Labo 4: Paramètres de tendance centrale: solution détaillée

Visseho Adjiwanou, PhD.

06 February 2023

PARTIE A

Voici les résultats obtenus au cours d'une enquête sur l'âge et le statut matrimonial des répondants.

Tableau 1: Distribution du statut matrimonial

Statut	Fréquence	Pourcentage	Pourcentage valide	Pourcentage cumulé
Marié	247	29.1	29.1	29.1
Veuf	3	.4	.4	29.4
Divorcé	36	4.2	4.2	33.6
Séparé	14	1.6	1.6	35.3
Jamais marié	550	64.7	64.7	100.0
Total	**850**	**100.0**	**100.0**	

Répondez aux questions suivantes:

1. Quel est le type de la variable étudiée?

```
Il s'agit d'une variable nominale
```

2. Quel est la valeur du mode?

```
Le mode est le score qui apparait le plus fréquemment. Il s'agit ici de marié
```

3. Si vous pouvez utiliser la médiane, indiquez sa valeur. Sinon, dites que ce n'est pas possible et expliquer votre réponse.

```
On ne peut pas calculer la médiane. Puisque la médiane exige que les scores soient ordonnés,
on ne peut trouver une médiane que pour des variables ordinales, ou d'intervalles/ratio.
```

4. Si vous pouvez utiliser la moyenne, indiquez sa valeur. Sinon, dites que ce n'est pas possible et expliquer votre réponse.

```
On ne peut pas calculer la moyenne car on ne peut faire la somme des modalités
```

5. Quel est le problème avec ce tableau? Quelle solution préconisez-vous?

```
1. Les effectifs sont trop faibles pour les catégories veuf, divorcé et séparé. On peut
considérer qu'elles reflètent la même réalité, des hommes qui ne vivent plus avec une
partenaire. On peut donc les regrouper ensemble dans la catégorie séparée. Bien sûr, vous
manquez de précisions

2. On ne devrait pas calculer les pourcentages cumulés sur les variables nominales. On ne
peut pas les interpréter. Mathématiquement, on peut faire le calcul, mais cela n'a pas de
sens.
```

Au cours de la même enquête, on a collecté les données sur le groupe d'âges des enquêtés. Les résultats sont présentés dans le tableau 2.

Tableau 2: Distribution du groupe d'âges

Groupe d'âge	Fréquence	Pourcentage	Pourcentage valide	Pourcentage cumulé
15 - 24	276	276/850*100	276/815*100	
25 - 34	199	199/850*100		
35 - 49	263	263/850*100		
50 et plus	77	77/850*100		
Non réponse	35	35/850*100		
Total	**850**	850/850*100		

1. Quel est le type de la variable étudiée?

`Il s'agit d'une variable qualitative ordinale`

2. Complétez le tableau

Groupe d'âge	Fréquence	Pourcentage	Pourcentage valide	Pourcentage cumulé
15 - 24	276	32.5	33.9	33.9
25 - 34	199	23.4	24.4	33.9 + 24.4 = 58.3
35 - 49	263	30.9	32.3	58.3 + 32.3 = 90.6
50 et plus	77	9.1	9.4	90.6 + 9. 4 = 100
Non réponse	35	4.1		
Total	**850**	100.0	100.0	

15- 24 : 10

15 16 17 18 19 20 21 22 23 24 1 1 1 1 1 1 1 1 1 1

(815 + 1)/2=408

3. Quel est la valeur du mode?

`La valeur du mode est 15-24 ans.`

4. Si vous pouvez utiliser la médiane, indiquez sa valeur. Sinon, dites que ce n'est pas possible et expliquer votre réponse.

Solution 1:

`Il faut exclure les données manquantes (NSP, pas de réponse, etc.) avant de calculer une médiane. De telles valeurs fournissent peu d'information. D'ailleurs, l'inclusion de ces données transformerait une variable ordinale ou d'intervalles/ratio en variable nominale puisqu'il n'y a pas d'ordre entre les valeurs des données manquantes.`

Pour calculer la médiane, on regarde le pourcentage cumulé. On peut se rendre compte que les 50% de la distribution se trouvent dans le groupe d'âges 25-34 ans. Et donc que la médiane est cette valeur, 25-34 ans.

Cependant, on peut affiner notre résultat. Si on ordonne tous les individus selon leur âge, on sait que la médiane est l'âge de l'individu situé à la (815+1)/2 (408e) position. On utilise 815 car j'ai exclu les non réponses.

On va supposer que les âges des 199 individus du groupe d'âges 25 - 34 ans sont répartis uniformément dans cet intervalle (10 ans). On peut aussi supposer que le premier individu de cet intervalle est l'individu numéro 277 et le dernier l'individu 276+199 = 475. On voit ainsi clairement que la 408e position est plus proche de 35 ans que de 25 ans. Si on positionne les âges avec les effectifs, on a:

25 26 27 28 29 30 31 M 32 33 34 35

277 408 475

Il y a une règle dite de “trois” qu’on peut appliquer pour trouver M

(M - 35) / (35 - 25) = (408 - 475) / (475 - 277)

Ainsi on a, X = 35 - (475-408)/(475 - 277)*(35-25)

X = 31,6 ans

Solution 2

Vous pouvez directement appliquer la formule du livre

$$M_d = L + (\frac{N/2 - F}{f})(i)$$

L = limite inférieure de l’intervalle contenant la médiane N = le nombre total de score F = la fréquence cumulative des scores inférieurs à l’intervalle contenant la médiane f = le nombre de scores que comprend l’intervalle contenant la médiane i = la largeur de l’intervalle contenant la médiane (c’est-à dire la limite supérieure de l’intervalle moins sa limite inférieure)

`Md = 25 + (815/2 - 276)/199*(34-25 + 1) soit, 31.6 ans.`

5. Si vous pouvez utiliser la moyenne, indiquez sa valeur. Sinon, dites que ce n’est pas possible et expliquer votre réponse.

`En règles générale: Il est dit dans le livre qu'on ne peut pas calculer la moyenne pour des variables ordinales. Cela peut se calculer dans certaines conditions (si les données ne sont pas des classes, en ce cas, ils sont traités comme des ratios discrets. L'interprétation est quand même délicate) sous quelques hypothèses (dans le cas des données groupées comme ici, il faut une hypothèse sur où se situe les données).`

`Une autre manière de regrouper les données est de penser que chaque intervalle peut être représenté par son centre. Ainsi les 276 individus du groupe d'âges 15-24 ans ont en moyenne 20 ans si nous supposons qu'ils ont des âges répartis uniformément dans cet intervalle. Ainsi, on a:` 20 ans –> 276 individus 30 ans –> 199 individus 42.5 ans –> 263 individus ? –> 77 individus.

Ceci ressemble à des données groupées dont on peut calculer aisément la moyenne qui vaut:

M = (276*20 + 199*30 +263*42.5 + 77*55)/815 = 33.0 ans (en supposant que le dernier groupe d’âge est 59 ans)

Si vous supposez que le dernier groupe d’âges est 79 ans, alors le dernier groupe d’âge peut se trouver au centre de cet intervalle qui vaut alors 65 ans. La moyenne devient:

M = (276*20 + 199*30 +263*42.5 + 77*65)/815 = 34.0 ans. Pas un grand changement, mais vous comprenez que la moyenne est affectée par les cas déviants. La médiane n’est pas du tout affectée.

PARTIE B

La solution technologique au changement climatique (exemple tiré de Krieg)

Beaucoup de gens pensent qu’en adoptant de nouvelles technologies, nous pouvons économiser à la fois de l’argent et protéger l’environnement en brûlant moins de combustibles fossiles. Cet exercice est tiré du livre

de krieg, “Statistics and data analysis for Social Science”.

1. Que pensez-vous de cette assertion?

Vrai et faux.

Vrai car elle permet d’améliorer l’efficacité des appareils. Faux car - les technologies produisent elles-mêmes des problèmes - elles sont coûteuses - ne valent pas grande chose sans un changement dans nos habitudes de consommations

2. En quoi n’est-elle pas valide?

Pour tester cette assertion, nous utilisons les données de 1994 et de 2009 sur les voitures les plus efficients entre les deux périodes. Le tableau suivant présente les vitesses (mile per gallon, mpg) pour les différentes marques de voitures pour leur circulation en ville et sur l’autoroute:

- **Pour 1994**

Marque et modèle	Ville (mpg)	Autoroute(mpg)
Mazda 626	23	31
Honda Accord	22	29
Chevrolet Corsica	22	28
Buick Century	22	28
Oldsmobile Cutlass Ciera	22	28
Oldsmobile Achieva	21	32
Pontiac Grand Am	21	32
Infiniti G20	21	29
Mitsubishi Galant	21	28
Dodge Spirit	21	27
Plymouth Acclaim	21	27
Subaru Legacy	20	28
Toyota Camry	20	27
Hyundai Sonata	19	26
Chrysler LeBaron	19	25
Ford Taurus	18	27
Mercury Sable	18	27
Eagle Vision	18	26

- **Pour 2009**

Marque et modèle	Ville (mpg)	Autoroute(mpg)
Toyota Prius Hybrid)	48	45
Nissan Altima (hybrid)	35	33
Toyota Camry (hybrid)	33	34
Chevrolet Malibu (hybrid)	26	34
Saturn Aura (hybrid)	26	34
Hyundai Elantra	25	33
Kia Spectra	24	32
Nissan Altima	23	32
Saturn Aura	22	33
Kia Optima	22	-
Hyundai Sonata	22	32

Marque et modèle	Ville (mpg)	Autoroute(mpg)
Honda Accord	22	31
Chevrolet Malibu	22	30
Toyota Camry	21	31
Volkswagen Passat	21	31
Mazda 6	21	30
Chrysler Sebring	21	30
Dodge Avenger	21	30
Ford Fusion	20	29
Mercury Milan	20	29
Mitsubishi Galant	20	27
Subaru Legacy	20	27
Nissan Maxima	19	260
Nissan Altima	19	26
Mercury Sable	18	28
Hyundai Azera	18	26
Buick LaCrosse/Allure	17	28

3. Quelle est la taille de chaque échantillon

échantillon 1994 = 18 échantillon 2009 = 27

4. Éfficacité gagnée en ville

Vous allez calculer le mode, la médiane et la moyenne pour la vitesse en **ville** en 1994 et 2009. Quelle conclusion tirez-vous? A cette étape de l'exercice, je vous demande de faire les calculs à la main.

- médiane: valeur au milieu de la distribution
  - 1989: valeur qui se situe entre la 18/2(9e) et la 10e position.
  - 2009 : valeur qui se situe à la 27+1/2 (14e) position
- Moyenne : somme des valeurs divisés par le nombre de cas

Paramètres	1994	2009
mode	21	21 et 22 (bimodale)
médiane	21	21
moyenne	$(23 + \ldots 18)/18 = 20.5$	23.2

5. Éfficacité sur autoroute

Le calcul que vous venez de faire est trop long. On peut présenter les données précédentes sous forme de données agrégées. C'est quoi encore les données agrégées?

4.1 Regrouper les données de la **vitesse sur l'autoroute** sous forme agrégée. Cela veut dire qu'il faut dénombrer le nombre de voitures pour chaque niveau de vitesse. Faite cela pour les données de 1994 et de 2009.

- Données de 1994

Vitesse	nombre de voiture (effectifs)	fréquence	Fréquence cumulée
23	1	1/18 = 5.6	1/18 = 5.6

Vitesse	nombre de voiture (effectifs)	fréquence	Fréquence cumulée
22	4	$4/18 = 22.2$	$1/18 + 4/18 = 27.8$
21	6	$6/18 = 33.3$	$1/18 + 4/18 + 6/18 = 61.1$
20	2	$2/18 = 11.1$	72.2
19	2	$2/18 = 11.1$	83.2
18	3	$3/18 = 16.7$	100
Total	18	100	

5.1 Présenter dans ce même tableau les fréquences, et les fréquences cumulées

5.2 Quelle représentation graphique vous semble la plus appropriée pour ces données?

- diagramme en bâton
- diagramme circulaire

Le diagramme en bâton est plus indiqué car il permet de représenter les deux données sur le même graphique.

5.3 Calculer à nouveau le mode, la médiane et la moyenne à partir de ses données groupées. Quelle conclusion tirez-vous?

Voir partie A.

6. Utilisation de R

Maintenant, nous allons utiliser R pour faire le même travail. Voici comment vous allez procéder.

1. Créer la base de données **donnee_1994** avec les variable suivantes:

- modele
- vitesse_ville et
- vitesse_autoroute

Vous comprenez que cette base de données contient donc 18 observations pour 3 variables. Quelle est la nature de chaque variable?

Réponse 1

- Pour 1994

```
# Créons d'abord des vecteurs

marque_1994 <- c("Mazda 626", "Honda Accord", "Chevrolet Corsica", "Buick Century", "Oldsmobile Cutlass
marque_1994
```

```
##  [1] "Mazda 626"                "Honda Accord"
##  [3] "Chevrolet Corsica"        "Buick Century"
##  [5] "Oldsmobile Cutlass Ciera" "Oldsmobile Achieva"
##  [7] "Pontiac Grand Am"         "Infiniti G20"
##  [9] "Mitsubishi Galant"        "Dodge Spirit"
## [11] "Plymouth Acclaim"         "Subaru Legacy"
## [13] "Toyota Camry"             "Hyundai Sonata"
## [15] "Chrysler LeBaron"         "Ford Taurus"
## [17] "Mercury Sable"            "Eagle Vision"
```

```
annee_pas_cool_1994 <- c(1994, 1994, 1994,1994, 1994, 1994, 1994, 1994, 1994, 1994, 1994, 1994, 1994, 19
annee_pas_cool_1994
```

```
##  [1] 1994 1994 1994 1994 1994 1994 1994 1994 1994 1994 1994 1994 1994 1994 1994
## [16] 1994 1994 1994
```

```
length(marque_1994)
```

```
## [1] 18
```

```
annee_un_peu_bon_1994 <- c(rep(1994, 18))
annee_un_peu_bon_1994
```

```
##  [1] 1994 1994 1994 1994 1994 1994 1994 1994 1994 1994 1994 1994 1994 1994 1994
## [16] 1994 1994 1994
```

```
annee_1994 <- c(rep(1994, length(marque_1994)))

annee_1994
```

```
##  [1] 1994 1994 1994 1994 1994 1994 1994 1994 1994 1994 1994 1994 1994 1994 1994
## [16] 1994 1994 1994
```

```
mpg_ville_1994 <- c(23, rep(22, 4), rep(21, 6), rep(20, 2), rep(19, 2), rep(18, 3))
mpg_ville_1994
```

```
##  [1] 23 22 22 22 22 21 21 21 21 21 21 20 20 19 19 18 18 18
```

```
mpg_autoroute_1994 <- c(31, 29, rep(28, 3), rep(32, 2), 29, 28, rep(27, 2), 28, 27, 26, 25, rep(27, 2),
mpg_autoroute_1994
```

```
##  [1] 31 29 28 28 28 32 32 29 28 27 27 28 27 26 25 27 27 26
```

Une fois qu'on a créé les vecteurs, on peut facilement créer les deux bases de données

```
donnee_1994 <- data.frame(marque = marque_1994, annee = annee_1994, vitesse_ville = mpg_ville_1994, vit

donnee_1994
```

```
##                      marque annee vitesse_ville vitesse_autoroute
## 1                 Mazda 626  1994            23                31
## 2              Honda Accord  1994            22                29
## 3         Chevrolet Corsica  1994            22                28
## 4             Buick Century  1994            22                28
## 5  Oldsmobile Cutlass Ciera  1994            22                28
## 6         Oldsmobile Achieva  1994            21                32
## 7           Pontiac Grand Am  1994            21                32
## 8              Infiniti G20  1994            21                29
## 9         Mitsubishi Galant  1994            21                28
## 10             Dodge Spirit  1994            21                27
## 11          Plymouth Acclaim  1994            21                27
## 12             Subaru Legacy  1994            20                28
## 13              Toyota Camry  1994            20                27
## 14            Hyundai Sonata  1994            19                26
## 15          Chrysler LeBaron  1994            19                25
## 16               Ford Taurus  1994            18                27
## 17             Mercury Sable  1994            18                27
## 18              Eagle Vision  1994            18                26
```

- Pour 2009

```
marque_2009 <- c("Toyota Prius (hybrid)", "Nissan Altima (hybrid)", "Toyota Camry (hybrid)", "Chevrolet


annee_2009 <- c(rep(2009, length(marque_2009)))

mpg_ville_2009 <- c(48, 35, 33, rep(26, 2), 25, 24, 23, rep(22, 5), rep(21, 5), rep(20, 4), rep(19, 2),

mpg_autoroute_2009 <- c(45, 33, rep(34,3), 33, rep(32,2), 33, NA , 32, 31, 30, rep(31,2), rep(30,3), re


donnee_2009 <- data.frame(marque = marque_2009, annee = annee_2009, vitesse_ville = mpg_ville_2009, vit

donnee_2009
```

```
##                       marque annee vitesse_ville vitesse_autoroute
## 1      Toyota Prius (hybrid)  2009            48                45
## 2     Nissan Altima (hybrid)  2009            35                33
## 3      Toyota Camry (hybrid)  2009            33                34
## 4  Chevrolet Malibu (hybrid)  2009            26                34
## 5       Saturn Aura (hybrid)  2009            26                34
## 6            Hyundai Elantra  2009            25                33
## 7                Kia Spectra  2009            24                32
## 8              Nissan Altima  2009            23                32
## 9                Saturn Aura  2009            22                33
## 10               Kia Optima  2009            22                NA
## 11            Hyundai Sonata  2009            22                32
## 12             Honda Accord  2009            22                31
## 13          Chevrolet Malibu  2009            22                30
## 14             Toyota Camry  2009            21                31
## 15         Volkswagen Passat  2009            21                31
## 16                 Mazda 6  2009            21                30
## 17          Chrysler Sebring  2009            21                30
## 18             Dodge Avenger  2009            21                30
## 19               Ford Fusion  2009            20                29
## 20             Mercury Milan  2009            20                29
## 21         Mitsubishi Galant  2009            20                27
## 22             Subaru Legacy  2009            20                27
## 23             Nissan Maxima  2009            19               260
## 24             Nissan Altima  2009            19                26
## 25             Mercury Sable  2009            18                28
## 26             Hyundai Azera  2009            18                26
## 27     Buick LaCrosse/Allure  2009            17                28
```

```
saveRDS(donnee_1994, "donnee1994.RDS")
```

2. Calculer la moyenne, la médiane et le mode des deux variables **vitesse_ville** et **vitesse_autoroute** à partir des données **donnee_1994**.

```
moyenne_1994 <- mean(donnee_1994$vitesse_ville)
moyenne_1994
```

```
## [1] 20.5
```

```
sum(donnee_1994[, "vitesse_ville"])/18
```

```
## [1] 20.5
```

```
getmode <- function(v) {
   uniqv <- unique(v)
   uniqv[which.max(tabulate(match(v, uniqv)))]
}

mode_1994 <- getmode(donnee_1994$vitesse_ville)
mode_1994
```

```
## [1] 21
```

- Commenter vos résultats. Si vous vous rappelez, pour calculer la moyenne et la médiane, il faut utiliser les fonctions **mean** et **mediane**.
- Cependant, il **N'existe PAS** de fonction **mode** pour calculer le mode. Je vous demande de faire quelques recherches et me venir avec une solution. Il est dès fois important de ne pas se focaliser pour comprendre ce que vous faites du moment où ça marche. Donnez-vous le temps de le comprendre plus tard.

3. Il y a plusieurs autres paramètres de tendance centrale que les trois que nous avons vus en classe. Vous avez le minimum, le maximum, le premier quartile, le 3e quartile et plus généralement les **ntiles**. Calculer ces différents paramètres sur les variables vitesse_ville et vitesse_autoroute. Commenter vos résultats.

- étendue (max- min)
- écart inter quartile (Q3 - Q1)
- Kurtosis (étalement)
- Skewness (asymetrie)

```
Q1_1994 <- quantile(donnee_1994$vitesse_ville, prob = 0.25)
Q1_1994
```

```
##   25%
## 19.25
```

```
Q3_1994 <- quantile(donnee_1994$vitesse_ville, prob = 0.75)
Q3_1994
```

```
##   75%
## 21.75
```

```
variance_1994 <- var(donnee_1994$vitesse_ville)
variance_1994
```

```
## [1] 2.382353
```

```
variance_2009_autoroute <- var(donnee_2009$vitesse_autoroute, na.rm = TRUE)
variance_2009_autoroute
```

```
## [1] 2030.962
```

4. La fonction **descr** de summarytools vous permet aussi de calculer ces paramètres de tendance centrale. Utiliser cette fonction pour calculer les paramètres calculer au 2 et 3.

```
library(summarytools)
```

```
## Warning in system2("/usr/bin/otool", c("-L", shQuote(DSO)), stdout = TRUE):
## running command ''/usr/bin/otool' -L '/Library/Frameworks/R.framework/Resources/
## library/tcltk/libs//tcltk.so'' had status 1
```

```
## For best results, restart R session and update pander using devtools:: or remotes::install_github('ra
```

```
descr(donnee_1994$vitesse_ville)
```

```
## Descriptive Statistics
## donnee_1994$vitesse_ville
## N: 18
##
##                     vitesse_ville
## ----------------- ---------------
##              Mean           20.50
##           Std.Dev            1.54
##               Min           18.00
##                Q1           19.00
##            Median           21.00
##                Q3           22.00
##               Max           23.00
##               MAD            1.48
##               IQR            2.50
##                CV            0.08
##          Skewness           -0.36
##       SE.Skewness            0.54
##          Kurtosis           -1.17
##           N.Valid           18.00
##         Pct.Valid          100.00
```

5. Maintenant, refaite la même chose avec les données de 2009.

6. Quelle conclusion tirez-vous sur la solution technologique au changement climatique?

annee parametre 1994
vitesse ville vitesse au moyenne mode médiane