Test de chi-deux

Visseho Adjiwanou, PhD.

27 March 2023

## 1. Tableau croisé et Chi-deux

A partir des données du tableau suivant:

| id | Sexe | Emploi | Age |
| --- | --- | --- | --- |
| 1 | Masculin | Temps plein | 27 |
| 2 | Féminin | Non | 34 |
| 3 | F | Partiel | 22 |
| 4 | F | Non | 38 |
| 5 | M | Partiel | 30 |
| 6 | F | Temps plein | 27 |
| 7 | M | Temps plein | 33 |
| 8 | M | Temps plein | 39 |
| 9 | F | Temps plein | 36 |
| 10 | F | Non | 25 |
| 11 | M | Partiel | 31 |
| 12 | M | Partiel | 28 |
| 13 | M | Non | 26 |
| 14 | F | Partiel | 21 |
| 15 | F | Temps plein | 29 |
| 16 | F | Temps plein | 34 |
| 17 | F | Non | 20 |
| 18 | F | Non | 32 |
| 19 | M | - | 27 |
| 20 | F | Partiel | 33 |
| 21 | M | Temps plein | 22 |
| 22 | M | Non | 38 |
| 23 | F | Non | 35 |
| 24 | M | Temps plein | 35 |
| 25 | F | Temps plein | 27 |
| 26 | M | Temps plein | 24 |
| 27 | F | Temps plein | 37 |
| 28 | F | Non | 38 |
| 29 | F | Temps plein | 29 |
| 30 | F | Temps plein | 32 |
| 31 | F | Non | 20 |
| 32 | M | Partiel | 24 |
| 33 | M | Non | 26 |
| 34 | M | Partiel | 31 |
| 35 | M | Temps plein | 32 |
| 36 | F | Partiel | 24 |

F H

TP TPa Pas

F\_TP  
F\_TPa 1 F\_Pas 11 H\_TP 1 T\_TPa H\_Pas

TP TPa Pas

F 13 H

1. Dire quel est le type de chaque variable

sexe : qualitative, nominale, dichotomique emploi: qualitative, ordinale age : quantitative, continue (discrete)

1. Conjecturer sur la question de recherche qui a poussé à la collecte de ces données
2. Dresser un tableau croisé liant les deux variables de sexe et d’emploi, préciser les informations importantes du tableau (nom des colonnes et des lignes, effectifs, et effectifs marginaux)
   * Non partiel Plein Total H 16 F 20 T 1 11 9 15 36

* Non partiel Plein Total
* H 15\*11/35 15 F 20 T 11 9 15 35

Supposon qu’il n’y a pas d’association entre le sexe et l’emploi, comment devrait être ce tableau? tableau de **fréquence attendues**

Non partiel Plein Total

H A B C 15 F D E F 20 T 11 9 15 35

A? 11/35 \* 15

parmi 35 personnes 11 ne travaillent pas parmi 15 personnes, combien ne travaillent pas (A)?

A = 15\*11/35 = 4.7

B = 15\*9/35 = 3.9

c = 15\*15/35 = 6.4

E = 20\*9/35 = 5.1

D = 20\*11/35 = 6.3

F = 20\*15/35 = 8.6

**Tableau attendue**

Non partiel Plein Total

H 4.7 3.9 6.4 15 F 6.3 5.1 8.6 20 T 11 9 15 35

Pourcentage

Non partiel Plein Total

H 31.3 26.0 42.7 100%  
F 31.3 26.0 42.7 100% T 11 9 15 35

Voyons maintenant les résultats obtenus de notre enquête

Chi-carré n’est que la distance entre les deux tableaux

fréquence observée (fo) frequence attendue (fa) fo-fa (f0-fa)^2 (f0-fa)^2/fa

3 4.7 3-4.7 (3-4.7)^2 (3-4.7)^2/4.7 5 3.9 5-3.9 (5-3.9)^2 (5-3.9)^2/3.9 7 6.4 7-6.4 (7-6.4)^2 (7-6.4)^2/6.4 8 6.3 8-6.3 (8-6.3)^2 (8-6.3)^2/6.3 4 5.1 4-5.1 (4-5.1)^2 (4-5.1)^2/5.1 8 8.6 8-8.6 (8-8.6)^2 (8-8.6)^2/8.6

chi-carré = (3-4.7)^2/4.7 + (5-3.9)^2/3.9 + (7-6.4)^2/6.4 + (8-6.3)^2/6.3 + (4-5.1)^2/5.1 + (8-8.6)^2/8.6

**chi-carré = 1.71**

1. Hypothèse (H0): Il n’y a pas d’association entre sexe et emploi
2. Niveau de signification: 0.05
3. degré de liberté: (2-1)\*(3-1) = 2 4 valeur MINIMALE du chi-carré: 5.99
4. Décision

Notre chi-carré vaut **1.71** qui est plus pétit que le chi-carré que nous avons lu, donc nous NE POUVONS PAS REJETER l’hypothèse. En conclusion, il n’existe pas de relation entre le sexe et l’emploi dans ma POPULATION.

1. Quelle conclusion tirez-vous sur ces deux variables? Quel est le problème avec cette conclusion?
2. Présenter un tableau de pourcentage marginal et commentez à nouveau.

* Non partiel Plein Total
* H 20% 33% 47%  
  F 40% 20% 40%  
  T 11 9 15 35

1. Quel est le type de relation que vous observez entre ces deux variables?
2. Diriez-vous que cette relation est forte, modérée, faible ou nulle?
3. Calculer le Chi-deux
4. Déterminer le degré de liberté
5. Quelle conclusion tirez-vous sur votre confiance au résultat?
6. Si vous devez étudier la relation entre l’âge et l’emploi, comment procéderiez-vous?

## 2. Test de chi-deux avec R

1. Entrer les données précédentes dans R

sexe <- c("M", "F", "F", "F", "M", "F", "M", "M", "F", "F", "M", "M", "M", "F", "F", "F", "F", "F", "M", "F", "M", "M", "F", "M", "F", "M", "F", "F", "F", "F", "F", "M", "M", "M", "M", "F")  
  
sexe

## [1] "M" "F" "F" "F" "M" "F" "M" "M" "F" "F" "M" "M" "M" "F" "F" "F" "F" "F" "M"  
## [20] "F" "M" "M" "F" "M" "F" "M" "F" "F" "F" "F" "F" "M" "M" "M" "M" "F"

sexe[5]

## [1] "M"

emploi <- c("Temps plein", "Non", "Partiel", "Non", "Partiel", "Temps plein", "Temps plein", "Temps plein", "Temps plein", "Non", "Partiel", "Partiel", "Non", "Partiel", "Temps plein", "Temps plein", "Non", "Non", "-", "Partiel", "Temps plein", "Non", "Non", "Temps plein", "Temps plein", "Temps plein", "Temps plein", "Non", "Temps plein", "Temps plein", "Non", "Partiel", "Non", "Partiel", "Temps plein", "Partiel")  
  
age <- c(27, 34, 22, 38, 30, 27, 33, 39, 36, 25, 31, 28, 26, 21, 29, 34, 20, 32, 27, 33, 22, 38, 35, 35, 27, 24, 37, 38, 29, 32, 20, 24, 26, 31, 32, 24)  
  
  
base\_donnee <- data.frame(sexe, emploi, age)  
  
write.csv(base\_donnee, file = "Emploi\_femme.csv")

1. Présenter le tableau de fréquence de la variable sexe, et commentez

library(tidyverse)

## ── Attaching core tidyverse packages ──────────────────────── tidyverse 2.0.0 ──  
## ✔ dplyr 1.1.0 ✔ readr 2.1.4  
## ✔ forcats 1.0.0 ✔ stringr 1.5.0  
## ✔ ggplot2 3.4.1 ✔ tibble 3.2.0  
## ✔ lubridate 1.9.2 ✔ tidyr 1.3.0  
## ✔ purrr 1.0.1   
## ── Conflicts ────────────────────────────────────────── tidyverse\_conflicts() ──  
## ✖ dplyr::filter() masks stats::filter()  
## ✖ dplyr::lag() masks stats::lag()  
## ℹ Use the ]8;;http://conflicted.r-lib.org/conflicted package]8;; to force all conflicts to become errors

library(summarytools)

##   
## Attaching package: 'summarytools'  
##   
## The following object is masked from 'package:tibble':  
##   
## view

freq(base\_donnee$sexe)

## Frequencies   
## base\_donnee$sexe   
## Type: Character   
##   
## Freq % Valid % Valid Cum. % Total % Total Cum.  
## ----------- ------ --------- -------------- --------- --------------  
## F 20 55.56 55.56 55.56 55.56  
## M 16 44.44 100.00 44.44 100.00  
## <NA> 0 0.00 100.00  
## Total 36 100.00 100.00 100.00 100.00

le mode de la variable sexe est F(Femme)

1. Présenter le tableau de fréquence de la variable emploi, et commentez

freq(base\_donnee$emploi)

## Frequencies   
## base\_donnee$emploi   
## Type: Character   
##   
## Freq % Valid % Valid Cum. % Total % Total Cum.  
## ----------------- ------ --------- -------------- --------- --------------  
## - 1 2.78 2.78 2.78 2.78  
## Non 11 30.56 33.33 30.56 33.33  
## Partiel 9 25.00 58.33 25.00 58.33  
## Temps plein 15 41.67 100.00 41.67 100.00  
## <NA> 0 0.00 100.00  
## Total 36 100.00 100.00 100.00 100.00

Le mode est Temps plein

1. Quelle est la relation entre le sexe et l’emploi?

ctable(base\_donnee$emploi, base\_donnee$sexe, "c")

## Cross-Tabulation, Column Proportions   
## emploi \* sexe   
## Data Frame: base\_donnee   
##   
## ------------- ------ ------------- ------------- -------------  
## sexe F M Total  
## emploi   
## - 0 ( 0.0%) 1 ( 6.2%) 1 ( 2.8%)  
## Non 8 ( 40.0%) 3 ( 18.8%) 11 ( 30.6%)  
## Partiel 4 ( 20.0%) 5 ( 31.2%) 9 ( 25.0%)  
## Temps plein 8 ( 40.0%) 7 ( 43.8%) 15 ( 41.7%)  
## Total 20 (100.0%) 16 (100.0%) 36 (100.0%)  
## ------------- ------ ------------- ------------- -------------

Je peux aussi alors lui demander de me calculer mon chi-carré

ctable(base\_donnee$emploi, base\_donnee$sexe, "no", chisq = TRUE)

## Warning in chisq.test(freq\_table\_min): Chi-squared approximation may be  
## incorrect

## Cross-Tabulation   
## emploi \* sexe   
## Data Frame: base\_donnee   
##   
##   
## ------------- ------ ---- ---- -------  
## sexe F M Total  
## emploi   
## - 0 1 1  
## Non 8 3 11  
## Partiel 4 5 9  
## Temps plein 8 7 15  
## Total 20 16 36  
## ------------- ------ ---- ---- -------  
##   
## ----------------------------  
## Chi.squared df p.value   
## ------------- ---- ---------  
## 3.0436 3 0.3849   
## ----------------------------

1. Cette relation est-elle statistiquement significative?
2. Créer une nouvelle variable d’emploi que vous notez **emploi2** qui prend les valeurs suivantes:

* Si emploi prend les valeurs “non”” ou “temps partiel”“, emploi2 prend la valeur”ne travaille pas”
* Si emploi prend la valeur “Temps plein”, emploi2 prend la valeur “Travaille”

# Mausvaise maniere  
  
base\_donnee2 <-  
 base\_donnee %>%   
 mutate(emploi2 = case\_when(  
 emploi == "Non" | emploi == "Partiel" ~ "Ne travaille pas",  
 emploi == "Temps plein" ~ "Travaille"  
 ))  
  
# Bonne manière  
  
base\_donnee <-  
 base\_donnee %>%   
 mutate(emploi2 = case\_when(  
 emploi == "Non" | emploi == "Partiel" ~ "Ne travaille pas",  
 emploi == "Temps plein" ~ "Travaille"  
 ))

Règles générale

* avec mutate, ne créez pas une nouvelle base de données
* Avec summarise, créez toujours une nouvelle base de données

resume\_age <-  
 base\_donnee %>%   
 summarise(moyenne = mean(age, na.rm = TRUE))

1. Créer aussi la nouvelle variable **groupe\_age** qui prends deux valeurs:

* Moins de 30 ans
* 30 ans et plus

1. Étudiez à nouveau la relation entre le sexe et l’emploi, cette fois-ci séparément pour les moins de 30 ans et les 30 ans et plus

Filter et mutate

Sélectionner - Filtrer (filter) - Selectionner (select)

Automatiquement,vous de nouvelles bases de données

# base avec sexe te emploi seulement  
  
sexe\_emploi <-  
 base\_donnee %>%   
 select(sexe, emploi)  
  
## base avec sexe et emploi mais chez les gens de 30 ans et moins  
  
sexe\_emploi\_30ans <-  
 base\_donnee %>%   
 select(sexe, emploi) %>%   
 filter(age < 30)