Lay ra khoang co tuan so cao nhat
+   for(i in 1:n){
+     if(max(x$F)== m[i]){
+       {
+         L<-gh[i]
+         yeuvi<- L+((m[i]-m[i-1])/(m[i]-m[i-1]+m[i]-m[i+1]))*dorongtb(x)
+       }
+     }
+   }
+   return(yeuvi)
+ }
```

Kết quả:

```
> Modekhoang(dl)
[1] 8.96
```

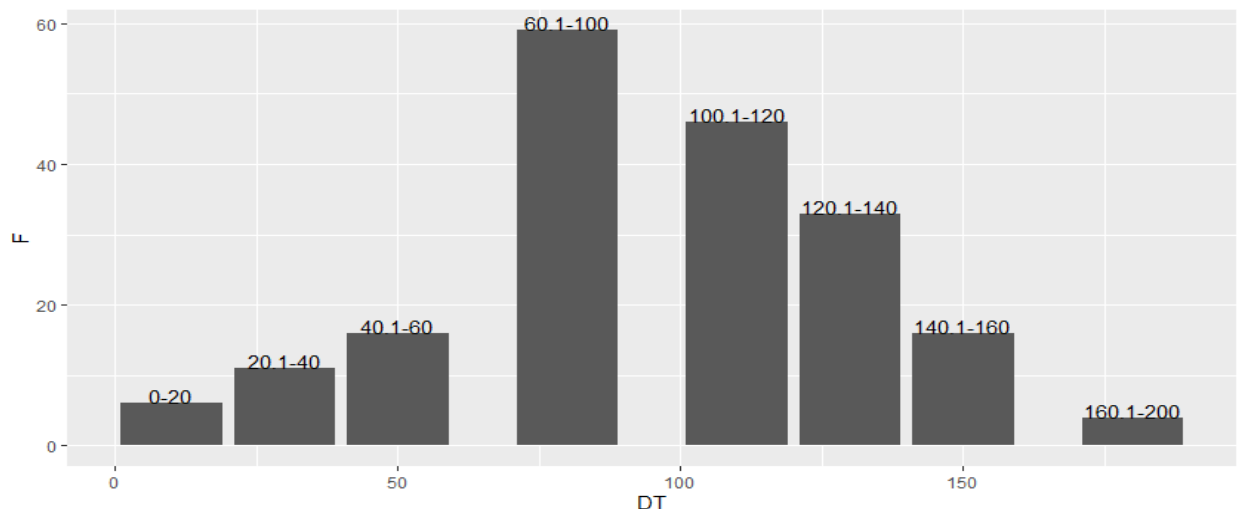
Bài 3.19: A study of the reliability of buses [“Large Sample Simultaneous Confidence Intervals for the Multinomial Probabilities on Transformations of the Cell Frequencies,” Technometrics(1980) 22:588] examined the reliability of 191 buses. The distance traveled (in 1,000s of miles) prior to the first major motor failure was classified into intervals. A modified form of the table follows.

Distance Traveled (1,000s of miles)	Frequency
0–20.0	6
20.1–40.0	11
40.1–60.0	16
60.1–100.0	59
100.1–120.0	46
120.1–140.0	33
140.1–160.0	16
160.1–200.0	4

- Sketch the relative frequency histogram for the distance data and describe its shape.
- Estimate the mode, median, and mean for the distance traveled by the 191 buses.
- What does the relationship among the three measures of center indicate about the shape of the histogram for these data?
- Which of the three measures would you recommend as the most appropriate representative of the distance traveled by one of the 191 buses? Explain your answer.

Giải:

Câu a: Vẽ biểu đồ tần số tương đối cho dữ liệu khoảng cách và mô tả hình dạng của nó.



- Trục DT là midpoint của mỗi khoảng.
- Cột F là tần số của khoảng.

Code R:

```
> dt1
  Ghdnoi Ghtren  F    label midpoint
1    0.0     20  6    0-20    10.00
2   20.1     40 11  20.1-40    30.05
3   40.1     60 16  40.1-60    50.05
4   60.1    100 59  60.1-100    80.05
5  100.1    120 46 100.1-120   110.05
6  120.1    140 33 120.1-140   130.05
7  140.1    160 16 140.1-160   150.05
8  160.1    200  4 160.1-200   180.05
>

library(ggplot2)
ggplot(dt1, aes(x = midpoint, y = F)) +
  geom_bar(stat = "identity") +
  geom_text(aes(label = label), vjust = 0)
```

Câu b: Xác định mode biến có tần số lớn nhất

- **Mode:** Yếu vị (khoảng có tần số lớn nhất)

Khoảng 60.1-100.0 có tần số là lớn nhất

Distance Travel	F	midpoint
0-20.0	6	10
20.1-40.0	11	30.05
40.1-60.0	16	50.05
60.1-100.0	59	80.05
100.1-120.0	46	110.05
120.1-140	33	130.05
140.1-160	16	150.05
160.1-200	4	180.05
Tổng	191	

$$\text{Mode} = L + \frac{f_m - f_{m-1}}{(f_m - f_{m-1}) + (f_m - f_{m+1})} * w = 60.1 + \frac{59 - 16}{(59 - 16) + (59 - 46)} * 24.29 = 78.75$$

L: giới hạn dưới của khoảng đã chọn

fm: tần số của khoảng

w: độ rộng giữa 2 mid-point

Code R:

```
#do rong tb w
dorongtb<-function(x){
  a<-x$midpoint
  b<-length(x$midpoint)
  dr<-0
  c<-0
  for(i in 1:(b-1)){
    dr<-a[i+1]-a[i]
    c<-c+dr
  }
  return (c/(b-1))
}

#Mode
Modekhoang<-function(x){
  m<-x$F
  gh<-x$Ghduoi
  n<-length(x$F)
  L<-0
  yeuvi<-0
  #lay ra khoang co tuan so cao nhat
  for(i in 1:n){
    if(max(x$F)== m[i])
    {
      L<-gh[i]
      yeuvi<- L+((m[i]-m[i-1])/(m[i]-m[i-1]+m[i]-m[i+1]))*dorongtb(x))
    }
  }
  return(yeuvi)
}
```

- **Mean:** Giá trị trung bình:

Distance Travel	Midpoint(\bar{y}_l)	F
0-20.0	10	6
20.1-40.0	30.05	11
40.1-60.0	50.05	16
60.1-100.0	80.05	59
100.1-120.0	110.05	46
120.1-140	130.05	33
140.1-160	150.05	16
160.1-200	180.05	4
Tổng		191

$$\text{Mean} = \frac{\sum_{i=1}^k f_x * \bar{y}_l}{n}$$

$$= \frac{10 * 6 + 30.05 * 11 + 50.05 * 16 + 80.05 * 59 + 110.05 * 46 + 130.05 * 33 + 150.05 * 16 + 180.05 * 4}{191} = 96.28$$

Code R:

```
#Mean gia tri trung binh
Meankhoang<-function(x){
  mean<-0
  y<-x$F
  mean<- sum(midpoint(x)*y)/sum(y)
  return (mean)
}
```

- **Median:** Trung vị là giá trị có 50% giá trị của dãy.

Distance Travel	F	Tần số tích lũy
0-20.0	6	6
20.1-40.0	11	17
40.1-60.0	16	33
60.1-100.0	59	92
100.1-120.0	46	138
120.1-140	33	171
140.1-160	16	187
160.1-200	4	191
Tổng	191	

Lấy khoảng giá trị khi $cf_b \geq n/2 \geq 191/2 \geq 95.5$

⇒ Phần tử cần lấy là 96th nhưng đây là 1 khoảng giá trị nên ta không thể lấy ra giá trị chính xác mà chỉ có thể dựa vào tần số tích lũy

Dựa vào tần số tích lũy ta lấy khoảng 100.1-120.0 có tần số tích lũy 138 > 95.5

$$\text{Median} = L + \frac{w}{f(m)} (0.5n - cf_b) = 100.1 + \frac{24.29}{46} (0.5 * 191 - 92) = 101.95$$

L: giới hạn dưới của khoảng đã chọn

fm: tần số của khoảng

cf_b: tần số tích lũy của khoảng bên dưới khoảng đã chọn

w: độ rộng giữa 2 midpoint

Code R:

```
## ##  
## Median trung vi  
## tích lũy tần số  
tichluytanso<-function(x){  
  a<-x$F  
  b<-rep(0,length(a))  
  for(i in 1:length(a))  
  {  
    if(i==1){  
      b[i]<-a[i]  
    }  
    else {  
      b[i]<-b[i-1]+a[i]  
    }  
  }  
  return(b)  
}  
dt1$tichluy<-tichluytanso(data)
```

```
mediankhoang<-function(x){  
  L<-0  
  z<-length(x$tichluy)  
  k<-sum(x$F)/2  
  #lay ra gia tri meadian  
  for(i in 1:z){  
    if(x$tichluy[i] >= k)  
    {  
      L<-x$Ghduoi[i]  
      f<-x$F[i]  
      cf<-x$tichluy[i-1]  
      w<-dorongtb(x)  
      m<- L+(w/f*(k-cf))  
      return(m)  
      break  
    }  
  }  
}
```

```
> Modekhoang(dt1)  
[1] 78.75344  
> Meankhoang(dt1)  
[1] 96.2788  
> mediankhoang(dt1)  
[1] 101.9484  
>
```

Câu c: Mối quan hệ giữa ba thước đo trung tâm cho biết gì về hình dạng của biểu đồ đối với những dữ liệu này?

Nhìn vào giá trị Mode $\cong 78.75$ thuộc khoảng 60.1-100.0 \Rightarrow khoảng có tần số lớn nhất (cột cao nhất)

Mean $\cong 96.28$ thuộc khoảng 60.1-100.0 chia đồ thị sang 2 bên của cột 60.1-100.0.

Còn Median $\cong 101.95$ thuộc khoảng 100.1-120.0 \Rightarrow hình dạng của đồ thị sẽ có xu hướng lệch sang phải.

Câu d: Bạn sẽ đề xuất độ đo nào trong ba độ đo thích hợp nhất cho quãng đường đi được của 1 trong 191 xe buýt? Giải thích câu trả lời của bạn.

Giá trị Mean thích hợp nhất để đề xuất quãng đường đi được của 191 xe buýt .

Mean không phụ thuộc vào tần số tích lũy và độ rộng trung bình.

Bài 3.20: In a study of 1,329 American men reported in American Statistician [(1974) 28:115–122], the men were classified by serum cholesterol and blood pressure. The group of 408 men who had blood pressure readings less than 127 mm Hg were then classified according to their serum cholesterol level.

Serum Cholesterol (mg/100cc)	Frequency
0.0–199.9	119
200.0–219.9	88
220.0–259.9	127
greater than 259	74

- Estimate the mode, median, and mean for the serum cholesterol readings (if possible).
- Which of the three summary statistics is most informative concerning a typical serum cholesterol level for the group of men? Explain your answer.

Yêu cầu:

Trong một nghiên cứu trên 1.329 đàn ông Mỹ được báo cáo trên tạp chí American Statistician [(1974) 28: 115–122], những người đàn ông được phân loại theo cholesterol huyết thanh và huyết áp. Nhóm 408 người đàn ông có chỉ số huyết áp dưới 127 mm Hg sau đó được phân loại theo mức cholesterol huyết thanh của họ.

Huyết thanh Cholesterol - Tần số

Serum Cholesterol (mg/100cc)	Frequency
0.0–199.9	119
200.0–219.9	88
220.0–259.9	127
greater than 259	74

- Ước tính yếu vị, trung vị và giá trị trung bình cho các chỉ số cholesterol huyết thanh (nếu có thể).
- Thống kê tóm tắt nào trong số ba thống kê tóm tắt có nhiều thông tin nhất liên quan đến mức cholesterol huyết thanh điển hình cho nhóm nam giới? Giải thích câu trả lời của bạn.

Giải:

a) - Tính mean:

Công thức để tính giá trị trung bình cho dữ liệu đã cho

$$\bar{y} = \frac{\sum_{i=1}^k f_i * y_i}{n}$$

Khoảng cách phân loại của Cholesterol huyết thanh	Tần số (fi)	yi
0.0-199.9	119	$\frac{0.0 + 199.9}{2} = 99.95$
200.0-219.9	88	$\frac{200.0 + 219.9}{2} = 209.95$
220.0-259.9	127	$\frac{220.0 + 259.9}{2} = 239.95$
Tốt hơn 259	74	$\frac{399 + 259}{2} = 329$
Tổng	408	

Giới hạn trên của 259 là :

$$(209.95-99.95) + (239.95-209.95) + 259 = 399$$

```
> dl
  Ghduoi  Ghtren    F    yi  tichluy
1      0   199.9  119   99.95     119
2    200   219.9   88  209.95     207
3    220   259.9  127  239.95     334
4    259   399.0   74  329.00     408
```

- Tìm Mode:

Hàm tìm độ rộng w :

```
> #do rong tb w
> dorongtb<-function(x){
+   a<-x$yi
+   b<-length(x$yi)
+   dr<-0
+   c<-0
+   for(i in 1:(b-1)){
+     dr<-a[i+1]-a[i]
+     c<-c+dr
+   }
+   return (c/(b-1))
+ }
```

$$\text{Mode} = L + \frac{f_m - f_{m-1}}{(f_m - f_{m-1}) + (f_m + f_{m+1})} * w = 220 + \frac{127 - 88}{(127 - 88) + (127 - 74)} * 76.35 = 252.3$$

```

> #tim khoang co tan so cao nhat lay ra vi tri
> Modekhoang<-function(x){
+   m<-x$F
+   gh<-x$Ghduoi
+   n<-length(x$F)
+   L<-0
+   yeuvi<-0
+   #lay ra khoang co tuan so cao nhat
+   for(i in 1:n){
+     if(max(x$F)== m[i])
+     {
+       L<-gh[i]
+       yeuvi<- L+((m[i]-m[i-1])/(m[i]-m[i-1]+m[i]-m[i+1]))*dorongtb(x)
+     }
+   }
+   return(yeuvi)
+ }

> Modekhoang(dl)
[1] 252.3658

```

$$\text{Mean} = \frac{\sum_{i=1}^k f_i \cdot \bar{y}_i}{n} = \frac{119 \cdot 99.95 + 88 \cdot 209.95 + 127 \cdot 239.95 + 74 \cdot 329}{408} = 208.79$$

```

> #Mean gia tri trung binh
> Meankhoang<-function(x){
+   mean<-0
+   y<-x$F
+   mean<- sum(midpoint(x)*y)/sum(y)
+   return (mean)
+ }

```

```

> Meankhoang(dl)
[1] 208.7973

```

- Tính Median: Trung vị là giá trị $\geq 50\%$ giá trị của dãy.

Khoảng cách phân loại của Cholesterol huyết thanh	Tần số (f _i)	cf_b
0.0-199.9	119	119
200.0-219.9	88	207
220.0-259.9	127	334
Tốt hơn 259	74	408
Tổng	408	

Lấy khoảng giá trị khi $cf_b \geq \frac{n}{2} > 204$

$$\text{Median} = L + \frac{w}{f(m)} (0.5n - cf_b) = 200 + \frac{76.35}{88} (0.5 * 408 - 119) = 273.74$$

```

> mediankhoang<-function(x){
+   L<-0
+   z<-length(x$tichluy)
+   k<-sum(x$F)/2
+   #lay ra gia tri meadian
+   for(i in 1:z){
+     if(x$tichluy[i] >= k)
+     {
+       L<-x$Ghduoi[i]
+       f<-x$F[i]
+       cf<-x$tichluy[i-1]
+       w<-dorongtb(x)
+       m<- L+(w/f*(k-cf))
+       return(m)
+       break
+     }
+   }
+ }

> mediankhoang(dl)
[1] 273.7472

```

b) Median là thống kê trong số ba thống kê tóm tắt có nhiều thông tin nhất liên quan đến mức cholesterol huyết thanh điển hình cho nhóm nam giới.

Bài 3.21:

The ratio of DDE (related to DDT) to PCB concentrations in bird eggs has been shown to have a number of biological implications. The ratio is used as an indication of the movement of contamination through the food chain. The paper “The Ratio of DDE to PCB Concentrations in Great Lakes Herring Gull Eggs and Its Use in Interpreting Contaminants Data” [Journal of Great Lakes Research (1998) 24(1):12–31] reports the following ratios for eggs collected at 13 study sites from the five Great Lakes. The eggs were collected from both terrestrial- and aquatic-feeding birds.

	DDE to PCB Ratio										
Terrestrial Feeders	76.50	6.03	3.51	9.96	4.24	7.74	9.54	41.70	1.84	2.50	1.54
Aquatic Feeders	0.27	0.61	0.54	0.14	0.63	0.23	0.56	0.48	0.16	0.18	

- Compute the mean and median for the 21 ratios, ignoring the type of feeder.
- Compute the mean and median separately for each type of feeder.
- Using your results from parts (a) and (b), comment on the relative sensitivity of the mean and median to extreme values in a data set.
- Which measure, mean or median, would you recommend as the more appropriate measure of the DDE to PCB level for both types of feeders? Explain your answer.

Yêu cầu:

Tỉ lệ DDE (liên quan đến DDT) so với nồng độ PCB trong trứng chim đã được chứng minh là có một số tác động sinh học. Tỉ lệ này được sử dụng như một dấu hiệu của sự di chuyển của ô nhiễm thông qua chuỗi thức ăn. Bài báo "The Ratio of DDE to PCB Concentration in Great Lakes Herring Gull Eggs and Its Use in Interpreting-SS Data" (Journal of Great Lakes Research (1998) 24(1):12–31) báo cáo các tỉ lệ sau đây cho trứng được thu thập tại 13 địa điểm nghiên cứu từ năm Hồ Lớn. Những quả trứng được thu thập từ cả chim ăn trên cạn và dưới nước.

	DDE to PCB Ratio										
Terrestrial Feeders	76.50	6.03	3.51	9.96	4.24	7.74	9.54	41.70	1.84	2.50	1.54
Aquatic Feeders	0.27	0.61	0.54	0.14	0.63	0.23	0.56	0.48	0.16	0.18	

Giải:

dl<-c(76.50,6.03,3.51,9.96,4.24,7.74,9.54,41.7,1.84,2.5,1.54,0.27,0.61,0.54,0.14,0.63,0.23,0.56,0.48,0.16,0.18)

a. Tính toán trung bình và trung bình cho 21 tỉ lệ, bỏ qua loại bộ nạp.

- Trung bình mean:

$$\bar{X} = \frac{76.50 + 6.03 + 3.51 + 9.96 + 4.24 + 7.74 + 9.54 + 41.7 + 1.84 + 2.5 + 1.54 + 0.27 + 0.61 + 0.54 + 0.14 + 0.63 + 0.23 + 0.56 + 0.48 + 0.16 + 0.18}{21}$$

$$= 8.042857$$

```
> dl<-c(0.14,0.16,0.18,0.23,0.27,0.48,0.54,0.56,0.61,0.63,1.54,1.84,2.5,3.51,4.24,6.03,7.74,9.54,9.96,41.7,76.5)
> mean(dl)
[1] 8.042857
```

- Trung vị median:

0.14, 0.16, 0.18, 0.23, 0.27, 0.48, 0.54, 0.56, 0.61, 0.63, 1.54, 1.84, 2.5, 3.51, 4.24, 6.03, 7.74, 9.54, 9.96, 41.7, 76.5

Ta nhận thấy, ở đây có 21 giá trị (quan sát), nên trung vị của dãy số này là giá trị số ở giữa, đó là các giá trị thứ 11.

$$\text{median}=1.54$$

```
> dl<-c(0.14,0.16,0.18,0.23,0.27,0.48,0.54,0.56,0.61,0.63,1.54,1.84,2.5,3.51,4.24,6.03,7.74,9.54,9.96,41.7,76.5)
> mean(dl)
[1] 8.042857
> median(dl)
[1] 1.54
```

b. Tính toán trung bình và trung bình riêng biệt cho từng loại máng nạp.

❖ Trên cạn:

$$\bar{X} = \frac{76.50 + 6.03 + 3.51 + 9.96 + 4.24 + 7.74 + 9.54 + 41.7 + 1.84 + 2.5 + 1.54}{11}$$

$$= 15.00909$$

```
> dll<-c(76.50,6.03,3.51,9.96,4.24,7.74,9.54,41.7,1.84,2.5,1.54)
> mean(dll)
[1] 15.00909
```

- Sắp xếp theo thứ tự tăng dần:

1.54, 1.84, 2.5, 3.51, 4.24, 6.03, 7.74, 9.54, 9.96, 41.7, 76.5

Ta nhận thấy, ở đây có 21 giá trị (quan sát), nên trung vị của dãy số này là giá trị số ở giữa, đó là các giá trị thứ 6.

$$\text{median}=6.03$$


```
> dl1<-c(76.50,6.03,3.51,9.96,4.24,7.74,9.54,41.7,1.84,2.5,1.54)
> mean(dl1)
[1] 15.00909
> median(dl1)
[1] 6.03
```

❖ Dưới nước:

$$\bar{X} = \frac{0.27 + 0.61 + 0.54 + 0.14 + 0.63 + 0.23 + 0.56 + 0.48 + 0.16 + 0.18}{10}$$

$$= 0.38$$

```
> dl2<-c(0.27,0.61,0.54,0.14,0.63,0.23,0.56,0.48,0.16,0.18)
> mean(dl2)
[1] 0.38
```

- Sắp xếp theo thứ tự tăng dần:

0.14,0.16,0.18,0.23,0.27,0.48,0.54,0.56,0.61,0.63

Ta nhận thấy, ở đây có 10 giá trị (quan sát), nên trung vị của dãy số này là trung bình cộng của 2 số ở giữa, đó là các giá trị thứ 5 và thứ 6.

$$\text{median} = (0.27+0.48) / 2 = 0.375$$

```
> dl2<-c(0.27,0.61,0.54,0.14,0.63,0.23,0.56,0.48,0.16,0.18)
> mean(dl2)
[1] 0.38
> median(dl2)
[1] 0.375
```

c. Sử dụng kết quả của bạn từ các bộ phận (a) và (b), nhận xét về độ nhạy tương đối của các giá trị trung bình và trung vị đến giá trị ngoại lệ trong một tập dữ liệu.

❖ Khi sử dụng đáp án câu A:

- Giá trị trung bình cộng là giá trị bình quân của tất cả các giá trị, vì thế giá trị trung bình cộng phụ thuộc rất nhiều vào giá trị cực biên(min = 0.14 và max =76.5). Từ đó, làm cho giá trị bình quân sai lệch rất nhiều so với các phần tử ngoại lệ . Cho nên giá trị trung bình không có ý nghĩa trong trường hợp này.

- Giá trị trung vị có rất nhiều giá trị khác biệt quá lớn so vs phần tử trung vị:

```
> #1. tính trung vị x:M
> #2. tính độ lệch các pt so vs M: d=|x-M|
> #3.tính trung vị của d: Md
> #4.tính tỷ số: t=d/Md
> #5. nhưng tp của x tương ứng với t>=4.5 ngoại lệ
> x<-c(0.14,0.16,0.18,0.23,0.27,0.48,0.54,0.56,0.61,0.63,1.54,1.84,2.5,3.51,4.24,6.03,7.74,9.54,9.96,41.7,76.5)
> M<-median(x)
> d<-abs(x-M)
> Md<-median(d)
> t<-d/Md
> x[which(t>=4.5)]
[1] 7.74 9.54 9.96 41.70 76.50
> |
```

Có 5 phần tử ngoại lệ so với phần tử trung vị. Chính vì thế trong trường hợp này giá trị trung vị cũng không có ý nghĩa.

❖ Khi sử dụng đáp án câu B:

- Trên cạn: TB = 15.00909, TV = 6.03. Tương tự trong trường hợp này, ta có thể thấy

Với chuỗi giá trị trên cạn x = 1.54, 1.84, 2.5, 3.51, 4.24, 6.03, 7.74, 9.54, 9.96, 41.7, 76.5

Thì giá trị trung bình cũng sai lệch quá nhiều so với giá trị cực biên (min = 1.54, max = 76.5) nên giá trị trung bình không có ý nghĩa trong trường hợp này.

Đồng thời, sau khi xây dựng hàm `ol()` tính giá trị khác biệt so với phần tử trung vị:

```
> ol<-function(x) {  
+ M<-median(x)  
+ d<-abs(x-M)  
+ Md<-median(d)  
+ t<-d/Md  
+ x[which(t>=4.5)] }
```

Thì kết quả trả về, là chuỗi giá trị có 2 phần tử ngoại lệ.

```
> y<-c(1.54,1.84,2.5,3.51,4.24,6.03,7.74,9.54,9.96,41.7,76.5)  
> ol(y)  
[1] 41.7 76.5
```

Như vậy, giá trị trung vị cũng không có ý nghĩa trong trường hợp này.

- Dưới nước: TB = 0.38, TV = 0.375

Với chuỗi giá trị trên cạn $x = 0.14, 0.16, 0.18, 0.23, 0.27, 0.48, 0.54, 0.56, 0.61, 0.63$

Ta có thể thấy, giá trị trung bình không sai lệch quá nhiều so với 2 giá trị cực biên (min = 0.14, max = 0.63) như vậy trong trường hợp này giá trị trung bình đã thể hiện được sự phân bố tập trung của dữ liệu.

Đồng thời, khi sử dụng hàm tìm phần tử ngoại lệ của chuỗi giá trị so với giá trị trung vị thì kết quả trả về là 0.

```
> z<-c(0.14,0.16,0.18,0.23,0.27,0.48,0.54,0.56,0.61,0.63)  
> ol(z)  
numeric(0)
```

Như vậy, giá trị trung vị trong trường hợp này là đã thể hiện được mức độ phân tán hay tập trung của dữ liệu.

d. Đo lường nào, trung bình hoặc trung vị, bạn sẽ đề xuất như là thước đo thích hợp hơn của mức DDE đến PCB cho cả hai loại thức ăn? Giải thích câu trả lời của bạn.

Với trên cạn:

Giá trị trung vị là thích hợp hơn bởi vì trong trường hợp này mặc dù tồn tại 2 giá trị ngoại lệ, tuy nhiên giá trị trung vị không phụ thuộc vào giá trị ngoại biên, giá trị nhiễu, còn giá trị trung bình cộng thì bị ảnh hưởng quá nhiều nếu giá trị cực biên chênh lệch nhau quá lớn.

Với dưới nước:

Giá trị trung bình là thích hợp hơn. Vì trong trường hợp này dữ liệu tương đồng nhau, giá trị cực biên không chênh lệch quá nhiều, thì giá trị trung bình phân tích chính xác nhất, thích hợp nhất để tóm tắt và cho biết đặc trưng của tập dữ liệu này.

Bài 3.22: A study of the survival times, in days, of skin grafts on burn patients was examined by Woolson and Lachenbruch [Biometrika (1980) 67:597–606]. Two of the patients left the study prior to the failure of their grafts. The survival time for these individuals is some number greater than the reported value.

Survival time (days): 37, 19, 57*, 93, 16, 22, 20, 18, 63, 29, 60*

(The “*” indicates that the patient left the study prior to failure of the graft; values given are for the day the patient left the study.)

a. Calculate the measures of center (if possible) for the 11 patients.

b. If the survival times of the two patients who left the study were obtained, how would these new values change the values of the summary statistics calculated in (a)?

Yêu cầu:

Một nghiên cứu về thời gian sống sót, trong nhiều ngày, ghép da trên bệnh nhân bỏng đã được Woolson và Lachenbruch [Biometrika (1980) kiểm tra 67:597-606]. Hai trong số các bệnh nhân đã rời khỏi nghiên cứu trước khi các ca ghép của họ thất bại. Thời gian sống sót của những cá nhân này lớn hơn một số so với giá trị được báo cáo.

Thời gian sinh tồn (days): 37, 19, 57*, 93, 16, 22, 20, 18, 63, 29, 60*

(Dấu "*" chỉ ra rằng bệnh nhân đã rời khỏi nghiên cứu trước khi ghép thất bại; Các giá trị được đưa ra là cho ngày bệnh nhân rời khỏi nghiên cứu.)

- Tính toán các biện pháp trung tâm (nếu có thể) cho 11 bệnh nhân.
- Nếu thời gian sống sót của hai bệnh nhân rời khỏi nghiên cứu được thu thập, các giá trị mới này sẽ thay đổi các giá trị của thống kê tóm tắt được tính trong (a) như thế nào?

Giải:

- a) Vì đối với 2 trong số 11 bệnh nhân, thời gian sống sót tối thiểu được đưa ra, do đó

❖ Tính Mean: $n = 11$ patients

$$\bar{x} > \frac{(37+19+57+93+16+22+20+18+63+29+60)}{11} > 39.5$$

```
> mean(d1)
[1] 39.45455
```

❖ Tính Median: Công thức tính M_e

$$M_e = \begin{cases} (x_{[\frac{n}{2}]} + x_{[\frac{(n+2)}{2}]}) & \text{nếu } n \text{ chẵn} \\ x_{[\frac{(n+1)}{2}]} & \text{nếu } n \text{ lẻ} \end{cases}$$

Sắp xếp lại : 16, 18, 19, 20, 22, 29, 37, 57*, 60*, 63, 93

- Nếu: Thời gian sinh tồn là đủ 11 bệnh nhân chưa rời khỏi nghiên cứu là số lẻ thì $M_e = 29$,là trung vị vì nằm ở giữa hay vị trí của nó là: $x_{[\frac{(n+1)}{2}]} = x_{[\frac{(11+1)}{2}]} = x_{[6]}$

```
> median(d1)
[1] 29
```

❖ Tính Mode: Mỗi thời gian xuất hiện đúng 1 lần nên không có Yếu vị.

Hàm tìm yếu vị:

```
# ham tim phan tu yeu vi

# ham dem
demts <- function(x,y) {
dem<-0
for(i in 1:length(y)) if (y[i] == x) dem <- dem + 1
return(dem) }

#hàm Mode
Mode <- function(x) {
kq<-unique(x)
tam<-rep(0,length(kq))
for(i in 1:length(kq))
tam[i] <- demts(kq[i],x)
if (min(tam) != max(tam))
for(i in 1:length(kq))
if (tam[i] == max(tam)) print(kq[i])
}

Mode(dl)|
```

b) Nhận xét:

Nếu ta thu thập cả hai bệnh nhân thì giá trị Median sẽ giữ nguyên;

Mean, Mode sẽ thay đổi vì 2 giá trị được thu thập (chỉ ra rằng bệnh nhân đã rời khỏi nghiên cứu trước khi ghép thất bại) có thể tăng lên một giá trị nào đó như thế sẽ ảnh hưởng đến kết quả được đưa ra ở (a).

Bài 3.23: A study of the reliability of diesel engines was conducted on 14 engines. The engines were run in a test laboratory. The time (in days) until the engine failed is given here. The study was terminated after 300 days. For those engines that did not fail during the study period, an asterisk is placed by the number 300. Thus, for these engines, the time to failure is some value greater than 300.

Failure time (days): 130, 67, 300*, 234, 90, 256, 87, 120, 201, 178, 300*, 106, 289, 74

a. Calculate the measures of center for the 14 engines.

b. What are the implications of computing the measures of center when some of the exact failure times are not known?

Yêu cầu:

Một nghiên cứu về độ tin cậy của động cơ diesel đã được thực hiện trên 14 động cơ. Các động cơ được chạy trong phòng thí nghiệm. Thời gian (tính bằng ngày) cho đến khi động cơ bị lỗi được đưa ra ở đây. Nghiên cứu đã kết thúc sau 300 ngày. Đối với những động cơ không hỏng hóc trong thời gian nghiên cứu, dấu hoa thị được đặt bằng số 300. Vì vậy, đối với những động cơ này, thời gian hỏng hóc là một giá trị nào đó lớn hơn 300.

Thời gian lỗi (ngày): 130, 67, 300 *, 234, 90, 256, 87, 120, 201, 178, 300 *, 106, 289, 74

a. Tính các giá trị trung tâm của 14 động cơ.

b. Việc tính toán các thước đo trung tâm có ý nghĩa gì khi không biết chính xác thời gian thất bại?

Giải:

a.

- Công thức trung bình cộng:

$$\bar{x} = \frac{\sum_{i=1}^n x_i}{n}$$

Trong đó: x_i : giá trị quan sát thứ i của mẫu
 n : kích thước mẫu.

+ Vì có 2 trong số 14 động cơ thời gian hỏng hóc lớn hơn 300 nên số đo tâm của 14 động cơ sẽ là:

$$\bar{x} > \frac{(130+67+300+234+90+256+87+120+201+178+300+106+289+74)}{14}$$

$$\bar{x} > \frac{2432}{14}$$

$$\bar{x} > 173,7$$

+ Ta dùng hàm mean() trong R để tính trung bình cộng:

```
> dc<-c(130,67,300,234,90,256,87,120,201,178,300,106,289,74)
> mean(dc)
[1] 173.7143
```

- Công thức tính trung vị:

$$M_e = \begin{cases} (x_{[n/2]} + x_{[(n+2)/2]})/2 & \text{nếu } n \text{ chẵn} \\ x_{[n+1]/2} & \text{nếu } n \text{ lẻ} \end{cases}$$

+ Sắp xếp lại: 67, 74, 87, 90, 106, 120, 130, 178, 201, 234, 256, 289, 300*, 300*

+ Ta có $n=14$ (chẵn), nên:

$$M_e = \frac{130 + 178}{2} = 154$$

+ Ta dùng hàm median() trong R để tính trung vị:

```
> median(dc)
[1] 154
```

- Yếu vị:

Quan sát dãy giá trị trên ta thấy có 1 yếu vị là 300*, nhưng theo đề bài giá trị này là giá trị lớn hơn 300 và cũng có thể bằng nhau nên chưa đủ điều kiện kết luận đây là giá trị yếu vị của 14 động cơ.

+ Trong R không có hàm tính yếu vị, xây dựng hàm:

```

> Mode<-function(x){
+ y<-sort(x)
+ tam<-as.vector(rep(1,length(x)))
+ for (i in 1:length(x)) tam[i]<-dem(y,y[i])
+ if (min(tam)!=max(tam)){
+   for (i in 1:length(y)) if (tam[i]==min(tam)) so<-y[i]
+   for (i in 1:length(y)) if ((tam[i]==max(tam))&&(so!=y[i])){
+     print(y[i]); so<-y[i]}}}

```

- b. Khi không biết chính xác thời gian thất bại, giá trị Median sẽ **không bị ảnh hưởng** vì trung vị không phụ thuộc vào giá trị biên, còn giá trị Mean và Mode sẽ **bị ảnh hưởng bởi giá trị 300*** không cụ thể và dẫn đến kết quả không có giá trị chính xác.

Bài 3.24: Effective tax rates (per \$100) on residential property for three groups of large cities, ranked by residential property tax rate, are shown in the following table.

- Compute the mean, median, and mode separately for the three groups.
- Compute the mean, median, and mode for the complete set of 30 measurements.
- What measure or measures best summarize the center of these distributions?

Explain.

Group 1	Rate	Group 2	Rate	Group 3	Rate
Detroit, MI	4.10	Burlington, VT	1.76	Little Rock, AR	1.02
Milwaukee, WI	3.69	Manchester, NH	1.71	Albuquerque, NM	1.01
Newark, NJ	3.20	Fargo, ND	1.62	Denver, CO	.94
Portland, OR	3.10	Portland ME	1.57	Las Vegas, NV	.88
Des Moines, IA	2.97	Indianapolis, IN	1.57	Oklahoma City, OK	.81
Baltimore, MD	2.64	Wilmington, DE	1.56	Casper, WY	.70
Sioux Falls, IA	2.47	Bridgeport, CT	1.55	Birmingham, AL	.70
Providence, RI	2.39	Chicago, IL	1.55	Phoenix, AZ	.68
Philadelphia, PA	2.38	Houston, TX	1.53	Los Angeles, CA	.64
Omaha, NE	2.29	Atlanta, GA	1.50	Honolulu, HI	.59

Source: Government of the District of Columbia, Department of Finance and Revenue, Tax Rates and Tax Burdens in the District of Columbia: A Nationwide Comparison (annual).

Yêu cầu:

Thuế suất hiệu dụng (trên 100 đô la) đối với bất động sản nhà ở cho ba nhóm thành phố lớn, được xếp hạng theo thuế suất bất động sản nhà ở, được thể hiện trong bảng sau.

Group 1	Rate	Group 2	Rate	Group 3	Rate
Detroit, MI	4.10	Burlington, VT	1.76	Little Rock, AR	1.02
Milwaukee, WI	3.69	Manchester, NH	1.71	Albuquerque, NM	1.01
Newark, NJ	3.20	Fargo, ND	1.62	Denver, CO	.94
Portland, OR	3.10	Portland ME	1.57	Las Vegas, NV	.88
Des Moines, IA	2.97	Indianapolis, IN	1.57	Oklahoma City, OK	.81
Baltimore, MD	2.64	Wilmington, DE	1.56	Casper, WY	.70
Sioux Falls, IA	2.47	Bridgeport, CT	1.55	Birmingham, AL	.70
Providence, RI	2.39	Chicago, IL	1.55	Phoenix, AZ	.68
Philadelphia, PA	2.38	Houston, TX	1.53	Los Angeles, CA	.64
Omaha, NE	2.29	Atlanta, GA	1.50	Honolulu, HI	.59

Source: Government of the District of Columbia, Department of Finance and Revenue, Tax Rates and Tax Burdens in the District of Columbia: A Nationwide Comparison (annual).

Giải:

B24

Group1	Rate1	Group2	Rate2	Group3	Rate3
Detroit,MI	4.10	Burling,VT	1.76	LittleRock	1.02
Milwaukee,WI	3.69	Manchester,VT	1.71	Albuquerque,NM	1.01
Newark,NJ	3.20	Fargo,ND	1.62	Denver,CO	0.94
Portland,OR	3.10	Portland,ME	1.57	Las Yega,NV	0.88
Des Moines,IA	2.97	Indianapolis	1.57	Oklahoma City,OK	0.81
Baltimore,MD	2.64	Wilming,DE	1.56	Casper,WY	0.70
Sioux Falls,IA	2.47	Bridgeport	1.55	Birmingham,AL	0.70
Providence,RI	2.39	Chiago,IL	1.55	Phoenix,AZ	0.68
Philadelphia,PA	2.38	Houston,TX	1.53	Los Angeles,CA	0.64
Omaha,NE	2.29	Atlanta,GA	1.50	Honolulu,HI	0.59

- a. Tính giá trị trung bình, trung vị và chế độ riêng biệt cho ba nhóm.

Group 1:

- **Trung bình(mean)**

$$GTTB = \frac{(4.1+3.69+3.2+3.10+2.97+2.64+2.47+2.39+2.38+2.29)}{10} = 2.923$$

10

- **Trung vị,(median)**

Sắp xếp: 2.29 2.38 2.39 2.47 2.64 2.97 3.1 3.2 3.69 4.1

Ta nhận thấy, ở đây có 10 giá trị (quan sát), nên trung vị của dãy số này là trung bình cộng của 2 số ở giữa, đó là các giá trị thứ 5 và thứ 6.

$$\text{median} = (2.64 + 2.97) / 2 = 2.805$$

- **Yếu vị(Mode): không có yếu vị**

Code R:

```
mean(B24$Rate1)
1] 2.923
median(B24$Rate1)
1] 2.805
Mode(B24$Rate1)
```

Group 2:

- **Trung bình(mean)**

$$GTTB = \frac{1.5+1.53+1.55+1.55+1.56+1.57+1.57+1.62+1.71+1.76}{10} = 1.592$$

10

- **Trung vị,(median)**

Sắp xếp theo thứ tự tăng dần: 1.5 1.53 1.55 1.55 1.56 1.57 1.57 1.62 1.71 1.76

Ta nhận thấy, ở đây có 10 giá trị (quan sát), nên trung vị của dãy số này là trung bình cộng của 2 số ở giữa, đó là các giá trị thứ 5 và thứ 6.

$$\text{median} = (1.56 + 1.57) / 2 = 1.565$$

- **Yếu vị(Mode): 1.55,1.57**

Code R:

```
> mean(B24$Rate2)
[1] 1.592
> median(B24$Rate2)
[1] 1.565
> Mode(B24$Rate2)
[1] 1.57
[1] 1.55
```

Group 3:

- Trung bình(mean)

$$\text{GTTB} = \frac{(0.59+0.64+0.68+0.70+0.70+0.81+0.88+0.94+1.01+1.02)}{10} = 0.797$$

10

- Trung vị,(median)

Sắp xếp theo thứ tự tăng dần: 0.59 0.64 0.68 0.70 0.70 0.81 0.88 0.94 1.01 1.02

Ta nhận thấy, ở đây có 10 giá trị (quan sát), nên trung vị của dãy số này là trung bình cộng của 2 số ở giữa, đó là các giá trị thứ 5 và thứ 6.

$$\text{median} = (0.70 + 0.81) / 2 = 0.755$$

- Yếu vị: 0.70

```
mean(B24$Rate3)
[1] 0.797
median(B24$Rate3)
[1] 0.755
Mode(B24$Rate3)
[1] 0.7
|
```

b. Tính giá trị trung bình, trung vị và chế độ cho toàn bộ 30 phép đo.

- Trung bình(mean)

GTTB

$$\begin{aligned} & 4.10 + 3.69 + 3.20 + 3.10 + 2.97 + 2.64 + 2.47 + 2.39 + 2.38 + 2.29 + 1.76 + 1.71 \\ & + 1.62 + 1.57 + 1.57 + 1.56 + 1.55 + 1.55 + 1.53 + 1.50 + 1.02 + 1.01 + \\ & = \frac{0.94 + 0.88 + 0.81 + 0.70 + 0.70 + 0.68 + 0.64 + 0.59}{30} \end{aligned}$$

$$= 1.770667$$

```
a<-c(B24$Rate1,B24$Rate2,B24$Rate3)
a
[1] 4.10 3.69 3.20 3.10 2.97 2.64 2.47 2.39 2.38 2.29 1.76 1.71 1.62 1.57 1.57
[6] 1.56 1.55 1.55 1.53 1.50 1.02 1.01 0.94 0.88 0.81 0.70 0.70 0.68 0.64 0.59
mean(a)
[1] 1.770667
|
```

- Trung vị (median)

Sắp xếp theo thứ tự tăng dần: 0.59 0.64 0.68 0.70 0.70 0.81 0.88 0.94 1.01 1.02 1.5 1.53 1.55 1.55 1.56 1.57 1.57 1.62 1.71 1.76 2.29 2.38 2.39 2.47 2.64 2.97 3.1 3.2 3.69 4.1

Ta nhận thấy, ở đây có 30 giá trị (quan sát), nên trung vị của dãy số này là trung bình cộng của 2 số ở giữa, đó là các giá trị thứ 15 và thứ 16.

$$(1.56+1.57)/2=1.565$$

```
median(a)
[1] 1.565
```

- **Yếu vị(Mode) : 0.7, 1.55 và 1.57**

```
> a<-c(B24$Rate1,B24$Rate2,B24$Rate3)
> a
[1] 4.10 3.69 3.20 3.10 2.97 2.64 2.47 2.39 2.38 2.29 1.76 1.71 1.62 1.57 1.57
[16] 1.56 1.55 1.55 1.53 1.50 1.02 1.01 0.94 0.88 0.81 0.70 0.70 0.68 0.64 0.59
> Mode(a)
[1] 1.57
[1] 1.55
[1] 0.7
> |
```

c. Biện pháp nào tóm tắt tốt nhất trung tâm của các phân bố này?

Biện pháp nào tóm tắt tốt nhất trung tâm của các phân bố này là biện pháp trung bình

- Giá trị cực biên(min = 0.59 và max =4.1).

- Vì giá trị cực biên không trên lệch quá nhiều và có nhiều, giá trị tương đồng khá nhiều nên biện pháp trung bình là thích hợp nhất để tóm tắt trung tâm phân bố này.

Bài 3.25: Refer to Exercise 3.24. Average the three group means, the three group medians, and the three group modes, and compare your results to those of part (b). Comment on your findings.

Yêu cầu:

Dựa vào bài 3.24. Trung bình ba nhóm trung bình, ba nhóm trung vị, và ba nhóm yếu vị, và so sánh kết quả với phần b. Nhận xét phần mình thấy.

Giải:

Trung bình nhóm 1 = 2.923

Trung bình nhóm 2 = 1.592

Trung bình nhóm 3 = 0.797

$$\begin{aligned}\text{Trung bình của ba nhóm trung bình} &= \frac{\text{Tổng ba nhóm trung bình}}{\text{Số nhóm}} \\ &= \frac{2.923+1.592+0.797}{3} \\ &= 1.7707\end{aligned}$$

```
> x<-c(2.923,1.592,0.797)
> mean(x)
[1] 1.770667
```

Trung vị nhóm 1 = 2.805

Trung vị nhóm 2 = 1.565

Trung vị nhóm 3 = 0.755

$$\begin{aligned}\text{Trung bình của ba nhóm trung vị} &= \frac{\text{Tổng ba nhóm trung vị}}{\text{Số nhóm}} \\ &= \frac{2.805+1.565+0.755}{3} \\ &= 1.7083\end{aligned}$$

```
> y<-c(2.805,1.565,0.755)
> mean(y)
[1] 1.708333
```

Yếu vị nhóm 1 : Không có

Yếu vị nhóm 2 = 1.55, 1.57

Yếu vị nhóm 3 = 0.70

$$\begin{aligned}\text{Trung bình của ba nhóm yếu vị} &= \frac{\text{Tổng ba nhóm yếu vị}}{\text{Số nhóm}} \\ &= \frac{0+(1.55+1.57)+0.70}{3} \\ &= 1.273\end{aligned}$$

```
> z<-c(0,3.12,0.7)
> mean(z)
[1] 1.273333
```

Nhận xét:

- Vì ba nhóm trung bình đều có 10 phần tử, nên khi trung bình của ba nhóm trung bình sẽ có 30 phần tử => Kết quả giống 3.24b.
- Vì ba nhóm trung vị có 3 giá trị trung vị khác nhau, nên khi trung bình của ba trung vị sẽ khác với trung vị của 30 phần tử => Kết quả khác 3.24b.
- Vì mỗi nhóm có một khoảng giá trị riêng, nên yếu vị của mỗi nhóm sẽ khác nhau. Khi gộp 30 phần tử, số lượng yếu vị cũng không thay đổi => Giá trị trung bình sẽ bằng nhau.

HẾT!