Labo 5: Paramètres de tendance centrale

Visseho Adjiwanou, PhD.

07 February 2023

# PARTIE A

## Question 1: (tiré de Krieg)

A partir des données du tableau suivant, calculer :

* le mode
* la médiane
* la moyenne
* l’étendue
* l’écart inter-quartile
* la variance et l’écart-type

| NE | Fréquence | Pourcentage | Pourcentage valide | Pourcentage cumulé |
| --- | --- | --- | --- | --- |
| 0 | 414 | 28.1 | 28.1 | 28.1 |
| 1 | 242 | 16.4 | 16.4 | 44.5 |
| 2 | 398 | 27.0 | 27.0 | 71.4 |
| 3 | 226 | 15.3 | 15.3 | 86.8 |
| 4 | 115 | 7.8 | 7.8 | 94.6 |
| 5 | 58 | 3.9 | 3.9 | 98.6 |
| 6 | 14 | .9 | .9 | 99.5 |
| 7 | 7 | .5 | .5 | 100.0 |

NE : Nombre d’enfants

**Taille de l’échantillon** N = 414 + 242 + … + 7

**Niveau de mesure de la variable**

Ratio

### Réponse

* Mode : 0
* Médiane : 44,5% des répondants ont au plus 1 enfant (c’est à dire qu’ils en ont 0 ou 1). 71,4% ont au plus 2 enfants (c’est à dire qu’ils en ont 0, ou 1 ou 2). Donc, on peut dire que la moitié de l’échantillon (50%) a entre 1 et 2 enfants. La médiane est alors l’intervalle [1,2]. On peut être plus précis en utilisant la formule de la fois passée:

1 médiane 2 44.5% 50% 71.4 414+242=656 1474/2=737 414+242+398=1054

(Médiane - 2)/(50 - 71.4) = (1 - 2)/(44.5 - 71.4)

Médiane = 2 + (1-2)/(44.5 - 71.4)\*(50 - 71.4)

Médiane = 1.2 enfants

Autre manière de le visualiser: On a au total 1474 répondants. La moitié est entre 737 et 737.5

0 1 md 2 414 242 737 398 658 1056

737 est plus proche de 658 (donc de 1) que de 1056 (2)

(mediane - 2)/(737 - 1056) = (1-2)/(658-1056) mediane = 2 + (1-2)/(658-1056)\*(737 - 1056) mediane = 1.20

* Moyenne

| NE | Fréquence | Pourcentage | Pourcentage valide | Pourcentage cumulé |
| --- | --- | --- | --- | --- |
| 0 | 414 | 28.1 | 28.1 | 28.1 |
| 1 | 242 | 16.4 | 16.4 | 44.5 |
| 2 | 398 | 27.0 | 27.0 | 71.4 |
| 3 | 226 | 15.3 | 15.3 | 86.8 |
| 4 | 115 | 7.8 | 7.8 | 94.6 |
| 5 | 58 | 3.9 | 3.9 | 98.6 |
| 6 | 14 | .9 | .9 | 99.5 |
| 7 | 7 | .5 | .5 | 100.0 |

414\*0 + 242\*1 + 298\*2 + 226\*3 + 115\*4 + 58\*5 + 14\*6 + 7\*7

## [1] 2399

414 + 242 + 398 + 226 + 115 + 58 + 14 +7

## [1] 1474

moyenne <- 2399 / 1474  
moyenne

## [1] 1.627544

mode < médiane < moyenne ==> asymétrique avec un étalement vers la droite.

# Parametres de dispersion

* étendue : 7 - 0 = 7 enfants
* EIQ:
  + 1e Quartile : 25% des gens ont 0 enfant donc 1eQ = 0

2 3eq 3  
71.4 75% 86.8%

3eq = 3 + (2-3)/(71.4 - 86.8)\*(75 - 86.8) = 2.2

* 3e quartile: 75% des gens ont entre 2 et 3 enfants = 2.2

donc EIQ = 2.2 - 0 = 2.2

Exemple simple pour calculer le Q1, Q2 et l’écart:

<https://www150.statcan.gc.ca/n1/edu/power-pouvoir/ch12/5214890-fra.htm#>:~:text=L’%C3%A9cart%20interquartile%20est%20une,comme%20mesure%20de%20la%20dispersion.&text=L’%C3%A9cart%20interquartile%20couvre%2050,le%20quartile%20le%20plus%20faible.

* Variance

| NE | Fréquence | Pourcentage | Pourcentage valide | Pourcentage cumulé |
| --- | --- | --- | --- | --- |
| 0 | 414 | 28.1 | 28.1 | 28.1 |
| 1 | 242 | 16.4 | 16.4 | 44.5 |
| 2 | 398 | 27.0 | 27.0 | 71.4 |
| 3 | 226 | 15.3 | 15.3 | 86.8 |
| 4 | 115 | 7.8 | 7.8 | 94.6 |
| 5 | 58 | 3.9 | 3.9 | 98.6 |
| 6 | 14 | .9 | .9 | 99.5 |
| 7 | 7 | .5 | .5 | 100.0 |

La formule de la varianlce est

414 personnes avec 0 enfant

# A = (0 - 1.6)^2 + ... + (0 - 1.6)^2, 414 fois  
  
A <- 414\*(0-1.6)^2  
B<- 242\*(1 - 1.6)^2  
C <- 398\*(2 - 1.6)^2  
D <- 226\*(3 -1.6)^2  
E <- 115\*(4 - 1.6)^2  
F <- 58\*(5 - 1.6)^2  
G <- 14\*(6 - 1.6)^2  
H <- 7\*(7 - 1.6)^2  
  
somme <- A + B + C + D + E + F + G + H  
  
variance <- somme / (1474 -1)

* écart-type

écart\_type <- sqrt(variance)

## Question 2: (tiré de Krieg)

Le graphique suivant présente l’histogramme de l’âge au premier mariage.

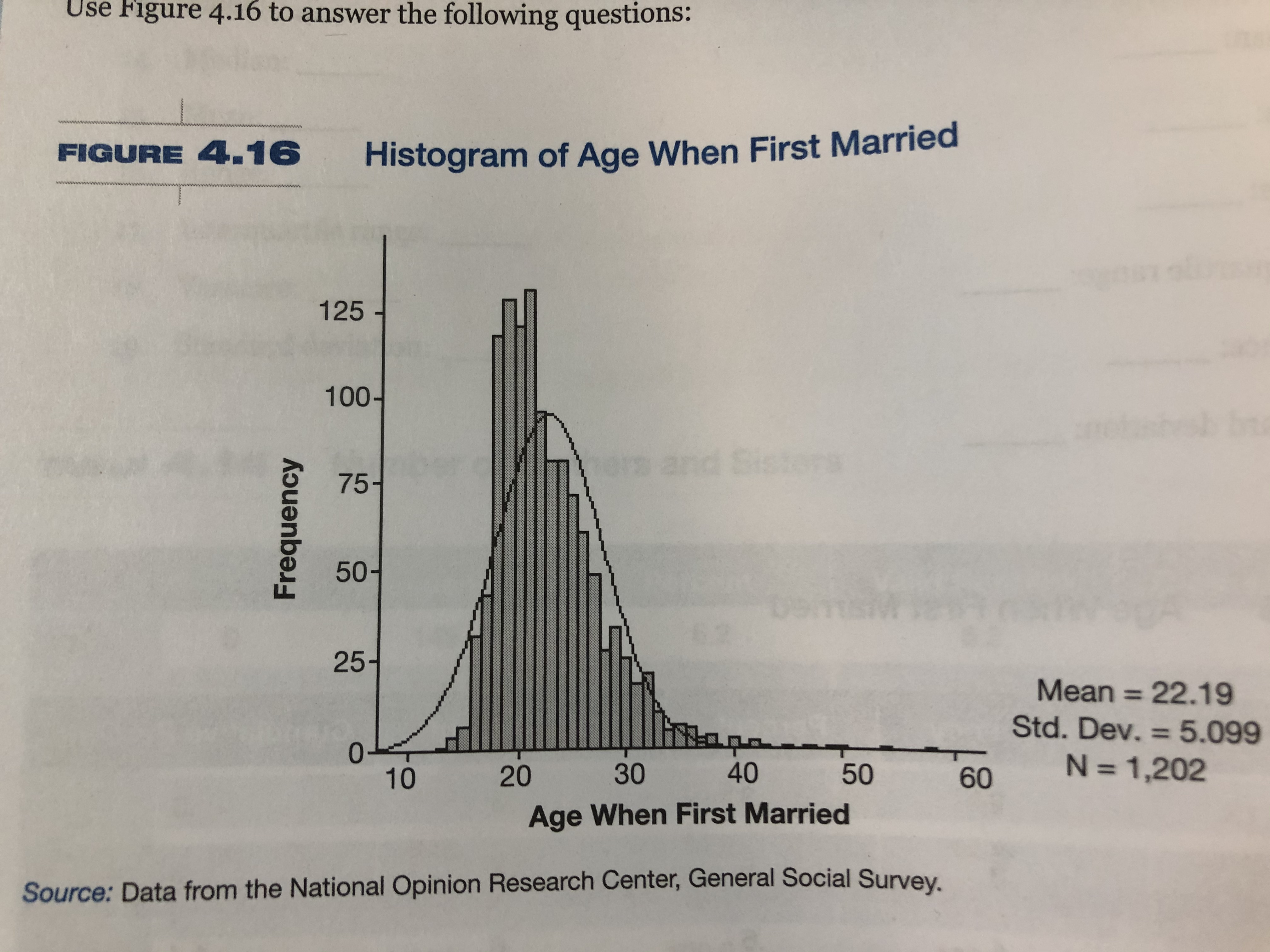


Figure 1: Histogramme de l’âge au premier mariage

A partir de ce graphique, répondez aux questions suivantes:

1. Quel est l’âge moyen des répondant.es à leur premier mariage?

22.19

1. Combien de répondant.es ont été enquêté.es? 1202
2. Quelle est la valeur de la variance?

variance <- 5.099^2

1. En se basant sur les propriétés de la courbe normale, nous pouvons dire que 68% des répondant.es se sont marié.es entre les âges 22.19 - 5.099 et 22.19 + 5.099, soit entre 17.10 et 27.29 ans.
2. En se basant sur les propriétés de la courbe normale, nous pouvons dire que 95% des répondant.es se sont marié.es entre les âges 22.19 - 2\*5.099 et 22.19 + 2\*5.099 soit entre 12.00 ans et 32.38 ans

## Question 4 - représentation graphique

Quelles sont les types de représentation graphique que l’on peut faire avec une variable quantitative (ratio ou intervalle) ?

* **Histogramme. l’histogramme est différent d’un diagramme de barre**
* **Diagramme de quartile encore appelé “boîte à moustache”**

## Exemple du cours

Score <- c(64, 68, 70, 71, 69, 66)  
Score

## [1] 64 68 70 71 69 66

Z\_score <- (Score - mean(Score))/sd(Score)  
Z\_score

## [1] -1.5339300 0.0000000 0.7669650 1.1504475 0.3834825 -0.7669650

# Moyenne  
mean(Z\_score)

## [1] 0

# Écart-type  
sd(Z\_score)

## [1] 1

# PARTIE B

# La solution technologique au changement climatique, suite et fin (exemple tiré de Krieg)

À partir de la base d données que vous avez crées et utilisées à partir des données sur les voitures, répondez aux questions suivantes en utilisant R:

* Pour 1994

# Créons d'abord des vecteurs  
  
marque\_1994 <- c("Mazda 626", "Honda Accord", "Chevrolet Corsica", "Buick Century", "Oldsmobile Cutlass Ciera", "Oldsmobile Achieva", "Pontiac Grand Am", "Infiniti G20", "Mitsubishi Galant", "Dodge Spirit", "Plymouth Acclaim", "Subaru Legacy", "Toyota Camry", "Hyundai Sonata", "Chrysler LeBaron", "Ford Taurus", "Mercury Sable", "Eagle Vision")  
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, 1994, 1994, 1994, 1994, 1994)  
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)) # Que fait la fonction rep?  
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))) # que fait la fonction length?  
  
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), 26)  
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, vitesse\_autoroute = mpg\_autoroute\_1994)  
  
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

Maintenant, on peut le sauvegarder si on veut pour ne pas être amené à refaire tout cela la prochaine fois. Remarquer que je l’ai enrégistré comme un fichier R (extension .RDS)

saveRDS(donnee\_1994, "donnee1994.RDS")

Je peux alors l’ouvrir et l’utiliser

d1994 <- readRDS("donnee1994.RDS") # Que signifie d1994?

1. Recalculer les paramètres de position sur les variables suivantes:

* la vitesse en ville en 1994
* la vitesse sur autoroute en 1994
* la vitesse en ville en 2009
* la vitesse sur autoroute en 2009

Je calcule la moyenne et la médiane pour les variables vitesse sur autorute et en ville en 1994.

# Moyenne  
  
moyenne\_vitesse\_ville\_1994 <- mean(d1994$vitesse\_ville)  
moyenne\_vitesse\_ville\_1994

## [1] 20.5

moyenne\_vitesse\_autoroute\_1994 <- mean(d1994$vitesse\_autoroute)  
moyenne\_vitesse\_autoroute\_1994

## [1] 28.05556

# Mis dans le même vecteur  
moy\_ville\_autoroute\_1994 <- c(moyenne\_vitesse\_ville\_1994, moyenne\_vitesse\_autoroute\_1994)  
moy\_ville\_autoroute\_1994

## [1] 20.50000 28.05556

#Mediane  
  
mediane\_vitesse\_ville\_1994 <- median(d1994$vitesse\_ville)  
mediane\_vitesse\_ville\_1994

## [1] 21

mediane\_vitesse\_autoroute\_1994 <- median(d1994$vitesse\_autoroute)  
mediane\_vitesse\_autoroute\_1994

## [1] 28

# Mis dans le même vecteur  
med\_ville\_autoroute\_1994 <- c(mediane\_vitesse\_ville\_1994, mediane\_vitesse\_autoroute\_1994)  
med\_ville\_autoroute\_1994

## [1] 21 28

# Mis ensemble dans un fichier  
  
lieu <- c("Ville", "Autoroute")  
  
Tendance\_1994 <- data.frame(lieu, moy\_ville\_autoroute\_1994, med\_ville\_autoroute\_1994)  
Tendance\_1994

## lieu moy\_ville\_autoroute\_1994 med\_ville\_autoroute\_1994  
## 1 Ville 20.50000 21  
## 2 Autoroute 28.05556 28

### Interprétation des résultats

On voit que qu’en moyenne, la vitesse est en moyenne plus grande sur l’autoroute qu’en ville. Les moyennes sont très proches des médianes.

1. Calculer la variance et l’écart-type des quatre variables précédentes.

# Variance  
variance\_vitesse\_ville\_1994 <- var(d1994$vitesse\_ville)  
variance\_vitesse\_ville\_1994

## [1] 2.382353

variance\_vitesse\_autoroute\_1994 <- var(d1994$vitesse\_autoroute)  
variance\_vitesse\_autoroute\_1994

## [1] 3.820261

var\_1994 <- c(variance\_vitesse\_ville\_1994, variance\_vitesse\_autoroute\_1994)  
  
# Écart-type  
  
  
ecart\_type\_vitesse\_ville\_1994 <- sd(d1994$vitesse\_ville)  
ecart\_type\_vitesse\_ville\_1994

## [1] 1.543487

ecart\_type\_vitesse\_autoroute\_1994 <- sd(d1994$vitesse\_autoroute)  
ecart\_type\_vitesse\_autoroute\_1994

## [1] 1.954549

ecart\_1994 <- c(ecart\_type\_vitesse\_ville\_1994, ecart\_type\_vitesse\_autoroute\_1994)  
  
Variation\_1994 <- data.frame(lieu, var\_1994, ecart\_1994)  
Variation\_1994

## lieu var\_1994 ecart\_1994  
## 1 Ville 2.382353 1.543487  
## 2 Autoroute 3.820261 1.954549

1. Comment ces résultats permettent-ils d’infirmer ou de renforcer la conclusion que vous avez tiré sur la solution technologique au changement climatique tiré au 1.

Pour répondre à cette question, j’ai besoin des données de 2009. Je vous laisse donc créer cette seconde base de données à partir des données du tableau du labo 4 et répondre à l’ensemble des questions. Bon travail.

## Solution de la question 4

laissez-moi vous entraîner dans la passion des statistiques.

1. La distribution d’échantillonnage n’est rien d’autre que la distribution d’un échantillon (Vrai ou Faux)

Réponse: Faux.

Une distribution d’échantillonnage correspond à une distribution de statistiques quelconques (moyennes, médianes, etc.) provenant de tous les échantillons possibles d’une taille donnée que l’on peut tirer d’une population déterminée. Une distribution d’un échantillon est la distribution des scores à l’intérieur d’un échantillon d’une taille donnée.

Voici une manière de comprendre la distribution d’échantillon. On va utiliser les données du marathon de Boston de 2012. Le tableau présente la description de cette base de données:

| Variables | Description |
| --- | --- |
| V1 | Nom |
| V2 | Sexe |
| V3 | Age |
| V4 | Division |
| V5 | Pays |
| V6 | Temps mis |

Pour estimer le temps moyen que ce marathon a été couru, nous n’avons pas besoin d’avoir forcément l’ensemble des résultats de la course. Un échantillon représentatif devrait suffire à nous fournir ce résultat. Donc, nous allons prendre un échantillon de 40 coureurs et calculer le temps moyen mis. Pour le moment, nous vivons comme si nous ne connaissons par le temps vrai (issu de la course).

Nous chargeons la base de données dans R.

rm(list = ls())  
marathon <- read.csv("bm\_results2012.txt", header = FALSE, quote = "")

On voit ainsi que 21541 personnes ont couru ce marathon.

Pour tirer un échantillon, nous allons utiliser la fonction **sample\_n** du package **dplyr**. Comme toujours, vous devez installer ce package (une seule fois), le charger (avant toute utilisation).

#install.packages("dplyr")  
library(tidyverse)

## ── Attaching packages ─────────────────────────────────────── tidyverse 1.3.2 ──  
## ✓ ggplot2 3.3.6 ✓ purrr 0.3.4  
## ✓ tibble 3.1.6 ✓ dplyr 1.0.8  
## ✓ tidyr 1.2.0 ✓ stringr 1.4.1  
## ✓ readr 2.1.2 ✓ forcats 0.5.2

## Warning: package 'tibble' was built under R version 3.6.2

## Warning: package 'tidyr' was built under R version 3.6.2

## Warning: package 'readr' was built under R version 3.6.2

## Warning: package 'purrr' was built under R version 3.6.2

## Warning: package 'dplyr' was built under R version 3.6.2

## ── Conflicts ────────────────────────────────────────── tidyverse\_conflicts() ──  
## x dplyr::filter() masks stats::filter()  
## x dplyr::lag() masks stats::lag()

set.seed(430)  
marathon\_S1 <- sample\_n(marathon, 40, replace = FALSE)

Nous voyons ainsi que marathon\_S1 a juste 40 observations.

**Remarques:**

1. dplyr fait partie d’une famille de package qui s’appelle **tidyverse** que nous allons voir lors du prochain cours.
2. L’option replace = FALSE signifie que quand nous tirons un élément de la population, nous ne le retournons pas dans la population avant de tirer le second.
3. Même si on choisit les individus de manières aléatoires, on veut quand même que d’autres personnes tirent exactement les mêmes individus que nous s’ils veulent vérifier nos résultats. set.seed(n’importe quel chiffre) permet de tirer chaque fois le même échantillon. N’utiliser pas cela quand vous faites vos calculs, cela va nous permettre de voir les différents résultats que nous obtenons tous.
4. Un rappel sur l’échantillonage : <https://www150.statcan.gc.ca/n1/edu/power-pouvoir/ch13/prob/5214899-fra.htm>

Ainsi, à partir de notre échantillon tiré, on peut calculer le temps moyen mis pour courir le marathon de

1, 2, 3, NA 1, 2, 3

temps\_moyen\_S1 <- mean(marathon\_S1$V6, na.rm = TRUE)  
temps\_moyen\_S1

## [1] 268.5303

sd(marathon\_S1$V6)

## [1] 45.23583

268 +- 1.96\*ecart-type/racine carré(N)

On voit ainsi qu’à partir de notre échantillon que le temps mis pour parcourir le marathon de New York est de 268.5 minutes.

Maintenant, jetons un coup d’œil sur les vrais résultats de ce marathon.

temps\_moyen\_vrai <- mean(marathon$V6, na.rm = TRUE)  
temps\_moyen\_vrai

## [1] 263.0493

Le temps moyen mis en vrai est de 263.0 minutes. Voyez comment est proche notre estimation à partir juste d’un échantillon aléatoire de 40 coureur.es.

Voua avez tous obtenus une réponse différente de la mienne. Une distribution d’échantillonnage n’est rien d’autre que l’ensemble des moyennes qu’on peut calculer à partir des millions d’échantillon de taille 40 de cette population de 21541 coureurs. En fait, on connaît exactement le nombre d’échantillons possible qu’on peut tirer. C’est cela qui s’appelle une combinaison avec la formule . Tenez-vous bien, ce nombre équivaut à :

253 077729093757612460294630785205954107797288308598032063927991250054643573537874160223994161945109975602007594 059 581 835 406 976

{A, B, C, D}

:{A, B}, {A, C}

J’ai utilisé ce générateur en ligne pour le calculer : <https://www.dcode.fr/combinaisons>

Bref, revenons à notre exercice. Aujourd’hui avec la puissance des machine, on peut déterminer en une fraction de secondes le nombre total d’échantillon. Mais, au fait, si on peut tirer tous ces échantillons, n’est-il juste pas plus simple de collecter l’information sur l’ensemble de la population ? Bien sûr. Dans les faits, on ne tirera jamais plus d’un échantillon.

Si je prenais par contre 100 échantillons de taille 40, la distribution que j’obtiens de la moyenne est ce qu’on appelle **une distribution d’échantillonnage**. Voyons ce que cela donne :

set.seed(123432)  
  
marat\_T40\_R100 <- bind\_rows(replicate(100, sample\_n(marathon, 40, replace = FALSE), simplify = F), .id = "Obs")

En lisant cette syntaxe de la droite vers la gauche, cela veut dire que : - **sample\_n(marathon, 40, replace = FALSE), simplify = F)**: je tire un échantillon de taille 40 dans la base de données **marathon** - **(replicate(100, sample\_n(marathon, 40, replace = FALSE)** : je replique cela 100 fois; - **bind\_rows(replicate(100, sample\_n(marathon, 40, replace = FALSE), simplify = F), .id = “Obs”)**: Je les colle ensemble (avec la fonction bind\_rows de dplyr) en les distinguant par leur numéro dans la nouvelle variable “Obs”

C’est tout. Vous voyez que cela me donne un échantillon de 4000 observations. La variable Obs distingue ainsi chaque unique échantillon.

Si nous calculons alors la moyenne de chaque échantillon on obtient la base de données avec les moyenne. On va utiliser la fonction stby de summarytools pour faire cela facilement. Remarquez une fois de plus que je charge ce package avant de l’utiliser.

#library(summarytools)  
  
#with(marat\_T40\_R100, stby(data = V6, INDICES = Obs,  
# FUN = descr, stats = c("mean")))

Summarytools ne semble pas fonctionner, je vais utiliser une autre option:

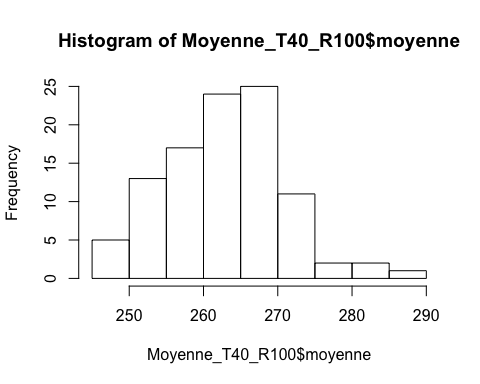
Moyenne\_T40\_R100 <-  
 marat\_T40\_R100 %>%   
 group\_by(Obs) %>%   
 summarise(moyenne = mean(V6, na.rm = TRUE))   
  
Moyenne\_T40\_R100

## # A tibble: 100 × 2  
## Obs moyenne  
## <chr> <dbl>  
## 1 1 259.  
## 2 10 253.  
## 3 100 265.  
## 4 11 259.  
## 5 12 267.  
## 6 13 267.  
## 7 14 271.  
## 8 15 269.  
## 9 16 255.  
## 10 17 261.  
## # … with 90 more rows

Ainsi, on a une nouvelle base de donnés avec les 100 moyennes issus de chaque marathon échantillonné.

Dressons la distribution de ces 100 moyennes, à l’aide de l’histogramme

hist(Moyenne\_T40\_R100$moyenne)



On voit ainsi que cela à l’allure d’une courbe normale.

Que pensez-vous que la moyenne de ces moyennes va nous donner?

moyenne\_moyenne <- mean(Moyenne\_T40\_R100$moyenne)  
moyenne\_moyenne

## [1] 263.0335

Cela nous donne exactement 263 minutes, ce qui est exactement la vraie moyenne à quelque centième de seconde près. Bingo, la moyenne des moyennes nous donne une estimation de la moyenne de la population.

Pour faire cet exercice, voici un peu toutes les recherches que j’ai faites sur le net:

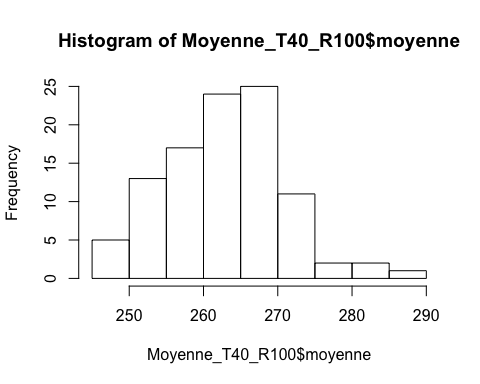


1. Énoncez et expliquer le théorème de la **limite centrale**

Ce théorème dit simplement que si N devient de plus en plus grand, la forme de la courbe tend de plus en plus vers une courbe normale. Voyons cela: A la place de 40 observations - Tirons plutôt 90; - Tirons à nouveau 100 échantillons; - Calculons la vitesse moyenne de chaque échantillon - Présentons l’histogramme de cette distribution

Vous pouvez le faire non? Essayer.

# Echantillon  
set.seed(12343)  
marat\_T90\_R100 <- bind\_rows(replicate(100, sample\_n(marathon, 90, replace = FALSE), simplify = F), .id = "Obs")  
  
# Moyenne  
  
Moyenne\_T90\_R100 <-  
 marat\_T90\_R100 %>%   
 group\_by(Obs) %>%   
 summarise(moyenne = mean(V6, na.rm = TRUE))  
  
# Histogramme  
  
hist(Moyenne\_T40\_R100$moyenne)



et la moyenne des moyennes vaut

mean(Moyenne\_T90\_R100$moyenne)

## [1] 262.8977

Vous verrez surtout que l’écart-type est plus faible.

sd(Moyenne\_T90\_R100$moyenne)

## [1] 5.428166

Alors que avec l’échantillon de 40, cela donnait:

sd(Moyenne\_T40\_R100$moyenne)

## [1] 7.796008

Quand la taille augmente, notre confiance dans les résultats augmente.

Alors, faites la même chose en choississant un échantillon de 150 coureur.es.

1. Il n’y a pas de différence entre l’**écart-type** et l’**erreur-type**

l’écart-type est ce que vous calculez à partir d’un seul échantillon, alors que l’erreur-type est ce que vous calculez à partir d’une distribution d’échantillonnage. Donc, en haut, j’ai plutôt calculé les erreurs-types.

Pouvez-vous me dire alors, comment sont calculés les erreurs-type?

**N’est-ce pas formidable les statistiques?**