

# HLMA408: Traitement des données

Estimations et tests: cas du cytomégalo virus

**Joseph Salmon**

<http://josephsalmon.eu>

Université de Montpellier



# Sommaire

Introduction

Modélisation probabiliste de la position des palindromes

Test d'adéquation à une loi

Test du  $\chi^2$  : schéma général

Estimation d'un paramètre

# Sommaire

## Introduction

Modélisation probabiliste de la position des palindromes

Test d'adéquation à une loi

Test du  $\chi^2$  : schéma général

Estimation d'un paramètre

# Étude du CytoMégaloVirus (CMV) humain<sup>(1)</sup>

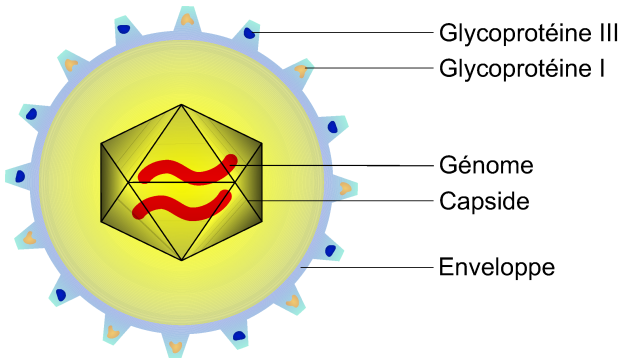
- ▶ Cytomégalovirus: famille des **herpesvirus**, comprenant le virus de l'herpès simplex, le virus d'Epstein-Barr, virus varicelle-zona, . . .
- ▶ Caractéristique du virus:
  - capacité à produire des **infections latentes et persistantes**
  - mydangereux pour **les fœtus** et les personnes avec **faibles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

---

<sup>(1)</sup>adapté de D. Nolan and T. P. Speed. *Stat labs: mathematical statistics through applications*. Springer Science & Business Media, 2001

# Structure de virus CMV<sup>(2)</sup>

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



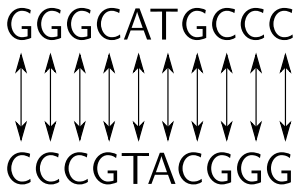
<sup>(2)</sup>source: <https://en.wikipedia.org/wiki/Cytomegalovirus>

# ADN et origine de la réplication

ADN : longue chaîne de lettres sur l'alphabet A, C, G, T  
complémentaires 2 à 2 (A  $\leftrightarrow$  T; G  $\leftrightarrow$  C, **paire de bases**: bp)

Origine de la réplication : les palindromes (complémentaires)  
semblent importants biologiquement

**Exemple:**



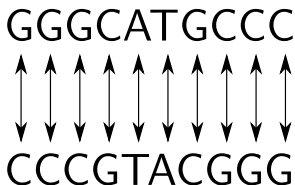
**Enjeu:** découvrir les zones de l'ADN où le nombre de palindromes  
est **anormalement** élevé

# ADN et origine de la réplication

ADN : longue chaîne de lettres sur l'alphabet A, C, G, T  
complémentaires 2 à 2 (A  $\leftrightarrow$  T; G  $\leftrightarrow$  C, **paire de bases**: bp)

Origine de la réplication : les palindromes (complémentaires)  
semblent importants biologiquement

**Exemple:**



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

# Données

Données du génome de CMV <sup>(3)</sup> : répertorie les positions de palindromes dans le génome de CMV

$$n_{\text{bp}} = 229\,354$$

lettres (paires de bases = bp)

$$n = 296$$

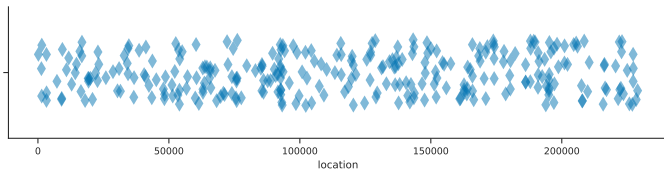
palindromes de longueur  $\geq 10$  bp

---

<sup>(3)</sup><http://www.stat.berkeley.edu/users/statlabs/data/hcmv.data>



# Position des palindromes



**TO DO: voir notebook `EstimationTest.ipynb`**

# Sommaire

Introduction

Modélisation probabiliste de la position des palindromes

Test d'adéquation à une loi

Test du  $\chi^2$  : schéma général

Estimation d'un paramètre

# Des palindromes positionnés au hasard ?

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

Sans région privilégiée = homogène

Le modèle = processus de Poisson homogène

# Modèle aléatoire

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

# Modèle aléatoire

- ▶ Considérer **tous les échantillons possibles** (même si en pratique un seul est un 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 un 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 un 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 : univers des possibles + probabilités + ...

- ▶ Dans le cas du modèle “échantillonnage aléatoire simple”, le modèle est complètement déterminé

---

<sup>(4)</sup> souvent notés par des lettres grecques



# Modélisation aléatoire

Modèle aléatoire : univers des possibles + probabilités + ...

- ▶ Dans le cas du modèle “échantillonnage aléatoire simple”, le modèle est complètement déterminé
- ▶ Certains modèles (et donc les probabilités d'événements) dépendent de **paramètres inconnus**<sup>(4)</sup>

---

<sup>(4)</sup> souvent notés par des lettres grecques

# Modélisation aléatoire

Modèle aléatoire : univers des possibles + probabilités + ...

- ▶ Dans le cas du modèle “échantillonnage aléatoire simple”, le modèle est complètement déterminé
- ▶ Certains modèles (et donc les probabilités d'événements) dépendent de **paramètres inconnus**<sup>(4)</sup>
- ▶ Ici, un modèle dont l'univers des possibles décrit les différentes façons de positionner les palindromes sur le génome

---

<sup>(4)</sup>souvent notés par des lettres grecques

# Modélisation aléatoire

Modèle aléatoire : univers des possibles + probabilités + ...

- ▶ Dans le cas du modèle “échantillonnage aléatoire simple”, le modèle est complètement déterminé
- ▶ Certains modèles (et donc les probabilités d'événements) dépendent de **paramètres inconnus**<sup>(4)</sup>
- ▶ Ici, un modèle dont l'univers des possibles décrit les différentes façons de positionner les palindromes sur le génome

---

<sup>(4)</sup>souvent notés par des lettres grecques

# Processus de Poisson homogène

Processus de Poisson homogène: modélise des palindromes répartis totalement au hasard, de façon homogène parmi les environ 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 une zone donnée

## Processus de Poisson homogène (2)

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 phénomène au cours du temps, e.g., passage d'un bus à un arrêt<sup>(5)</sup>

Propriétés :

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

---

<sup>(5)</sup> cf. <http://jakevdp.github.io/blog/2018/09/13/waiting-time-paradox/> pour une analyse fine

## Processus de Poisson homogène (2)

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 phénomène au cours du temps, e.g., passage d'un bus à un arrêt<sup>(5)</sup>

Propriétés :

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

---

<sup>(5)</sup> cf. <http://jakevdp.github.io/blog/2018/09/13/waiting-time-paradox/> pour une analyse fine

## Processus de Poisson homogène (2)

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 phénomène au cours du temps, e.g., passage d'un bus à un arrêt<sup>(5)</sup>

Propriétés :

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

---

<sup>(5)</sup> cf. <http://jakevdp.github.io/blog/2018/09/13/waiting-time-paradox/> pour une analyse fine

# Propriétés du processus de Poisson

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  $N$

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

Rem: on peut alors vérifier que  $\mathbb{E}(N) = \theta$

**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  $\theta = \lambda L$ ,  $N \sim \mathcal{P}(\lambda L)$

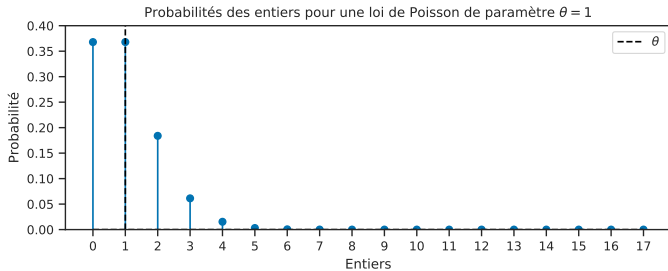
$\lambda$  : paramètre d'**intensité** représentant le taux d'apparition du phénomène (ici les palindromes)

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

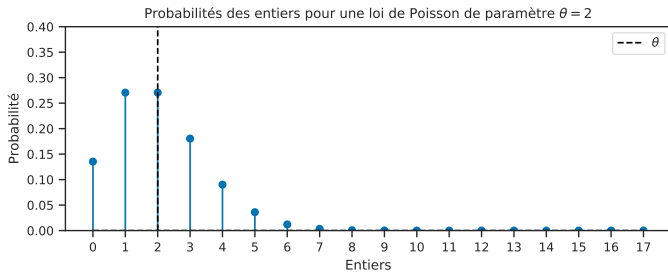
Unité de  $\lambda$  : homogène à l'inverse d'une longueur (taux)



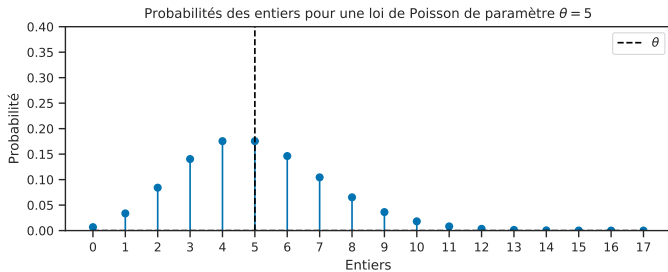
# Loi de Poisson



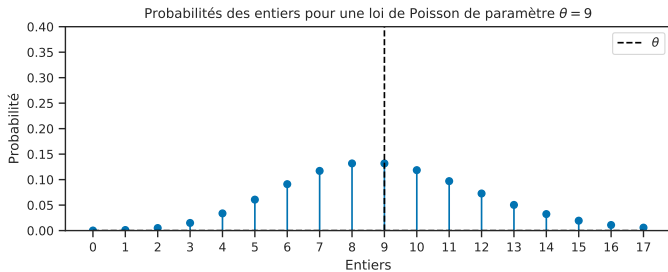
# Loi de Poisson



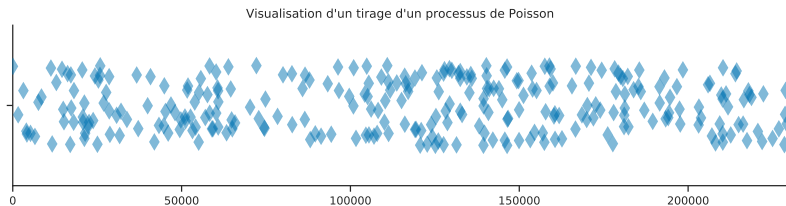
# Loi de Poisson



# Loi de Poisson

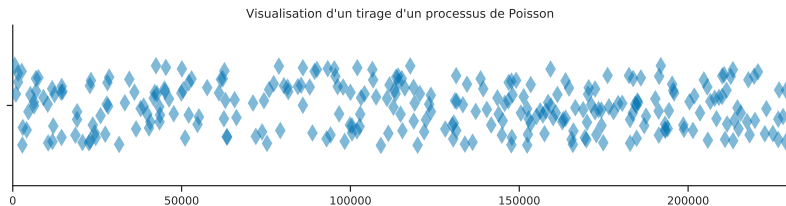


# Quatre réalisations d'un processus de Poisson



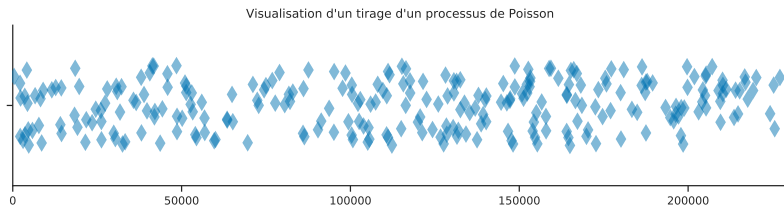
Rem: paramètres “similaires” à ceux des données CMV

# Quatre réalisations d'un processus de Poisson



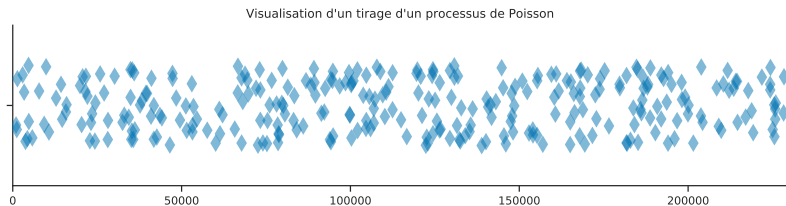
Rem: paramètres “similaires” à ceux des données CMV

# Quatre réalisations d'un processus de Poisson



Rem: paramètres “similaires” à ceux des données CMV

# Quatre réalisations d'un processus de Poisson



Rem: paramètres “similaires” à ceux des données CMV



# Estimation du taux $\lambda$



:  $\lambda$  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 **vraisemblable**

## Estimation de $\lambda$

Ici, on choisira comme estimateur de  $\lambda$  :

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

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

Rem: dans le cas présent les deux méthodes  
(moments/vraisemblance) coïncident

# Sommaire

Introduction

Modélisation probabiliste de la position des palindromes

Test d'adéquation à une loi

Test du  $\chi^2$  : schéma général

Estimation d'un paramètre

# Problème général

“Ce qui est simple est toujours faux. Ce qui ne l'est pas est inutilisable.”<sup>(6)</sup>

“All models are wrong but some are useful”<sup>(7)</sup>

---

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.

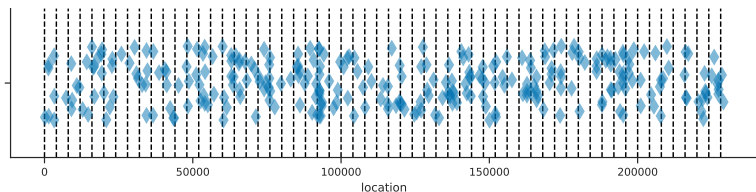
---

<sup>(6)</sup> P. Valéry. *Mauvaises pensées et autres*. Gallimard, 1942.

<sup>(7)</sup> 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<sup>(8)</sup>
- ▶ compter dans chacune des régions l'occurrence de palindromes:



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																					

---

<sup>(8)</sup>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

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

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

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

## 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 de  $\hat{\theta}$  voisine de 5.16, on obtient un effectif attendu de :

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

etc.

**TO DO: cf. notebook associé pour les autres calculs (en Python)**



## 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

—→ pour mesurer la différence entre les deux colonnes, on introduit **une statistique de test**

# Statistique de test

Application numérique:

$$\frac{(7-6.4)^2}{6.4} + \frac{(8-7.5)^2}{7.5} + \frac{(10-9.7)^2}{9.7} + \frac{(9-10.0)^2}{10.0} + \frac{(8-8.6)^2}{8.6} + \frac{(5-6.3)^2}{6.3} + \frac{(4-4.1)^2}{4.1} + \frac{(6-4.5)^2}{4.5} \approx 1.0$$

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

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

$$\mathbb{P}(\chi^2(6) \geq 1.0) \approx 0.98$$

# Statistique de test

Application numérique:

$$\frac{(7-6.4)^2}{6.4} + \frac{(8-7.5)^2}{7.5} + \frac{(10-9.7)^2}{9.7} + \frac{(9-10.0)^2}{10.0} + \frac{(8-8.6)^2}{8.6} + \frac{(5-6.3)^2}{6.3} + \frac{(4-4.1)^2}{4.1} + \frac{(6-4.5)^2}{4.5} \approx 1.0$$

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

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

$$\mathbb{P}(\chi^2(6) \geq 1.0) \approx 0.98$$

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

# Statistique de test

Application numérique:

$$\frac{(7-6.4)^2}{6.4} + \frac{(8-7.5)^2}{7.5} + \frac{(10-9.7)^2}{9.7} + \frac{(9-10.0)^2}{10.0} + \frac{(8-8.6)^2}{8.6} + \frac{(5-6.3)^2}{6.3} + \frac{(4-4.1)^2}{4.1} + \frac{(6-4.5)^2}{4.5} \approx 1.0$$

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

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

$$\mathbb{P}(\chi^2(6) \geq 1.0) \approx 0.98$$

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

# 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  $\chi^2$

# 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  $\chi^2$

3. Pour distinguer dans quel cas on est, on regarde si la valeur observée (ici  $\approx 1.0$ ) 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  $\chi^2$

3. Pour distinguer dans quel cas on est, on regarde si la valeur observée (ici  $\approx 1.0$ ) 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  $\chi^2$

3. Pour distinguer dans quel cas on est, on regarde si la valeur observée (ici  $\approx 1.0$ ) 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



# Un peu de vocabulaire

On vient de comparer deux hypothèses :

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

vs.

- ▶ l'**hypothèse alternative**, notée  $\mathcal{H}_1$ , sous laquelle on doit connaître le comportement 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

- ▶ l'**hypothèse nulle**

$\mathcal{H}_0$  : “vous êtes enceinte!”

vs.

- ▶ l'**hypothèse alternative**

$\mathcal{H}_1$ : “vous **N**'êtes **PAS** enceinte!”

Réalité:

$\mathcal{H}_0$

Diagnostic:

$\mathcal{H}_1$



Erreur de type 1 = erreur de première espèce  
= vrai négatif = rejet à tort

Réalité:

$\mathcal{H}_1$



Diagnostic:

$\mathcal{H}_0$

Erreur de type 2 = erreur de seconde espèce  
= faux positif = fausse alarme

Réalité:

$\mathcal{H}_0$

$\mathcal{H}_1$

Diagnostic:

$\mathcal{H}_0$



$\mathcal{H}_1$



Les quatre situations possibles:  
de la réalité au diagnostic

## Autres exemples

- Contexte militaire / guerre froide (historique):

$\mathcal{H}_0$  : “un missile arrive sur nous!”

vs.

$\mathcal{H}_1$  : “il n’y a pas de missile”

Rem: le vocabulaire **fausse alarme** vient de ce contexte pour l’erreur de 2<sup>nde</sup> espèce

- *Le Garçon qui criait au loup*<sup>(9)</sup>:

$\mathcal{H}_0$  : “le loup est dans la bergerie!”

vs.

$\mathcal{H}_1$  : “le loup n’est pas dans la bergerie ”

---

<sup>(9)</sup> fable d’Ésope (VII<sup>e</sup>-VI<sup>e</sup> siècle av. J.-C.) connue sous le nom de “Le Berger mauvais plaisant”  
[https://fr.wikisource.org/wiki/Le\\_Berger\\_mauvais\\_plaisant](https://fr.wikisource.org/wiki/Le_Berger_mauvais_plaisant)

# Choix des hypothèses

En pratique, comment choisir laquelle des deux hypothèses doit être nommée hypothèse nulle  $\mathcal{H}_0$ ?

Plusieurs heuristiques:

- ▶ Choisir comme  $\mathcal{H}_0$  l'hypothèse que l'on cherche à rejeter :  
**Exemple:** test de médicament,  $\mathcal{H}_0$  : “le médicament n'est pas efficace”  
**Exemple:** test de VIH,  $\mathcal{H}_0$  “la personne a la virus”  
**Exemple:** test de grossesse,  $\mathcal{H}_0$  “la femme est enceinte”
- ▶ Si l'une de deux hypothèses est plus simple ou “de dimension plus petite” que l'autre, on la choisit pour  $\mathcal{H}_0$   
**Exemple:** :  $\mathcal{H}_0 : \theta = 5$  ,  $\mathcal{H}_1 : \theta \neq 5$
- ▶ Souvent :  $\mathcal{H}_0$  plus “importante” ou plus “mydangereuse” que  $\mathcal{H}_1$   
**Exemple:** de la détection de missile  $\mathcal{H}_0$ : “il y a un missile”

# Deux types d'erreur dans un test

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



- **Erreur de 2<sup>de</sup> 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  $\alpha$  et  $\beta$  :

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

Plus ces quantités sont petites, mieux c'est !

# Erreurs extrêmes

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

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

**Exemple:** diagnostiquer “vous êtes enceinte” tout le temps

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

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

---

Conclusion: besoin d'un **compromis**

# Théorie classique des tests

Classiquement: l'**utilisateur** fixe  $\alpha$ , la probabilité d'erreur de 1<sup>re</sup> espèce maximale souhaitée (probabilité de rejeter à tort  $\mathcal{H}_0$ )

Valeurs classiques de  $\alpha$ : 0.10, 0.05 ou 0.01 (selon le contexte)

Rappel: on s'intéresse à des  $\alpha$  petits

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

# Sommaire

Introduction

Modélisation probabiliste de la position des palindromes

Test d'adéquation à une loi

Test du  $\chi^2$  : schéma général

Estimation d'un paramètre

# Test d'adéquation à une loi

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

**But**: tester si l'échantillon  $x_1, \dots, x_n$  provient d'une loi donnée

- ▶ 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 conserve  $\mathcal{H}_0$ , alors rien dans l'échantillon ne semble 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  $\mathcal{H}_0$
4. Regrouper les classes pour que les eff. espérés soient  $\geq 5$  et conserver uniquement cette nouvelle table.
5. Pour cette table avec  $K$  classes et  $n$  observations, calculer:


$$\begin{aligned}\chi_{obs}^2 &:= \sum_{\text{différentes classes}} \frac{(\text{Eff. observé} - \text{Eff. espéré dans la classe})^2}{\text{Eff. espéré dans la classe}} \\ &= \sum_{k=1}^K \frac{(\hat{f}_k - f_k)^2}{f_k}; p_k : \text{probabilité théorique de la classe } k \\ &= \sum_{k=1}^K \frac{(\hat{f}_k - np_k)^2}{np_k}\end{aligned}$$

## Pourquoi diviser par les effectifs espérés?

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

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<sup>nde</sup> statistique donnerait un poids trop grand aux petites valeurs, e.g., même contribution si  $f_1 = 10$ ,  $\hat{f}_1 = 5$  et si  $f_2 = 500$ ,  $\hat{f}_1 = 505$

**MAIS**: en relatif cela représente 50% de variation ou 1%...

Conclusion: atténuer les petites variations des grands effectifs en multipliant par  $\frac{1}{f_k}$

# Comportement de la statistique de test

---

## Théorème<sup>(10)</sup>

---

- ▶ Si  $\mathcal{H}_0$  est vraie, la statistique de test  $\chi_{obs}^2$  suit une loi du  $\chi^2$  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  $\chi_{obs}^2$  est grande, de l'ordre de  $n \times$  distance entre loi réelle et loi prescrite par  $\mathcal{H}_0$
- 

Rem: preuve techniquement difficile<sup>(11), (12)</sup>

---

<sup>(10)</sup> 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.

<sup>(11)</sup> A. W. van der Vaart. *Asymptotic Statistics*. Cambridge Series in Statistical and Probabilistic Mathematics. Cambridge University Press, 2000.

<sup>(12)</sup> E. Benhamou and V. Melot. *Seven proofs of the Pearson Chi-squared independence test and its graphical interpretation*. Tech. rep. 2018.



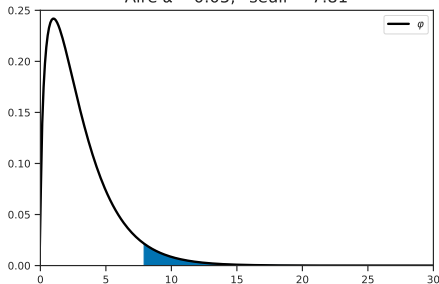
# Valeur du nombre de paramètres estimés

- ▶ Cas où la loi théorique est connue:  $D = 0$   
**Exemple:** c'était le cas quand on connaît les paramètres de la loi sous  $\mathcal{H}_0$  (e.g., pour un lancé de pièces ou de dés)
- ▶ Cas où la loi théorique est inconnue et qu'elle dépend d'un seul paramètre:  $D = 1$   
**Exemple:** cas du modèle de Poisson dont le taux est inconnu
- ▶ Cas où la loi théorique est inconnue et qu'elle dépend de deux paramètres (inconnus et à estimer):  $D = 2$   
**Exemple:** cas du modèle gaussien avec  $\mu$  et  $\sigma^2$  inconnus

# Conclusion à niveau $\alpha$ fixé

Distribution du  $\chi^2$  à  $\ell$  degrés de liberté:

Visualisation d'un quantile de  $\chi^2(3)$   
Aire  $\alpha = 0.05$ , seuil = 7.81



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

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

**Décision:**

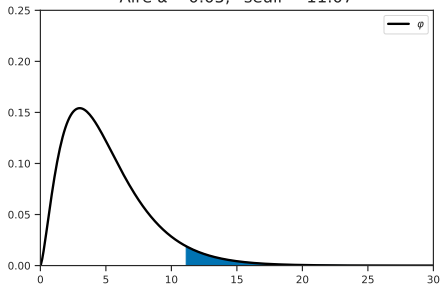
Rejeter  $\mathcal{H}_0$  si  $\chi_{obs}^2 \geq q_{\chi^2}(1 - \alpha)$

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 $\alpha$ fixé

Distribution du  $\chi^2$  à  $\ell$  degrés de liberté:

Visualisation d'un quantile de  $\chi^2(5)$   
Aire  $\alpha = 0.05$ , seuil = 11.07



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

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

**Décision:**

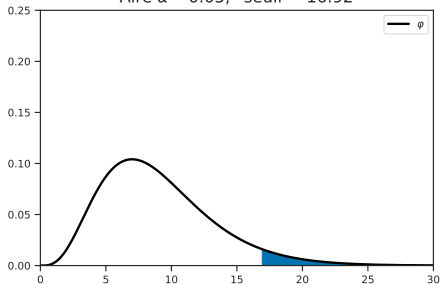
Rejeter  $\mathcal{H}_0$  si  $\chi_{obs}^2 \geq q_{\chi^2}(1 - \alpha)$

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 $\alpha$ fixé

Distribution du  $\chi^2$  à  $\ell$  degrés de liberté:

Visualisation d'un quantile de  $\chi^2(9)$   
Aire  $\alpha = 0.05$ , seuil = 16.92



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

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

**Décision:**

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

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

## 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  $\alpha$  pour laquelle on rejette  $\mathcal{H}_0$  sur l'échantillon observé

---

---

### Interprétation:

“La  $p$ -valeur est la probabilité sous  $\mathcal{H}_0$  que l'on observe 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  $\chi^2$ : la  $p$ -valeur vaut  $\mathbb{P}\left(\chi^2(K-1-D) \geq \chi_{obs}^2\right)$

## *p*-valeur et exemples

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

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

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 est donc pas du tout surpris, et on conserve (non rejet) l'hypothèse de processus de Poisson

Rem: plus de lecture: <https://www.statisticsonewrong.com/>

# Sommaire

Introduction

Modélisation probabiliste de la position des palindromes

Test d'adéquation à une loi

Test du  $\chi^2$  : schéma général

Estimation d'un paramètre

# La méthode des moments

$x_1, \dots, x_n$ : échantillon *i.i.d.* selon une loi dépendant d'un paramètre **inconnu**  $\theta$

**Méthode des moments**<sup>(13)</sup>:

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

---

<sup>(13)</sup>le moment d'ordre  $k$  d'une v.a.  $X$  est  $\mathbb{E}(X^k)$



# La méthode des moments

$x_1, \dots, x_n$ : échantillon *i.i.d.* selon une loi dépendant d'un paramètre **inconnu**  $\theta$

**Méthode des moments**<sup>(13)</sup>:

1. Calculer  $\mathbb{E}(x)$  quand  $x$  suit la loi de paramètre  $\theta$
2. À partir du calcul précédent, exprimer  $\theta$  en fonction de  $\mathbb{E}(x)$

---

<sup>(13)</sup>le moment d'ordre  $k$  d'une v.a.  $X$  est  $\mathbb{E}(X^k)$

# La méthode des moments

$x_1, \dots, x_n$ : échantillon *i.i.d.* selon une loi dépendant d'un paramètre **inconnu**  $\theta$

**Méthode des moments**<sup>(13)</sup>:

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

---

<sup>(13)</sup>le moment d'ordre  $k$  d'une v.a.  $X$  est  $\mathbb{E}(X^k)$

# La méthode des moments

$x_1, \dots, x_n$ : échantillon *i.i.d.* selon une loi dépendant d'un paramètre **inconnu**  $\theta$

**Méthode des moments**<sup>(13)</sup>:

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

**Exemple:** (modèle de Poisson), loi :  $\mathcal{P}(\theta)$ , ainsi

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

---

<sup>(13)</sup>le moment d'ordre  $k$  d'une v.a.  $X$  est  $\mathbb{E}(X^k)$

# La méthode des moments

$x_1, \dots, x_n$ : échantillon *i.i.d.* selon une loi dépendant d'un paramètre **inconnu**  $\theta$

**Méthode des moments**<sup>(13)</sup>:

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

**Exemple:** (modèle de Poisson), loi :  $\mathcal{P}(\theta)$ , ainsi

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

---

<sup>(13)</sup>le moment d'ordre  $k$  d'une v.a.  $X$  est  $\mathbb{E}(X^k)$

## Méthode des moments (suite)

- ▶ S'il y a plus d'un paramètre
- ▶ Si  $\mathbb{E}(x)$  ne dépend pas de  $\theta$

↪ Faire la même chose en utilisant aussi le moment d'ordre 2

Méthode des moments pour deux paramètres:  $(\theta_1, \theta_2)$

1. Calculer  $\mathbb{E}(x)$  et  $\mathbb{E}(x^2)$  en fonction de  $\theta_1$  et  $\theta_2$

## Méthode des moments (suite)

- ▶ S'il y a plus d'un paramètre
- ▶ Si  $\mathbb{E}(x)$  ne dépend pas de  $\theta$

↪ Faire la même chose en utilisant aussi le moment d'ordre 2

Méthode des moments pour deux paramètres:  $(\theta_1, \theta_2)$

1. Calculer  $\mathbb{E}(x)$  et  $\mathbb{E}(x^2)$  en fonction de  $\theta_1$  et  $\theta_2$
2. Résoudre le système de deux équations à deux inconnues donnant  $\theta_1$  et  $\theta_2$  en fonction de  $\mathbb{E}(x)$  et  $\mathbb{E}(x^2)$

## Méthode des moments (suite)

- ▶ S'il y a plus d'un paramètre
- ▶ Si  $\mathbb{E}(x)$  ne dépend pas de  $\theta$

↪ Faire la même chose en utilisant aussi le moment d'ordre 2

Méthode des moments pour deux paramètres:  $(\theta_1, \theta_2)$

1. Calculer  $\mathbb{E}(x)$  et  $\mathbb{E}(x^2)$  en fonction de  $\theta_1$  et  $\theta_2$
2. Résoudre le système de deux équations à deux inconnues donnant  $\theta_1$  et  $\theta_2$  en fonction de  $\mathbb{E}(x)$  et  $\mathbb{E}(x^2)$
3. Remplacer  $\mathbb{E}(x)$  par  $\bar{x}_n$  et  $\mathbb{E}(x^2)$  par  $\frac{1}{n} \sum x_i^2$

## Méthode des moments (suite)

- ▶ S'il y a plus d'un paramètre
- ▶ Si  $\mathbb{E}(x)$  ne dépend pas de  $\theta$

↪ Faire la même chose en utilisant aussi le moment d'ordre 2

Méthode des moments pour deux paramètres:  $(\theta_1, \theta_2)$

1. Calculer  $\mathbb{E}(x)$  et  $\mathbb{E}(x^2)$  en fonction de  $\theta_1$  et  $\theta_2$
2. Résoudre le système de deux équations à deux inconnues donnant  $\theta_1$  et  $\theta_2$  en fonction de  $\mathbb{E}(x)$  et  $\mathbb{E}(x^2)$
3. Remplacer  $\mathbb{E}(x)$  par  $\bar{x}_n$  et  $\mathbb{E}(x^2)$  par  $\frac{1}{n} \sum x_i^2$



## Vraisemblance: variable continue

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


---

---

### Définition

---

---


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

---

---

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

$$L(\theta) := f_{\theta}(x_1) \times \dots \times f_{\theta}(x_n)$$

 : on dit **vraisemblance d'un paramètre au vue 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  $\theta$ , la même formule pour la vraisemblance est encore valable!

**Exemple:** on cherche le paramètre  $\theta$  d'une loi de Poisson à l'aide des observations  $x_1 = 7, x_2 = 1, \dots, x_{57} = 6$  (cf. diapo n° 22)

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

# Maximum de vraisemblance

( : *Maximum Likelihood Estimator, MLE*)

---

---

## Définition

---

---

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

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

---

---

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

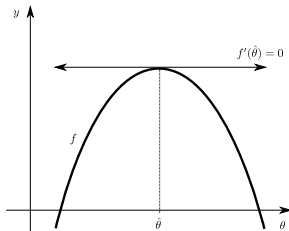
Rem:  $\arg \max$  signifie “le point qui atteint le maximum”

# Optimisation et résolution

## Règle de Fermat<sup>(14)</sup>

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

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



<sup>(14)</sup> on appelle aussi parfois cette propriété la "condition nécessaire du 1<sup>er</sup> ordre"

# Maximum de vraisemblance et cas Poisson

$$\hat{\theta}^{\text{MLE}} = \arg \max_{\theta} L(\theta)$$
$$\iff \hat{\theta}^{\text{MLE}} = \arg \max_{\theta} \log L(\theta)$$

---

<sup>(15)</sup> on admettra que c'est bien un maximum (et non un minimum ou un point selle)

# Maximum de vraisemblance et cas Poisson

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

$$\iff \hat{\theta}^{\text{MLE}} = \arg \max_{\theta} \log L(\theta)$$

$$\iff (\log L)'(\hat{\theta}^{\text{MLE}}) = \frac{d}{d\theta} \left( \log L(\hat{\theta}^{\text{MLE}}) \right) = 0$$

---

<sup>(15)</sup> on admettra que c'est bien un maximum (et non un minimum ou un point selle)

# Maximum de vraisemblance et cas Poisson

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

$$\iff \hat{\theta}^{\text{MLE}} = \arg \max_{\theta} \log L(\theta)$$

$$\iff (\log L)'(\hat{\theta}^{\text{MLE}}) = \frac{d}{d\theta} (\log L(\hat{\theta}^{\text{MLE}})) = 0$$

Dans le cas Poisson (cf. diapo n° 50):

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

---

<sup>(15)</sup> on admettra que c'est bien un maximum (et non un minimum ou un point selle)

# Maximum de vraisemblance et cas Poisson

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

$$\iff \hat{\theta}^{\text{MLE}} = \arg \max_{\theta} \log L(\theta)$$

$$\iff (\log L)'(\hat{\theta}^{\text{MLE}}) = \frac{d}{d\theta} (\log L(\hat{\theta}^{\text{MLE}})) = 0$$

Dans le cas Poisson (cf. diapo n° 50):

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

Cette dérivée s'annule en  $\hat{\theta}^{\text{MLE}} = \frac{x_1 + \cdots + x_n}{n}$  et alors <sup>(15)</sup>:

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

---

<sup>(15)</sup> on admettra que c'est bien un maximum (et non un minimum ou un point selle)



# 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.