Statistique descriptive en logiciel R

Exercice corrigé

EXERCICE Cette série statistique représente le nombre d'enfants par familles pour un l'échantillon de 100 famille marocaines :

1 - 4 - 4 - 3 - 1 - 5 - 2 - 3 - 2 - 2 - 2 - 3 - 2 - 6 - 4 - 1 - 3 - 1 - 6 - 2 - 1 - 4 - 1 - 1 - 6 - 4 - 1 - 5 - 5 - 1 - 3 -
3 - 7 - 2 - 4 - 3 - 2 - 3 - 3 - 5 - 2 - 3 - 6 - 4 - 4 - 3 - 2 - 2 - 1 - 5 - 3 - 6 - 2 - 0 - 5 - 1 - 7 - 4 - 3 - 5 - 4 - 3 -
6 - 5 - 2 - 7 - 7 - 3 - 3 - 1 - 2 - 0 - 2 - 3 - 5 - 3 - 3 - 7 - 6 - 7 - 2 - 3 - 1 - 3 - 3 - 3 - 5 - 3 - 2 - 5 - 6 - 4 - 1 -
0 - 2 - 6 - 5 - 1 - 0 – 1.

1. Déterminer la population étudiée, le caractère statistique étudié

2. Chargement des données

```
> nbEnfant = c(1, 4, 4, 3, 1, 5, 2, 3, 2, 2, 2, 3, 2, 6, 4, 1, 3, 1, 6, 2, 1, 4, 1, 1, 6, 4, 1, 5, 5,
1, 3, 3, 7, 2, 4, 3, 2, 3, 3, 5, 2, 3, 6, 4, 4, 3, 2, 2, 1, 5, 3, 6, 2, 0, 5, 1, 7, 4, 3, 5, 4, 3, 6, 5,
2, 7, 7, 3, 3, 1, 2, 0, 2, 3, 5, 3, 3, 7, 6, 7, 2, 3, 1, 3, 3, 3, 5, 3, 2, 5, 6, 4, 1, 0, 2, 6, 5, 1, 0,
1) > nbEnfant [1] 1 4 4 3 1 5 2 3 2 2 2 3 2 6 4 1 3 1 6 2 1 4 1 1 6 4 1 5 5 1 3 3 7 2 [35] 4
3 2 3 3 5 2 3 6 4 4 3 2 2 1 5 3 6 2 0 5 1 7 4 3 5 4 3 6 5 2 7 7 3 [69] 3 1 2 0 2 3 5 3 3 7
6 7 2 3 1 3 3 3 5 3 2 5 6 4 1 0 2 6 5 1 0 1>
```

3. Taille de la population

```
> length(nbEnfant)
[1] 100
```

4. Type du caractère statistique

```
> mode(nbEnfant)
[1] "numeric"
```

5. Structure de la distribution statistique taille

```
> str(nbEnfant)
 num [1:100] 1 4 4 3 1 5 2 3 2 2 ...
```

6. Liste de modalités

```
> unique(nbEnfant)
[1] 1 4 3 5 2 6 7 0
```

7. L'effectif de la modalité 4

```
> length(nbEnfant[nbEnfant==4])
[1] 11
```

8. Fréquence de la modalité 4

```
> length(nbEnfant[nbEnfant==4])/length(nbEnfant)
[1] 0.11
```

9. L'effectif cumulée croissante de 4

```
> length(nbEnfant[nbEnfant<=4])
[1] 73
```

10. Fréquence cumulée croissante de 4

```
> length(nbEnfant[nbEnfant<=4])/length(nbEnfant)
[1] 0.73
```

11. Table des effectifs

```
> table(nbEnfant)
nbEnfant
 0 1 2 3 4 5 6 7 4 16 18 24
 11 12 9 6
```

12. Table des effectifs cumulés

```
> cumsum(table(nbEnfant))
          7 4 20 38 62 73 85 94 100
```

13. Table des Fréquences cumulées

```
> cumsum(table(nbEnfant))/length(nbEnfant)
    7 0.04 0.20 0.38 0.62 0.73 0.85 0.94 1.00
```

Mesures de tendance centrale :

14. Le mode

En R, il n'existe pas de fonction native directe pour calculer le mode d'une distribution statistique. Cependant, vous pouvez utiliser l'une des deux syntaxes suivantes :

```
> as.numeric(names(sort(table(nbEnfant),decreasing=TRUE)[1])) [1]
3
```

ou encore

```
> as.numeric(names(table(nbEnfant)[table(nbEnfant)==max(table(nbEnfant))]))
[1] 3
```

15. La moyenne

```
> mean(nbEnfant)
[1] 3.24
```

16. La médiane

```
> median(nbEnfant)
[1] 3
```

Mesures de dispersion

17. L'étendue

```
> range(nbEnfant)
[1] 0 7
```

18. Les quartiles

```
> quantile(nbEnfant)
   0% 25% 50% 75% 100% 7
    0       2       3       5
```

19. Utilisation de la fonction quantile pour calculer la médiane

```
> quantile(nbEnfant,0.5)
 50%
 3
```

20. Les déciles

```
> quantile(nbEnfant,0:10/10)
   0% 10% 20% 30% 40% 50% 60% 70% 80% 90% 100%
```

```
0.0 1.0 1.8 2.0 3.0 3.0 3.0 4.0 5.0 6.0 7.0
```

21. La variance

On distincts deux concepts de la variance : **la variance théorique** et **la variance corrigée**. Ces deux mesures statistiques évaluent la dispersion des données par rapport à leur moyenne. La variance théorique, également désignée sous le nom de **variance totale**, évalue la variabilité globale dans un ensemble de données en calculant la moyenne des carrés des écarts entre chaque point de données et la moyenne générale de l'ensemble. Elle permet de comprendre la variabilité naturelle des valeurs présentes dans la population ou l'échantillon. En contraste, la variance corrigée, souvent appelée **variance de l'échantillon**, ajuste cette mesure pour tenir compte de l'estimation de la moyenne à partir d'un échantillon plutôt que de la population entière. La variance corrigée utilise le degré de liberté corrigé (n-1) au dénominateur pour compenser le biais potentiel résultant de l'utilisation de l'échantillon pour estimer la moyenne. Elle est fréquemment employée dans l'estimation de la variance d'une population à partir d'un échantillon, offrant ainsi une mesure plus précise de la dispersion des données dans la population.

Variance corrigé : fonction var()

```
> var(nbEnfant)
[1] 3.517576
```

Variance théorique en utilisant la formule de la variance :

```
> mean(nbEnfant ^2) - mean(nbEnfant)^2
[1] 3.4824
```

Variance théorique à partir de la variance corrigée :

```
> var(nbEnfant)*(length(nbEnfant)-1)/ length(nbEnfant) [1]
3.4824
```

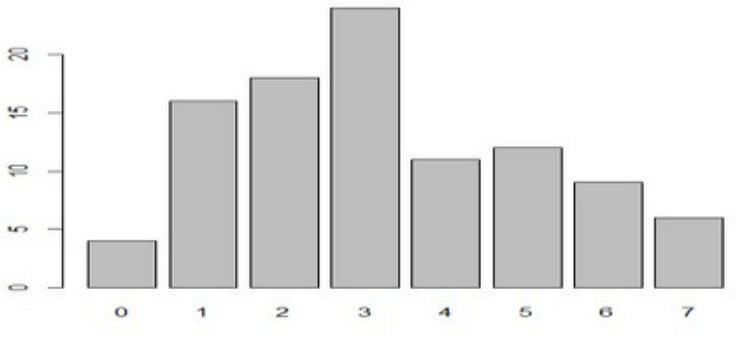
23. L'écart type

```
> sd(nbEnfant)
[1] 1.87552
```

Représenter graphiquement

24. Histogramme

```
> barplot(table(nbEnfant))
```



25. Boite à moustache

```
boxplot(nbEnfant,horizontal = T)
```