TP statistiques bivariées

Eric Marcon

23 février 2024

Comptages

Données

TP statistiques bivariées

Eric Marcon

Comptages

Variables continues

Enquête de vie 2003 de l'INSEE

```
library("questionr")
data(hdv2003)
```

Tableaux de contingence

TP statistiques bivariées

Eric Marcon

Comptages

Variables continues

Tableau croisé de comptage.

```
(tab_x <- table(hdv2003$sexe, hdv2003$cuisine))
```

```
##
##          Non Oui
##   Homme 629 270
##   Femme 490 611
```

Test du chi^2

Test de l'indépendance des lignes et des colonnes.

Hypothèse nulle : la fréquence relative de chaque cellule du tableau est le produit des fréquences marginales.

```
n <- sum(tab_x)
sexe_f <- rowSums(tab_x) / n
cuisine_f <- colSums(tab_x) / n
outer(sexe_f, cuisine_f, `*`) * n
```

```
##             Non      Oui
## Homme 502.9905 396.0095
## Femme 616.0095 484.9905
```

La somme des carrés des écarts des effectifs divisés par la valeur attendue suit une loi du χ_2 à $(I-1) \times (J-1)$ degrés de liberté (I et J sont les nombres de lignes et colonnes)

Test du chi^2


```
chisq.test(tab_x)
```

```
##
##  Pearson's Chi-squared test with Yates'
##  continuity correction
##
## data:  tab_x
## X-squared = 129.15, df = 1, p-value <
## 2.2e-16
```

Les écarts sont significatifs avec une p-value proche de 0.

Graphique

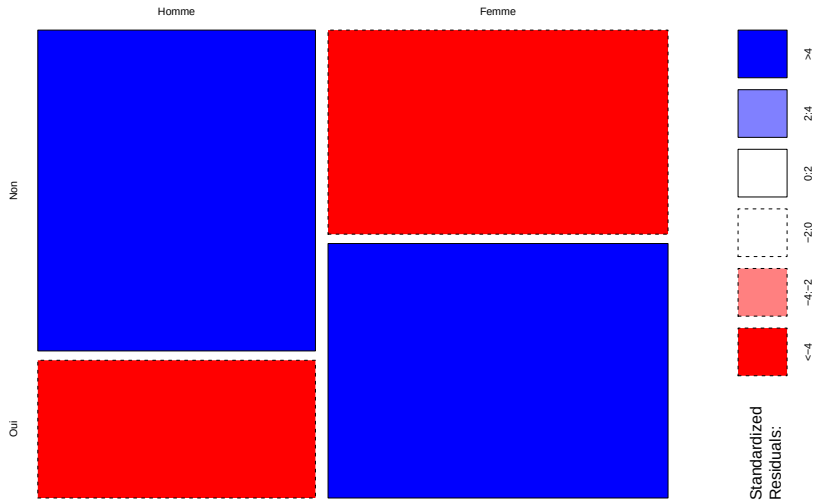
TP statistiques bivariées

Eric Marcon

Comptages

Variables continues

```
mosaicplot(tab_x, shade = TRUE, main = "")
```



L'argument `shade = TRUE` affiche les résidus du test qui suivent approximativement une loi normale centrée réduite (la valeur critique 2 correspond à 95% de confiance).

Variables continues

Covariance

TP statistiques bivariées

Eric Marcon

Comptages

Variables continues

La covariance entre X et Y, deux variables aléatoires, est

$$\mathrm{Cov}(X,Y) = \mathbb{E}[(X - \mathbb{E}(X))(Y - \mathbb{E}(Y))]$$

donc

$$\mathrm{var}(X) = \mathrm{Cov}(X,X)$$

Empiriquement :

$$\hat{\mathrm{Cov}}(X,Y) = \frac{1}{n-1}\sum_i (x_i - \bar{x})(y_i - \bar{y})$$

Estimation

Données Ventoux.

```
read_csv2("data/Inv_GEEFT_Ventoux_09-2020.csv") |>
  rename(
    espece = Espèce,
    diametre = `Diamètre (cm)`,
    hauteur = `Hauteur réelle (m)`
  ) -> ventoux
```

La hauteur des arbres covarie positivement avec le diamètre.

```
with(ventoux, cov(hauteur, diametre))
```

```
## [1] 75.31186
```

Corrélation de Pearson

TP statistiques bivariées
Eric Marcon
Comptages
Variables continues

Pour simplifier l'interprétation, on normalise la covariance par le produit des écarts-types :

$$\mathrm{Cor}(X,Y) = \frac{\mathrm{Cov}(X,Y)}{\sqrt{(\mathrm{var}(X)\mathrm{var}(Y))}}$$

Donc $\mathrm{Cor}(X,X) = 1$ et $\mathrm{Cor}(X,-X) = -1$.

La corrélation est comprise entre -1 et 1.

```
with(ventoux, cor(hauteur, diametre))
```

```
## [1] 0.8427001
```

Les données sont très corrélées (le test viendra plus tard).

Corrélation de Spearman

Les valeurs des données sont remplacées par leurs rangs.

```
with(ventoux, cor(hauteur, diametre, method = "spearman"))
```

```
## [1] 0.8490534
```

Remarquer la proximité des valeurs.

AgroParisTech

TP statistiques bivariées

Eric Marcon

Comptages

Variables continues