

TP statistiques bivariées

Eric Marcon

17 février 2024

Comptages

Enquête de vie 2003 de l'INSEE

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

Tableau croisé de comptage.

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

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

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
```

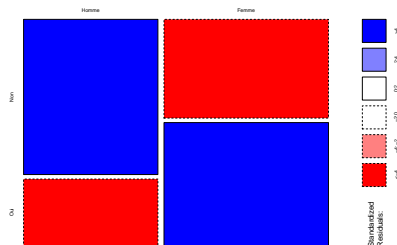
La somme des carrés des écarts des effectifs divisés par la valeur attendue suit une loi du χ^2 à $(I - 1) \times (J - 1)$ degrés de liberté (I et J sont les nombres de lignes et colonnes)

```
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.

```
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

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

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

donc

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

Empiriquement :

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

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
```

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

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

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

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).

Les valeurs des données sont remplacées par leurs rangs.

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

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

Remarquer la proximité des valeurs.

TP
statistiques
bivariées

Eric Marcon

Comptages

Variables
continues