HLMA408: Traitement des données

Estimations et tests: cas du cytomégalovirus

Joseph Salmon

http://josephsalmon.eu

Université de Montpellier

Sommaire

Sommaire

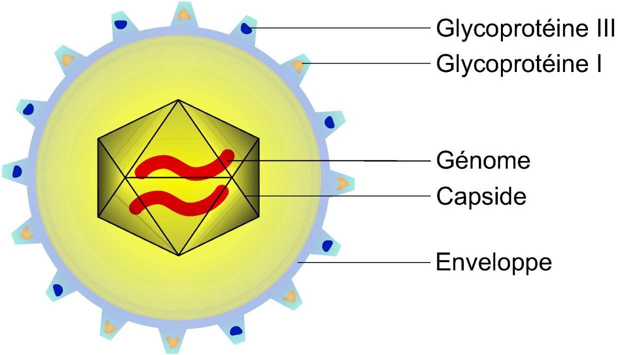
Étude du CytoMégaloVirus (CMV) humain[1]

- Cytomégalovirus: famille des **herpesvirus**, comprenant le virus de l'herpès simplex, le virus d'Epstein-Barr, virus varicelle-zona,...
- Caractéristiques du virus:
  - capacité à produire des **infections latentes et persistantes**
  - dangereux pour **les fœtus** et les personnes avec **faibles défenses immunitaires**

[1] adapté de D. Nolan and T. P. Speed. *Stat labs: mathematical statistics through applications*. Springer Science & Business Media, 2001

Structure de virus CMV[2]

- **génome**
- capside
- enveloppe recouverte de glycoprotéines
- ...



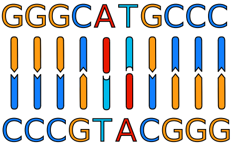
[2] source: https://en.wikipedia.org/wiki/Cytomegalovirus

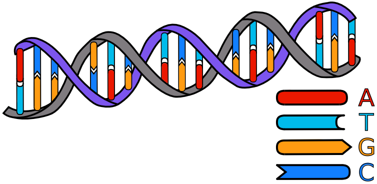
ADN et origine de la réplication

ADN: double hélice composée de nucléotides, chacun caractérisé par sa base azotée (Adénine, Cytosine, Guanine, Thymine); on raisonne souvent en paires de bases (A $\leftrightarrow$ T; G $\leftrightarrow$ C) par complémentarité entre les deux brins

Origine de la réplication : les **palindromes**[3] (complémentaires[4]) semblent importants biologiquement

Exemple :





Enjeu : découvrir les zones de l'ADN où le nombre de palindromes est **anormalement** élevé

---

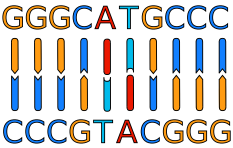
[3] *e.g.*, *In girum imus nocte et consumimur igni*

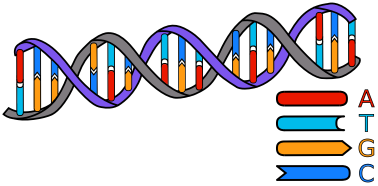
[4] https://fr.wikipedia.org/wiki/Séquence_palindromique

ADN et origine de la réplication

ADN: double hélice composée de nucléotides, chacun caractérisé par sa base azotée (Adénine, Cytosine, Guanine, Thymine); on raisonne souvent en paires de bases (A $\leftrightarrow$ T; G $\leftrightarrow$ C) par complémentarité entre les deux brins

Origine de la réplication : les **palindromes**[3] (complémentaires[4]) semblent importants biologiquement

Exemple : 



Enjeu : découvrir les zones de l'ADN où le nombre de palindromes est **anormalement** élevé

---

[3] *e.g.*, *In girum imus nocte et consumimur igni*

[4] https://fr.wikipedia.org/wiki/Séquence_palindromique

Données

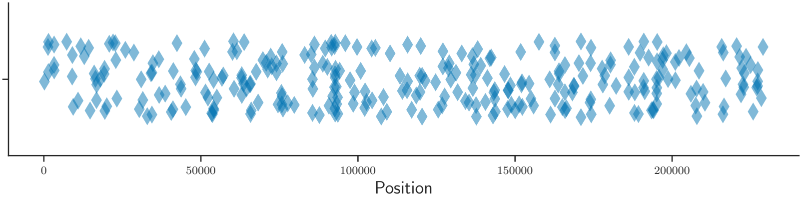
Données du génome de CMV [5] : répertorie les positions de palindromes dans le génome de CMV

$$n_{\rm bp} = 229\ 354 \quad \text{: nombre de nucléotides (paires de bases = bp)}$$
$$n = 296 \quad \text{: nombre de palindromes de longueur} \geq 10 \text{ bp}$$

Rem. : les palindromes trop courts sont exclus

[5] http://www.stat.berkeley.edu/users/statlabs/data/hcmv.data

Position des palindromes



TO DO: voir notebook `EstimationTest.ipynb`

Sommaire

Des palindromes positionnés au hasard ?

Comment modélise-t-on des palindromes positionnés au hasard sur le génome, sans région privilégiée ?

Sans région privilégiée : homogénéité

Le modèle probabiliste : processus de Poisson homogène

Modèle aléatoire

- Considérer **tous les** échantillons **possibles** (même si en pratique un seul est observé)

Modèle aléatoire

- Considérer **tous les** échantillons **possibles**
(même si en pratique un seul est observé)

- Associer un poids à chacun des échantillons possibles, qui représente sa **probabilité**
Exemple : poids égaux (échantillons **équiprobables**) valant $\frac{1}{n}$ (n : nombre d'échantillons)

Modèle aléatoire

- Considérer **tous les** échantillons **possibles** (même si en pratique un seul est observé)
- Associer un poids à chacun des échantillons possibles, qui représente sa **probabilité**
  Exemple : poids égaux (échantillons **équiprobables**) valant $\frac{1}{n}$ (n : nombre d'échantillons)
- Avec le **calcul des probabilités**, on peut en déduire des propriétés intéressantes.
  Exemple : intervalle de confiance

Modèle aléatoire

- Considérer **tous les** échantillons **possibles**
(même si en pratique un seul est observé)

- Associer un poids à chacun des échantillons possibles, qui représente sa **probabilité**
Exemple : poids égaux (échantillons **équiprobables**) valant $\frac{1}{n}$ (n : nombre d'échantillons)

- Avec le **calcul des probabilités**, on peut en déduire des propriétés intéressantes.
Exemple : intervalle de confiance

Modélisation aléatoire

Modèle aléatoire : composé par

- l'univers des possibles : description des événements
- les probabilités : description de la probabilité des événements
- . . . (paramètres potentiels)

(6) souvent notés par des lettres grecques

Modélisation aléatoire

Modèle aléatoire : composé par

- l'univers des possibles : description des événements
- les probabilités : description de la probabilité des événements
- ... (paramètres potentiels)

Rem. : dans le cas du modèle "échantillonnage aléatoire simple", le modèle est complètement déterminé (pas de paramètre)

(6) souvent notés par des lettres grecques

Modélisation aléatoire

Modèle aléatoire : composé par

- l'univers des possibles : description des événements
- les probabilités : description de la probabilité des événements
- ... (paramètres potentiels)

Rem. : dans le cas du modèle "échantillonnage aléatoire simple", le modèle est complètement déterminé (pas de paramètre)

Certains modèles (et donc les probabilités d'événements) dépendent de **paramètres inconnus**[6]

[6] souvent notés par des lettres grecques

Modélisation aléatoire

Modèle aléatoire : composé par

- l'univers des possibles : description des événements
- les probabilités : description de la probabilité des événements
- ... (paramètres potentiels)

Rem. : dans le cas du modèle "échantillonnage aléatoire simple", le modèle est complètement déterminé (pas de paramètre)

Certains modèles (et donc les probabilités d'événements) dépendent de **paramètres inconnus**[6]

Exemple : pour les palindromes sur le génome, le modèle doit décrire la probabilité de leur apparition

[6] souvent notés par des lettres grecques

Processus de Poisson homogène

Processus de Poisson homogène : utilisé pour modéliser des palindromes répartis totalement au hasard, de façon homogène parmi $\approx 200\,000$ positions possibles

- Modèle basique: génome modélisé comme une demi-droite, les palindromes sont des points sur cet ensemble
- Univers des possibles: ensemble des façons de placer des points sur cette demi-droite (grand, $\approx$ infini)
- Écart à ce modèle dans une zone donnée
  $\iff$ palindromes anormalement fréquents dans la zone

Processus de Poisson homogène (2)

Processus de Poisson homogène : naturel pour construire un modèle probabiliste de points distribués au hasard dans l'espace ou dans un intervalle de temps

Exemple : modèle courant pour l'arrivée de phénomènes au cours du temps (passage de bus à un arrêt, nombre d'appels à un serveur téléphonique, etc.)[7]

Propriétés :

- les nombres de points apparaissant dans deux régions disjointes sont **indépendants** (pas de **mémoire**)

[7] *cf.* http://jakevdp.github.io/blog/2018/09/13/waiting-time-paradox/ pour une analyse fine

Processus de Poisson homogène (2)

Processus de Poisson homogène : naturel pour construire un modèle probabiliste de points distribués au hasard dans l'espace ou dans un intervalle de temps

Exemple : modèle courant pour l'arrivée de phénomènes au cours du temps (passage de bus à un arrêt, nombre d'appels à un serveur téléphonique, etc.)[7]

Propriétés :

- les nombres de points apparaissant dans deux régions disjointes sont **indépendants** (pas de **mémoire**)
- taux λ: taux avec lequel les points apparaissent dans des régions de même taille (**homogénéité**)

[7] *cf.* http://jakevdp.github.io/blog/2018/09/13/waiting-time-paradox/ pour une analyse fine

Processus de Poisson homogène (2)

Processus de Poisson homogène : naturel pour construire un modèle probabiliste de points distribués au hasard dans l'espace ou dans un intervalle de temps

Exemple : modèle courant pour l'arrivée de phénomènes au cours du temps (passage de bus à un arrêt, nombre d'appels à un serveur téléphonique, etc.)[7]

Propriétés :

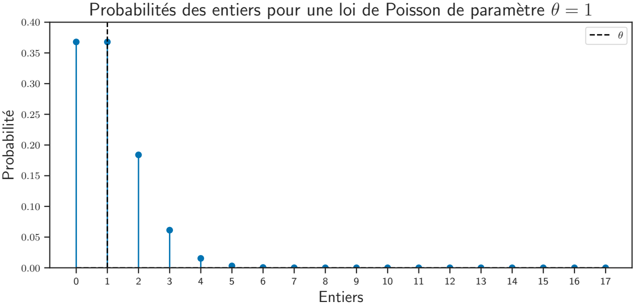
- les nombres de points apparaissant dans deux régions disjointes sont **indépendants** (pas de **mémoire**)
- taux λ: taux avec lequel les points apparaissent dans des régions de même taille (**homogénéité**)

[7] *cf.* `http://jakevdp.github.io/blog/2018/09/13/waiting-time-paradox/` pour une analyse fine

Loi de Poisson (distribution discrète)

Rappel : une variable aléatoire N suit une loi de Poisson de paramètre $\theta > 0$, ce que l'on note $N \sim \mathcal{P}(\theta)$, si

$$\boxed{\mathbb{P}(N = k) = e^{-\theta} \frac{\theta^k}{k!}} \quad \text{pour tout entier } k \in \mathbb{N}$$

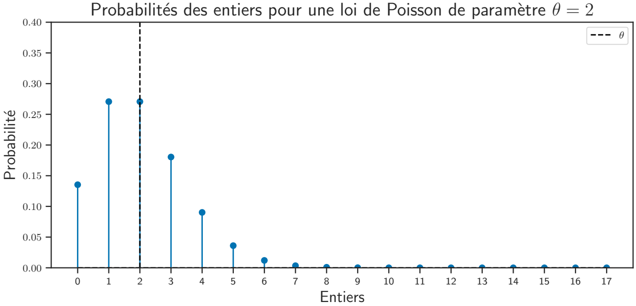



```
from scipy.stats import poisson
```

Loi de Poisson (distribution discrète)

Rappel : une variable aléatoire N suit une loi de Poisson de paramètre $\theta > 0$, ce que l'on note $N \sim \mathcal{P}(\theta)$, si

$$\boxed{\mathbb{P}(N = k) = e^{-\theta}\frac{\theta^k}{k!}} \quad \text{pour tout entier } k \in \mathbb{N}$$

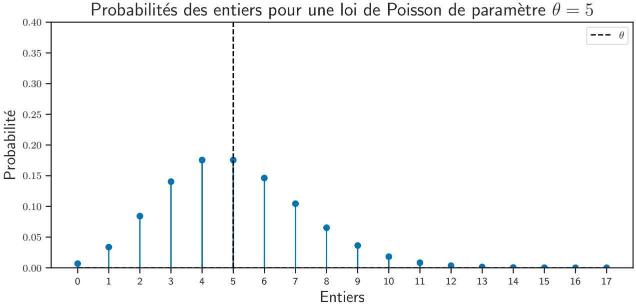



```
from scipy.stats import poisson
```

Loi de Poisson (distribution discrète)

Rappel : une variable aléatoire N suit une loi de Poisson de paramètre $\theta > 0$, ce que l'on note $N \sim \mathcal{P}(\theta)$, si

$$\boxed{\mathbb{P}(N = k) = e^{-\theta}\frac{\theta^k}{k!}} \quad \text{pour tout entier } k \in \mathbb{N}$$

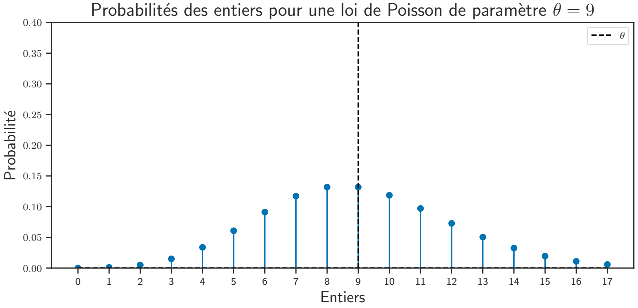



```python
from scipy.stats import poisson
```

Loi de Poisson (distribution discrète)

Rappel : une variable aléatoire N suit une loi de Poisson de paramètre $\theta > 0$, ce que l'on note $N \sim \mathcal{P}(\theta)$, si

$$\boxed{\mathbb{P}(N = k) = e^{-\theta}\frac{\theta^k}{k!}} \qquad \text{pour tout entier } k \in \mathbb{N}$$




```python
from scipy.stats import poisson
```

Processus de Poisson

- λ : paramètre d'**intensité**, taux d'apparition/d'occurrence du phénomène (*e.g.*, les palindromes dans notre exemple)
- L : longueur de l'intervalle

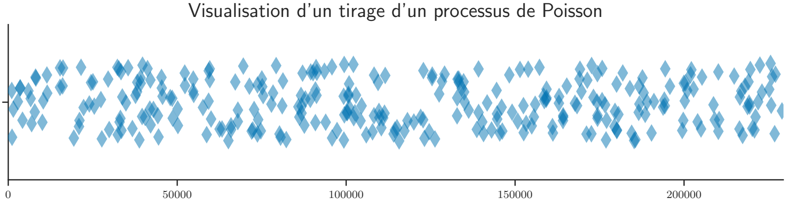
Processus de Poisson : le nombre de points (*i.e.*, la variable N) tombant dans un intervalle de longueur L suit une loi de Poisson de paramètre $\theta = \lambda L$

$$N \sim \mathcal{P}(\lambda L)$$

Interprétation : en espérance, sur un intervalle de taille L, il y a $\lambda \times L$ occurrences

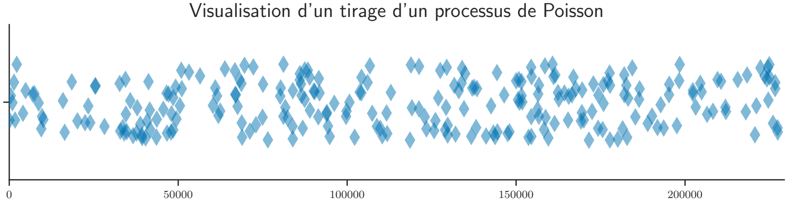
Unité de λ : homogène à l'inverse d'une longueur (taux)

Quatre réalisations d'un processus de Poisson



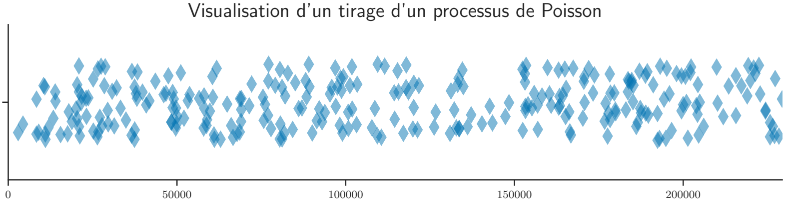
Rem. : paramètres utilisés "similaires" à ceux des données CMV

Quatre réalisations d’un processus de Poisson



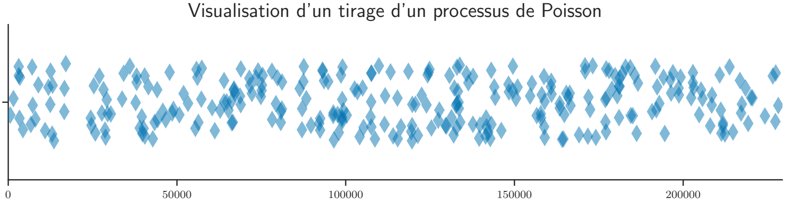
Rem. : paramètres utilisés “similaires” à ceux des données CMV

Quatre réalisations d'un processus de Poisson



Rem. : paramètres utilisés "similaires" à ceux des données CMV

Quatre réalisations d'un processus de Poisson



<u>Rem.</u> : paramètres utilisés “similaires” à ceux des données CMV

Estimation du taux λ

⚠ : λ généralement inconnu, doit être estimé!

Méthodes populaires d'estimation :

1. **méthode des moments** : utilise la loi des grands nombres, approchant l'espérance par la moyenne

2. **méthode du maximum de vraisemblance** : consiste à choisir parmi tous les modèles de Poisson celui dont le paramètre est le plus raisonnable ou **vraisemblable**

Estimation de λ

Ici, on choisira comme estimateur du taux d'occurrence λ :

$$\hat{\lambda} = \frac{\text{nombre de palindromes observés}}{\text{longueur de l'ADN dans l'unité choisie}}$$

Application numérique : $\hat{\lambda} = \frac{296}{229354} \approx 0.0013$

Rem. : on admet que dans le cas présent les méthodes des moments et du maximum de vraisemblance coïncident

Sommaire

Problème général

"*Ce qui est simple est toujours faux. Ce qui ne l'est pas est inutilisable.*"[8]
(: *"All models are wrong but some are useful"*)[9]

Hypothèse de modélisation : les observations sont des réalisations de variables aléatoires indépendantes, de loi connue (*e.g.*, Poisson)

Modèle probabiliste: jamais exact, mais souvent décrit suffisamment bien le caractère aléatoire du phénomène

Néanmoins, on peut chercher à vérifier cette hypothèse.

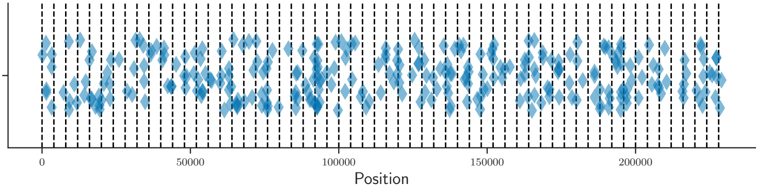
---

[8] P. Valéry. *Mauvaises pensées et autres*. Gallimard, 1942.

[9] G. E. P. Box. "Robustness in the strategy of scientific model building". In: *Robustness in statistics*. Elsevier, 1979, pp. 201–236.

Test d'adéquation et processus de Poisson

- découper l'ADN de CMV en 57 régions qui ne se recouvrent pas, de longueur $L = 4000$ bp[10]
- compter l'occurrence de palindromes par région:



Occurrences :

7 1 5 3 8 6 1 4 5 3 6 2 5 8 2 9 6 4 9 4 1 7 7 14 4 4 4 3 5
5 3 6 5 3 9 9 4 5 6 1 7 6 7 5 3 4 4 8 11 5 3 6 3 1 4 8 6

[10] attention la dernière région ne fait pas la même taille, mais on passera cette difficulté sous silence

Présentation synthétique des données

Comptage de palindrome	Effectifs
0 – 2	7
3	8
4	10
5	9
6	8
7	5
8	4
9 et plus	6
Total	57

Ainsi, il y a 10 régions de notre découpage de l'ADN dans lesquelles on observe exactement 4 palindromes. . .

Estimation des comptages attendus

Ici, l'unité de longueur est 4000 bp. Le nombre moyen de palindromes dans les 57 régions est

$$\hat{\theta} = \frac{1 \times 7 + 3 \times 8 + \cdots + 9 \times 6}{57} \simeq 5.16 \ (\simeq 4000\hat{\lambda})$$

Pour une loi $\mathcal{P}(5.16)$, les effectifs attendus sur 57 tirages sont

Comptage de palindrome	Eff. observés	Eff. théoriques
0 – 2	7	6.4
3	8	7.5
4	10	9.7
5	9	10.0
6	8	8.6
7	5	6.3
8	4	4.1
9 et plus	6	4.5
Total	57	57

Qu'est-ce qu'un effectif attendu ?

Si N suit une loi de Poisson de paramètre $\hat{\theta}$, $\mathbb{P}(N = 3) = \frac{e^{-\hat{\theta}}\hat{\theta}^3}{3!}$

Si on réalise 57 copies indépendantes de N, on s'attend à voir $N = 3$ se réaliser un nombre de fois égal à

$$57 \times \mathbb{P}(N = 3) = 57\frac{e^{-\hat{\theta}}\hat{\theta}^3}{3!}$$

Avec une estimation $\hat{\theta} \approx 5.16$, on obtient un effectif attendu de:

$$57 \times e^{-5.16}\frac{(5.16)^3}{3!} \approx 7.5$$

etc.

TO DO: *cf.* notebook associé pour les autres calculs

La différence entre attendu et observé est-elle importante ?

Comptage de palindrome	Eff. observés	Eff. attendus
0 – 2	7	6.4
3	8	7.5
4	10	9.7
5	9	10.0
6	8	8.6
7	5	6.3
8	4	4.1
9 et plus	6	4.5
Total	57	57

$\longrightarrow$ pour mesurer la différence entre les deux colonnes, on introduit **une statistique de test**

Statistique de test

Application numérique :

$$\frac{(7-6.4)^2}{6.4} + \frac{(8-7.5)^2}{7.5} + \frac{(10-9.7)^2}{9.7} + \frac{(9-10.0)^2}{10.0} + \frac{(8-8.6)^2}{8.6} + \frac{(5-6.3)^2}{6.3} + \frac{(4-4.1)^2}{4.1} + \frac{(6-4.5)^2}{4.5} \approx 1.02$$

Si le modèle aléatoire est vrai, cette statistique est distribuée suivant une loi du khi-deux[11] à 6 degrés de liberté, notée $\chi^2(6)$

Cette valeur est-telle exceptionnellement grande pour une variable aléatoire qui suit une telle distribution ?

$$\mathbb{P}(\chi^2(6) \geq 1.02) \approx 0.985$$

[11] https://fr.wikipedia.org/wiki/Loi_du_χ^2

Statistique de test

Application numérique :

$$\frac{(7-6.4)^2}{6.4} + \frac{(8-7.5)^2}{7.5} + \frac{(10-9.7)^2}{9.7} + \frac{(9-10.0)^2}{10.0} + \frac{(8-8.6)^2}{8.6} + \frac{(5-6.3)^2}{6.3} + \frac{(4-4.1)^2}{4.1} + \frac{(6-4.5)^2}{4.5} \approx 1.02$$

Si le modèle aléatoire est vrai, cette statistique est distribuée suivant une loi du khi-deux[11] à 6 degrés de liberté, notée $\chi^2(6)$

Cette valeur est-telle exceptionnellement grande pour une variable aléatoire qui suit une telle distribution ?

$$\mathbb{P}(\chi^2(6) \geq 1.02) \approx 0.985$$

Pour notre jeu de données: **valeur non exceptionnelle**

[11] https://fr.wikipedia.org/wiki/Loi_du_χ^2

Statistique de test

Application numérique :

$$\frac{(7-6.4)^2}{6.4} + \frac{(8-7.5)^2}{7.5} + \frac{(10-9.7)^2}{9.7} + \frac{(9-10.0)^2}{10.0} + \frac{(8-8.6)^2}{8.6} + \frac{(5-6.3)^2}{6.3} + \frac{(4-4.1)^2}{4.1} + \frac{(6-4.5)^2}{4.5} \approx 1.02$$

Si le modèle aléatoire est vrai, cette statistique est distribuée suivant une loi du khi-deux[11] à 6 degrés de liberté, notée $\chi^2(6)$

Cette valeur est-telle exceptionnellement grande pour une variable aléatoire qui suit une telle distribution ?

$$\mathbb{P}\big(\chi^2(6) \geq 1.02\big) \approx 0.985$$

Pour notre jeu de données: **valeur non exceptionnelle**

[11] https://fr.wikipedia.org/wiki/Loi_du_χ^2

Qu'avons nous fait ?

1. Calculer la valeur observée d'une **statistique de test**, dont la formule générale est :

$$\sum_{\text{différentes classes}} \frac{(\text{Eff. observé} - \text{Eff. espéré dans la classe})^2}{\text{Eff. espéré dans la classe}}$$

2. Constater que
   - si le modèle est mauvais, cette statistique est très grande
   - si le modèle est bon, cette statistique suit une loi du χ^2

Qu'avons nous fait ?

1. Calculer la valeur observée d'une **statistique de test**, dont la formule générale est :

$$\sum_{\text{différentes classes}} \frac{(\text{Eff. observé} - \text{Eff. espéré dans la classe})^2}{\text{Eff. espéré dans la classe}}$$

2. Constater que
   - si le modèle est mauvais, cette statistique est très grande
   - si le modèle est bon, cette statistique suit une loi du χ^2

3. Pour distinguer dans quel cas on est, on regarde si la valeur observée (ici ≈ 1.02) est anormalement grande pour la loi de la statistique quand le modèle est exact

Qu'avons nous fait ?

1. Calculer la valeur observée d'une **statistique de test**, dont la formule générale est :

$$\sum_{\text{différentes classes}} \frac{(\text{Eff. observé} - \text{Eff. espéré dans la classe})^2}{\text{Eff. espéré dans la classe}}$$

2. Constater que
   - si le modèle est mauvais, cette statistique est très grande
   - si le modèle est bon, cette statistique suit une loi du χ^2

3. Pour distinguer dans quel cas on est, on regarde si la valeur observée (ici ≈ 1.02) est anormalement grande pour la loi de la statistique quand le modèle est exact

4. Conclusion :
   - si anormalement grand, on conclut que le modèle est mauvais
   - sinon, on peut conserver le modèle

Qu'avons nous fait ?

1. Calculer la valeur observée d'une **statistique de test**, dont la formule générale est :

$$\sum_{\text{différentes classes}} \frac{(\text{Eff. observé} - \text{Eff. espéré dans la classe})^2}{\text{Eff. espéré dans la classe}}$$

2. Constater que
   - si le modèle est mauvais, cette statistique est très grande
   - si le modèle est bon, cette statistique suit une loi du χ^2

3. Pour distinguer dans quel cas on est, on regarde si la valeur observée (ici ≈ 1.02) est anormalement grande pour la loi de la statistique quand le modèle est exact

4. Conclusion :
   - si anormalement grand, on conclut que le modèle est mauvais
   - sinon, on peut conserver le modèle

Terminologie

On vient de comparer deux hypothèses :

- **hypothèse nulle** : notée $\mathcal{H}_0$, sous laquelle on doit connaître la loi de la statistique de test

vs.

- **hypothèse alternative** : notée $\mathcal{H}_1$, sous laquelle on ne connaît pas en général la loi de la statistique de test

Exemple :
$\mathcal{H}_0$: "les palindromes suivent un processus de Poisson"
$\mathcal{H}_1$: "les palindromes **NE** suivent **PAS** un processus de Poisson"

Rem. : une statistique de test doit avoir un comportement différent sous les deux hypothèses pour distinguer les deux cas

Erreurs d'un test: exemple du test de grossesse

- **Hypothèse nulle** :
$\mathcal{H}_0$:"vous **N'**êtes **PAS** enceinte!"

(taux d'hormone béta-hCG $= 0$)

vs.

- **Hypothèse alternative** :
$\mathcal{H}_1$:"vous êtes enceinte!"

(taux d'hormone béta-hCG > 0)

$\mathcal{H}_0$: hypothèse raisonnable jusqu'à preuve du contraire apportée par les données ("non coupable jusqu'à preuve du contraire")

Réalité :

$\mathcal{H}_1$

$\mathcal{H}_1$

Diagnostic :

Le test permet de rejeter $\mathcal{H}_0$ et il a raison

Réalité :

$\mathcal{H}_0$

Diagnostic :

$\mathcal{H}_0$

Le test ne permet pas de rejeter $\mathcal{H}_0$ et il a raison

Réalité :

$\mathcal{H}_0$

Diagnostic :

$\mathcal{H}_1$

Erreur de type 1 = erreur de première espèce
= **rejet à tord de** $\mathcal{H}_0$

Réalité :

$\mathcal{H}_1$

Diagnostic :

$\mathcal{H}_0$

Erreur de type 2 $=$ erreur de seconde espèce

$=$ non-rejet à tord de $\mathcal{H}_0$

Réalité :

$\mathcal{H}_1$ $\mathcal{H}_0$

Diagnostic :

$\mathcal{H}_1$

$\mathcal{H}_0$

Les quatre situations possibles :
de la réalité au diagnostic

Autres exemples

- Contexte militaire / guerre froide (historique):
  $\mathcal{H}_0$: "il n'y a pas de missile"
  vs.
  $\mathcal{H}_1$: "un missile arrive sur nous!"

- *Le Garçon qui criait au loup*[12] :
  $\mathcal{H}_0$: "le loup n'est pas dans la bergerie"
  vs.
  $\mathcal{H}_1$: "le loup est dans la bergerie!"

- Widget Covid19 sur la page du cours

[12] fable d'Ésope (VIIe-VIe siècle av. J.-C.) connue sous le nom de "Le Berger mauvais plaisant", https://fr.wikisource.org/wiki/Le_Berger_mauvais_plaisant

Deux types d'erreur dans un test

- **Erreur de 1^{re} espèce** : décider en faveur de $\mathcal{H}_1$ alors que $\mathcal{H}_0$ est vraie

- **Erreur de 2^{nde} espèce** : décider en faveur de $\mathcal{H}_0$ alors que $\mathcal{H}_1$ est vraie

Les erreurs d'un test

⚠ : les erreurs des tests sont asymétriques!

- Deux mesures d'erreurs: notées classiquement α et β :

$$\begin{cases} \alpha = \mathbb{P}\left(\text{décider en faveur de } \mathcal{H}_1 \middle| \mathcal{H}_0\right) : \textbf{1}^{\textbf{re}} \textbf{ espèce} \\ \beta = \mathbb{P}\left(\text{décider en faveur de } \mathcal{H}_0 \middle| \mathcal{H}_1\right) : \textbf{2}^{\textbf{nde}} \textbf{ espèce} \end{cases}$$

Plus ces quantités sont petites, mieux c'est !

Les cas extrêmes

$$\begin{cases} \alpha = \mathbb{P}\left(\text{décider en faveur de } \mathcal{H}_1 \middle| \mathcal{H}_0\right) & : \textbf{1}^{\textbf{re}} \textbf{ espèce} \\ \beta = \mathbb{P}\left(\text{décider en faveur de } \mathcal{H}_0 \middle| \mathcal{H}_1\right) & : \textbf{2}^{\textbf{nde}} \textbf{ espèce} \end{cases}$$

- Décider toujours en faveur de $\mathcal{H}_0 \iff \alpha = 0$ et $\beta = 1$

Exemple : diagnostiquer “vous n’êtes pas enceinte ” tout le temps

- Décider toujours en faveur de $\mathcal{H}_1 \iff \alpha = 1$ et $\beta = 0$

Exemple : diagnostiquer “vous êtes enceinte” tout le temps

Conclusion : besoin d’un **compromis**(13)

(13) sur ce thème on verra notamment la courbe ROC en TP

Choix des hypothèses (II)

Exemple : tester l'effet d'un médicament avant mise sur le marché
Hypothèse nulle : **toujours** $\mathcal{H}_0$ "*le médicament est inefficace*"

Conséquences : α contrôle l'erreur suivante

"*J'ai annoncé l'efficacité alors qu'il n'a pas d'effet*"

Théorie classique des tests

Classiquement: l'**utilisateur** fixe α, la probabilité d'erreur de 1^{re} espèce maximale souhaitée (probabilité de rejeter à tord $\mathcal{H}_0$)

Valeurs classiques de α selon le contexte:

- $\alpha = 0.10$
- $\alpha = 0.05$ (le plus commun)
- $\alpha = 0.01$
- etc.

Rappel : on s'intéresse à des α petits

Conséquence : la valeur $1 - \beta$ (la **puissance** du test) est entièrement déterminée et peut être évaluée dans les cas standards

Exercice

TO DO: detailing https://mobile.twitter.com/taaltree/status/1248467731545911296

Sommaire

Test d'adéquation à une loi

Modèle aléatoire de mesures répétées $x_1, \ldots, x_n$, supposées indépendantes et de même loi (*i.i.d.*)

But : tester si l'échantillon $x_1, \ldots, x_n$ provient d'une loi "cible"

- Si le test rejette $\mathcal{H}_0$, alors il est peu vraisemblable que la loi soit celle prescrite par $\mathcal{H}_0$

- Si le test ne rejette pas $\mathcal{H}_0$, alors l'échantillon ne semble pas en contradiction avec l'hypothèse nulle

Statistique de test

1. Faire une table de contingence classe / effectif observé
2. Estimer le (les) paramètre(s) de la famille de loi (si besoin)
3. Calculer les effectifs espérés sous $\mathcal{H}_0$ pour n observations
4. Regrouper les classes pour que les eff. espérés soient ≥ 5 et conserver uniquement cette nouvelle table.
5. Pour cette table avec K classes et n observations, calculer:

$$\chi^2_{obs} := \sum_{\text{différentes classes}} \frac{(\text{Eff. observé} - \text{Eff. espéré dans la classe})^2}{\text{Eff. espéré dans la classe}}$$

$$= \sum_{k=1}^{K} \frac{(\hat{f}_k - f_k)^2}{f_k}$$

$$= \sum_{k=1}^{K} \frac{(\hat{f}_k - np_k)^2}{np_k}; \; p_k : \text{probabilité théorique de la classe } k$$

Pourquoi diviser par les effectifs espérés?

$$\boxed{\chi^2_{obs} := \sum_{\text{différentes classes}} \frac{(\text{Eff. observé} - \text{Eff. espéré dans la classe})^2}{\text{Eff. espéré dans la classe}}}$$

Sans correction au dénominateur on prendrait l'erreur quadratique:

$$\sum_{\text{différentes classes}} (\text{Eff. observé} - \text{Eff. espéré dans la classe})^2$$

⚠: la 2^{nde} statistique donne un poids trop grand aux petites valeurs, *e.g.*, même contribution pour $f_1 = 10, \hat{f}_1 = 5$ et $f_2 = 500, \hat{f}_2 = 505$ **MAIS** représente 50% de variation vs. 1%!

<u>Conclusion</u> : diminuer la contribution des petites variations des grands effectifs en divisant par f_k (effectif théorique)

Comportement de la statistique de test

Théorème[14]

- Si $\mathcal{H}_0$ est vraie, la statistique de test χ^2_{obs} suit une loi du χ^2 à $(K-1-D)$ degrés de liberté, notée $\chi^2(K-1-D)$, où
  - K : nombre de classes (après regroupements éventuels)
  - D : nombre de paramètres estimés
- Si $\mathcal{H}_0$ est fausse, la statistique de test χ^2_{obs} est grande: $\approx n \times$ distance(loi réelle ; loi prescrite par $\mathcal{H}_0$)

Rem. : preuve techniquement difficile[15],[16]

[14] K. Pearson. "On the criterion that a given system of deviations from the probable in the case of a correlated system of variables is such that it can be reasonably supposed to have arisen from random sampling". In: *The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science* 50.302 (1900), pp. 157–175.

[15] A. W. van der Vaart. *Asymptotic Statistics*. Cambridge Series in Statistical and Probabilistic Mathematics. Cambridge University Press, 2000.

[16] E. Benhamou and V. Melot. *Seven proofs of the Pearson Chi-squared independence test and its graphical interpretation*. Tech. rep. 2018.

Valeur du nombre de paramètres estimés

- Cas où la loi théorique est connue: $D = 0$
  Exemple : lancé de pièces, de dés, loi uniforme, etc.

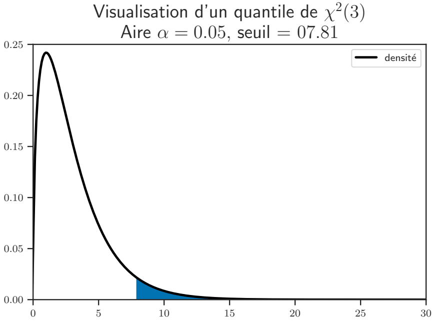
- Cas où la loi théorique est inconnue et dépend d'un unique paramètre (réel): $D = 1$
  Exemple : cas de la loi de Poisson de taux inconnu

- Cas où la loi théorique est inconnue et dépend de deux paramètres (réels), inconnus et à estimer: $D = 2$
  Exemple : cas du modèle gaussien avec μ et σ^2 inconnus

- etc.

Conclusion à niveau α fixé

$\chi^2(\ell)$: distribution du χ^2 à ℓ degrés de liberté



Fixer $q_{\chi^2}(1-\alpha)$ tel que:

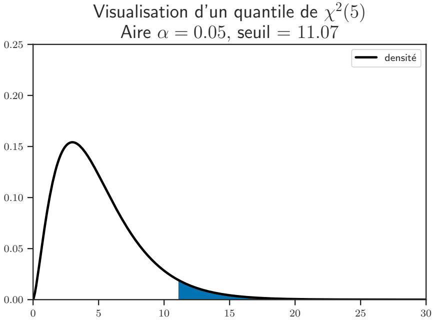
$$\mathbb{P}\left(\chi^2(\ell) \geq q_{\chi^2_\ell}(1-\alpha)\right) = \alpha$$

Décision :
Rejeter $\mathcal{H}_0$ si $\chi^2_{obs} \geq q_{\chi^2_\ell}(1-\alpha)$

<u>Rem.</u> : si $x_1, \ldots x_n$ sont *i.i.d.* $x_i \sim \mathcal{N}(0,1)$, alors $\sum_{i=1}^{n} x_i^2 \sim \chi^2(n)$

Conclusion à niveau α fixé

$\chi^2(\ell)$: distribution du χ^2 à ℓ degrés de liberté



Fixer $q_{\chi^2}(1-\alpha)$ tel que:

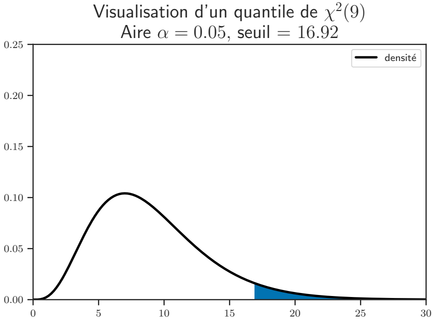
$$\mathbb{P}\left(\chi^2(\ell) \geq q_{\chi^2_\ell}(1-\alpha)\right) = \alpha$$

Décision :
Rejeter $\mathcal{H}_0$ si $\chi^2_{obs} \geq q_{\chi^2_\ell}(1-\alpha)$

Rem. : si $x_1, \dots x_n$ sont *i.i.d.* $x_i \sim \mathcal{N}(0,1)$, alors $\sum_{i=1}^{n} x_i^2 \sim \chi^2(n)$

Conclusion à niveau α fixé

$\chi^2(\ell)$: distribution du χ^2 à ℓ degrés de liberté



Fixer $q_{\chi^2}(1-\alpha)$ tel que:

$$\mathbb{P}\left(\chi^2(\ell) \geq q_{\chi^2_\ell}(1-\alpha)\right) = \alpha$$

Décision :
Rejeter $\mathcal{H}_0$ si $\chi^2_{obs} \geq q_{\chi^2_\ell}(1-\alpha)$

Rem. : si $x_1, \ldots x_n$ sont *i.i.d.* $x_i \sim \mathcal{N}(0,1)$, alors $\sum_{i=1}^{n} x_i^2 \sim \chi^2(n)$

Alternative: p-valeur

Rappel : Si $\alpha = 0$, on conserve toujours $\mathcal{H}_0$

Définition

La p-**valeur** (: p-*value*) est la plus petite valeur de α pour laquelle on rejette $\mathcal{H}_0$ sur l'échantillon observé

Interprétation :

"La p-**valeur** est la probabilité sous $\mathcal{H}_0$ d'observer un résultat aussi surprenant sur les données juste par hasard"

- p-valeur petite: on rejette l'hypothèse $\mathcal{H}_0$
- p-valeur grande: on ne rejette pas l'hypothèse $\mathcal{H}_0$

Cas du test du χ^2 : la p-valeur vaut $\mathbb{P}\left(\chi^2(K-1-D) \geq \chi^2_{obs}\right)$

p-valeur et exemples

Si $0.00 < p$-valeur< 0.05 $\begin{cases} \text{on rejette } \mathcal{H}_0 & \text{au niveau } 95\% \\ \text{on conserve } \mathcal{H}_0 & \text{au niveau } 100\% \end{cases}$

Si $0.05 < p$-valeur< 0.10 $\begin{cases} \text{on rejette } \mathcal{H}_0 & \text{au niveau } 90\% \\ \text{on conserve } \mathcal{H}_0 & \text{au niveau } 95\% \end{cases}$

Retour sur l'application numérique: $\mathbb{P}(\chi^2(6) \geq 1.0) \approx 0.98$

Conclusion : la probabilité d'observer une statistique aussi grande par hasard vaut 98%; on n'est donc pas du tout surpris et on conserve (non rejet) l'hypothèse de processus de Poisson

Rem. : p-valeur et tests de permutation (avec des alpagas)
https://www.jwilber.me/permutationtest/

Rem. : plus de lecture
https://www.statisticsdonewrong.com/

Sommaire

La méthode des moments: cas d'un paramètre inconnu

$x_1, \ldots, x_n$: *i.i.d.* selon une loi avec un paramètre **inconnu** θ

Méthode des moments[17] :

1. Calculer $\mathbb{E}(x)$ quand x suit la loi de paramètre θ

[17] $\mathbb{E}(X^k)$: moment d'ordre k d'une v.a. X, *i.e.*, $\int x^k f(x)dx$ (cas continu), $\sum_i i^k \mathbb{P}(X = i)$ (cas discret)

La méthode des moments: cas d'un paramètre inconnu

$x_1, \ldots, x_n$: *i.i.d.* selon une loi avec un paramètre **inconnu** θ

Méthode des moments[17] :

1. Calculer $\mathbb{E}(x)$ quand x suit la loi de paramètre θ
2. Exprimer le paramètre θ en fonction de l'espérance $\mathbb{E}(x)$

[17] $\mathbb{E}(X^k)$: moment d'ordre k d'une v.a. X, *i.e.*, $\int x^k f(x)dx$ (cas continu), $\sum_i i^k \mathbb{P}(X = i)$ (cas discret)

La méthode des moments: cas d'un paramètre inconnu

$x_1, \ldots, x_n$: *i.i.d.* selon une loi avec un paramètre **inconnu** θ

Méthode des moments[17] :

1. Calculer $\mathbb{E}(x)$ quand x suit la loi de paramètre θ
2. Exprimer le paramètre θ en fonction de l'espérance $\mathbb{E}(x)$
3. Remplacer l'espérance $\mathbb{E}(x)$ par la moyenne $\bar{x}_n$ dans la formule donnant θ et obtenir un estimateur $\hat{\theta}^{\text{moment}}$ de θ

[17] $\mathbb{E}(X^k)$: moment d'ordre k d'une v.a. X, *i.e.*, $\int x^k f(x)dx$ (cas continu), $\sum_i i^k \mathbb{P}(X = i)$ (cas discret)

La méthode des moments: cas d'un paramètre inconnu

$x_1, \ldots, x_n$: *i.i.d.* selon une loi avec un paramètre **inconnu** θ

Méthode des moments[17] :

1. Calculer $\mathbb{E}(x)$ quand x suit la loi de paramètre θ
2. Exprimer le paramètre θ en fonction de l'espérance $\mathbb{E}(x)$
3. Remplacer l'espérance $\mathbb{E}(x)$ par la moyenne $\bar{x}_n$ dans la formule donnant θ et obtenir un estimateur $\hat{\theta}^{\text{moment}}$ de θ

Exemple du modèle de Poisson : $\mathcal{P}(\theta)$

$$\mathbb{E}(x) = \theta \implies \boxed{\hat{\theta}^{\text{moment}} = \bar{x}_n}$$

[17] $\mathbb{E}(X^k)$: moment d'ordre k d'une v.a. X, *i.e.*, $\int x^k f(x)dx$ (cas continu), $\sum_i i^k \mathbb{P}(X = i)$ (cas discret)

La méthode des moments: cas d'un paramètre inconnu

$x_1, \ldots, x_n$: *i.i.d.* selon une loi avec un paramètre **inconnu** θ

Méthode des moments[17] :

1. Calculer $\mathbb{E}(x)$ quand x suit la loi de paramètre θ
2. Exprimer le paramètre θ en fonction de l'espérance $\mathbb{E}(x)$
3. Remplacer l'espérance $\mathbb{E}(x)$ par la moyenne $\bar{x}_n$ dans la formule donnant θ et obtenir un estimateur $\hat{\theta}^{\text{moment}}$ de θ

Exemple du modèle de Poisson : $\mathcal{P}(\theta)$

$$\mathbb{E}(x) = \theta \implies \boxed{\hat{\theta}^{\text{moment}} = \bar{x}_n}$$

[17] $\mathbb{E}(X^k)$: moment d'ordre k d'une v.a. X, *i.e.*, $\int x^k f(x)dx$ (cas continu), $\sum_i i^k \mathbb{P}(X = i)$ (cas discret)

La méthode des moments: cas de deux paramètres inconnus

$x_1, \ldots, x_n$: *i.i.d.* selon une loi avec 2 paramètres **inconnus** θ_1, θ_2

$\hookrightarrow$ Faire la même chose en utilisant aussi le moment d'ordre 2

Méthode des moments pour deux paramètres (θ_1, θ_2) :

1. Calculer $\mathbb{E}(x)$ et $\mathbb{E}(x^2)$ en fonction de θ_1 et θ_2

La méthode des moments: cas de deux paramètres inconnus

$x_1, \ldots, x_n$: *i.i.d.* selon une loi avec 2 paramètres **inconnus** θ_1, θ_2

$\hookrightarrow$ Faire la même chose en utilisant aussi le moment d'ordre 2

Méthode des moments pour deux paramètres (θ_1, θ_2) :

1. Calculer $\mathbb{E}(x)$ et $\mathbb{E}(x^2)$ en fonction de θ_1 et θ_2
2. Résoudre le système de deux équations à deux inconnues donnant θ_1 et θ_2 en fonction de $\mathbb{E}(x)$ et $\mathbb{E}(x^2)$

La méthode des moments: cas de deux paramètres inconnus

$x_1, \ldots, x_n$: *i.i.d.* selon une loi avec 2 paramètres **inconnus** θ_1, θ_2

$\hookrightarrow$ Faire la même chose en utilisant aussi le moment d'ordre 2

Méthode des moments pour deux paramètres (θ_1, θ_2) :

1. Calculer $\mathbb{E}(x)$ et $\mathbb{E}(x^2)$ en fonction de θ_1 et θ_2
2. Résoudre le système de deux équations à deux inconnues donnant θ_1 et θ_2 en fonction de $\mathbb{E}(x)$ et $\mathbb{E}(x^2)$
3. Remplacer $\mathbb{E}(x)$ par $\bar{x}_n$ et remplacer $\mathbb{E}(x^2)$ par $\frac{1}{n}\sum_i x_i^2$

La méthode des moments: cas de deux paramètres inconnus

$x_1, \ldots, x_n$: *i.i.d.* selon une loi avec 2 paramètres **inconnus** θ_1, θ_2

$\hookrightarrow$ Faire la même chose en utilisant aussi le moment d'ordre 2

Méthode des moments pour deux paramètres (θ_1, θ_2) :

1. Calculer $\mathbb{E}(x)$ et $\mathbb{E}(x^2)$ en fonction de θ_1 et θ_2
2. Résoudre le système de deux équations à deux inconnues donnant θ_1 et θ_2 en fonction de $\mathbb{E}(x)$ et $\mathbb{E}(x^2)$
3. Remplacer $\mathbb{E}(x)$ par $\bar{x}_n$ et remplacer $\mathbb{E}(x^2)$ par $\frac{1}{n}\sum_i x_i^2$

Exemple du modèle Gaussien : voir TD/TP

La méthode des moments: cas de deux paramètres inconnus

$x_1, \ldots, x_n$: *i.i.d.* selon une loi avec 2 paramètres **inconnus** θ_1, θ_2

$\hookrightarrow$ Faire la même chose en utilisant aussi le moment d'ordre 2

Méthode des moments pour deux paramètres (θ_1, θ_2) :

1. Calculer $\mathbb{E}(x)$ et $\mathbb{E}(x^2)$ en fonction de θ_1 et θ_2
2. Résoudre le système de deux équations à deux inconnues donnant θ_1 et θ_2 en fonction de $\mathbb{E}(x)$ et $\mathbb{E}(x^2)$
3. Remplacer $\mathbb{E}(x)$ par $\bar{x}_n$ et remplacer $\mathbb{E}(x^2)$ par $\frac{1}{n}\sum_i x_i^2$

Exemple du modèle Gaussien : voir TD/TP

Vraisemblance: variable continue

On note $f_\theta(\cdot)$ la densité (continue) de la loi de paramètre θ, et on suppose qu'on observe $x_1, \ldots, x_n$ *i.i.d.* suivant cette loi

Définition

La vraisemblance (: *Likelihood*) du paramètre θ est la densité de $(x_1, \ldots, x_n)$ vue comme une fonction de θ.

- cas $n = 1$: la vraisemblance de θ (au vu de x_1) est $f_\theta(x_1)$
- cas n quelconque: la vraisemblance de θ, notée $L(\theta)$, est le produit des vraisemblances :

$$L(\theta) := f_\theta(x_1) \times \cdots \times f_\theta(x_n) = \prod_{i=1}^{n} f_\theta(x_i)$$

⚠: on dit **vraisemblance d'un paramètre au vu des données**; les données ne sont pas vraisemblables, elles sont ce qu'elles sont!

Vraisemblance: variable discrète

Notant $f_\theta(x) := \mathbb{P}(X = x)$ lorsque X suit la loi de paramètre θ, la même formule pour la vraisemblance est encore valable!

Exemple : on cherche le paramètre θ d'une loi de Poisson

$$\begin{aligned} L(\theta) &= \mathrm{e}^{-\theta}\frac{\theta^{x_1}}{x_1!} \times \cdots \times \mathrm{e}^{-\theta}\frac{\theta^{x_n}}{x_n!} \\ &= \mathrm{e}^{-n\theta}\frac{\theta^{x_1+x_2+\ldots+x_n}}{\text{ne dépend pas de } \theta} \\ &= \frac{\mathrm{e}^{-n\theta+(x_1+x_2+\ldots+x_n)\log(\theta)}}{\text{ne dépend pas de } \theta} \end{aligned}$$

<u>Rem.</u> : pour plus tard (Cste : constante qui ne dépend pas de θ)

$$\log(L(\theta)) = -n\theta + (x_1 + x_2 + \ldots + x_n)\log(\theta) + \text{Cste}$$

Maximum de vraisemblance
(: *Maximum Likelihood Estimator, MLE*)

Définition

L'estimateur $\hat{\theta}^{\mathrm{MLE}}$ du **maximum de vraisemblance** est l'estimateur qui maximise la fonction de vraisemblance L, *i.e.*,

$$\hat{\theta}^{\mathrm{MLE}} = \arg\max_{\theta} L(\theta) \ .$$

Rem. : mathématiquement il est plus simple de maximiser $\log(L)$ que L car **on dérive alors des sommes plutôt que des produits**

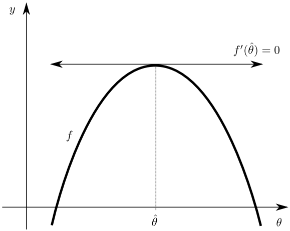
Rem. : $\arg\max$ signifie “le point qui atteint le maximum”

Optimisation et résolution

Règle de Fermat (18)

Soit $f : \begin{cases} \mathbb{R} & \mapsto \mathbb{R} \\ \theta & \to f(\theta) \end{cases}$ une fonction dérivable qui atteint son maximum au point $\hat{\theta}$, alors la dérivée de f est nulle en $\hat{\theta}$, *i.e.*,

$$\boxed{f'(\hat{\theta}) = 0}$$



(18) on appelle aussi parfois cette propriété la “condition nécessaire du 1^{er} ordre”

Maximum de vraisemblance et cas Poisson

$$\begin{aligned} & \hat{\theta}^{\mathrm{MLE}} = \arg\max_{\theta} L(\theta) \\ \iff \quad & \hat{\theta}^{\mathrm{MLE}} = \arg\max_{\theta} \log L(\theta) \end{aligned}$$

(19) on admettra que c'est bien un maximum (et non un minimum ou un point selle)

Maximum de vraisemblance et cas Poisson

$$
\begin{aligned}
& \hat{\theta}^{\mathrm{MLE}} = \arg\max_{\theta} L(\theta) \\
\iff \quad & \hat{\theta}^{\mathrm{MLE}} = \arg\max_{\theta} \log L(\theta) \\
\implies \quad & (\log L)'\,(\hat{\theta}^{\mathrm{MLE}}) = \frac{d\,(\log L)}{d\theta}(\hat{\theta}^{\mathrm{MLE}}) = 0
\end{aligned}
$$

(19) on admettra que c'est bien un maximum (et non un minimum ou un point selle)

Maximum de vraisemblance et cas Poisson

$$
\begin{aligned}
& \hat{\theta}^{\mathrm{MLE}} = \arg\max_{\theta} L(\theta) \\
\iff \quad & \hat{\theta}^{\mathrm{MLE}} = \arg\max_{\theta} \log L(\theta) \\
\implies \quad & (\log L)'\,(\hat{\theta}^{\mathrm{MLE}}) = \frac{d\,(\log L)}{d\theta}(\hat{\theta}^{\mathrm{MLE}}) = 0
\end{aligned}
$$

Dans le cas de la loi de Poisson (*cf.* diapo $n^\circ\, 15$) :

$$\forall \theta > 0, \quad (\log L)'(\theta) = \frac{x_1 + \cdots + x_n}{\theta} - n$$

[19] on admettra que c'est bien un maximum (et non un minimum ou un point selle)

Maximum de vraisemblance et cas Poisson

$$
\begin{aligned}
& \hat{\theta}^{\mathrm{MLE}} = \arg\max_{\theta} L(\theta) \\
\iff \quad & \hat{\theta}^{\mathrm{MLE}} = \arg\max_{\theta} \log L(\theta) \\
\implies \quad & (\log L)'\,(\hat{\theta}^{\mathrm{MLE}}) = \frac{d\,(\log L)}{d\theta}(\hat{\theta}^{\mathrm{MLE}}) = 0
\end{aligned}
$$

Dans le cas de la loi de Poisson (*cf.* diapo $n^\circ\,15$) :

$$\forall \theta > 0, \quad (\log L)'(\theta) = \frac{x_1 + \cdots + x_n}{\theta} - n$$

Cette dérivée s'annule en $\hat{\theta}^{\mathrm{MLE}} = \frac{x_1+\cdots+x_n}{n}$ et alors [19] :

$$\boxed{\hat{\theta}^{\mathrm{MLE}} = \bar{x}_n}$$

[19] on admettra que c'est bien un maximum (et non un minimum ou un point selle)

Biographie du jour : Ronald Aylmer Fisher[20]

- biologiste et statisticien britannique (1890-1962)
- un des fondateurs de la génétique moderne et un grand continuateur de Darwin
- Il a introduit le maximum de vraisemblance, l'analyse de la variance (ANOVA), le test éponyme
- ...

[20] https://fr.wikipedia.org/wiki/Ronald_Aylmer_Fisher

Bibliographie I

- Benhamou, E. and V. Melot. *Seven proofs of the Pearson Chi-squared independence test and its graphical interpretation*. Tech. rep. 2018.
- Box, G. E. P. “Robustness in the strategy of scientific model building”. In: *Robustness in statistics*. Elsevier, 1979, pp. 201–236.
- Nolan, D. and T. P. Speed. *Stat labs: mathematical statistics through applications*. Springer Science & Business Media, 2001.
- Pearson, K. “On the criterion that a given system of deviations from the probable in the case of a correlated system of variables is such that it can be reasonably supposed to have arisen from random sampling”. In: *The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science* 50.302 (1900), pp. 157–175.
- Valéry, P. *Mauvaises pensées et autres*. Gallimard, 1942.

Bibliographie II

- van der Vaart, A. W. *Asymptotic Statistics*. Cambridge Series in Statistical and Probabilistic Mathematics. Cambridge University Press, 2000.