

# Statistique

L. Rouvière

*laurent.rouviere@univ-rennes2.fr*

SEPTEMBRE 2020

## Présentation

- *Objectifs* : Comprendre le problème de la modélisation statistique et acquérir les premières notions fondamentales de la théorie de l'estimation.
- *Pré-requis* : théorie des probabilités, variables aléatoires discrètes et continues.
- *Enseignant* : Laurent Rouvière *laurent.rouviere@univ-rennes2.fr*
  - **Recherche** : statistique non paramétrique, apprentissage statistique
  - **Enseignements** : statistique et probabilités (Université, école d'ingénieur et de commerce, formation continue).
  - **Consulting** : énergie, finance, marketing, sport.

## Programme

- 40h : 20h *CM* + 20h *TD*.
- *Matériel* : slides + feuilles d'exercices. Disponible à l'url : [https://lrouviere.github.io/doc\\_cours/](https://lrouviere.github.io/doc_cours/)
- 5 parties :
  1. La modélisation
  2. Théorie de l'estimation
  3. Convergences stochastiques
  4. Critères de performance asymptotique et estimation par intervalles
  5. Introduction à l'approche non paramétrique

## Table des matières

<b>I</b>	<b>La modélisation statistique</b>	<b>4</b>
<b>1</b>	<b>Un exemple de modèle</b>	<b>4</b>
<b>2</b>	<b>Quelques exemples de problèmes statistiques</b>	<b>6</b>
<b>3</b>	<b>Modèle statistique</b>	<b>9</b>
<b>4</b>	<b>Quelques rappels de probabilités</b>	<b>11</b>
4.1	Variable aléatoire réelle . . . . .	11
4.2	Vecteurs aléatoires . . . . .	14

<b>5</b>	<b>Bibliographie</b>	<b>15</b>
<b>II</b>	<b>Théorie de l'estimation</b>	<b>16</b>
<b>1</b>	<b>Modèle - estimateur</b>	<b>16</b>
<b>2</b>	<b>Biais, variance, risque quadratique</b>	<b>20</b>
<b>3</b>	<b>Quelques méthodes d'estimation</b>	<b>22</b>
3.1	La méthode des moments . . . . .	22
3.2	La méthode du maximum de vraisemblance . . . . .	23
<b>4</b>	<b>Information de Fisher</b>	<b>24</b>
<b>5</b>	<b>Annexe : La famille exponentielle</b>	<b>26</b>
<b>6</b>	<b>Bibliographie</b>	<b>27</b>
<b>III</b>	<b>Convergences stochastiques</b>	<b>28</b>
<b>1</b>	<b>Les différents modes de convergence</b>	<b>29</b>
1.1	Convergence presque sûre ou convergence forte . . . . .	29
1.2	La convergence en probabilité . . . . .	30
1.3	La convergence en moyenne d'ordre $p$ . . . . .	31
1.4	La convergence en loi . . . . .	32
<b>2</b>	<b>Lois des grands nombres et Théorème Central Limite</b>	<b>36</b>
2.1	Lois des grands nombres . . . . .	36
2.2	Le théorème central limite . . . . .	37
<b>3</b>	<b>Bibliographie</b>	<b>40</b>
<b>IV</b>	<b>Critères de performance asymptotiques, intervalles de confiance et estimation multivariée</b>	<b>41</b>
<b>1</b>	<b>Critères asymptotiques</b>	<b>41</b>
<b>2</b>	<b>Estimation par intervalles</b>	<b>42</b>
<b>3</b>	<b>Estimation multivariée</b>	<b>49</b>
3.1	Biais, variance, risque quadratique . . . . .	49
3.2	Critères asymptotiques . . . . .	50
3.3	Borne de Cramer-Rao . . . . .	51

<b>V</b>	<b>Approche paramétrique vs non paramétrique pour les modèles de densité et de régression</b>	<b>54</b>
<b>1</b>	<b>Le modèle de densité</b>	<b>57</b>
1.1	Approche paramétrique : le modèle Gaussien . . . . .	57
1.2	Approche non paramétrique : l'estimateur à noyau . . . . .	58
<b>2</b>	<b>Le modèle de régression</b>	<b>62</b>
2.1	Approche paramétrique : le modèle de régression linéaire . . . . .	64
2.2	Approche non paramétrique : l'estimateur à noyau . . . . .	67
<b>3</b>	<b>Bibliographie</b>	<b>71</b>

## Première partie

# La modélisation statistique

## 1 Un exemple de modèle

### Statistique (version Wikipedia)

La statistique est l'étude de la collecte de *données*, leur analyse, leur traitement, l'interprétation des résultats et leur présentation afin de rendre les données compréhensibles par tous.

### Conséquence

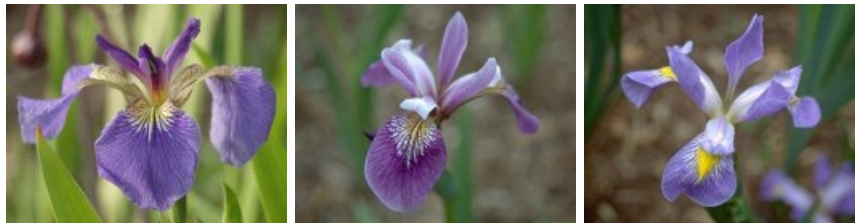
Plusieurs étapes :

1. Collecte des données
2. Analyse et vérification des données (*statistiques descriptives*)
3. Traitement (*modélisation*)
4. Interprétation des résultats (ou du *modèle*)
5. Présentation des résultats (*visualisation*)

### Un exemple célèbre : les iris de Fisher

### Question

Pour 3 espèces d'iris différentes, est-il possible d'*expliquer* (ou de *prédire*) l'appartenance à une des espèces connaissant les longueurs et largeurs de sépales ?

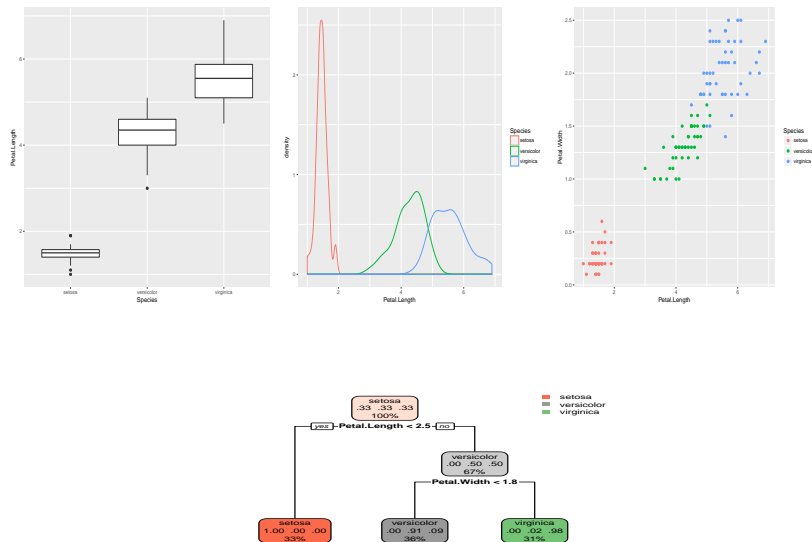


### Collecte des données

— On a mesuré sur  $n = 150$  iris les quantités d'intérêts.

```
> data(iris)
> head(iris)
  Sepal.Length Sepal.Width Petal.Length Petal.Width Species
1          5.1         3.5          1.4         0.2  setosa
2          4.9         3.0          1.4         0.2  setosa
3          4.7         3.2          1.3         0.2  setosa
4          4.6         3.1          1.5         0.2  setosa
5          5.0         3.6          1.4         0.2  setosa
6          5.4         3.9          1.7         0.4  setosa
```

```
> summary(iris)
  Sepal.Length      Sepal.Width      Petal.Length      Petal.Width      Species
Min.   :4.300   Min.   :2.000   Min.   :1.000   Min.   :0.100   setosa   :50
1st Qu.:5.100   1st Qu.:2.800   1st Qu.:1.600   1st Qu.:0.300   versicolor:50
Median :5.800   Median :3.000   Median :4.350   Median :1.300   virginica :50
Mean   :5.843   Mean   :3.057   Mean   :3.758   Mean   :1.199
3rd Qu.:6.400   3rd Qu.:3.300   3rd Qu.:5.100   3rd Qu.:1.800
Max.   :7.900   Max.   :4.400   Max.   :6.900   Max.   :2.500
```



## Statistiques descriptives

- Indicateurs *numériques* et *graphiques* permettant de mieux comprendre le problème.

```
> library(ggplot2)
> ggplot(iris)+aes(x=Species,y=Petal.Length)+geom_boxplot()
> ggplot(iris)+aes(x=Petal.Length,color=Species)+geom_density()
> ggplot(iris)+aes(x=Petal.Length,y=Petal.Width,color=Species)+geom_point()
```

## Modélisation

- *Modéliser* = créer un objet qui permette d'**expliquer l'espèce** à partir des 4 variables quantitatives.
- On utilise ici un *arbre de classification*

```
> library(rpart)
> model <- rpart(Species~.,data=iris)
```

- que l'on peut visualiser

```
> library(rpart.plot)
> rpart.plot(model)
```

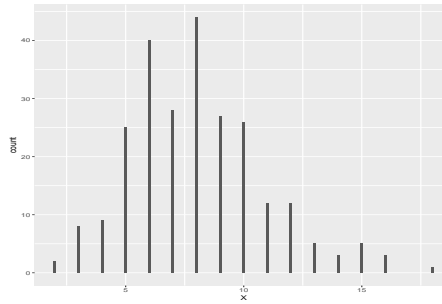
## Prévisions

- On dispose de 5 nouveaux iris sur lesquels on a mesuré les longueurs et largeurs de pétales et sépales.

```
> iris_prev
  Sepal.Length Sepal.Width Petal.Length Petal.Width
          5.0         3.6         1.4         0.2
          5.5         2.4         3.7         1.0
          5.8         2.7         5.1         1.9
          5.1         3.5         1.4         0.3
          6.3         2.9         5.6         1.8
```

- On souhaite connaître (*prédire, estimer...*) l'espèce de chacun.
- On utilise le *modèle* (l'arbre) pour faire ces prévisions.
- Prévisions des *probabilités* d'appartenance aux espèces :

```
> predict(model,newdata=iris_prev)
  setosa versicolor virginica
      1      0.000      0.000
```



0	0.907	0.093
0	0.022	0.978
1	0.000	0.000
0	0.022	0.978

— Prévisions des *espèces* :

```
> predict(model,newdata=iris_prev,type="class")
      setosa versicolor virginica  setosa  virginica
Levels: setosa versicolor virginica
```

— Chacune de ces étapes est *primordiale* pour le succès d'une étude statistique.

### Dans ce cours

- On va s'intéresser à la phase de **modélisation mathématique** d'un problème.
- On supposera les données **collectées** (c'est en grande partie une affaire de praticien). Elles seront souvent notées  $x_1, \dots, x_n$ .
- Les phases d'**interprétation** et de **visualisation** des résultats seront abordées plus tard.

## 2 Quelques exemples de problèmes statistiques

### Nombre de voitures à un feu rouge

- Afin de mieux gérer la circulation, on s'intéresse au *nombre de voitures* à un feu rouge sur un créneau donné.
- *Expérience* : on compte le nombre de voitures dans la file d'attente à chaque fois que le feu passe au vert.
- On récolte  $n = 250$  observations

5 9 9 9 11 9
--------------

### Question

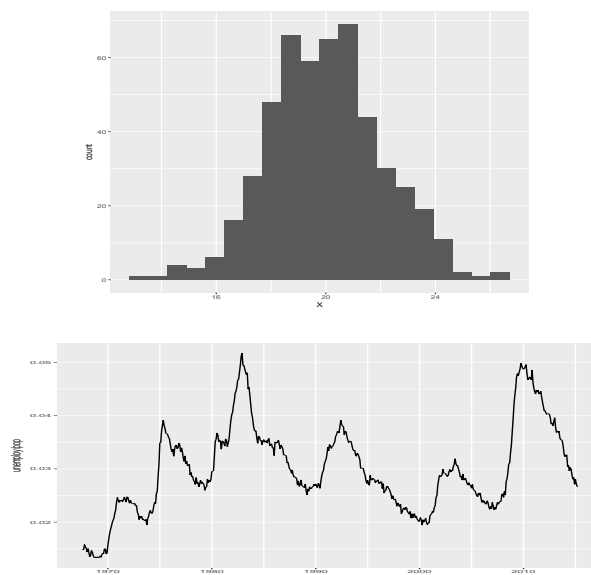
Comment **utiliser au mieux ces données** pour gérer le feu ?

### Quantité d'intérêt

- Il serait intéressant d'avoir de l'information sur la *loi de probabilité* du nombre de voitures arrêtées au feu à ce créneau.
- On dispose juste de mesures, cette loi est donc *inconnue*.
- Le *travail statistique* va donc consister à essayer de reconstruire au mieux cette loi (discrète) à partir des mesures effectuées.

### Durée d'un trajet

- J'ai une réunion à mon travail à 8h, à quelle heure dois-je partir pour "*avoir de grandes chances*" d'être à l'heure ?
- *Expérience* : je mesure la durée de trajet domicile/travail pendant plusieurs jours.
- Je récolte  $n = 100$  observations



20.87 22.12 20.90 21.33 17.73

### Question

Comment **utiliser au mieux ces données** pour gérer mon heure de départ ?

### Quantité d'intérêt

- Il serait intéressant d'avoir de l'information sur la *loi de probabilité* de la durée de trajet domicile/travail.
- On dispose juste de mesures, cette loi est donc *inconnue*.
- Le *travail statistique* va donc consister à essayer de reconstruire au mieux cette loi (continue) à partir des mesures effectuées.

### Séries temporelles

- On s'intéresse au *taux de chômage* d'une population entre deux dates  $t_0$  et  $t_1$ . On souhaite prédire le taux de chômage futur.
- *Expérience* : on mesure le taux de chômage entre les deux dates

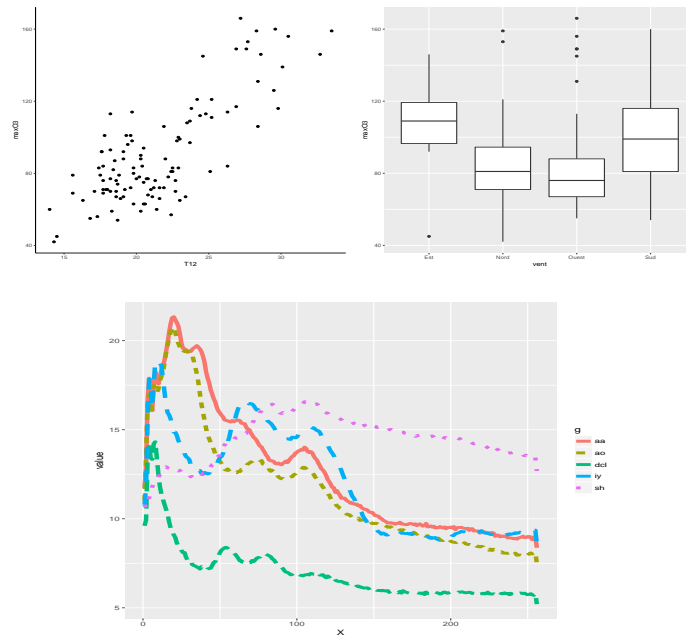
```
> head(economics)
# A tibble: 6 x 6
  date       pce    pop psavert uempmed unemploy
<date> <dbl> <int> <dbl> <dbl> <int>
1 1967-07-01 507.4 198712  12.5   4.5   2944
2 1967-08-01 510.5 198911  12.5   4.7   2945
3 1967-09-01 516.3 199113  11.7   4.6   2958
4 1967-10-01 512.9 199311  12.5   4.9   3143
5 1967-11-01 518.1 199498  12.5   4.7   3066
6 1967-12-01 525.8 199657  12.1   4.8   3018
```

### Question

Comment **utiliser au mieux ces données** pour prédire le taux de chômage en 2012 ?

### Quantité d'intérêt

- Il serait intéressant d'avoir de l'information sur la *loi de probabilité* du taux de chômage à l'instant  $t$  sachant le taux de chômage avant  $t$ .
- On dispose juste de mesures, cette loi est donc *inconnue*.
- Le *travail statistique* va donc consister à essayer de reconstruire au mieux cette loi (continue) à partir des mesures effectuées.



## Prévision ozone

- On s'intéresse à la *prévision de la concentration en ozone* dans l'air.
- *Expérience* : on mesure la **concentration en ozone** dans l'air ainsi d'**autres variable** (météo) qui pourraient potentiellement expliquer cette quantité.

```
> head(ozone)
      maxO3   T9  T12  T15 Ne9 Ne12 Ne15   Vx9   Vx12   Vx15 maxO3v vent pluie
20010601   87 15.6 18.5 18.4   4   4   8  0.6946 -1.7101 -0.6946   84 Nord  Sec
20010602   82 17.0 18.4 17.7   5   5   7 -4.3301 -4.0000 -3.0000   87 Nord  Sec
20010603   92 15.3 17.6 19.5   2   5   4  2.9544  1.8794  0.5209   82 Est   Sec
20010604  114 16.2 19.7 22.5   1   1   0  0.9848  0.3473 -0.1736   92 Nord  Sec
20010605   94 17.4 20.5 20.4   8   8   7 -0.5000 -2.9544 -4.3301  114 Ouest Sec
20010606   80 17.7 19.8 18.3   6   6   7 -5.6382 -5.0000 -6.0000   94 Ouest Pluie
```

```
> ggplot(ozone)+aes(x=T12,y=maxO3)+geom_point()
> ggplot(ozone)+aes(x=vent,y=maxO3)+geom_boxplot()
```

## Question

Comment **utiliser au mieux ces données** pour prédire la concentration en ozone **sachant les variables météo** ?

## Quantité d'intérêt

- Il serait intéressant d'avoir de l'information sur la *loi conditionnelle de probabilité* de la concentration en ozone sachant les variables météo.

## Reconnaissance de la voix

- On souhaite développer une procédure *automatique* permettant de reconnaître un son.
- *Expérience* : on prononce 5 sons un certain nombre de fois et on considère la courbe temporelle associé au son dans la base de Fourier.
- On dispose de  $n = 4509$  courbes, chacune étant associée à un son.

## Question

Comment **utiliser au mieux ces données** pour identifier un son à partir d'une courbe ?

## Quantité d'intérêt



- Il serait intéressant d'avoir de l'information sur la *loi conditionnelle de probabilité* de la variable son sachant la courbe.

## Bilan

- Pour chacun de ces problèmes on cherche à reconstruire (ou *estimer*) des probabilités (ou plus généralement des *lois de probabilité*).
- Les probabilités sont cependant *différentes* : la nature des quantités qui interviennent diffèrent
  - **discrètes** (voitures)
  - **continues** (durée de trajet)
  - **conditionnelles** (ozone, phonèmes)
- Les *objets mesurés* sont également de nature *différente* (entiers, réel, vecteurs, courbes...).

## Conséquence importante

Il va être primordial d'introduire un formalisme (*mathématique*) précis pour représenter (*modéliser*) ces problèmes.

- Ces problèmes peuvent être appréhendés à l'aide d'un *modèle statistique*.

## Modèle statistique

- **Définition avec des mots** : vision simplifiée de la réalité.
- **Définition mathématique** : triplet  $(F, \mathcal{H}, \{P, P \in \mathcal{P}\})$  où
  - $F$  est un ensemble (l'espace des observations)
  - $\mathcal{H}$  est une tribu sur  $F$
  - $\{P, P \in \mathcal{P}\}$  est une famille de lois de probabilité.

## Question importante

Quel est le *lien* entre ces deux définitions ?

## 3 Modèle statistique

- On suppose que des *données* ont été collectées.
- Ces données sont le résultat d'une *expérience répétée  $n$  fois*.
- On va les noter  $x_1, \dots, x_n$ .

## Exemple des durées de trajet

- Données :

20.87 22.12 20.90 21.33 17.73
-------------------------------

- $x_1 = 20.87, x_2 = 22.12 \dots$

## Hasard, aléa...

## Question

- Sur les  $n = 100$  trajet, on obtient une *moyenne* de 20.02 minutes.
- Peut-on en *conclure* que le durée moyenne du trajet domicile/travail est de 20.02 minutes ?
- Le résultat dépend des *conditions* de l'expérience.
- Si on *re-mesure 100 fois le trajet*, il est fort possible qu'on n'obtienne *pas la même durée moyenne*.

## Conséquence

- Nécessité de prendre en compte que le résultat observé **dépend des** conditions expérimentales.
- Ces dernières vont être **difficiles à caractériser précisément**.
- On dit souvent que le **hasard ou l'aléa** intervient dans ces conditions.

## Variable aléatoire

### Un outil spécifique

L'outil mathématique permettant de prendre en compte l'aléa dans l'expérience est la **variable aléatoire**.

### Définition

Une *variable aléatoire réelle* (v.a.r.) est une application  $X : \Omega \rightarrow \mathbb{R}$  et une *réalisation de X* est une valeur  $X(\omega)$  pour une éventualité  $\omega \in \Omega$ .

- *Remarque* : la définition d'une v.a. est étrange et ne présente un intérêt que si on comprend son **utilité dans la modélisation**.

### V.a. et modélisation

- $x_1, \dots, x_n$  représentent le *résultat de l'expérience*. On suppose que  $x_i \in \mathbb{R}, i = 1, \dots, n$ .
- Pour prendre en compte l'aléa de l'expérience, on va considérer des variables aléatoires *réelles* (v.a.r.).

### Lien observation/v.a.r.

Les  $x_i$  sont des réalisations de v.a.r.  $X_i$ . C'est-à-dire

$$\forall i = 1, \dots, n \exists \omega_i \in \Omega \quad \text{tel que } x_i = X_i(\omega_i).$$

- On suppose donc qu'il existe  $n$  v.a.r.  $X_1, \dots, X_n$  et des éléments  $\omega_1, \dots, \omega_n$  tels que

$$x_1 = X_1(\omega_1), \dots, x_n = X_n(\omega_n).$$

### Question

Que représentent les  $\omega_i$  ?

### Réponse

- $\omega_i$  représente les *conditions expérimentales* associées à la  $i^{\text{e}}$  mesure, c'est-à-dire toutes les conditions qui permettent "d'expliquer" qu'on a obtenu  $x_i$ .
- Cette quantité n'est généralement *pas caractérisable* (on sait qu'elle existe mais on ne peut pas en dire plus).

### Exemple : durée de trajet

- $x_1 = 20.87, x_2 = 22.12, x_3 = 20.90, x_4 = 21.33, x_5 = 17.73, \dots$
- $X_1, \dots, X_n$  définies sur  $(\Omega, \mathcal{A})$ ,  $n$  v.a.r. telles que  $X_i(\omega_i) = x_i$ .

### Interprétation

- On dit que  $X_i$  est la v.a.r. représentant le  **$i^{\text{e}}$  temps de trajet**.
- L'ensemble  $\Omega$  contient **toutes les conditions expérimentales** possibles... C'est-à-dire tout ce qui peut se produire sur le trajet (feux, passant qui traverse, vitesse à laquelle on roule...).
- $\omega_i$  correspondant à ce qui **s'est produit sur le  $i^{\text{e}}$  trajet**.
- Par exemple  $\omega_1$  représente tout ce qui s'est passé sur le trajet permettant **d'expliquer qu'on a mis 20.87 minutes**.

### Remarque

On voit bien sur cet exemple qu'il est *difficile de caractériser mathématiquement*  $\Omega$  et les  $\omega_i, i = 1, \dots, n$ .

## Récapitulatif

- $n$  observations  $x_1, \dots, x_n$  telles que  $x_i \in \mathbb{R}$ .
- Les  $n$  valeurs observées  $x_1, \dots, x_n$  sont des *réalisations de variables aléatoires*  $X_1, \dots, X_n$  à valeurs dans  $\mathbb{R}$ .

### Attention

$X_i$  est une *variable aléatoire*, c'est-à-dire une fonction, et  $x_i$  est une *réalisation* de cette variable, c'est-à-dire une quantité déterministe.

### Remarque

- Les v.a.  $X_1, \dots, X_n$  n'ont **pas forcément un grand intérêt** dans la modélisation.
- La quantité qui va nous intéresser est la **loi de probabilité** associée à ces v.a.
- C'est cette loi qui nous permettra d'**apporter des réponses** au problème posé.

## Loi de probabilité

### Loi de probabilité

La *loi de probabilité* d'une v.a.r. est représentée par les probabilités  $\mathbf{P}(X \in [a, b])$  avec  $a \leq b$ .

### Intérêt

- La loi de probabilité permet de **mesurer tous les évènements** dans l'espace d'arrivée.
- C'est elle qui va nous intéresser pour **comprendre le phénomène** qui nous intéresse.

## 4 Quelques rappels de probabilités

### 4.1 Variable aléatoire réelle

#### Fonction de répartition

- La loi de probabilité telle qu'elle est définie précédemment n'est *pas facile à manipuler*.
- Nécessité de trouver des outils mathématiques qui permettent de la *caractériser ou de l'identifier*.

#### Définition

Soit  $X$  une v.a.r. On appelle *fonction de répartition* de  $X$  la fonction  $F_X : \mathbb{R} \rightarrow [0, 1]$  définie par

$$F_X(x) = \mathbf{P}(X \leq x).$$

#### Propriété

La fonction de répartition  $F_X$  d'une v.a.r.  $X$  satisfait les **propriétés suivantes** :

1.  $\forall x \in \mathbb{R}, 0 \leq F_X(x) \leq 1$  ;
2.  $F_X$  est une fonction croissante, continue à droite en tout point  $x \in \mathbb{R}$  ;
3.  $\lim_{x \rightarrow -\infty} F_X(x) = 0$  et  $\lim_{x \rightarrow +\infty} F_X(x) = 1$ .

#### Propriété

La fonction de répartition **caractérise la loi de probabilité** d'une variable aléatoire réelle.

- $F_X$  permet de *caractériser la loi de n'importe quelle v.a.r.*
- Il existe d'*autres outils* pour caractériser les lois qui peuvent dépendre de la **nature de la variable**.
  - *Cas discret* : fonction de masse.
  - *Cas continue* : densité.

## Cas discret

### Définition

- On dit qu'une v.a.r  $X$  est *discrète* si son support  $\mathcal{S}_X$  est fini ou dénombrable.
- La *fonction de masse* définie par

$$\begin{aligned}\pi_X : \mathcal{S}_X &\rightarrow [0, 1] \\ x &\mapsto \mathbf{P}(X = x)\end{aligned}$$

- *Exemples* : Bernoulli, binomiale, Poisson...

### Propriété

La fonction de masse **caractérise la loi de probabilité d'une v.a.r discrète**.

## Cas continu

- Généralement pour des v.a.r qui prennent leurs valeurs sur un *intervalle* de  $\mathbb{R}$  ou une *réunion d'intervalles* de  $\mathbb{R}$ .

### Définition

Une v.a.r  $X$  est dite de *loi à densité* si il existe une **densité**  $f_X : \mathbb{R} \rightarrow \mathbb{R}^+$  telle que pour tous  $a, b$  avec  $a \leq b$  on a

$$\mathbf{P}(a \leq X \leq b) = \int_a^b f_X(x) dx.$$

- *Exemples* : Gaussienne, exponentielle...

### Propriété

La densité **caractérise la loi de probabilité d'une v.a.r continue**.

## Quelques propriétés

- Toute fonction  $f$  positive, continue et qui intègre à 1 est une *densité*.
- Lien fonction de répartition densité :  $f_X = F'_X$  sur l'ensemble où  $F_X$  est dérivable.
- Une v.a.r n'est pas forcément discrète ou continue, ça peut aussi être un *mélange des deux*...

## Espérance d'un v.a.r.

### Définition

Soit  $X$  une v.a.r.  $\mathbf{P}$ -intégrable. On appelle *espérance mathématique* de  $X$ , notée  $\mathbf{E}[X]$  l'intégrale de  $X$  par rapport à  $\mathbf{P}$  :

$$\mathbf{E}[X] = \int X d\mathbf{P} = \int_{\Omega} X(\omega) d\mathbf{P}(\omega).$$

### Interprétation

- L'espérance revient à **intégrer les valeurs de la v.a.r.  $X$**  pour chaque événement  $\omega$  **pondéré** par la mesure de probabilité  $\mathbf{P}$ .
- D'où l'interprétation de **valeur moyenne** prise par  $X$ .
- *Problème* : l'espérance dépend de  $\Omega$  que l'on ne peut généralement pas caractériser !
- Le **théorème de transfert** permet de pallier à cette difficulté.

## Calcul en pratique

- On déduit du théorème de transfert un moyen "simple" pour *calculer l'espérance* dans les cas **discret et continu**.

### Propriété

- **Cas discret** :

$$\mathbf{E}[X] = \sum_{x \in \mathcal{S}_X} x \pi_x(x).$$

- **Cas continu** :

$$\mathbf{E}[X] = \int_{\mathbb{R}} x f_X(x) d\lambda(x).$$

$\Rightarrow$  l'espérance s'obtient en calculant une **somme** ou une **intégrale**.

## Variance

### Définition

- Le moment centré d'ordre 2 de  $X$  est appelé la *variance* de  $X$  et est noté  $\mathbf{V}[X]$  :

$$\mathbf{V}[X] = \mathbf{E}[(X - E[X])^2] = \mathbf{E}[X^2] - (\mathbf{E}[X])^2.$$

- Sa racine carrée positive est appelée l'*écart-type* de  $X$ , noté  $\sigma[X]$ .

### Interprétation

- La variance est un réel **positif**.
- Elle mesure l'écart entre les valeurs prises par  $X$  et l'espérance (moyenne) de  $X \Rightarrow$  interprétation en terme de **dispersion**.

## Exemples

1. Loi de Bernoulli  $\mathcal{B}(p)$  :  $\mathbf{V}[X] = p(1 - p)$  ;
2. Loi uniforme sur  $[0, 1]$  :  $\mathbf{V}[X] = 1/12$  ;
3. Loi uniforme sur  $[1/4, 3/4]$  :  $\mathbf{V}[X] = 1/48$ .

## Quelques propriétés

### Espérance

1.  $\forall (a, b) \in \mathbb{R}^2, \mathbf{E}[aX + b] = a\mathbf{E}[X] + b$  ;
2.  $\mathbf{E}[X_1 + X_2] = \mathbf{E}[X_1] + \mathbf{E}[X_2]$
3. *Jensen* : Soit  $X$  à valeurs dans  $]a, b[$  et  $\varphi$  une fonction réelle convexe sur  $]a, b[$

$$\varphi(\mathbf{E}[X]) \leq \mathbf{E}[\varphi(X)].$$

### Variance

1.  $\forall \alpha \in \mathbb{R}, \mathbf{V}[\alpha X] = \alpha^2 \mathbf{V}[X]$  ;
2.  $\forall a \in \mathbb{R}, \mathbf{V}[a + X] = \mathbf{V}[X]$  ;
3.  $\mathbf{V}[X] = 0$  si et seulement si  $X$  est une v.a.r. presque sûrement constante ( $X = \mathbf{E}[X]$  p.s.).

## Inégalités sur les moments

### Markov

Si  $X$  est une v.a.r. positive, on a pour tout réel  $a > 0$

$$\mathbf{P}(X \geq a) \leq \frac{\mathbf{E}[X]}{a}.$$

### Bienaymé-Chebychev

Si  $\mathbf{E}[X^2] < +\infty$ , alors on a pour tout réel  $a > 0$

$$\mathbf{P}(|X - \mathbf{E}[X]| > a) \leq \frac{\mathbf{V}[X]}{a^2}.$$

## 4.2 Vecteurs aléatoires

— On se restreindra à la notion de *couple aléatoire*.

### Définitions

— Un *couple* de v.a.r. est une application :

$$(X, Y) : \Omega \rightarrow \mathbb{R}^2 \\ \omega \mapsto (X(\omega), Y(\omega))$$

— La *loi* de  $(X, Y)$  est représentée par les probabilités

$$\mathbf{P}((X, Y) \in [a, b] \times [c, d]) = \mathbf{P}(X \in [a, b] \text{ et } Y \in [c, d])$$

pour tous  $a \leq b$  et  $c \leq d$ .

— Les v.a.r.  $X$  et  $Y$  sont les *marginale*s du couple  $(X, Y)$ .

— Les notions vues pour les v.a.r. se *généralisent* aux couples aléatoires.

### Exemple

— **Fonction de répartition** :

$$F_{X,Y}(x, y) = \mathbf{P}(X \leq x, Y \leq y).$$

— **Densité** (si elle existe) : fonction  $f_{X,Y} : \mathbb{R}^2 \rightarrow \mathbb{R}^+$  telle que

$$\mathbf{P}((X, Y) \in [a, b] \times [c, d]) = \int_a^b \int_c^d f_{X,Y}(x, y) \, dy \, dx.$$

— **Densités marginales** (si elles existent) :

$$F_X(x) = \int_{\mathbb{R}} f_{X,Y}(x, y) \, dy \quad \text{et} \quad F_Y(y) = \int_{\mathbb{R}} f_{X,Y}(x, y) \, dx.$$

## Calcul d'espérance

— *Question* : étant donné un couple  $(X, Y)$  et une fonction  $g : \mathbb{R}^2 \rightarrow \mathbb{R}$ , que vaut  $\mathbf{E}[g(X, Y)]$  ?

### Théorème de transfert

Si  $\int_{\mathbb{R}^2} |g(x, y)| f_{X,Y}(x, y) \, dx \, dy < +\infty$  alors  $g(X, Y)$  est intégrable et

$$\mathbf{E}[g(X, Y)] = \int_{\mathbb{R}^2} g(x, y) f_{X,Y}(x, y) \, dx \, dy.$$

— On déduit la *linéarité de l'espérance* : soient  $a$  et  $b$  dans  $\mathbb{R}$  alors

$$\mathbf{E}[aX + bY] = a\mathbf{E}[X] + b\mathbf{E}[Y].$$

## Covariance

### Définitions

— *Covariance* entre  $X$  et  $Y$  :

$$\mathbf{cov}(X, Y) = \mathbf{E}([X - \mathbf{E}[X]](Y - \mathbf{E}[Y])) = \mathbf{E}[XY] - \mathbf{E}[X]\mathbf{E}[Y].$$

— *Matrice de variance covariance* : matrice  $2 \times 2$

$$\Sigma_{X,Y} = \begin{pmatrix} \mathbf{V}[X] & \mathbf{cov}(X, Y) \\ \mathbf{cov}(Y, X) & \mathbf{V}[Y] \end{pmatrix}$$

### Propriétés

- $\mathbf{cov}(X, Y) = \mathbf{cov}(Y, X)$  ;
- $\mathbf{cov}(aX + b, Y) = a\mathbf{cov}(X, Y)$  ;
- $\mathbf{V}[aX + bY] = a^2\mathbf{V}[X] + b^2\mathbf{V}[Y] + 2ab\mathbf{cov}(X, Y)$ .

## Indépendance

### Définition

Soit  $(X, Y)$  un couple aléatoire.  $X$  et  $Y$  sont *indépendantes* si pour tous  $a \leq b$  et  $c \leq d$  on a

$$\mathbf{P}(a \leq X \leq b, c \leq Y \leq d) = \mathbf{P}(a \leq X \leq b)\mathbf{P}(c \leq Y \leq d).$$

### En pratique

Si  $(X, Y)$  admet pour densité  $f_{X,Y}$  alors  $X$  et  $Y$  sont indépendantes si et seulement si

$$F_{X,Y}(x, y) = f_X(x)f_Y(y) \quad \text{pour tous } x, y \in \mathbb{R}.$$

### Propriété

Soient  $X$  et  $Y$  2 v.a.r **indépendantes**. Alors

1.  $\mathbf{E}[XY] = \mathbf{E}[X]\mathbf{E}[Y]$  et donc  $\mathbf{cov}(X, Y) = 0$
2.  $\mathbf{V}[X + Y] = \mathbf{V}[X] + \mathbf{V}[Y]$ .

— *Attention* : les réciproques sont fausses !

## 5 Bibliographie

### Références

#### Bibliol

[Jacod and Protter, 2003] Jacod, J. and Protter, P. (2003). *L'essentiel en théorie des probabilités*. Cassini.

[Lejeune, 2004] Lejeune, M. (2004). *Statistique. La théorie et ses applications*. Springer.

[Rouvière, 2015] Rouvière, L. (2015). Probabilités générales. Polycopié de cours, <https://perso.univ-rennes2.fr/laurent.rouviere>.

## Deuxième partie

# Théorie de l'estimation

### Rappels

- $n$  observations  $x_1, \dots, x_n$ .
- Ces observations sont des *réalisations de variables aléatoires*  $X_1, \dots, X_n \implies \exists \omega_i$  tel que

$$X_i(\omega_i) = x_i, \quad i = 1, \dots, n.$$

### Hypothèse

- On va supposer que les variables  $X_i$  sont **indépendantes et de même loi** de probabilité (**inconnue**)  $\mathbf{P}$ .

### Le problème de l'estimation

Il consiste à trouver (*estimer*) la loi  $\mathbf{P}$  à partir de l'échantillon  $X_1, \dots, X_n$ .

## 1 Modèle - estimateur

- Poser un *modèle* revient à **supposer** que la loi de probabilité **inconnue**  $\mathbf{P}$  appartient à une famille de lois  $\mathcal{P}$ .

### Définition

On appelle **modèle statistique** tout *triplet*  $(\mathcal{H}, \mathcal{A}, \mathcal{P})$  où

- $\mathcal{H}$  est l'*espace des observations* (l'ensemble dans lequel les observations prennent valeurs) ;
- $\mathcal{A}$  est une *tribu* sur  $\mathcal{H}$  ;
- $\mathcal{P}$  est une *famille de probabilités* définies sur  $(\mathcal{H}, \mathcal{A})$ .

### Remarque

- $\mathcal{H}$  et  $\mathcal{A}$  ne sont généralement pas difficile à caractériser.
- Le statisticien ou le praticien doit par contre **choisir une famille de loi de probabilité** susceptible de contenir la loi inconnue  $\mathbf{P}$ .

### Exemple

- On souhaite tester l'efficacité d'un nouveau traitement à l'aide d'un essai clinique.
- On traite  $n = 100$  patients atteints de la pathologie.
- A l'issue de l'étude, 72 patients sont guéris.

### Modélisation

- On note  $x_i = 1$  si le  $i^{\text{ème}}$  patient a guéri, 0 sinon.
- On suppose que  $x_i$  est la réalisation d'une variable aléatoire  $X_i$  de loi de bernoulli de paramètre **inconnu**  $p \in [0, 1]$ .
- Si les individus sont choisis de manière indépendante et ont tous la même probabilité de guérir (ce qui peut revenir à dire qu'ils en sont au même stade de la pathologie), il est alors raisonnable de supposer que les variables aléatoires  $X_1, \dots, X_n$  sont **indépendantes**.



## Spécification du triplet

### Le triplet pour l'exemple

- $\mathcal{H}$  : pas le choix  $\mathcal{H} = \{0, 1\}$ .
- $\mathcal{A}$  : pas le choix  $\mathcal{A} =$  ensemble des parties de  $\{0, 1\}$ .
- $\mathcal{P} = \{\text{lois de Bernoulli de paramètre } p \in [0, 1]\} = \{B(p) : p \in [0, 1]\}$ .
- A travers ce *modèle*, on **suppose** que la variable aléatoire  $X_i$  qui représente la réaction du  $i^{\text{e}}$  patient au traitement suit une loi de Bernoulli de paramètre **inconnu**  $p \in [0, 1]$ .
- *Le problème statistique* : reconstruire ou **estimer ce paramètre** à l'aide de l'échantillon  $X_1, \dots, X_n$ .

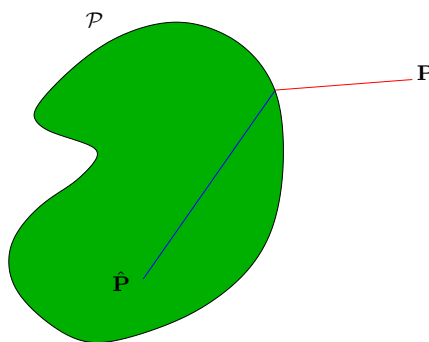
## Autres exemples

- *Exemple 1* : Traitement.
- *Exemple 2* : Nombre de voitures au feu rouge.
- *Exemple 3* : Durée de trajet domicile/travail.

	$\mathcal{H}$	$\mathcal{A}$	$\mathcal{P}$
Exemple 1	$\{0, 1\}$	$\mathcal{P}(\{0, 1\})$	$\{B(p), p \in [0, 1]\}$
Exemple 2	$\mathbb{N}$	$\mathcal{P}(\mathbb{N})$	$\{\mathcal{P}(\lambda), \lambda > 0\}$
Exemple 3	$\mathbb{R}$	$\mathcal{B}(\mathbb{R})$	$\{\mathcal{N}(\mu, \sigma^2), \mu \in \mathbb{R}, \sigma \in \mathbb{R}^+\}$

## 2 types d'erreur

- *Poser un modèle* = choisir une famille de lois  $\mathcal{P}$  candidates pour  $\mathbf{P}$ .



On distingue deux types d'erreurs :

- **Erreur d'estimation** : erreur commise par le choix d'une loi dans  $\mathcal{P}$  par rapport au meilleur choix.
- **Erreur d'approximation** : erreur commise par le choix de  $\mathcal{P}$ .
- Ces deux termes évoluent généralement en *sens inverse*.

### Exemple des durées de trajet

- $\mathcal{M}_1 : \mathcal{P} = \{\mathcal{N}(\mu, \sigma^2), \mu \in \mathbb{R}, \sigma \in \mathbb{R}^+\}$ .
- $\mathcal{M}_2 : \mathcal{P} = \{\text{Lois à densités continues}\}$ .

- $\mathcal{M}_2$  est **plus flexible** que  $\mathcal{M}_1$ . On a même  $\mathcal{M}_1 \subset \mathcal{M}_2$ .
- La théorie montrera qu'il est *plus difficile de bien estimer* dans  $\mathcal{M}_2$  que dans  $\mathcal{M}_1$ .

### Conséquence

- Le travail du statisticien consistera **toujours** à essayer de trouver le **meilleur compromis** entre ces deux erreurs.
- Dans ce cours, nous étudierons essentiellement l'**erreur d'estimation** dans les *modèles paramétriques*.

## Paramétrique versus non paramétrique

### Définition

- Si  $\mathcal{P} = \{\mathbf{P}_\theta, \theta \in \Theta\}$  où  $\Theta \in \mathbb{R}^d$  alors on parle de **modèle paramétrique** et  $\Theta$  est l'espace des paramètres.
- Si  $\mathcal{P} = \{\mathbf{P}, \mathbf{P} \in \mathcal{F}\}$  où  $\mathcal{F}$  est de dimension infinie, on parle de **modèle non paramétrique**.

### Exemple : modèle de densité

- $\mathcal{P} = \{\mathcal{N}(\mu, \sigma^2), (\mu, \sigma^2) \in \mathbb{R} \times \mathbb{R}^+\}$  est un modèle *paramétrique*.
- $\mathcal{P} = \{\text{densités } f \text{ 2 fois dérivables}\}$  est un modèle *non paramétrique*.

Le problème statistique sera d'*estimer*  $(\mu, \sigma^2)$  ou  $f$  à partir de l'échantillon  $X_1, \dots, X_n$ .

## Le problème de régression

- *Données* :  $(x_1, y_1), \dots, (x_n, y_n)$ . On veut expliquer les sorties  $y_i \in \mathbb{R}$  par les entrées  $x_i \in \mathbb{R}^p$ .
- Les données sont des *réalisations de variables aléatoires*  $(X_1, Y_1), \dots, (X_n, Y_n)$  telles qu'il existe une fonction **inconnue**  $m : \mathbb{R}^p \rightarrow \mathbb{R}$  vérifiant

$$Y_i = m(X_i) + \varepsilon_i, \quad i = 1, \dots, n$$

où les  $\varepsilon_i$  sont i.i.d de loi  $\mathcal{N}(0, \sigma^2)$ .

### Le problème statistique

Il consiste à estimer la *fonction inconnue*  $m$  à l'aide de l'échantillon  $(X_1, Y_1), \dots, (X_n, Y_n)$ .

## Régression paramétrique vs non paramétrique

### Modèle linéaire (paramétrique)

- On **suppose**  $m(x) = \beta_0 + \beta_1 x_1 + \dots + \beta_p x_p$ .
- Le problème est d'**estimer**  $\beta = (\beta_0, \dots, \beta_p) \in \mathbb{R}^{p+1}$  à l'aide de  $(X_1, Y_1), \dots, (X_n, Y_n)$ .
- Paramètre à estimer de **dimension finie**  $\implies$  modèle **paramétrique**.

### Un modèle non paramétrique

- On suppose que  $m : \mathbb{R}^p \rightarrow \mathbb{R}$  est une fonction continue.
- Le problème est d'estimer  $m$  à l'aide de  $(X_1, Y_1), \dots, (X_n, Y_n)$ .
- Paramètre à estimer de **dimension infinie**  $\implies$  modèle **non paramétrique**.

## Objectifs

### Estimer...

Etant donné un **modèle**  $(\mathcal{H}, \mathcal{A}, \mathcal{P})$  :

- Trouver des **procédures (automatiques)** permettant de sélectionner une loi  $\hat{\mathbf{P}}$  dans  $\mathcal{P}$  à partir d'un  $n$ -échantillon  $X_1, \dots, X_n$ .
- Etudier les **performances** des lois choisies.

### Paramétrique

- Dans la suite, on va considérer uniquement des *modèles paramétriques*  $\mathcal{P} = \{\mathbf{P}_\theta, \theta \in \Theta\}$  avec  $\Theta$  de dimension finie (typiquement  $\mathbb{R}^p$ ).
- Choisir une loi reviendra donc à *choisir un paramètre*  $\hat{\theta}$  à partir de l'échantillon  $X_1, \dots, X_n$ .
- Les modèles que nous allons considérer auront pour *espace d'observations* un ensemble dénombrable  $\Omega$  ou  $\mathbb{R}^d$  et seront munis des tribus  $\mathcal{P}(\Omega)$  ou  $\mathcal{B}(\mathbb{R}^d)$ .
- Dans la suite, on se donne un modèle  $\mathcal{M} = (\mathcal{H}, \mathcal{P} = \{\mathbf{P}_\theta, \theta \in \Theta\})$ .

### Echantillon

Un **échantillon** de taille  $n$  est une suite  $X_1, \dots, X_n$  de  $n$  **variables aléatoires indépendantes et de même loi**  $\mathbf{P}_\theta$ , pour  $\theta \in \Theta$ .

### Identifiabilité

- Si  $\theta \mapsto \mathbf{P}_\theta$  est injective, le modèle est dit *identifiable*.
- L'identifiabilité implique
  - 2 paramètres *différents* correspondent à deux lois *différentes*.
  - 2 paramètres **identiques** correspondent à deux lois **identiques**.
- Elle permet donc d'identifier une loi à un *unique* paramètre et est *capitale* pour savoir ce que l'on doit estimer.

## La démarche statistique

1. On récolte  $n$  *observations* ( $n$  valeurs)  $x_1, \dots, x_n$  qui sont les résultats de  $n$  expériences aléatoires indépendantes.
2. **Modélisation** : on suppose que les  $n$  valeurs sont des *réalisations de  $n$  variables aléatoires indépendantes*  $X_1, \dots, X_n$  et de même loi  $\mathbf{P}_\theta$ . Ce qui nous amène à définir le modèle  $\mathcal{M} = (\mathcal{H}, \{\mathbf{P}_\theta\}, \theta \in \Theta)$ .
3. **Estimation** : chercher dans le modèle une loi  $\mathbf{P}_{\hat{\theta}}$  qui soit *la plus proche possible* de  $\mathbf{P}_\theta \implies$  chercher un **estimateur**  $\hat{\theta}$  de  $\theta$ .

## Estimateurs

### Définitions

- Une *statistique* est une application (mesurable) définie sur  $\mathcal{H}^n$ .
- Un *estimateur* (de  $\theta$ ) est une fonction (mesurable) de  $(X_1, \dots, X_n)$  indépendante de  $\theta$  à valeurs dans un sur-ensemble de  $\Theta$ .

### Exemple 1 (modèle de Bernoulli)

Les variables aléatoires  $\hat{p}_1 = X_1$  et  $\hat{p}_2 = \bar{X}_n = \frac{1}{n} \sum_{i=1}^n X_i$  sont des estimateurs de  $p$ .

### Remarque

- Un estimateur  $\hat{\theta} = \hat{\theta}(X_1, \dots, X_n)$  : c'est une *variable aléatoire*.
- *Démarche* :
  1. Chercher le "meilleur" **estimateur**  $\hat{\theta}(X_1, \dots, X_n)$ .
  2. A la fin, calculer **l'estimation**  $\hat{\theta}(x_1, \dots, x_n)$  (renvoyé par le logiciel).

## Estimateurs vs estimation...

- Donner une bonne réponse au problème posé nécessite de se placer dans un premier temps dans un cadre *abstrait*.
- On cherche alors la *meilleure* fonction  $\hat{\theta}(X_1, \dots, X_n)$  vis à vis de *critères* à définir.
- Une fois cette fonction trouvée, il faut donner une *réponse* (qui ne doit pas être abstraite!)... On applique la fonction trouvée aux données observées  $\hat{\theta}(x_1, \dots, x_n)$ .

## Abus de notation

Malheureusement on note souvent de la *même façon* l'estimateur et l'estimation :

- on écrit  $\hat{\theta}$  pour l'**estimateur**  $\hat{\theta}(X_1, \dots, X_n)$  ;
- on écrit  $\hat{\theta}$  pour l'**estimation**  $\hat{\theta}(x_1, \dots, x_n)$  ;
- Il est donc nécessaire de faire soi-même la *distinction entre ces deux objets* lorsque on lit ou écrit  $\hat{\theta}$ .

## Exemple : réponse à un traitement

- Les données

$x_1$	$x_2$	$x_3$	$x_4$	$x_5$	$x_6$	$x_7$	$x_8$
1	0	0	0	1	0	1	0

- *Modèle* : les  $x_i$  sont des réalisations de v.a.  $X_i$  indépendantes et de loi de Bernoulli de paramètre  $p$  (inconnu).
- *Problème statistique* : estimer  $p$ .
- *Estimateur* :

$$\hat{p} = \hat{p}(X_1, \dots, X_n) = \frac{1}{n} \sum_{i=1}^n X_i.$$

- *Estimation* :

$$\hat{p} = \hat{p}(x_1, \dots, x_n) = \frac{1}{n} \sum_{i=1}^n x_i = \frac{3}{8}.$$

## 2 Biais, variance, risque quadratique

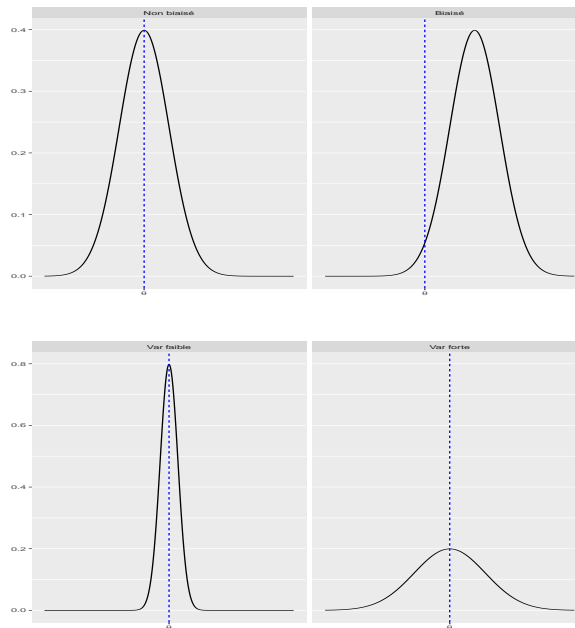
- $X_1, \dots, X_n$  i.i.d de loi  $\mathbf{P}_\theta$  avec  $\theta \in \Theta$  *inconnu*.
- On cherche un *estimateur*  $\hat{\theta} = \hat{\theta}(X_1, \dots, X_n)$ .
- Un estimateur est donc une **variable aléatoire**. Il va donc (le plus souvent) posséder
  - une loi de probabilité
  - une espérance
  - une variance...

## Espérance d'un estimateur

- On représente ci-dessous les lois de probabilité de 2 *estimateurs* de  $\theta$ .

## Commentaires

- L'estimateur de gauche semble être **préférable** à celui de droite.
- Sa loi de probabilité est en effet **centrée sur le paramètre inconnu**  $\implies \mathbf{E}[\hat{\theta}] \approx \theta$ .



## Biais d'un estimateur

- Dans la suite, pour un modèle de famille de loi  $\{\mathbf{P}_\theta, \theta \in \Theta\}$ , on désigne par  $\mathbf{E}$  et  $\mathbf{V}$  les variables sous la loi  $\mathbf{P}_\theta$ .

### Définition

Soit  $\hat{\theta}$  un estimateur d'ordre 1 (l'espérance existe).

1. Le **biais** de  $\hat{\theta}$  en  $\theta$  est  $\mathbf{E}(\hat{\theta}) - \theta$ .
2.  $\hat{\theta}$  est **sans biais** lorsque son biais est nul.
3.  $\hat{\theta}$  est **asymptotiquement sans biais** si  $\lim_{n \rightarrow \infty} \mathbf{E}(\hat{\theta}) = \theta$ .

### Exemple 1

Les estimateurs  $\hat{p}_1$  et  $\hat{p}_2$  sont *sans biais*.

## Variance d'un estimateur

- Mesurer le biais n'est *pas suffisant*, il faut également mesurer la *dispersion* des estimateurs.
- Les deux estimateurs sont **sans biais**.
- L'estimateur de gauche semble être **préférable** à celui de droite.
- Sa variance est plus faible :  $\implies \mathbf{V}[\hat{\theta}_1] \leq \mathbf{V}[\hat{\theta}_2]$ .

## Risque quadratique

- *Objectif* : trouver des estimateurs ayant un **biais** et une **variance faibles**.
- Le *risque quadratique* prend en compte simultanément ces deux critères.

### Définition

Soit  $\hat{\theta}$  un estimateur d'ordre 2.

1. Le *risque quadratique* de  $\hat{\theta}$  de  $\theta \in \mathbb{R}$  :

$$\mathcal{R}(\theta, \hat{\theta}) = \mathbf{E}(\hat{\theta} - \theta)^2$$

2. Soit  $\hat{\theta}'$  un autre estimateur d'ordre 2. On dit que  $\hat{\theta}$  est *préférable* à  $\hat{\theta}'$  si

$$\mathcal{R}(\theta, \hat{\theta}) \leq \mathcal{R}(\theta, \hat{\theta}') \quad \forall \theta \in \Theta.$$

### Exemple (Bernoulli)

$\hat{p}_2$  est *préférable* à  $\hat{p}_1$ .

## Estimateur VUMSB

### Propriété décomposition biais variance

Si  $\hat{\theta}$  est d'ordre 2, on a la décomposition

$$\mathcal{R}(\theta, \hat{\theta}) = (\mathbf{E}[\hat{\theta}] - \theta)^2 + \mathbf{E}(\hat{\theta} - \mathbf{E}[\hat{\theta}])^2 = b^2(\hat{\theta}) + \mathbf{V}[\hat{\theta}].$$

### Définition

Si  $\hat{\theta}$  est sans biais, on dit qu'il est de *variance uniformément minimum parmi les estimateurs sans biais (VUMSB)* si il est préférable à tout autre estimateur sans biais d'ordre 2 :

$$\hat{\theta} \text{ VUMSB} \iff \begin{cases} \mathbf{E}[\hat{\theta}] = \theta \\ \forall \tilde{\theta} \text{ tel que } \mathbf{E}[\tilde{\theta}] = \theta, \mathbf{V}[\hat{\theta}] \leq \mathbf{V}[\tilde{\theta}] \end{cases}$$

### Exemple

Dans le modèle de Bernoulli  $\mathcal{B}(p)$  nous montrerons que  $\hat{p}_2$  est *VUMSB*.

## 3 Quelques méthodes d'estimation

- $X_1, \dots, X_n$  i.i.d de loi  $\mathbf{P}_\theta$  avec  $\theta \in \Theta$  *inconnu*.
- Le *biais* et la *variance* permettent de mesurer la performance d'un estimateur  $\hat{\theta}$ .

### Question

Comment construire un estimateur (que l'on espère) *performant* ?

### Construction d'estimateurs

- Il existe des procédures *automatiques* qui permettent de construire des estimateurs.
- Nous présentons dans cette partie la méthode des *moments* et du *maximum de vraisemblance*.

### 3.1 La méthode des moments

- C'est une approche *intuitive* qui repose sur le fait que pour de nombreux modèles les moments *empiriques* doivent être *proches* des moments *théoriques*.
- En effet, on a d'après la *LFGN* que pour de nombreux modèles :

$$\bar{X}_n = \frac{1}{n} \sum_{i=1}^n X_i \approx \mathbf{E}[X_1].$$

### Définition

L'*estimateur des moments*  $\hat{\theta}_m$ , si il existe, est la solution en  $\theta$  de l'équation

$$\frac{1}{n} \sum_{i=1}^n X_i = \mathbf{E}[X_1].$$

### Remarque

- L'estimateur des moments *n'existe pas toujours*.
- Même lorsqu'il existe, il n'est *pas toujours performant* (voir TD).

Bernoulli $\mathcal{B}(p)$	$\hat{p}_m = \bar{X}_n$
Poisson $\mathcal{P}(\lambda)$	$\hat{\lambda}_m = \bar{X}_n$
Uniforme $\mathcal{U}_{[0,\theta]}$	$\hat{\theta}_m = 2\bar{X}_n$
Exponentielle $\mathcal{E}(\lambda)$	$\hat{\lambda}_m = 1/\bar{X}_n$

## 3.2 La méthode du maximum de vraisemblance

### Retour à l'exemple 1

- $X_1, \dots, X_n$  i.i.d.  $X_1 \sim \mathcal{B}(p)$ .
- $x_1, \dots, x_n$  réalisations de  $X_1, \dots, X_n$ .

#### Idée

1. La quantité  $L(x_1, \dots, x_n; p) = \mathbf{P}(X_1 = x_1, \dots, X_n = x_n)$  peut être vue comme une mesure de la **probabilité d'observer les données observées**.
2. Choisir le paramètre  $p$  qui **maximise** cette probabilité.

#### Notion de vraisemblance

- $L(x_1, \dots, x_n; p)$  est appelée **vraisemblance** (elle mesure la vraisemblance des réalisations  $x_1, \dots, x_n$  sous la loi  $\mathbf{P}_p$ ).
- L'approche consiste à choisir  $p$  qui "rend ces réalisations *les plus vraisemblables possible*".

### Vraisemblance

#### Cas discret

La **vraisemblance** du paramètre  $\theta$  pour la réalisation  $(x_1, \dots, x_n)$  est l'application  $L : \mathcal{H}^n \times \Theta$  définie par

$$L(x_1, \dots, x_n; \theta) = \mathbf{P}(X_1 = x_1, \dots, X_n = x_n) = \prod_{i=1}^n \mathbf{P}(X_i = x_i).$$

#### Cas absolument continu

Soit  $f(\cdot, \theta)$  la densité associée à  $\mathbf{P}_\theta$ . La **vraisemblance** du paramètre  $\theta$  pour la réalisation  $(x_1, \dots, x_n)$  est l'application  $L : \mathcal{H}^n \times \Theta$  définie par

$$L(x_1, \dots, x_n; \theta) = \prod_{i=1}^n f(x_i, \theta).$$

Bernoulli $\mathcal{B}(p)$	$\hat{p}_{MV} = \bar{X}_n$
Poisson $\mathcal{P}(\lambda)$	$\hat{\lambda}_{MV} = \bar{X}_n$
Uniforme $\mathcal{U}_{[0,\theta]}$	$\hat{\theta}_{MV} = \max_{1 \leq i \leq n} X_i$

## L'estimateur du maximum de vraisemblance

### Définition

Un **estimateur du maximum de vraisemblance (EMV)** est une statistique  $g$  qui maximise la vraisemblance, c'est-à-dire  $\forall (x_1, \dots, x_n) \in \mathcal{H}^n$

$$L(x_1, \dots, x_n; g(x_1, \dots, x_n)) = \sup_{\theta \in \Theta} L(x_1, \dots, x_n; \theta).$$

L'EMV s'écrit alors  $\hat{\theta} = g(X_1, \dots, X_n)$ .

### Exemples

## 4 Information de Fisher

- $X_1, \dots, X_n$  i.i.d de loi  $\mathbf{P}_\theta$  avec  $\theta$  **inconnu** dans  $\mathbb{R}$ .

### Objectif

Montrer que sous certaines hypothèses de régularité l'EMV est **asymptotiquement VUMSB** :

1.  $\hat{\theta}$  est asymptotiquement **sans biais**.
2. il existe une fonction  $r(n, \theta)$  telle que pour tout estimateur  $T$  sans biais de  $\theta$ , on a  $\mathbf{V}(T) \geq r(n, \theta)$ .
3. la **variance asymptotique** de l'EMV vaut  $r(n, \theta)$ .

### Information de Fisher

- Considérons pour l'instant 1 seule observation  $X$  de loi  $\mathbf{P}_\theta$ .
- On désigne par  $L_1(\cdot; \theta)$  la vraisemblance associée.

### Définition

Si elle existe (c'est-à-dire si la dérivée par rapport à  $\theta$  de la log-vraisemblance est de carré intégrable), l'*information de Fisher* associée à l'observation  $X$  est définie par :

$$I : \Theta \rightarrow \mathbb{R}^+ \\ \theta \mapsto \mathbf{E} \left[ \left( \frac{\partial}{\partial \theta} \log(L(X, \theta)) \right)^2 \right]$$

### Interprétation

L'**information de Fisher** peut s'interpréter comme :

- la **quantité d'information** apportée par l'observation  $X$  pour estimer le paramètre inconnu.
- une **mesure du pouvoir de discrimination** du modèle entre deux valeurs proches du paramètre  $\theta$  :
  - $I(\theta)$  grand : il sera "facile" d'identifier quel paramètre est le meilleur.
  - $I(\theta)$  petit : l'identification sera plus difficile.

### Propriété



- Si elle existe, l'information de Fisher vérifie

$$I(\theta) = -\mathbf{E} \left[ \frac{\partial^2}{\partial \theta^2} \log(L(X, \theta)) \right] = \mathbf{V} \left[ \frac{\partial}{\partial \theta} \log(L(X, \theta)) \right].$$

- On a de plus

$$I(\theta) \geq 0 \text{ et } I(\theta) = 0 \Leftrightarrow f(x, \theta) = f(x).$$

### Exemple

- On considère le modèle de *Bernoulli* :  $X \sim \mathcal{B}(p)$ .
- On a alors

$$L(x, p) = p^x(1-p)^{1-x}$$

et

$$\frac{\partial^2}{\partial p^2} \log(L(x, p)) = -\frac{x}{p^2} - \frac{1-x}{(1-p)^2}.$$

- D'où

$$I(p) = -\mathbf{E} \left[ -\frac{X}{p^2} - \frac{1-X}{(1-p)^2} \right] = \frac{1}{p(1-p)}.$$

### Fisher pour $n$ observations

- On considère maintenant  $n$  observations  $X_1, \dots, X_n$  de loi  $\mathbf{P}_\theta$ .
- On désigne par  $L_1(\cdot; \theta)$  la *vraisemblance* associée.

### Définition

Si elle existe (c'est-à-dire si la dérivée par rapport à  $\theta$  de la log-vraisemblance est de carré intégrable), l'*information de Fisher* associée à l'échantillon  $X_1, \dots, X_n$  est définie par :

$$I_n : \Theta \rightarrow \mathbb{R}^+ \\ \theta \mapsto \mathbf{E}_\theta \left[ \left( \frac{\partial}{\partial \theta} \log(L(X_1, \dots, X_n, \theta)) \right)^2 \right]$$

### Propriété d'additivité

L'information de Fisher est *additive* :

$$I_n(\theta) = nI(\theta).$$

### Modèle de *Bernoulli*

- $X_1, \dots, X_n$  i.i.d de loi de Bernoulli  $\mathcal{B}(p)$ .
- On a

$$I_n(p) = \frac{n}{p(1-p)}.$$

## Cramér-Rao

### Proposition

Soit  $\hat{\theta}$  un estimateur de  $\theta$  de biais  $b(\theta) = \mathbf{E}_{\theta}[\hat{\theta}] - \theta$ . Alors sous certaines hypothèses de régularité (voir [Guyader, 2017]), on a

$$\mathcal{R}(\theta, \hat{\theta}) = \mathbf{E}[(\hat{\theta} - \theta)^2] \geq b(\theta)^2 + \frac{(1 + b'(\theta))^2}{I_n(\theta)}.$$

### Corollaire : Inégalité de Cramér-Rao

On déduit que si  $\hat{\theta}$  est un estimateur **sans biais** de  $\theta$  alors

$$\mathbf{V}[\hat{\theta}] \geq \frac{1}{nI(\theta)}.$$

- La quantité  $\frac{1}{I_n(\theta)}$  est appelée *borne de Cramer-Rao*.
- Si un estimateur sans biais  $\hat{\theta}$  atteint la borne de Cramer-Rao, il est *VUMSB*. On dit aussi qu'il est *efficace*.

### Exemple : modèle de Bernoulli

- $X_1, \dots, X_n$  i.i.d. de loi de Bernoulli  $\mathcal{B}(p)$ .
- On a vu que  $I_n(p) = \frac{n}{p(1-p)}$ . La **borne de Cramér-Rao** vaut donc  $\frac{p(1-p)}{n}$ .
- On considère l'**estimateur**  $\hat{p} = \bar{X} = \frac{1}{n} \sum_{i=1}^n X_i$ .
- Il est facile de voir que

$$\mathbf{E}[\hat{p}] = p \quad \text{et} \quad \mathbf{V}[\hat{p}] = \frac{p(1-p)}{n}.$$

- On conclut donc que  $\hat{p}$  est **VUMSB** ou **efficace**.

## 5 Annexe : La famille exponentielle

### La classe exponentielle

#### Définition

Soit un **famille de lois** admettant des densités (cas continu) ou des fonctions de masse (cas discret)  $\{f(x, \theta), \theta \in \Theta \subseteq \mathbb{R}\}$ . On dit qu'elle appartient à la *famille ou classe exponentielle* de lois si  $f(x, \theta)$  peut s'écrire

$$f(x, \theta) = a(\theta)b(x) \exp(c(\theta)d(x))$$

pour tout  $x \in \mathbb{R}$ .

- La plupart des lois *standards* appartiennent à la **famille exponentielle**.

### Exemples

- Loi de *Bernoulli*  $\mathcal{B}(p)$  :

$$f(x, p) = p^x(1-p)^{1-x} = (1-p) \exp\left(x \log \frac{p}{1-p}\right).$$

- Loi de *Poisson*  $\mathcal{P}(\lambda)$  :

$$f(x, \lambda) = \frac{\lambda^x \exp(-\lambda)}{x!} = \exp(-\lambda) \frac{1}{x!} \exp(x \log \lambda).$$

### Mais aussi

Lois exponentielle, normale, gamma...

- Il est possible de montrer que les lois de la *famille exponentielle* possèdent de **bonnes propriétés**.
- Notamment pour l'*estimateur du maximum de vraisemblance*.
- Ces propriétés seront étudiés au S2, on pourra aussi consulter [Lejeune, 2004].

## 6 Bibliographie

### Références

#### Biblio2

- [Cadre and Vial, 2012] Cadre, B. and Vial, C. (2012). *Statistique mathématique, cours et exercices corrigés*. Ellipses.
- [Guyader, 2017] Guyader, A. (2017). Statistique mathématique. Polycopié de cours, [http ://www.lsta.upmc.fr/guyader/index.html](http://www.lsta.upmc.fr/guyader/index.html).
- [Lejeune, 2004] Lejeune, M. (2004). *Statistique. La théorie et ses applications*. Springer.

## Troisième partie

# Convergences stochastiques

### Motivations

- $X_1, \dots, X_n$  i.i.d. de loi  $\mathbf{P}_\theta$  avec  $\theta$  **inconnu** dans  $\Theta$ .
- *Un estimateur* : une fonction  $\hat{\theta}(X_1, \dots, X_n)$ .
- Le paramètre  $n$  représente souvent le *nombre de mesures* que l'on peut voir d'une certaine façon comme une **quantité d'information** à disposition pour **bien estimer**  $\theta$ .

### Conséquence

- Plus on a d'information, plus on doit être **précis**.
- Plus  $n$  est grand, plus  $\hat{\theta}(X_1, \dots, X_n)$  doit être **proche** de  $\theta$ .
- On a donc envie de traduire cela par  **$\lim_{n \rightarrow \infty} \hat{\theta}(X_1, \dots, X_n) = \theta$** .

### Problème

Que signifie cette *notion de limite* ?

### Retour vers les probabilités

- *Cadre* :  $(X_n)_n$  une suite de variables aléatoires réelles et  $X$  une variable aléatoire réelle.
- On cherche à définir la *notion de limite* :  $\lim_{n \rightarrow \infty} X_n = X$ .

### Première idée

- Une variable aléatoire réelle est une **fonction** qui va de  $\Omega$  dans  $\mathbb{R}$ .
- Utiliser les **modes de convergence** réservés aux fonctions.

### Exemple

On pourrait dire que  $(X_n)_n$  *converge simplement vers*  $X$  si pour tout  $\omega \in \Omega$  la *suite réelle*  $(X_n(\omega))_n$  converge vers  $X(\omega)$  :

$$\forall \omega \in \Omega, \quad \lim_{n \rightarrow \infty} X_n(\omega) = X(\omega).$$

- Bien que naturelle, cette définition est, de manière surprenante, à peu près *inutile* en probabilités.

### Exemple du pile ou face

- On joue  $n$  fois à pile ou face avec une pièce **équilibrée**.
- $X_i$  : v.a.r. qui vaut 1 si face au  $i^{\text{e}}$  jet, 0 sinon.  $X_i \sim \mathcal{B}(1/2)$ .
- Lorsque  $n$  est **grand**, la proportion de faces après  $n$  lancers "doit" tendre vers  $1/2$ . On a donc envie d'écrire

$$\lim_{n \rightarrow \infty} \frac{X_1(\omega) + \dots + X_n(\omega)}{n} = \frac{1}{2}.$$

- Ceci est pourtant **faux**, si on utilise la définition précédente : il suffit de considérer l'évènement  **$\omega_0 = \{f, f, f, f, f, \dots\}$**  (obtenir que des faces)

$$\lim_{n \rightarrow \infty} \frac{X_1(\omega_0) + \dots + X_n(\omega_0)}{n} = 1.$$

- Il est donc *nécessaire* de définir des modes de convergence *spécifiques aux v.a.*

# 1 Les différents modes de convergence

## 1.1 Convergence presque sûre ou convergence forte

### Exemple du pile ou face (retour)

- Il est facile de voir que l'évènement  $\omega_0$  est assez **invraisemblable** lorsque  $n$  est grand. En effet  $\mathbf{P}(\{\omega_0\}) = 1/2^n$ .
- On peut même montrer qu'il en est de même pour **tous les évènements où on n'a pas convergence**, on a donc

$$\mathbf{P} \left( \left\{ \omega : \lim_{n \rightarrow \infty} \frac{1}{n} \sum_{i=1}^n X_i(\omega) = \frac{1}{2} \right\} \right) = 1.$$

- **Conclusion** : l'ensemble des évènements où la convergence ne se produit pas est de **probabilité nulle**. On parle de convergence **presque sûre**.

### Définition

On dit que  $(X_n)_n$  converge *presque sûrement* vers une variable aléatoire  $X$  si l'ensemble  $N$  des  $\omega$  tels que la suite numérique  $(X_n(\omega))_n$  ne converge pas vers  $X(\omega)$  est négligeable (c'est-à-dire vérifie  $\mathbf{P}(N) = 0$ ). On note

$$\lim_{n \rightarrow \infty} X_n = X \quad \text{p.s.} \quad \text{ou} \quad X_n \xrightarrow{p.s.} X.$$

### Remarque

On peut aussi dire que  $X_n \xrightarrow{p.s.} X$  si et seulement si

$$\mathbf{P} \left( \left\{ \omega \in \Omega : \lim_{n \rightarrow \infty} X_n(\omega) \neq X(\omega) \right\} \right) = 0$$

ou encore

$$\mathbf{P} \left( \left\{ \omega \in \Omega : \lim_{n \rightarrow \infty} X_n(\omega) = X(\omega) \right\} \right) = 1.$$

### Proposition : opérations sur la cv ps

1. Si  $X_n \xrightarrow{p.s.} X$  et si  $\varphi : \mathbb{R} \rightarrow \mathbb{R}$  est une fonction **continue** sur  $\mathbb{R}$  alors  $\varphi(X_n) \xrightarrow{p.s.} \varphi(X)$ .
2. Si  $X_n \xrightarrow{p.s.} X$  et  $Y_n \xrightarrow{p.s.} Y$  alors
  - pour tout réels  $a$  et  $b$ ,  $aX_n + bY_n \xrightarrow{p.s.} aX + bY$  ;
  - $X_n Y_n \xrightarrow{p.s.} XY$ .
  - $X_n/Y_n \xrightarrow{p.s.} X/Y$  si  $\mathbf{P}(Y = 0) = 0$ .

### Conclusion

Les opérations classiques sur les limites sont *conservées* par la convergence presque sûre.

### Comment montrer une convergence ps

- On utilise *rarement la définition* pour montrer la convergence presque sûre. On a souvent recourt à l'un des *critères* suivants.

### Théorème

La suite de v.a.r.  $(X_n)_n$  converge presque sûrement vers  $X$  si et seulement si pour tout  $\varepsilon > 0$ ,

$$\lim_{n \rightarrow \infty} \mathbf{P}(\sup_{m \geq n} |X_m - X| > \varepsilon) = 0.$$

### Lemme de Borel-Cantelli

Si pour tout  $\varepsilon > 0$ ,

$$\sum_{n \in \mathbb{N}} \mathbf{P}(|X_n - X| > \varepsilon) < +\infty$$

alors  $X_n \xrightarrow{p.s.} X$ .

### Exemple

- $(X_n)_n$  suite de v.a.r. i.i.d telle que  $\mathbf{P}(X_n = 1) = \mathbf{P}(X_n = -1) = \frac{1}{2}$ .
- *Question* : est-ce que

$$\frac{1}{n^2} \sum_{i=1}^n X_i \xrightarrow{p.s.} 0 ?$$

- On a d'après **B.T.**

$$\mathbf{P} \left( \left| \frac{1}{n^2} \sum_{i=1}^n X_i \right| > \varepsilon \right) \leq \frac{1}{n^3 \varepsilon^2}.$$

- On a donc

$$\frac{1}{n^2} \sum_{i=1}^n X_i \xrightarrow{p.s.} 0.$$

## 1.2 La convergence en probabilité

### Définition

On dit que  $(X_n)_{n \in \mathbb{N}}$  converge en probabilité vers  $X$  si pour tout  $\varepsilon > 0$ , on a

$$\lim_{n \rightarrow \infty} \mathbf{P}(|X_n - X| > \varepsilon) = 0.$$

On note  $X_n \xrightarrow{\mathbf{P}} X$ .

### Exemple

- Soit  $X_1, \dots, X_n, n \geq 1$  des v.a.r. indépendantes telles que  $\mathbf{E}[X_n] = 0$  et  $\mathbf{V}(X_n) = \sigma^2$ . On note  $\bar{X}_n = \frac{1}{n} \sum_{i=1}^n X_i$ .
- D'après *Bienaymé-Tchebychev*, on a

$$\mathbf{P}(|\bar{X}_n| > \varepsilon) \leq \frac{1}{n^2 \varepsilon^2} \mathbf{V} \left( \sum_{i=1}^n X_i \right) = \frac{\sigma^2}{n \varepsilon^2}.$$

- On a donc  $\bar{X}_n \xrightarrow{\mathbf{P}} 0$ .

### Exemple

- Soit  $(X_n)_{n \in \mathbb{N}}$  une suite de variables aléatoires dont la loi est définie par

$$\mathbf{P}(X_n = \sqrt{n}) = \frac{1}{n} \quad \text{et} \quad \mathbf{P}(X_n = 0) = 1 - \frac{1}{n}.$$

- On a pour  $\varepsilon > 0$  fixé,

$$\begin{aligned} \mathbf{P}(|X_n| > \varepsilon) &= \mathbf{P}(|X_n| > \varepsilon \cap X_n = \sqrt{n}) + \mathbf{P}(|X_n| > \varepsilon \cap X_n = 0) \\ &= \mathbf{P}(|X_n| > \varepsilon \cap X_n = \sqrt{n}). \end{aligned}$$

- Or, pour  $n$  assez grand,  $\{|X_n| > \varepsilon\} = \{X_n = \sqrt{n}\}$ , donc

$$\lim_{n \rightarrow \infty} \mathbf{P}(|X_n| > \varepsilon) = \lim_{n \rightarrow \infty} 1/n = 0.$$

- On déduit  $X_n \xrightarrow{\mathbf{P}} 0$ .

- Les opérations sur les limites présentées pour la convergence presque sûre sont également vraies pour la convergence en probabilité.

**Proposition : opérations sur la cv en proba**

1. Si  $X_n \xrightarrow{\mathbf{P}} X$  et si  $\varphi : \mathbb{R} \rightarrow \mathbb{R}$  est une **fonction continue** sur  $\mathbb{R}$  alors  $\varphi(X_n) \xrightarrow{\mathbf{P}} \varphi(X)$ .
2. Si  $X_n \xrightarrow{\mathbf{P}} X$  et  $Y_n \xrightarrow{\mathbf{P}} Y$  alors
  - pour tout réels  $a$  et  $b$ ,  $aX_n + bY_n \xrightarrow{\mathbf{P}} aX + bY$ ;
  - $X_n Y_n \xrightarrow{\mathbf{P}} XY$ .
  - $X_n/Y_n \xrightarrow{\mathbf{P}} X/Y$  si  $\mathbf{P}(Y = 0) = 0$ .

**Théorème**

Si  $X_n \xrightarrow{p.s.} X$  alors  $X_n \xrightarrow{\mathbf{P}} X$ .

- *Attention* : la réciproque est **fausse** ! Une contre exemple est donné dans [Jacod and Protter, 2003], page 152.

### 1.3 La convergence en moyenne d'ordre $p$

**Définition**

Soit  $p > 0$ . On dit que  $(X_n)_{n \in \mathbb{N}}$  converge en moyenne d'ordre  $p$  (ou dans  $L_p$ ) vers  $X$  si les  $X_n$  et  $X$  sont dans  $L_p$  ( $\mathbf{E}[|X_n|^p] < +\infty$  et  $\mathbf{E}[|X|^p] < +\infty$ ), et si on a

$$\lim_{n \rightarrow \infty} \mathbf{E}[|X_n - X|^p] = 0.$$

On note  $X_n \xrightarrow{L_p} X$ .

- Les cas les plus importants sont  $p = 1$  (convergence en *moyenne*) et  $p = 2$  (convergence en *moyenne quadratique*).
- *Convergence en moyenne (dans  $L_1$ )* : si  $X_n \xrightarrow{L_1} X$ , alors

$$\lim_{n \rightarrow \infty} \mathbf{E}[X_n] = \mathbf{E}[X] \quad \text{et} \quad \lim_{n \rightarrow \infty} \mathbf{E}[|X_n|] = \mathbf{E}[|X|].$$

#### Convergence dans $L_2$

- Il est facile de voir que

$$\mathbf{E}[(X_n - a)^2] = (\mathbf{E}[X_n] - a)^2 + \mathbf{V}[X_n].$$

- On déduit

$$X_n \xrightarrow{L_2} a \iff \begin{cases} \lim_{n \rightarrow \infty} \mathbf{E}[X_n] = a \\ \lim_{n \rightarrow \infty} \mathbf{V}[X_n] = 0 \end{cases}$$

#### Application en statistique

Si  $\hat{\theta}_n \xrightarrow{L_2} \theta$  alors

- le **biais** de  $\hat{\theta}_n$  tend vers 0.
- la **variance** tend vers 0.
- On a d'après l'*inégalité de Jensen*

$$\mathbf{E}|X_n - X| = \mathbf{E}\sqrt{(X_n - X)^2} \leq \sqrt{\mathbf{E}|X_n - X|^2}.$$

- On déduit la propriété suivante.

**Propriété**

$$X_n \xrightarrow{L_2} X \implies X_n \xrightarrow{L_1} X.$$

**Théorème**

Si  $X_n \xrightarrow{L_p} X$  alors  $X_n \xrightarrow{\mathbf{P}} X$ .

- *Attention* : la réciproque est **fausse**!
- On peut comme *contre-exemple* utiliser pour  $p = 2$  la suite de v.a.r. de loi

$$\mathbf{P}(X_n = \sqrt{n}) = \frac{1}{n} \quad \text{et} \quad \mathbf{P}(X_n = 0) = 1 - \frac{1}{n}.$$

## 1.4 La convergence en loi

- Bien que différent, les trois modes de convergence vus précédemment sont de même nature et peuvent être abordés comme des *variantes de la convergence habituelle*.
- Il existe un autre mode de convergence, différent des précédents mais *très utile en probabilité* : la convergence **en loi**, ou convergence **faible** ou encore convergence **étroite**.
- Dans cette partie, nous donnons la *définition* ainsi que les *principales propriétés* de ce nouveau mode de convergence. Pour plus de détails, ainsi que pour les preuves des résultats, on pourra consulter [Jacod and Protter, 2003].

### L'idée

- La loi de  $X_n$  se **rapproche de la loi** de  $X$  lorsque  $n$  est grand.
- Définir la **convergence en loi** par quelque chose du genre

$$X_n \xrightarrow{\mathcal{L}} X \iff \begin{cases} \text{pour } n \text{ grand } \mathcal{L}(X_n) \approx \mathcal{L}(X) \\ \text{ou} \\ \forall A \in \mathcal{B}(\mathbb{R}), \lim_{n \rightarrow \infty} \mathbf{P}(X_n \in A) = \mathbf{P}(X \in A) \\ \text{ou} \\ \forall x \in \mathbb{R}, \lim_{n \rightarrow \infty} F_{X_n}(x) = F_X(x) \end{cases} \quad (1)$$

Mais...

Cette définition n'est cependant *pas totalement satisfaisante*.

### (Contre) exemple

- $(X_n)_n$  de loi uniforme sur  $] -1/n; 1/n[$  et  $X = 0$  p.s.

### Cv p.s., proba, $L_p$

- On a pour tout  $\varepsilon > 0$

$$\begin{aligned} \mathbf{P}(|X_n| > \varepsilon) &= 1 - \mathbf{P}(-\varepsilon < X_n < \varepsilon) \\ &= 1 - \frac{n}{2} \left[ \min\left(\frac{1}{n}, \varepsilon\right) - \max\left(-\frac{1}{n}, -\varepsilon\right) \right] \\ &= 0 \text{ pour } n \text{ assez grand.} \end{aligned}$$

- *Conclusion* :  $X_n \xrightarrow{\mathbf{P}} X$  (mais aussi p.s. et dans  $L_p$ ).

### Remarque

- Cependant

$$\begin{cases} \mathbf{P}(X_n \leq 0) = \frac{1}{2} \neq 1 = \mathbf{P}(X \leq 0) \\ \mathbf{P}(X_n > 0) = \frac{1}{2} \neq 0 = \mathbf{P}(X > 0) \end{cases}$$

- **Conséquence** :  $(X_n)_n$  ne converge pas en loi vers  $X$  au **sens de la définition (1)**.

### Remarque

- Pour tout intervalle  $[a, b]$  avec  $a \neq 0$  et  $b \neq 0$ , on a

$$\lim_{n \rightarrow \infty} \mathbf{P}(X_n \in [a, b]) = \mathbf{P}(X \in [a, b]).$$

- On a également pour  $x \neq 0$   $\lim_{n \rightarrow \infty} F_{X_n}(x) = F_X(x)$ .
- Les problèmes de la définition (1) se situent lorsque  $x = 0$ , c'est-à-dire en l'*unique point de discontinuité* de la fonction de répartition de  $F_X$ .



## Convergence en loi

### Définition

On dit que la suite  $(X_n)_{n \in \mathbb{N}}$  converge en loi vers  $X$  si, en tout point de continuité de  $F_X$ , on a  $\lim_{n \rightarrow \infty} F_{X_n}(x) = F(x)$ . On note  $X_n \xrightarrow{\mathcal{L}} X$ .

### Exemple

— Sur l'exemple précédent on a

$$F_{X_n}(x) = \begin{cases} 0 & \text{si } x \leq -1/n \\ n/2(x + 1/n) & \text{si } -1/n < x \leq 1/n \\ 1 & \text{si } x > 1/n. \end{cases}$$

— Ainsi,

$$\begin{cases} \lim_{n \rightarrow \infty} F_{X_n}(x) = 0 & \text{si } x < 0 \\ \lim_{n \rightarrow \infty} F_{X_n}(x) = 1 & \text{si } x > 0. \end{cases}$$

— Comme  $F_X$  est discontinue en 0, on conclut que  $X_n \not\xrightarrow{\mathcal{L}} X$ .

## Attention

### Remarque

- Les opérations conservées par les cv en probabilités et presque sûre ne le sont pas forcément par la convergence en loi !
- Par exemple,  $X_n \xrightarrow{\mathcal{L}} X$  n'implique pas
  - $\mathbf{P}(X_n \in A) \rightarrow \mathbf{P}(X \in A), \forall A$  (déjà vu) ;
  - $\mathbf{E}[X_n] \rightarrow \mathbf{E}[X]$ . Il suffit de prendre  $\mathcal{L}(X_n) = \frac{1}{n}\delta_{\{n\}} + (1 - 1/n)\delta_{\{0\}}$  ;
  - $X_n - X \xrightarrow{\mathcal{L}} 0$ . Il suffit de prendre  $\mathcal{L}(X) = N(0, 1)$  et  $X_n = (-1)^n X$ .

## Fonctions caractéristiques

— Très souvent utilisées pour montrer des convergences en loi.

### Définition

On appelle *fonction caractéristique* de  $X$  la fonction  $\varphi_X : \mathbb{R} \rightarrow \mathbb{C}$  définie comme la transformée de Fourier de sa loi de probabilité

$$\varphi_X(t) = \mathbf{E}[e^{itX}].$$

### Calcul en pratique

— Si  $X$  est discrète de support  $\mathcal{S}$  et de fonction de masse  $\pi_X$  alors

$$\varphi_X(t) = \sum_{x \in \mathcal{S}} \pi_X(x) e^{itx}.$$

— Si  $X$  est absolument continue de densité  $f_X$  alors

$$\varphi_X(t) = \int_{\mathbb{R}} e^{itx} f_X(x) dx.$$

Loi	Fonction caractéristique
Bernoulli $\mathcal{B}(p)$	$pe^{it} + (1-p)$
Binomiale $\mathcal{B}(n, p)$	$(pe^{it} + (1-p))^n$
Poisson $\mathcal{P}(\lambda)$	$e^{\lambda(e^{it}-1)}$
Géométrique $\mathcal{G}(p)$	$pe^{it}/(1-(1-p)e^{it})$
Uniforme $\mathcal{U}([-a, a])$	$\sin(at)/(at)$
Exponentielle $\xi(\lambda)$	$\lambda/(\lambda - it)$
Gaussienne $(m, \sigma^2)$	$e^{im}e^{-\sigma^2 t^2/2}$

## Exemple

### Proposition

1.  $\varphi_X$  est définie et continue pour tout nombre réel  $t$  ;
2.  $\varphi_X$  est bornée et  $\forall t \ |\varphi_X(t)| \leq 1$  ;
3.  $\forall (a, b) \in \mathbb{R}^2, \varphi_{aX+b}(t) = e^{ibt}\varphi_X(at)$  ;
4. Si la loi de  $X$  est **symétrique** alors  $\varphi_X$  est une fonction **réelle paire** ;
5.  $\varphi_X$  **caractérise** la loi de  $X$ .

### Proposition

Si  $X$  et  $Y$  sont deux v.a.r. **indépendantes** alors on a pour tout  $t$

$$\varphi_{X+Y}(t) = \varphi_X(t)\varphi_Y(t).$$

— **Exercice** : calculer la fonction caractéristique de la loi Binomiale  $B(n, p)$  en utilisant la propriété précédente.

## Fonction caractéristique et moments

— En plus de caractériser la loi, la fonction caractéristique permet de *calculer les moments* d'une v.a.r. (lorsqu'ils existent).

### Théorème

Si il existe  $n \in \mathbb{N}^*$  tel que  $\mathbf{E}[|X|^n] < \infty$ , alors

1.  $\varphi_X$  est continument dérivable jusqu'à l'ordre  $n$  inclu ;
2.  $\forall k = 0, 1, \dots, n, \varphi_X^{(k)}(0) = i^k \mathbf{E}[X^k]$ .
3. On a le développement

$$\varphi_X(t) = \sum_{k=0}^n \frac{(it)^k}{k!} \mathbf{E}[X^k] + o(|t|^n)$$

lorsque  $t \rightarrow 0$ .

## Retour à la convergence en loi

— La fonction caractéristique est très souvent utilisée pour *montrer des convergences en loi* grâce au théorème suivant.

### Théorème

Les trois assertions suivantes sont équivalentes :

1.  $X_n \xrightarrow{\mathcal{L}} X$  ;
2. Pour toute fonction  $f : \mathbb{R} \rightarrow \mathbb{R}$  continue bornée, on a  $\lim_{n \rightarrow \infty} \mathbf{E}[f(X_n)] = \mathbf{E}[f(X)]$ .
3. Pour tout  $t \in \mathbb{R}$ , on a  $\lim_{n \rightarrow \infty} \varphi_{X_n}(t) = \varphi_X(t)$ .

— La dernière assertion est une conséquence directe du **théorème de Paul Levy** (voir [Jacod and Protter, 2003]).

## Exemples

### Binomiale vers Poisson

1. Soit  $(X_n)_{n \in \mathbb{N}}$  une suite de variable aléatoire de loi  $\mathcal{B}(n, p_n)$  telle  $np_n \rightarrow \lambda$  lorsque  $n \rightarrow \infty$ . On a lorsque  $n \rightarrow \infty$  (faire un DL)

$$\varphi_{X_n}(t) = [p_n e^{it} + (1 - p_n)]^n \sim [1 + (e^{it} - 1)p_n]^n \rightarrow e^{\lambda(e^{it} - 1)}.$$

2. On déduit  $X_n \xrightarrow{\mathcal{L}} X$  avec  $X$  qui suit une loi de Poisson de paramètre  $\lambda$ . On note  $X_n \xrightarrow{\mathcal{L}} \mathcal{P}(\lambda)$ .

### Poisson vers normale

- Soit  $(X_n)_{n \in \mathbb{N}}$  une suite de variables aléatoires de loi de *Poisson* de paramètre  $\lambda_n$  avec  $\lambda_n \rightarrow \infty$  lorsque  $n \rightarrow \infty$ .
- De la même manière que dans l'exemple précédent on montre que

$$\frac{X_n - \lambda_n}{\sqrt{\lambda_n}} \xrightarrow{\mathcal{L}} \mathcal{N}(0, 1).$$

## Convergence en loi et densités

- Dans les cas discret et absolument continue, la convergence en loi peut également se montrer à partir des *fonctions de masse et de densité*.

### Théorème

1. Soit  $X_n$  et  $X$  des v.a.r. à valeurs dans un espace  $E$  **fini ou dénombrable**. Alors  $X_n \xrightarrow{\mathcal{L}} X$  si et seulement si

$$\forall j \in E, \quad \lim_{n \rightarrow \infty} \mathbf{P}(X_n = j) = \mathbf{P}(X = j).$$

2. Soit  $X_n$  et  $X$  des v.a.r. dont les lois admettent pour densité (par rapport à la mesure de Lebesgue)  $f_n$  et  $f$ . Si pour presque tout  $x$  de  $\mathbb{R}$  on a  $\lim_{n \rightarrow \infty} f_n(x) = f(x)$ , alors  $X_n \xrightarrow{\mathcal{L}} X$ .

- La convergence en loi est préservée par *certaines opérations arithmétiques*.

### Théorème (Slutsky)

Soit  $(X_n)_{n \in \mathbb{N}}$  et  $(Y_n)_{n \in \mathbb{N}}$  deux suites de v.a.r.,  $X$  une v.a.r. et  $a$  un **réel**. On a :

1. Si  $X_n \xrightarrow{\mathcal{L}} X$  et  $Y_n \xrightarrow{\mathcal{L}} a$  alors

$$X_n + Y_n \xrightarrow{\mathcal{L}} X + a, \quad X_n Y_n \xrightarrow{\mathcal{L}} aX \quad \text{et} \quad \frac{X_n}{Y_n} \xrightarrow{\mathcal{L}} \frac{X}{a} \quad (\text{si } a \neq 0).$$

2. Si  $g : \mathbb{R} \rightarrow \mathbb{R}$  est **continue** en tout point de  $\mathbb{R}$  alors  $g(X_n) \xrightarrow{\mathcal{L}} g(X)$ .

- **Attention** : les résultats ne sont plus vraies si  $Y_n$  converge vers une *variable aléatoire*  $Y$ .

## Relation entre les convergences

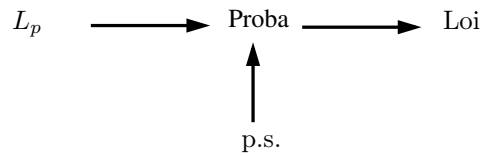
### Théorème

Si  $X_n \xrightarrow{\mathbf{P}} X$  alors  $X_n \xrightarrow{\mathcal{L}} X$ .

- *Réciproque fautive* : il suffit de prendre  $X \sim \mathcal{N}(0, 1)$  et  $X_n = (-1)^n X$ .
- La réciproque devient vraie lorsque  $X_n$  converge en loi vers une *variable constante*  $a$ . On a

$$X_n \xrightarrow{\mathcal{L}} a \iff X_n \xrightarrow{\mathbf{P}} a.$$

- On peut résumer les *relations entre les différents modes de convergence* par le diagramme suivant :



## 2 Lois des grands nombres et Théorème Central Limite

### Présentation

- $X_1, \dots, X_n$  i.i.d. admettant une espérance  $\mu = \mathbf{E}[X_1]$ .
- Intuitivement, lorsque  $n$  *augmente* la moyenne empirique

$$\bar{X}_n = \frac{1}{n} \sum_{i=1}^n X_i$$

doit se "*rapprocher*" de  $\mu$ .

- Les *lois des grands nombres* et le *théorème central limite* permettent de **préciser rigoureusement ce rapprochement**.

### 2.1 Lois des grands nombres

#### Un exemple

- Soit  $X_1, \dots, X_n$   $n$  v.a.r. indépendantes de loi Bernoulli de paramètre  $p$ .
- *Question* : est-ce que  $\bar{X}_n$  converge en probabilité vers  $p$  ?
- On a d'après **Bienaymé-Chebychev**  $\forall \varepsilon > 0$

$$\mathbf{P}(|\bar{X}_n - p| \geq \varepsilon) \leq \frac{p(1-p)}{n\varepsilon^2} \rightarrow 0 \quad \text{quand } n \rightarrow \infty.$$

- *Réponse* :  $\bar{X}_n \xrightarrow{\mathbf{P}} p$ .

#### Lois faibles et fortes

- Les lois des grands nombres permettent de généraliser ce type de résultats à d'autres lois que la loi de Bernoulli.
- On parle de lois **faibles** des grands nombres pour des convergences en probabilité. Pour des convergences presque sûre, on parlera de lois **fortes** des grands nombres.

### 2 lois faibles des grands nombres

#### Loi faible dans $L_1$

Soit  $(X_n)_{n \in \mathbb{N}}$  une suite de v.a.r. 2 à 2 *indépendantes*, de même loi et qui admettent une *espérance*. On note  $\mathbf{E}[X_1] = \mu$ . On a

$$\frac{1}{n} \sum_{i=1}^n X_i \xrightarrow{L_1} \mu.$$

#### Loi faible dans $L_2$

Soit  $(X_n)_{n \in \mathbb{N}}$  une suite de v.a.r. 2 à 2 *non corrélées*, de même loi et qui admettent une *variance*. On note  $\mathbf{E}[X_1] = \mu$ . On a

$$\frac{1}{n} \sum_{i=1}^n X_i \xrightarrow{L_2} \mu.$$

- On pourra consulter [**Foata and Fuchs, 2003**], chapitre 17, pour la preuve de ces résultats.

## Loi forte des grands nombres

- Elle s'obtient en supposant l'*indépendance mutuelle*.

### Loi forte des grands nombres

Soit  $(X_n)_{n \in \mathbb{N}}$  une suite de v.a.r. **indépendantes**, de **même loi** et qui admettent une **espérance**. On note  $\mathbf{E}[X_1] = \mu$ . On a

$$\frac{1}{n} \sum_{i=1}^n X_i \xrightarrow{p.s.} \mu.$$

### Application

- $X_1, \dots, X_n$  i.i.d de loi  $\mathcal{E}(\lambda)$  avec  $\lambda > 0$  (*inconnu*).
- $LFGN \implies \bar{X}_n \xrightarrow{p.s.} 1/\lambda$ .
- *Opérations sur les convergences p.s.* :  $1/\bar{X}_n \xrightarrow{p.s.} \lambda$ .

## Méthode de Monte-Carlo

- Soit  $f : ]0, 1[ \rightarrow \mathbb{R}$  intégrable. On cherche à *approcher*  $I = \int_0^1 f(x) dx$ .
- Pour  $X$  de loi uniforme sur  $[0, 1]$ , on a

$$I = \int_0^1 f(x) dx = \mathbf{E}[f(X)].$$

- *LFGN* : Soit  $(X_n)_n$  une suite de v.a.r i.i.d de loi uniforme sur  $[0, 1]$ . Alors  $(f(X_n))_n$  une suite de v.a.r i.i.d et on a

$$\frac{1}{n} \sum_{i=1}^n f(X_i) \xrightarrow{p.s.} \mathbf{E}[f(X)] = I.$$

### Algorithme de Monte-Carlo

1. Générer  $n$  (grand) observations suivant une loi uniforme sur  $[0, 1]$  ;
2. Approcher  $I$  par  $\frac{1}{n} \sum_{i=1}^n f(X_i)$ .

## 2.2 Le théorème central limite

### Présentation

- Soit  $(X_n)_{n \in \mathbb{N}}$  une suite de v.a.r. indépendantes et de même loi  $\mathcal{N}(\mu, \sigma^2)$ .
- On rappelle que

$$\sqrt{n} \frac{\bar{X}_n - \mu}{\sigma} \sim \mathcal{N}(0, 1).$$

- *Interprétation* :  $\mathcal{L}(\bar{X}_n) = \mathcal{N}(\mu, \sigma^2/n)$ .

### Approche TCL

- Le théorème central limite stipule que, sous des **hypothèses très faibles**, on peut étendre ce résultat (pour  $n$  grand) à **"n'importe quelle" suite de variables aléatoires indépendantes**.
- C'est l'un des résultats les plus **impressionnants et les plus utilisés** en probabilités et statistique.

## Le TCL

### Théorème Central Limite (TCL)

Soit  $(X_n)_{n \in \mathbb{N}}$  une suite de variables aléatoires indépendantes, de même loi, et telles que  $\mathbf{E}[X_i^2] < +\infty$ . On note  $\mathbf{E}[X_i] = \mu$ ,  $\mathbf{V}[X_i] = \sigma^2$  et  $\bar{X}_n = \frac{1}{n} \sum_{i=1}^n X_i$ . On a alors

$$\sqrt{n} \frac{\bar{X}_n - \mu}{\sigma} \xrightarrow{\mathcal{L}} \mathcal{N}(0, 1) \quad \text{quand } n \rightarrow \infty.$$

- Les hypothèses sont *faibles* : on demande juste des v.a.r i.i.d. qui admettent une variance.
- **Conséquence** : si  $n$  est suffisamment grand, on pourra approcher la loi de  $\bar{X}_n$  par la loi  $\mathcal{N}(\mu, \sigma^2/n)$ .
- On pourra écrire  $\mathcal{L}(\bar{X}_n) \approx \mathcal{N}(\mu, \sigma^2/n)$  mais *pas*

$$\mathcal{L}(\bar{X}_n) \xrightarrow{\mathcal{L}} \mathcal{N}(\mu, \sigma^2/n).$$

### Éléments de preuve

- Bien que ce résultat soit impressionnant, on peut voir la preuve comme un "simple" exercice sur les *fonctions caractéristiques* (voir [Jacod and Protter, 2003] pour des compléments).
- On note  $\varphi$  la fonction caractéristique des variables aléatoires  $X_i - \mu$  et

$$Y_n = \sqrt{n} \frac{\bar{X}_n - \mu}{\sigma}.$$

- On obtient des propriétés de la fonction caractéristique

$$\varphi_{Y_n}(t) = \left( \varphi \left( \frac{t}{\sigma \sqrt{n}} \right) \right)^n.$$

- De plus

$$\varphi(0) = 1, \quad \varphi'(0) = 0 \quad \text{et} \quad \varphi''(0) = -\sigma^2.$$

- On déduit

$$\varphi(u) = 1 - \frac{\sigma^2 u^2}{2} + o(u^2)$$

et

$$\varphi_{Y_n}(t) = \exp \left( n \log(1 - t^2/2n + o(1/n)) \right).$$

- Par conséquent

$$\lim_{n \rightarrow \infty} \varphi_{Y_n}(t) = \exp(-t^2/2)$$

et  $t \mapsto \exp(-t^2/2)$  est la fonction caractéristique de la loi  $\mathcal{N}(0, 1)$ .

- D'après le théorème de Paul Levy, on conclut  $Y_n \xrightarrow{\mathcal{L}} \mathcal{N}(0, 1)$ .

### Exemple : loi de Bernoulli

- Soit  $(X_n)_{n \in \mathbb{N}}$  une suite de v.a.r. i.i.d. de loi de Bernoulli de paramètre  $p \in ]0, 1[$ .
- On a d'après la *loi forte des grands nombres*

$$\bar{X}_n \xrightarrow{p.s.} p \quad \text{quand } n \rightarrow \infty$$

et d'après le *théorème central limite*

$$\sqrt{n} \frac{\bar{X}_n - p}{\sqrt{p(1-p)}} \xrightarrow{\mathcal{L}} \mathcal{N}(0, 1) \quad \text{quand } n \rightarrow \infty.$$

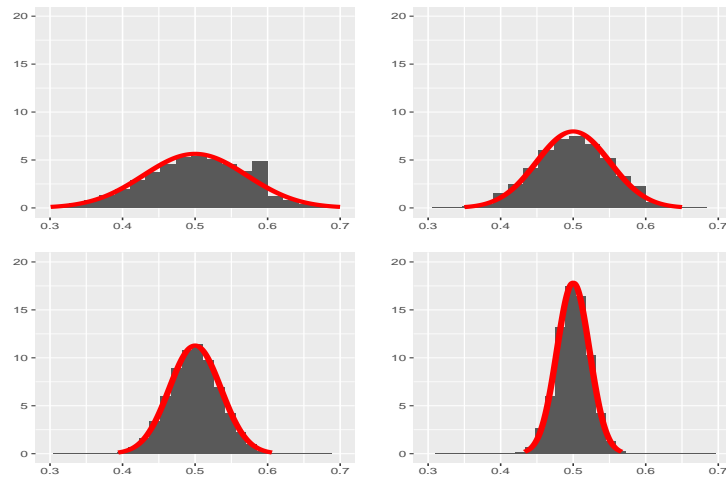


FIGURE 1 – Approximation TCL pour le modèle de Bernoulli  $\mathcal{B}(1/2)$  avec  $n = 50, 100, 200, 500$ .

## Illustration

### Slutsky

— Par *continuité*, on a

$$\sqrt{(\bar{X}_n)(1 - \bar{X}_n)} \xrightarrow{\mathbf{P}} \sqrt{p(1 - p)},$$

et donc

$$\frac{\sqrt{p(1 - p)}}{\sqrt{(\bar{X}_n)(1 - \bar{X}_n)}} \xrightarrow{\mathbf{P}} 1.$$

— On obtient donc d'après *Slutsky*

$$\sqrt{n} \frac{\bar{X}_n - p}{\sqrt{\bar{X}_n(1 - \bar{X}_n)}} = \sqrt{n} \frac{\bar{X}_n - p}{\sqrt{p(1 - p)}} \times \frac{\sqrt{p(1 - p)}}{\sqrt{(\bar{X}_n)(1 - \bar{X}_n)}} \xrightarrow{\mathcal{L}} \mathcal{N}(0, 1).$$

### Remarque importante

Ce type de raisonnement est très souvent utilisé pour trouver des **intervalles de confiance asymptotique**.

### Exemple : loi exponentielle

— Soit  $(X_n)_{n \in \mathbb{N}}$  une suite de v.a.r. i.i.d. de loi exponentielle de paramètre  $\lambda > 0$ .

— On a d'après la *loi forte des grands nombres*

$$\bar{X}_n \xrightarrow{p.s.} \frac{1}{\lambda} \quad \text{et} \quad \frac{1}{\bar{X}_n} \xrightarrow{p.s.} \lambda \quad \text{quand } n \rightarrow \infty$$

et d'après le *théorème central limite*

$$\sqrt{n} \frac{\bar{X}_n - 1/\lambda}{1/\lambda} \xrightarrow{\mathcal{L}} \mathcal{N}(0, 1) \quad \text{quand } n \rightarrow \infty.$$

### Problème

— Comment obtenir un TCL pour  $1/\bar{X}_n$  ?

— La *delta méthode* permet d'y parvenir.

## Delta méthode

- Elle permet (notamment) d'étendre le *TCL* à des estimateurs  $g(\bar{X}_n)$  qui s'écrivent comme une *fonction de la moyenne empirique*.

### *Théorème (Delta méthode)*

Soit  $(X_n)_n$  une suite de v.a.r. et  $(v_n)$  une suite de réels qui tend vers  $+\infty$ . On suppose qu'il existe un réel  $a$  et une variable  $X$  tels que

$$v_n(X_n - a) \xrightarrow{\mathcal{L}} X.$$

Si  $g$  est une *fonction dérivable au point  $a$* , alors

$$v_n(g(X_n) - g(a)) \xrightarrow{\mathcal{L}} g'(a)X.$$

En particulier, si  $X \sim \mathcal{N}(0, \sigma^2)$  et  $g'(a) \neq 0$ , alors

$$v_n(g(X_n) - g(a)) \xrightarrow{\mathcal{L}} N(0, (\sigma g'(a))^2).$$

## Application : loi exponentielle

- Pour le modèle exponentiel, on a montré grâce au *TCL*

$$\sqrt{n}(\bar{X}_n - 1/\lambda) \xrightarrow{\mathcal{L}} \mathcal{N}\left(0, \frac{1}{\lambda^2}\right) \quad \text{quand } n \rightarrow \infty.$$

- On applique la *delta méthode* avec  $g(u) = 1/u$  :

$$\sqrt{n}\left(\frac{1}{\bar{X}_n} - \lambda\right) \xrightarrow{\mathcal{L}} \mathcal{N}(0, \lambda^2) \quad \text{quand } n \rightarrow \infty,$$

ou encore

$$\frac{\sqrt{n}}{\lambda}\left(\frac{1}{\bar{X}_n} - \lambda\right) \xrightarrow{\mathcal{L}} \mathcal{N}(0, 1) \quad \text{quand } n \rightarrow \infty.$$

- Donc, en note  $\hat{\lambda} = 1/\bar{X}_n$ , d'après *Slutsky*,

$$\frac{\sqrt{n}}{\hat{\lambda}}(\hat{\lambda} - \lambda) \xrightarrow{\mathcal{L}} \mathcal{N}(0, 1) \quad \text{quand } n \rightarrow \infty.$$

## 3 Bibliographie

### Références

#### Biblio3

- [Foata and Fuchs, 2003] Foata, D. and Fuchs, A. (2003). *Calcul des probabilités*. Dunod, 2<sup>e</sup> édition.
- [Jacod and Protter, 2003] Jacod, J. and Protter, P. (2003). *L'essentiel en théorie des probabilités*. Cassini.
- [Rouvière, 2015] Rouvière, L. (2015). Probabilités générales. Polycopié de cours, <https://perso.univ-rennes2.fr/laurent.rouviere>.



## Quatrième partie

# Critères de performance asymptotiques, intervalles de confiance et estimation multivariée

### Rappel

- $X_1, \dots, X_n$  i.i.d de loi  $\mathbf{P}_\theta$  avec  $\theta \in \Theta$  univarié.
- $\hat{\theta} = \hat{\theta}(X_1, \dots, X_n) = \hat{\theta}_n$  un estimateur de  $\theta$ .
- Critère de performance pour  $\hat{\theta}_n$  : biais, variance, risque quadratique, VUMSB...

### Dans cette partie

- Critères de performance asymptotiques ;
- Estimation par intervalles ;
- Estimation multivariée ( $\theta \in \mathbb{R}^p$ ).

## 1 Critères asymptotiques

### Pourquoi ?

#### Postulat

On veut définir des estimateurs qui soient de *plus en plus précis* lorsque la *quantité d'information augmente*.

- La **quantité d'information** à disposition du statisticien peut être représentée par le *nombre d'observations*  $n$ .
- On cherche donc des estimateurs *de plus en plus précis* lorsque  $n$  augmente.
- Mathématiquement, on va donc **chercher des estimateurs**  $\hat{\theta}_n$  qui **convergent** (en probabilité, presque sûrement, en loi...) vers  $\theta$ .

### Consistance

#### Définition

On dit que l'estimateur  $\hat{\theta}_n$  est **consistant** (ou **convergent**) si  $\hat{\theta} \xrightarrow{\mathbf{P}} \theta$ , c'est-à-dire

$$\forall \varepsilon > 0 \quad \lim_{n \rightarrow \infty} \mathbf{P}_\theta(|\hat{\theta}_n - \theta| \geq \varepsilon) = 0.$$

#### Définition

Soit  $(v_n)_n$  une suite de réels positifs telle que  $v_n \rightarrow \infty$ . On dit que  $\hat{\theta}_n$  est **asymptotiquement normal**, de vitesse  $v_n$  si

$$v_n(\hat{\theta}_n - \theta) \xrightarrow{\mathcal{L}} \mathcal{N}(0, \sigma_\theta)$$

où  $\sigma_\theta$  est positif.

## Outils consistance

- Bienaymé-Tchebychev.
- Loi forte des grands nombres.
- Opérations sur les convergences en probabilité.

### Exemple

- *Modèle de Bernoulli* :  $\hat{p}_n = \bar{X}_n$  est consistant.
- *Modèle exponentiel* :  $\hat{\lambda}_n = 1/\bar{X}_n$  est consistant.

## Outils normalité asymptotique

- Théorème central limite.
- Delta méthode.

### Exemple

- *Modèle de Bernoulli* :  $\hat{p}_n = \bar{X}_n$  est asymptotiquement normal à la vitesse  $\sqrt{n}$  :

$$\sqrt{n}(\hat{p}_n - p) \xrightarrow{\mathcal{L}} \mathcal{N}(0, p(1-p)).$$

- *Modèle exponentiel* :  $\hat{\lambda}_n = 1/\bar{X}_n$  est asymptotiquement normal à la vitesse  $\sqrt{n}$  :

$$\sqrt{n}(\hat{\lambda}_n - \lambda) \xrightarrow{\mathcal{L}} \mathcal{N}(0, \lambda^2).$$

## 2 Estimation par intervalles

### Motivations

- Donner une *seule valeur* pour estimer un paramètre peut se révéler trop ambitieux.
- *Exemple* : on traite 100 patients à l'aide d'un traitement. 72 guérissent. Affirmer que la performance est de 72% lorsque on prend le traitement (alors qu'on ne l'a *testé que sur 100 athlètes*) est un peu fort.
- Il peut parfois être plus raisonnable de donner une réponse dans le genre, la performance se trouve dans l'*intervalle* [70%, 74%] avec une *confiance* de 90%.

### Intervalle de confiance

- $X_1, \dots, X_n$  un échantillon i.i.d. de loi  $\mathbf{P}_\theta$  avec  $\theta \in \Theta$  **inconnu**.

#### Définition

Soit  $\alpha \in ]0, 1[$ . On appelle **intervalle de confiance** pour  $\theta$  tout intervalle de la forme  $[A_n, B_n]$ , où  $A_n$  et  $B_n$  sont des fonctions telles que :

$$\mathbf{P}(\theta \in [A_n, B_n]) = 1 - \alpha.$$

Si  $\lim_{n \rightarrow \infty} \mathbf{P}(\theta \in [A_n, B_n]) = 1 - \alpha$ , on dit que  $[A_n, B_n]$  est un *intervalle de confiance asymptotique* pour  $\theta$  au niveau  $1 - \alpha$ .

#### Remarque importante

- $A_n = A_n(X_1, \dots, X_n)$  et  $B_n = B_n(X_1, \dots, X_n)$  sont **aléatoires** !
- Les logiciels renverront les **réels**  $a_n = A_n(x_1, \dots, x_n)$  et  $b_n = B_n(x_1, \dots, x_n)$ .

## Construction d'un IC

- Inégalité de *Bienaymé-Tchebychev* (intervalle de confiance par excès) :

$$\mathbf{P}(\theta \in [A_n, B_n]) \geq 1 - \alpha.$$

- Utilisation d'une **fonction pivotable pour le paramètre  $\theta$**  : fonction mesurable des observations et du paramètre inconnu mais dont la loi ne dépend pas de  $\theta$ .

### Méthode

1. se donner un niveau  $1 - \alpha$ .
2. trouver un **estimateur**  $\hat{\theta}_n$  de  $\theta$  dont **on connaît la loi** afin de construire une fonction pivotable.

## Construction d'IC

- Un *intervalle de confiance* pour un paramètre inconnu  $\theta$  se construit généralement à partir d'un **estimateur de  $\theta$  dont on connaît la loi**.
- A partir de la loi de  $\hat{\theta}$ , on cherche deux bornes  $A_n$  et  $B_n$  telles que

$$\mathbf{P}(\theta \in [A_n, B_n]) = 1 - \alpha.$$

### Remarque

A priori, plus  $\alpha$  est *petit*, plus l'intervalle aura une *grande amplitude*.

## Exemple

- $X_1, \dots, X_n$  i.i.d. de loi normale  $\mathcal{N}(\mu, 1)$ .
- On suppose la variance connue et on cherche un IC pour  $\mu$ .

### Construction de l'IC

- **Estimateur** :  $\hat{\mu} = \bar{X}_n$ .
- **Loi de l'estimateur** :  $\mathcal{L}(\hat{\mu}) = \mathcal{N}(\mu, 1/n)$ .
- On déduit

$$\mathbf{P}\left(\hat{\mu} - q_{1-\alpha/2} \frac{1}{\sqrt{n}} \leq \mu \leq \hat{\mu} + q_{1-\alpha/2} \frac{1}{\sqrt{n}}\right) = 1 - \alpha.$$

- Un **intervalle de confiance de niveau  $1 - \alpha$**  est donc donné par

$$\left[\hat{\mu} - q_{1-\alpha/2} \frac{1}{\sqrt{n}}, \hat{\mu} + q_{1-\alpha/2} \frac{1}{\sqrt{n}}\right].$$

## Quantiles

- $q_{1-\alpha/2}$  désigne le *quantile d'ordre  $1 - \alpha/2$*  de la loi normale  $\mathcal{N}(0, 1)$  défini par

$$\mathbf{P}(X \leq q_{1-\alpha/2}) = 1 - \frac{\alpha}{2}.$$

### Définition

Plus généralement, le *quantile d'ordre  $\alpha$*  d'une variable aléatoire  $X$  est défini par le **réel**  $q_\alpha$  vérifiant

$$q_\alpha = \inf_x \{x : F(x) \geq \alpha\}.$$

- Les quantiles sont généralement renvoyés par les *logiciels statistiques* :

```
> c(qnorm(0.975), qnorm(0.95), qnorm(0.5))
[1] 1.959964 1.644854 0
```

## Exemple

- $n = 50$  observation issues d'une loi  $\mathcal{N}(\mu, 1)$  :

```
> head(X)
[1] 3.79 5.28 6.08 2.65 5.43 5.51
```

- *Estimation de  $\mu$*  :

```
> mean(X)
[1] 4.55
```

- *Intervalle de confiance de niveau 95%* :

```
> binf <- mean(X) - qnorm(0.975) * 1/sqrt(50)
> bsup <- mean(X) + qnorm(0.975) * 1/sqrt(50)
> c(binf, bsup)
[1] 4.269766 4.824128
```

## Intervalle de confiance pour une proportion

- $X_1, \dots, X_n$  i.i.d. de loi  $\mathcal{B}(p)$ .
- On cherche un *intervalle de confiance asymptotique* pour  $p$ .

### Construction de l'IC

- **Estimateur** :  $\hat{p}_n = \bar{X}_n$ .
- **Loi asymptotique de l'estimateur** :

$$\sqrt{n}(\hat{p}_n - p) \xrightarrow{\mathcal{L}} \mathcal{N}(0, p(1-p)).$$

- On déduit

$$\mathbf{P} \left( \hat{p}_n - q_{1-\alpha/2} \sqrt{\frac{p(1-p)}{n}} \leq \mu \leq \hat{p}_n + q_{1-\alpha/2} \sqrt{\frac{p(1-p)}{n}} \right) \rightarrow 1 - \alpha.$$

### Première version de l'IC

- Un **intervalle de confiance asymptotique de niveau  $1 - \alpha$**  est donc donné par

$$\left[ \hat{p}_n - q_{1-\alpha/2} \sqrt{\frac{\hat{p}_n(1-\hat{p}_n)}{n}}, \hat{p}_n + q_{1-\alpha/2} \sqrt{\frac{\hat{p}_n(1-\hat{p}_n)}{n}} \right].$$

- **Problème** : l'IC dépend de  $p$  qui est inconnu !
- **Solution** : Slutsky  $\Rightarrow$

$$\sqrt{n} \frac{\hat{p}_n - p}{\sqrt{\hat{p}_n(1-\hat{p}_n)}} \xrightarrow{\mathcal{L}} \mathcal{N}(0, 1).$$

## Conclusion

Un *intervalle de confiance asymptotique de niveau  $1 - \alpha$*  est donné par

$$\left[ \hat{p}_n - q_{1-\alpha/2} \sqrt{\frac{\hat{p}_n(1-\hat{p}_n)}{n}}, \hat{p}_n + q_{1-\alpha/2} \sqrt{\frac{\hat{p}_n(1-\hat{p}_n)}{n}} \right].$$

- $n = 500$  observation issues d'une loi  $\mathcal{B}(p)$ .
- *Estimation de  $p$*  :

```
> phat <- mean(X)
> phat
[1] 0.756
```

— *Intervalle de confiance asymptotique* de niveau 95% :

```
> binf <- phat - qnorm(0.975)*sqrt(phat*(1-phat)/n)
> bsup <- phat + qnorm(0.975)*sqrt(phat*(1-phat)/n)
> c(binf,bsup)
[1] 0.718354 0.793646
```

### Fonction *prop.test*

On peut récupérer un IC plus précis à l'aide de la fonction **prop.test** :

```
> prop.test(sum(X),n,correct=FALSE)$conf.int
[1] 0.7164952 0.7916011
attr(,"conf.level")
[1] 0.95
```

### Loi normale (cas réel)

- $X_1, \dots, X_n$  i.i.d de loi  $\mathcal{N}(\mu, \sigma^2)$ .
- On a vu qu'un IC pour  $\mu$  est donné par

$$\left[ \hat{\mu} - q_{1-\alpha/2} \frac{\sigma}{\sqrt{n}}, \hat{\mu} + q_{1-\alpha/2} \frac{\sigma}{\sqrt{n}} \right].$$

### Problème

- Dans la vraie vie,  $\sigma$  est **inconnu** !
- L'intervalle de confiance **n'est donc pas calculable**.

### Idée

1. Estimer  $\sigma^2$  par

$$\widehat{\sigma^2} = \frac{1}{n} \sum_{i=1}^n (X_i - \bar{X}_n)^2$$

2. Et considérer l'IC :

$$\left[ \hat{\mu} - q_{1-\alpha/2} \frac{\widehat{\sigma}}{\sqrt{n}}, \hat{\mu} + q_{1-\alpha/2} \frac{\widehat{\sigma}}{\sqrt{n}} \right]. \quad (2)$$

### Problème

- On a bien

$$\mathcal{L} \left( \sqrt{n} \frac{\bar{X}_n - \mu}{\sigma} \right) = \mathcal{N}(0, 1)$$

- mais

$$\mathcal{L} \left( \sqrt{n} \frac{\bar{X}_n - \mu}{\widehat{\sigma}} \right) \neq \mathcal{N}(0, 1)$$

- Pour avoir la loi de

$$\sqrt{n} \frac{\bar{X}_n - \mu}{\widehat{\sigma}} \neq \mathcal{N}(0, 1)$$

avec

$$\widehat{\sigma^2} = \frac{1}{n} \sum_{i=1}^n (X_i - \bar{X}_n)^2$$

- il faut définir d'autres *lois de probabilité*.

## La loi normale (Rappel)

### Définition

- Une v.a.r  $X$  suit une loi *normale* de paramètres  $\mu \in \mathbb{R}$  et  $\sigma^2 > 0$  admet pour densité

$$f(x) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right).$$

### Propriétés

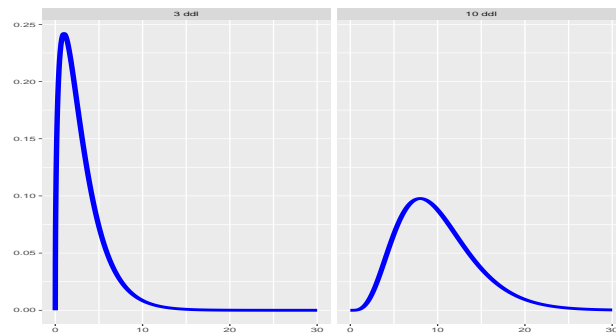
- $\mathbf{E}[X] = \mu$  et  $\mathbf{V}[X] = \sigma^2$ .
- Si  $X \sim \mathcal{N}(\mu, \sigma^2)$  alors

$$\frac{X - \mu}{\sigma} \sim \mathcal{N}(0, 1).$$

## Loi du $\chi^2$

### Définition

- Soit  $X_1, \dots, X_n$   $n$  variables aléatoires réelles indépendantes de loi  $\mathcal{N}(0, 1)$ . La variable  $Y = X_1^2 + \dots + X_n^2$  suit une loi du *Chi-Deux* à  $n$  degrés de liberté. Elle est notée  $\chi^2(n)$ .
- $\mathbf{E}[Y] = n$  et  $\mathbf{V}[Y] = 2n$ .



## Loi de Student

### Définition

- Soient  $X$  et  $Y$  deux v.a.r. *indépendantes* de loi  $\mathcal{N}(0, 1)$  et  $\chi^2(n)$ . Alors la v.a.r.

$$T = \frac{X}{\sqrt{Y/n}}$$

suit une *loi de student* à  $n$  degrés de liberté. On note  $\mathcal{T}(n)$ .

- $\mathbf{E}[T] = 0$  et  $\mathbf{V}[T] = n/(n-2)$ .
- Lorsque  $n$  est grand la loi de student à  $n$  degrés de liberté peut être *approchée par la loi*  $\mathcal{N}(0, 1)$ .

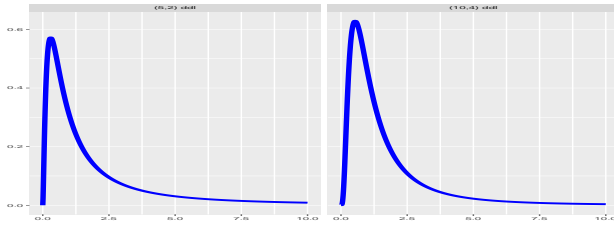
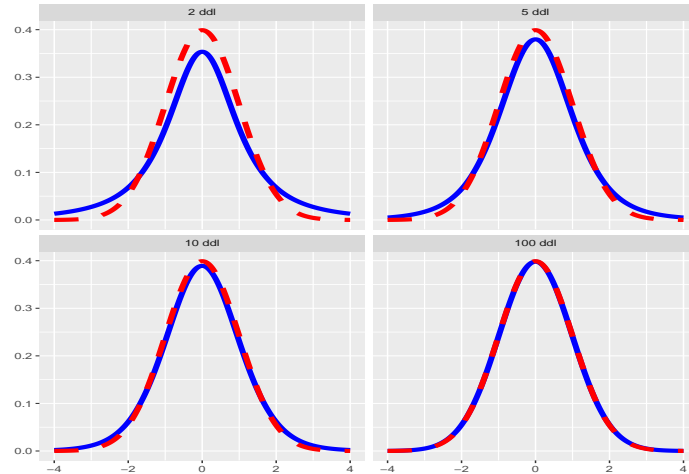


FIGURE 2 – Densités  $\mathcal{F}(5, 2)$  et  $\mathcal{F}(10, 4)$



### Légende

Densités des lois de student à 2, 5, 10 et 100 degrés de liberté (bleu) et densité de la loi  $\mathcal{N}(0, 1)$  (rouge).

### Loi de Fisher

#### Définition

— Soient  $X$  et  $Y$  deux v.a.r *indépendantes* de lois  $\chi^2(m)$  et  $\chi^2(n)$ . Alors la v.a.r

$$F = \frac{X/m}{Y/n}$$

suit une *loi de Fisher* à  $m$  et  $n$  degrés de liberté. On note  $\mathcal{F}(m, n)$ .

— Si  $F \sim \mathcal{F}(m, n)$  alors  $1/F \sim \mathcal{F}(n, m)$ .

### Théorème de Cochran

—  $X_1, \dots, X_n$  i.i.d. de loi  $\mathcal{N}(\mu, \sigma^2)$ .

— On note

$$S^2 = \frac{1}{n-1} \sum_{i=1}^n (X_i - \bar{X}_n)^2.$$

### Théorème de Cochran

On a alors

1.  $(n-1) \frac{S^2}{\sigma^2} \sim \chi^2(n-1)$ .

2.  $\bar{X}_n$  et  $S^2$  sont indépendantes.
3. On déduit

$$\sqrt{n} \frac{\bar{X}_n - \mu}{S} \sim \mathcal{T}(n-1).$$

### Remarque

Les résultats 1 et 3 sont très importants pour construire des *IC*.

### IC pour la loi gaussienne

#### IC pour $\mu$

On déduit du résultat précédent qu'un **IC de niveau  $1 - \alpha$**  pour  $\mu$  est donné par

$$\left[ \bar{X}_n - t_{1-\alpha/2} \frac{S}{\sqrt{n}}, \bar{X}_n + t_{1-\alpha/2} \frac{S}{\sqrt{n}} \right],$$

où  $t_{1-\alpha/2}$  est le quantile d'ordre  $1 - \alpha/2$  de la loi de Student à  $n - 1$  ddl.

#### IC pour $\sigma^2$

Un **IC de niveau  $1 - \alpha$**  pour  $\sigma^2$  est donné par

$$\left[ \frac{(n-1)S^2}{\chi_{1-\alpha/2}}, \frac{(n-1)S^2}{\chi_{\alpha/2}} \right]$$

où  $\chi_{1-\alpha/2}$  et  $\chi_{\alpha/2}$  sont les quantiles d'ordre  $1 - \alpha/2$  et  $\alpha/2$  de loi  $\chi^2(n-1)$ .

### Exemple : modèle Gaussien - IC pour $\mu$

- $n = 50$  observations issues d'une loi  $\mathcal{N}(\mu, \sigma^2)$  :

```
> head(X)
[1] 3.79 5.28 6.08 2.65 5.43 5.51
```

- *Estimation* de  $\mu$  :

```
> mean(X)
[1] 4.55
```

- *Estimation* de  $\sigma^2$  :

```
> S <- var(X)
> S
[1] 0.783302
```

- *Intervalle de confiance* de niveau 95% :

```
> binf <- mean(X) - qt(0.975, 49) * sqrt(S) / sqrt(50)
> bsup <- mean(X) + qt(0.975, 49) * sqrt(S) / sqrt(50)
> c(binf, bsup)
[1] 4.295420 4.798474
```

- On peut obtenir directement l'intervalle de confiance à l'aide de la fonction *t.test*

```
> t.test(X)$conf.int
[1] 4.295420 4.798474
attr(,"conf.level")
[1] 0.95
```



### Exemple : modèle gaussien - IC pour $\sigma^2$

- On obtient l'IC pour  $\sigma^2$  à l'aide de la formule

$$\left[ \frac{(n-1)S^2}{\chi_{1-\alpha/2}}, \frac{(n-1)S^2}{\chi_{\alpha/2}} \right]$$

- On peut donc le calculer sur  $R$  :

```
> binf <- 49*S/qchisq(0.975,49)
> bsup <- 49*S/qchisq(0.025,49)
> c(binf,bsup)
[1] 0.5465748 1.2163492
```

## 3 Estimation multivariée

### Jusqu'à présent

- $X_1, \dots, X_n$  i.i.d de loi  $\mathbf{P}_\theta$  avec  $\theta \in \mathbb{R}$ .
- La loi  $\mathbf{P}_\theta$  dépend donc d'un *seul paramètre* (à estimer).
- Dans de nombreux problèmes *concrets*, les choses sont plus *complexes*.
- Il faut donc envisager le cas où on dispose de **plus d'un paramètre**.

### Cadre

- Pour simplifier on se place dans le cas d'un *paramètre bivarié*.
- $X_1, \dots, X_n$  i.i.d de loi  $\mathbf{P}_\theta$  avec  $\theta = (\theta_1, \theta_2)$  **inconnu dans  $\mathbb{R}^2$** .

### Estimateur

Un estimateur  $\hat{\theta} = (\hat{\theta}_1, \hat{\theta}_2)$  est une fonction mesurable de  $X_1, \dots, X_n$  indépendante de  $\theta$  à valeurs dans  $\mathbb{R}^2$ .

### Exemple : le modèle gaussien

- $\theta = (\mu, \sigma^2)$
- $\hat{\theta} = (\hat{\mu}, S^2)$  tels que

$$\hat{\mu} = \bar{X}_n \quad \text{et} \quad S^2 = \frac{1}{n-1} \sum_{i=1}^n (X_i - \bar{X}_n)^2.$$

### 3.1 Biais, variance, risque quadratique

- Pour le *biais*, on travaille composante par composante :

$$\mathbf{E}[\hat{\theta}] = \begin{pmatrix} \mathbf{E}[\hat{\theta}_1] \\ \mathbf{E}[\hat{\theta}_2] \end{pmatrix} \quad \text{et} \quad b(\hat{\theta}) = \mathbf{E}[\hat{\theta}] - \theta = \begin{pmatrix} b(\hat{\theta}_1) \\ b(\hat{\theta}_2) \end{pmatrix}.$$

- $\hat{\theta} = (\hat{\theta}_1, \hat{\theta}_2)$  est un *vecteur* aléatoire ! Il ne va donc pas posséder de variance mais une **matrice de variance covariance** :

$$\Sigma_{\hat{\theta}} = \begin{pmatrix} \mathbf{V}[\hat{\theta}_1] & \text{cov}(\hat{\theta}_1, \hat{\theta}_2) \\ \text{cov}(\hat{\theta}_2, \hat{\theta}_1) & \mathbf{V}[\hat{\theta}_2] \end{pmatrix}.$$

### Exemple : le modèle gaussien

- $\theta = (\mu, \sigma^2)$  et  $\hat{\theta} = (\bar{X}_n, S^2)$ .
- On a  $b(\hat{\theta}) = (0, 0)$ .
- D'après *Cochran*, on déduit

$$\Sigma_{\hat{\theta}} = \begin{pmatrix} \frac{\sigma^2}{n} & 0 \\ 0 & \frac{2\sigma^4}{n-1} \end{pmatrix}.$$

### Risque quadratique

- Il existe également un *risque quadratique* en **estimation multivariée**.

#### Définition

On appelle *risque quadratique* de  $\hat{\theta} = (\hat{\theta}_1, \hat{\theta}_2)$  le réel positif

$$\mathcal{R}(\theta, \hat{\theta}) = \mathbf{E}_{\theta} \|\hat{\theta} - \theta\|^2$$

#### Propriété

$$\mathcal{R}(\theta, \hat{\theta}) = \|\mathbf{E}_{\theta}(\hat{\theta}) - \theta\|^2 + \mathbf{E}_{\theta} \|\hat{\theta} - \mathbf{E}_{\theta} \hat{\theta}\|^2.$$

- On a toujours une *décomposition "biais/variance"*.

## 3.2 Critères asymptotiques

### Consistance

#### Définition

On dit que l'estimateur  $\hat{\theta}$  est *consistant* (ou *convergent*) si  $\hat{\theta} \xrightarrow{\mathbf{P}} \theta$ , c'est-à-dire

$$\forall \varepsilon > 0 \quad \lim_{n \rightarrow \infty} \mathbf{P}_{\theta}(\|\hat{\theta} - \theta\| \geq \varepsilon) = 0.$$

- La valeur absolue est juste remplacée par la *norme euclidienne*.
- En pratique, ce n'est pas difficile : en effet  $\hat{\theta} \xrightarrow{\mathbf{P}} \theta$  si et seulement si  $\hat{\theta}_1 \xrightarrow{\mathbf{P}} \theta_1$  et  $\hat{\theta}_2 \xrightarrow{\mathbf{P}} \theta_2$ .

#### Exemple : le modèle gaussien

$\hat{\theta} = (\bar{X}_n, S^2)$  est consistant.

### Normalité asymptotique

#### Définition

Soit  $(v_n)_n$  une suite de réels positifs telle que  $v_n \rightarrow \infty$ . On dit que  $\hat{\theta} = (\hat{\theta}_1, \hat{\theta}_2)$  est *asymptotiquement normal*, de vitesse  $v_n$  si

$$v_n(\hat{\theta} - \theta) \xrightarrow{\mathcal{L}} \mathcal{N}(0, \Sigma_{\theta})$$

où  $\Sigma_{\theta}$  est une matrice symétrique  $2 \times 2$  définie positive.

- La loi limite est une loi gaussienne *multivariée*.
- Il existe une version *multivariée* du **TCL** et de la **delta méthode**. Ce sont les principaux outils pour montrer la normalité asymptotique d'estimateurs multivariés.

## Vecteurs gaussiens (rappels)

### Définition

- $X = (X_1, X_2)$  est un *vecteur aléatoire gaussien* si toute combinaison linéaire de ses marginales  $\alpha_1 X_1 + \alpha_2 X_2$  est une variable aléatoire réelle gaussienne.
- On note  $X \sim \mathcal{N}(\mu, \Sigma)$  où  $\mu \in \mathbb{R}^2$  est l'*espérance* de  $X$  et  $\Sigma$  est la matrice  $(2 \times 2)$  de *variance covariance* de  $X$ .

### Propriété

Soit  $X$  un vecteur gaussien de loi  $\mathcal{N}(\mu, \Sigma)$ . Alors  $X$  admet une *densité* si et seulement si  $\det(\Sigma) \neq 0$ . Elle est donnée par

$$f(x) = \frac{1}{2\pi\sqrt{\det(\Sigma)}} \exp\left(-\frac{1}{2}(x - \mu)' \Sigma^{-1}(x - \mu)\right).$$

## TCL et delta méthode multivariés

### TCL

Soit  $(X_n)_n$  une suite de *vecteurs aléatoires* i.i.d. d'espérance  $\mu \in \mathbb{R}^2$  et de matrice de variance covariance  $(2 \times 2)$   $\Sigma$ , alors

$$\sqrt{n}(\bar{X}_n - \mu) \xrightarrow{\mathcal{L}} \mathcal{N}(0, \Sigma).$$

### Delta méthode

Si  $v_n(\hat{\theta} - \theta) \xrightarrow{\mathcal{L}} X \sim \mathcal{N}(0, \Sigma)$  et si  $h : \mathbb{R}^d \rightarrow \mathbb{R}^m$  admet des dérivées partielles au point  $\theta$ , alors

$$v_n(h(\hat{\theta}) - h(\theta)) \xrightarrow{\mathcal{L}} Dh_\theta X$$

où  $Dh_\theta$  est la matrice  $m \times d$  de terme  $(Dh_\theta)_{ij} = \frac{\partial h_i}{\partial \theta_j}(\theta)$ .

## 3.3 Borne de Cramer-Rao

### Rappels - cas univarié

- $X_1, \dots, X_n$  i.i.d de loi  $\mathbf{P}_\theta$  avec  $\theta \in \mathbb{R}$ .

### Inégalité de Cramér-Rao

Si  $\hat{\theta}$  est un estimateur *sans biais* de  $\theta$  alors

$$\mathbf{V}_\theta[\hat{\theta}] \geq \frac{1}{nI(\theta)}$$

où

$$I(\theta) = \mathbf{E}_\theta \left[ \left( \frac{\partial}{\partial \theta} \log(L(X, \theta)) \right)^2 \right].$$

### Retour au cas multivarié

- $X_1, \dots, X_n$  i.i.d de loi  $\mathbf{P}_\theta$  avec  $\theta = (\theta_1, \theta_2) \in \mathbb{R}^2$ .
- On désigne par  $L(x, \theta)$  la vraisemblance de  $\theta$  pour une observation  $x$ .

### Exemple : le modèle gaussien

- $\theta = (\mu, \sigma^2)$ .
- La vraisemblance est

$$L(x, \mu, \sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(x - \mu)^2}{2\sigma^2}\right).$$

## Matrice d'information de Fisher

### Définition

La *matrice d'information de Fisher* (si elle existe) au point  $\theta$  est la matrice de dimension  $2 \times 2$  de terme général

$$\begin{aligned} I(\theta)_{i,j} &= \mathbf{E}_\theta \left[ \frac{\partial}{\partial \theta_i} \log(L(X, \theta)) \frac{\partial}{\partial \theta_j} \log(L(X, \theta)) \right] \\ &= - \mathbf{E}_\theta \left[ \frac{\partial^2}{\partial \theta_i \partial \theta_j} \log(L(X, \theta)) \right] \end{aligned}$$

avec  $1 \leq i, j \leq 2$ .

### Exemple

Pour le modèle gaussien, la *matrice d'information de Fisher* est donnée par

$$I(\theta) = \begin{pmatrix} \frac{1}{\sigma^2} & 0 \\ 0 & \frac{1}{2\sigma^4} \end{pmatrix} \quad \text{avec } \theta = (\mu, \sigma^2).$$

## Borne de Cramer Rao

—  $X_1, \dots, X_n$  i.i.d de loi  $\mathbf{P}_\theta$  avec  $\theta = (\theta_1, \theta_2) \in \mathbb{R}^2$ .

### Théorème

Si elle existe, la *borne de Cramer-Rao* du modèle précédent est  $\frac{1}{n}I(\theta)^{-1}$ . C'est-à-dire que *pour tout estimateur sans biais*  $\hat{\theta}$  de  $\theta$ , on a

$$\Sigma_{\hat{\theta}} \geq_{sdp} \frac{1}{n} I(\theta)^{-1}.$$

### Remarques

— L'inégalité est à prendre au sens des *matrices semi définies positives* :

$$\forall u \in \mathbb{R}^2, \quad u' \Sigma_{\hat{\theta}} u \geq u' \left( \frac{1}{n} I(\theta)^{-1} \right) u.$$

— Interprétation *similaire au cas univarié* : la BCR vue comme une *matrice de variance covariance optimale* pour un estimateur *sans biais*.

## Retour au modèle gaussien

—  $\hat{\theta} = (\bar{X}_n, S_n^2)$  est *sans biais*.

— Sa matrice de *variance covariance* est donnée par

$$\Sigma_{\hat{\theta}} = \begin{pmatrix} \frac{\sigma^2}{n} & 0 \\ 0 & \frac{2\sigma^4}{n-1} \end{pmatrix}.$$

— La *BCR* vaut

$$\frac{1}{n} I(\theta)^{-1} = \frac{1}{n} \begin{pmatrix} \frac{1}{\sigma^2} & 0 \\ 0 & \frac{1}{2\sigma^4} \end{pmatrix}^{-1} = \begin{pmatrix} \frac{\sigma^2}{n} & 0 \\ 0 & \frac{2\sigma^4}{n} \end{pmatrix}.$$

— *Conclusion* :  $\hat{\theta}$  n'est pas *VUMSB* (mais il n'est pas loin).

## Retour à l'emv

- L'emv possède, sous certaines hypothèses, de bonnes propriétés.

### "Propriété"

Sous certaines hypothèses de régularité sur la loi  $\mathbf{P}_\theta$ , l'emv  $\hat{\theta}_{MV}$  de  $\theta$  est

- consistant ;
- asymptotiquement normal :

$$\sqrt{n}(\hat{\theta}_{MV} - \theta) \xrightarrow{\mathcal{L}} \mathcal{N}(0, I(\theta)^{-1}).$$

### En pratique...

- Les hypothèses de ce résultat sont *techniques* et généralement *difficiles à vérifier*.
- Il est souvent plus simple d'obtenir ce résultat en **travaillant sur l'emv** (c'est ce qu'il faudra faire).

## Cinquième partie

# Approche paramétrique vs non paramétrique pour les modèles de densité et de régression

### Dans ce chapitre

- Nous étudions deux problèmes classiques de la *théorie de l'estimation* : la **densité** et la **régression**.
- A travers ces deux problèmes, nous étudions le compromis entre les erreurs d'*estimation* et d'*approximation*.
- Ce compromis sera notamment étudié en confrontant l'approche *paramétrique* à l'approche *non paramétrique*.

### L'estimation de densité.

- Les données  $x_1, \dots, x_n$  telles que  $x_i \in \mathbb{R}$ .
- L'échantillon :  $X_1, \dots, X_n$  i.i.d. de loi **P inconnue**.
- On suppose que **P** admet une densité  $f$  (qui est donc **inconnue**).

### Le problème

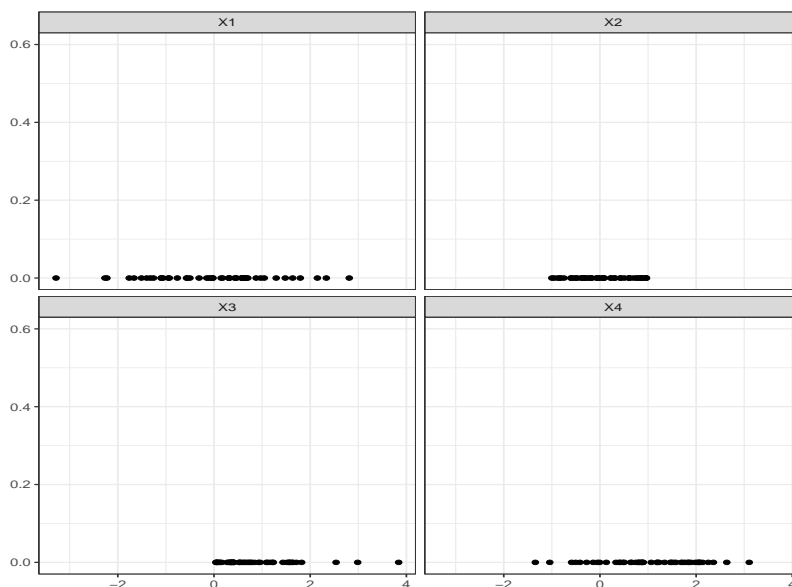
Estimer  $f$ .

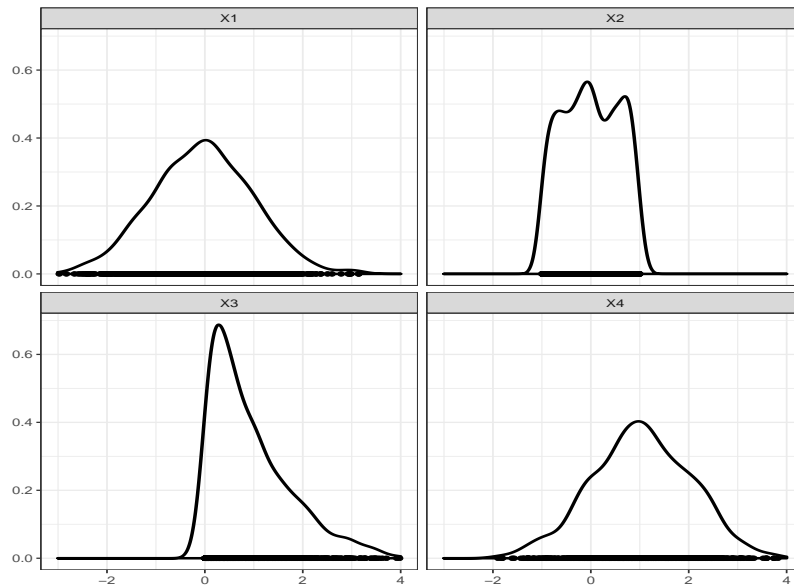
### Performance d'un estimateur

On mesurera la performance d'un estimateur  $\hat{f}(\cdot) = \hat{f}(\cdot, X_1, \dots, X_n)$  par son *risque quadratique ponctuel* :

$$\mathcal{R}(\hat{f}(x)) = \mathbf{E}((\hat{f}(x) - f(x))^2) = b^2(\hat{f}(x)) + \mathbf{V}(\hat{f}(x)).$$

### Exemple





## Le problème de la régression

- *Données* :  $(x_1, y_1), \dots, (x_n, y_n)$ . On veut expliquer les sorties  $y_i \in \mathbb{R}$  par les entrées  $x_i \in \mathbb{R}^p$ .
- Les données sont des *réalisations de v.a.*  $(X_1, Y_1), \dots, (X_n, Y_n)$  i.i.d. telles qu'il existe une fonction **inconnue**  $m : \mathbb{R}^p \rightarrow \mathbb{R}$  vérifiant

$$Y_i = m(X_i) + \varepsilon_i, \quad i = 1, \dots, n$$

où les  $\varepsilon_i$  sont i.i.d de loi  $\mathcal{N}(0, \sigma^2)$ .

### Le problème

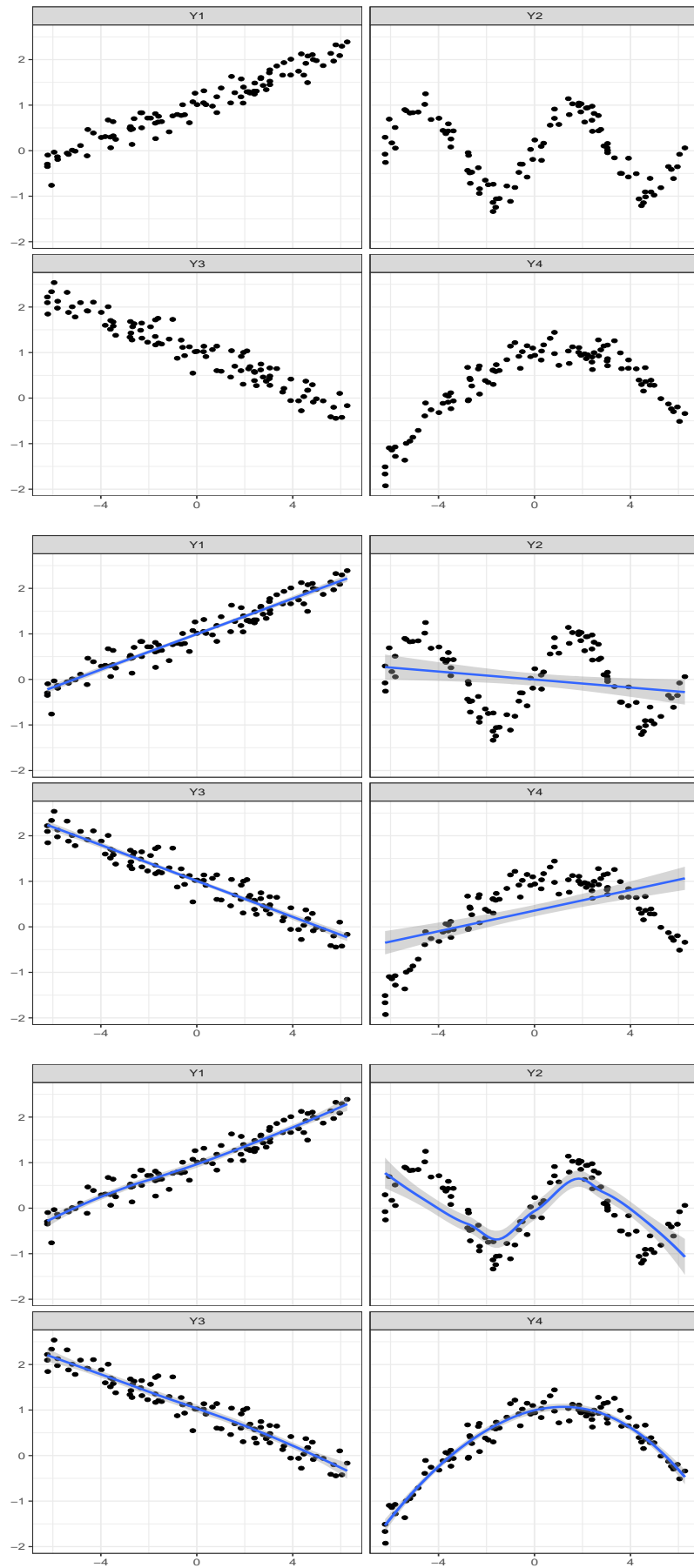
Estimer  $m$ .

### Performance d'un estimateur

On mesurera la performance d'un estimateur  $\hat{m}(\cdot) = \hat{m}(\cdot, X_1, \dots, X_n)$  par son *risque quadratique ponctuel* :

$$\mathcal{R}(\hat{m}(x)) = \mathbf{E}((\hat{m}(x) - m(x))^2) = b^2(\hat{m}(x)) + \mathbf{V}(\hat{m}(x)).$$

## Exemple





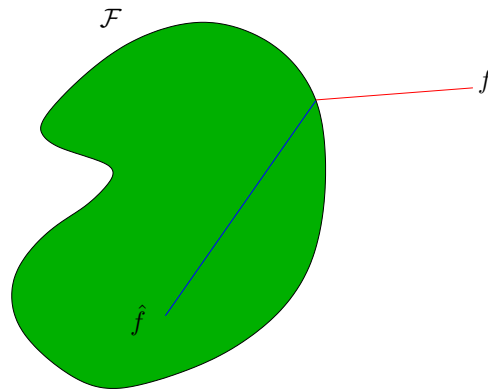
- Dans les deux cas, le problème est d'estimer une *fonction*.
- Poser un *modèle* revient à supposer que cette fonction appartient à un certain espace  $\mathcal{F}$ .

### Définition

- Si  $\mathcal{F}$  est de dimension **finie**, le modèle est *paramétrique*.
- Si  $\mathcal{F}$  est de dimension **infinie**, le modèle est *non paramétrique*.

### A priori

- *Non paramétrique* : plus flexible mais précision d'estimation plus faible.
- *Paramétrique* : meilleure précision d'estimation mais plus rigide.



- **Erreur d'estimation** : erreur commise par le choix d'une loi dans  $\mathcal{P}$  par rapport au meilleur choix.
- *Erreur d'approximation* : erreur commise par le choix de  $\mathcal{P}$ .

### Commentaire

Ces deux termes varient généralement *en sens inverse*.

## 1 Le modèle de densité

### 1.1 Approche paramétrique : le modèle Gaussien

- $X_1, \dots, X_n$  i.i.d. de densité  $f$  **inconnue**.
- On suppose que  $f \in \mathcal{F} = \{f_\theta, \theta \in \Theta\}$  avec  $\Theta$  de dimension *finie*.

#### Exemple : le modèle Gaussien

- On suppose  $f \in \mathcal{F} = \{f_{\mu, \sigma^2}, \mu \in \mathbb{R}, \sigma^2 > 0\}$ .
- *Le problème* : estimer  $\mu$  et  $\sigma^2$ .
- On peut estimer ces paramètres par *maximum de vraisemblance* :

$$\hat{\mu} = \bar{X}_n \quad \text{et} \quad \widehat{\sigma^2} = \frac{1}{n} \sum_{i=1}^n (X_i - \bar{X}_n)^2.$$

- On montre "facilement" que

$$\mathbf{E}[(\hat{\mu} - \mu)^2] = O\left(\frac{1}{n}\right) \quad \text{et} \quad \mathbf{E}[(\widehat{\sigma^2} - \sigma^2)^2] = O\left(\frac{1}{n}\right).$$

- En notant  $\theta = (\mu, \sigma^2)$ , on déduit

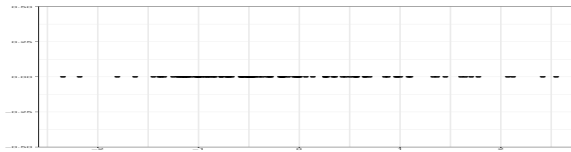
$$\mathbf{E}[\|\hat{\theta} - \theta\|^2] = O\left(\frac{1}{n}\right).$$

### Remarque

$1/n$  est la *vitesse paramétrique* classique pour l'erreur quadratique.

## Exemple

```
> df <- data.frame(X=rnorm(100))
> ggplot(df)+aes(x=X,y=0)+geom_point()+theme_bw()
```

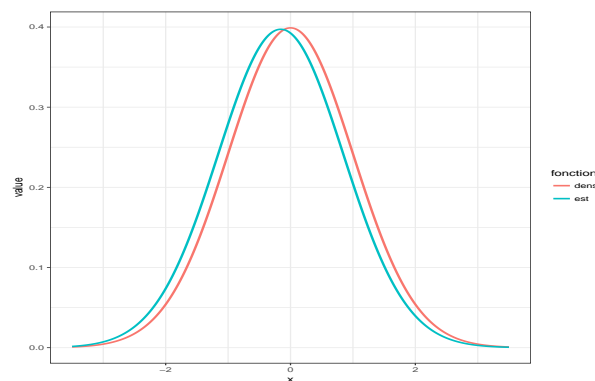


— On *estime*  $\mu$  et  $\sigma^2$  :

```
> theta <- c(mean(df$X),var(X))
> theta
[1] -0.1567617  1.0088300
```

— On trace l'*estimateur* et on le compare à la densité à estimer :

```
> x <- seq(-3.5,3.5,by=0.01); dens <- dnorm(x,mean=0,sd=1)
> est <- dnorm(x,mean=theta[1],sd=sqrt(theta[2]))
> df1 <- data.frame(x,dens,est); df2 <- melt(df1,id.vars="x")
> names(df2)[2] <- "fonction"
> ggplot(df2)+aes(x=x,y=value,color=fonction)+geom_line(size=1)+theme_bw()
```



## 1.2 Approche non paramétrique : l'estimateur à noyau

### Des moyennes locales

- En l'absence d'hypothèse paramétrique forte, on se base sur ce qui se passe au *voisinage de x* pour estimer  $f(x)$ .
- L'*histogramme* est un estimateur non paramétrique bien connu.

### L'*histogramme*

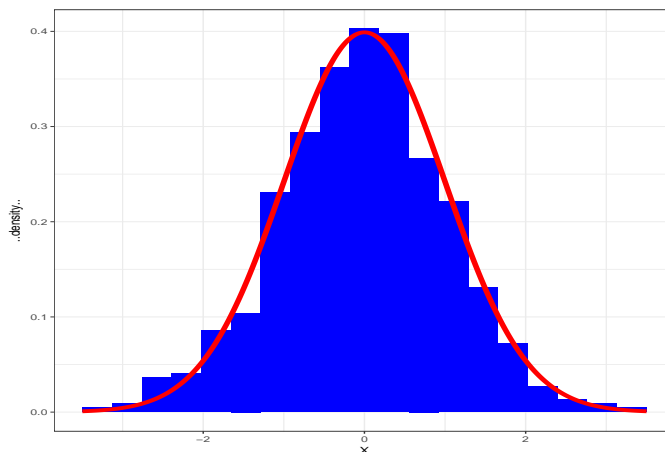
- $\mathcal{P} = \{I_1, \dots, I_K\}$  une **partition** de  $\mathbb{R}$  en  $K$  intervalles.
- L'**histogramme** est défini par

$$\hat{f}(x) = \frac{1}{n\lambda(I(x))} \sum_{i=1}^n \mathbf{1}_{X_i \in I(x)},$$

où  $I(x)$  désigne l'intervalle qui contient  $x$  et  $\lambda(I)$  la longueur de l'intervalle  $I$ .

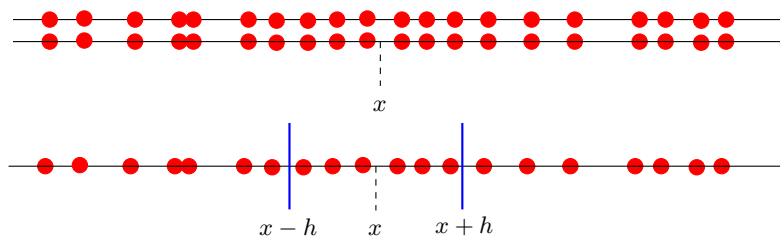
## Exemple

```
> ggplot(df)+aes(x=X,y=..density..)+geom_histogram(bins=20,fill="blue")+  
  geom_line(data=df1,aes(x=x,y=dens),color="red",size=2)+theme_bw()
```



## Estimateurs à noyau

- L'histogramme n'est *pas continu*.
- L'*estimateur à noyau* permet de pallier à ce problème en *ne fixant pas de partition*.
- L'idée est d'utiliser une *fenêtre glissante*.
- $n = 20$  observations.
- On veut estimer la densité en  $x$ .
- On considère une fenêtre  $[x - h, x + h]$ .



- On fait comme pour l'histogramme

$$\hat{f}(x) = \frac{1}{2nh} \sum_{i=1}^n \mathbf{1}_{X_i \in [x-h, x+h]}.$$

- On peut réécrire cet *estimateur*

$$\begin{aligned} \hat{f}(x) &= \frac{1}{2nh} \sum_{i=1}^n \mathbf{1}_{X_i \in [x-h, x+h]} = \frac{1}{nh} \sum_{i=1}^n \frac{1}{2} \mathbf{1}_{-1 \leq \frac{x-X_i}{h} \leq 1} \\ &= \frac{1}{nh} \sum_{i=1}^n K\left(\frac{x-X_i}{h}\right) \end{aligned}$$

avec  $K(u) = \frac{1}{2} \mathbf{1}_{[-1,1]}(u)$ .

## Estimateur à noyau de la densité

### Définition [Parzen, 1962]

Etant donné  $h > 0$  et  $K : \mathbb{R} \rightarrow \mathbb{R}$  intégrable et tel que  $\int K(u) du = 1$ , l'estimateur à noyau de la densité est défini par

$$\hat{f}(x) = \frac{1}{nh} \sum_{i=1}^n K\left(\frac{x - X_i}{h}\right).$$

### Remarque

L'utilisateur doit *choisir deux paramètres* : un réel positif  $h$  et un noyau  $K$

### Exemples de noyau

Les noyaux suivants sont les plus utilisés :

— *Uniforme* :

$$K(u) = \frac{1}{2} \mathbf{1}_{[-1,1]}(u).$$

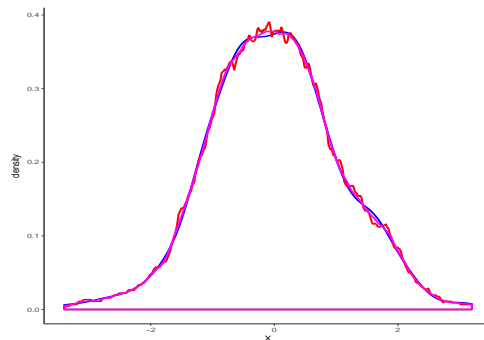
— *Gaussien* :

$$K(u) = \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{u^2}{2}\right).$$

— *Epanechnikov* :

$$K(u) = \frac{3}{4}(1 - u^2) \mathbf{1}_{[-1,1]}(u).$$

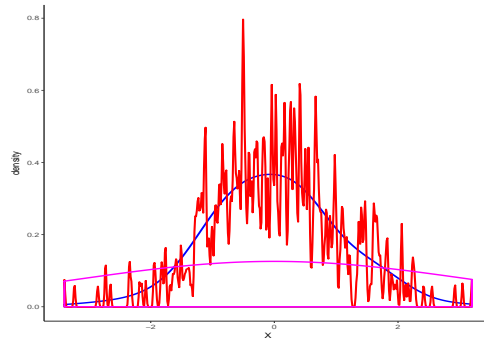
```
> X <- rnorm(500)
> df <- data.frame(X)
> ggplot(df)+aes(X)+geom_density(kernel=c("gaussian"),color="blue",size=1)+
  geom_density(kernel=c("rectangular"),color="red",size=1)+
  geom_density(kernel=c("epanechnikov"),color="black",size=1)+theme_classic()
```



### Conclusion

Le choix du noyau n'est généralement *pas primordial* sur la performance de l'estimateur.

```
> X <- rnorm(500)
> df <- data.frame(X)
> ggplot(df)+aes(X)+geom_density(bw=0.4,color="blue",size=1)+
  geom_density(bw=0.01,color="red",size=1)+
  geom_density(bw=3,color="magenta",size=1)+theme_classic()
```



### Conclusion

Le choix de la fenêtre  $h$  est *crucial* sur la performance de l'estimateur.

### Choix de $h$

- $h$  grand : fenêtre grande  $\implies$  beaucoup d'observations dans les fenêtres  $\implies$  densités proches  $\forall x \implies$  *biais fort*, *variance faible*.
- $h$  petit : fenêtre petite  $\implies$  peu d'observations dans les fenêtres  $\implies$  densités instables  $\forall x \implies$  *biais faible*, *variance forte*.

### Conclusion

- Le paramètre  $h$  régule le *compromis biais/variance* de l'estimateur à noyau.
- On sait le quantifier mathématiquement.

### Contrôle de la variance

#### Théorème

On suppose que :

- $f$  est bornée.
- $K$  est tel que  $\int K(u) du = 1$ ,  $\int uK(u) du = 0$  et  $\int K(u)^2 du < +\infty$ .

On a alors  $\forall x \in \mathbb{R}, \forall h > 0$  et  $\forall n \geq 1$

$$\mathbf{V}[\hat{f}(x)] = O\left(\frac{1}{nh}\right).$$

#### Remarque

On retrouve bien que la *variance* est *faible* lorsque  $h$  est *grand* et réciproquement.

### Contrôle du biais

- Pour le terme de biais, il faut supposer un peu de *régularité* sur la densité à estimer.

#### Théorème

On suppose que

- la densité  $f$  est *dérivable* et que sa dérivée est *Lipschitzienne* :

$$|f'(x) - f'(y)| \leq L|x - y|, \quad \forall x, y \in \mathbb{R};$$

- $K$  est tel que  $\int u^2 K(u) du < +\infty$ .

On a alors  $\forall x \in \mathbb{R}$

$$|b(\hat{f}(x))| = O(h^2).$$

#### Remarque

On retrouve bien le *biais* est *faible* lorsque  $h$  est *petit* et réciproquement.

Modèle	param	non-param
Vitesse	$n^{-1}$	$n^{-\frac{4}{5}}$

## Risque quadratique

### Corollaire (convergence $L_2$ )

Sous les hypothèses des deux théorèmes précédents, on déduit que si  $h \rightarrow 0$  et  $nh \rightarrow +\infty$  alors le risque quadratique de  $\hat{f}(x)$  tend vers 0 (**convergence en moyenne d'ordre 2**).

### Corollaire (choix de $h$ )

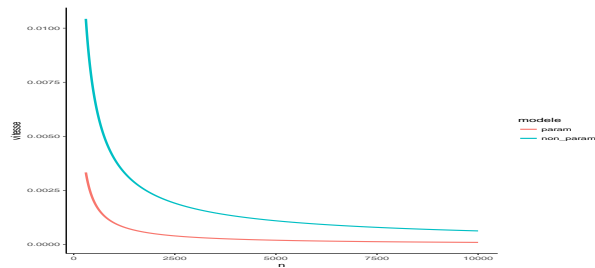
Le  $h^*$  qui **minimise l'erreur quadratique** vérifie

$$h^* = Cn^{-\frac{1}{5}}.$$

Pour cette valeur de  $h$ , on a

$$\mathcal{R}(\hat{f}(x)) = \mathbf{E}[(\hat{f}(x) - f(x))^2] = O\left(n^{-\frac{4}{5}}\right).$$

### Remarque importante



### Conclusion

- La convergence est *moins rapide* dans les modèles **non-paramétrique**.
- C'est le *prix à payer* pour **plus de flexibilité**.
- La théorie nous dit que le  $h$  optimal est

$$h^* = Cn^{-\frac{1}{5}}.$$

- Ce résultat n'est quasiment d'*aucune utilité pratique*.
- En pratique, il existe un grand nombre de *procédures automatiques* (plus ou moins performantes selon les cas) permettant de sélectionner  $h$ .

## 2 Le modèle de régression

### Présentation du modèle

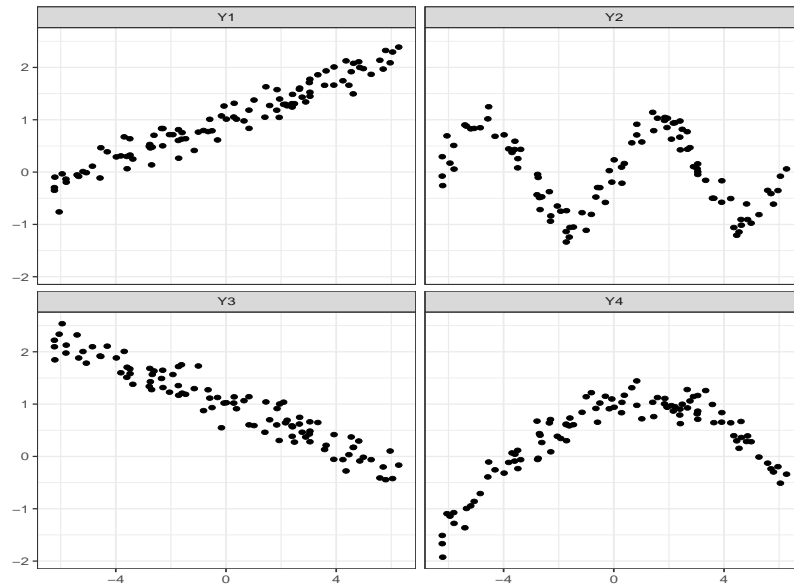
- *Les données* :  $(x_1, y_1), \dots, (x_n, y_n)$  où  $y_i \in \mathbb{R}$  et  $x_i \in \mathbb{R}$  (pour simplifier).
- *L'échantillon*  $(x_1, Y_1), \dots, (x_n, Y_n)$  i.i.d. (on suppose que les  $x_i$  sont déterministes).
- *Le problème* : expliquer les sorties  $Y_i$  par les entrées  $X_i$ .
- *La fonction de régression* : c'est la fonction  $m : \mathbb{R} \rightarrow \mathbb{R}$  telle que

$$Y_i = m(x_i) + \varepsilon_i$$

où les termes d'erreurs  $\varepsilon_i$  sont i.i.d. de loi  $\mathcal{N}(0, \sigma^2)$ .

- *Le problème statistique* : estimer  $m$ .

## Exemples



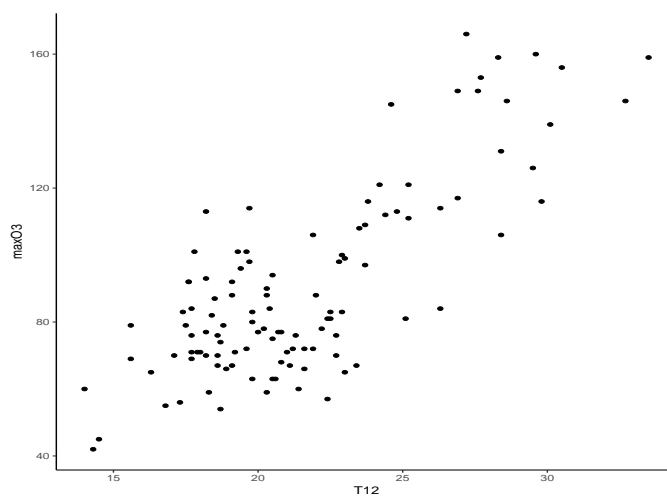
## Un exemple concret

- On souhaite expliquer la *concentration en ozone* par la température à 12h.
- $n = 112$  observations :

```
> ozone %>% select(maxO3,T12) %>% head()
      maxO3  T12
20010601   87 18.5
20010602   82 18.4
20010603   92 17.6
20010604  114 19.7
20010605   94 20.5
20010606   80 19.8
```

## Représentation du nuage

```
> ggplot(ozone)+aes(x=T12,y=maxO3)+geom_point()+theme_classic()
```



## 2.1 Approche paramétrique : le modèle de régression linéaire

### Le modèle linéaire

- On fait l'hypothèse que la fonction de régression est *linéaire* :

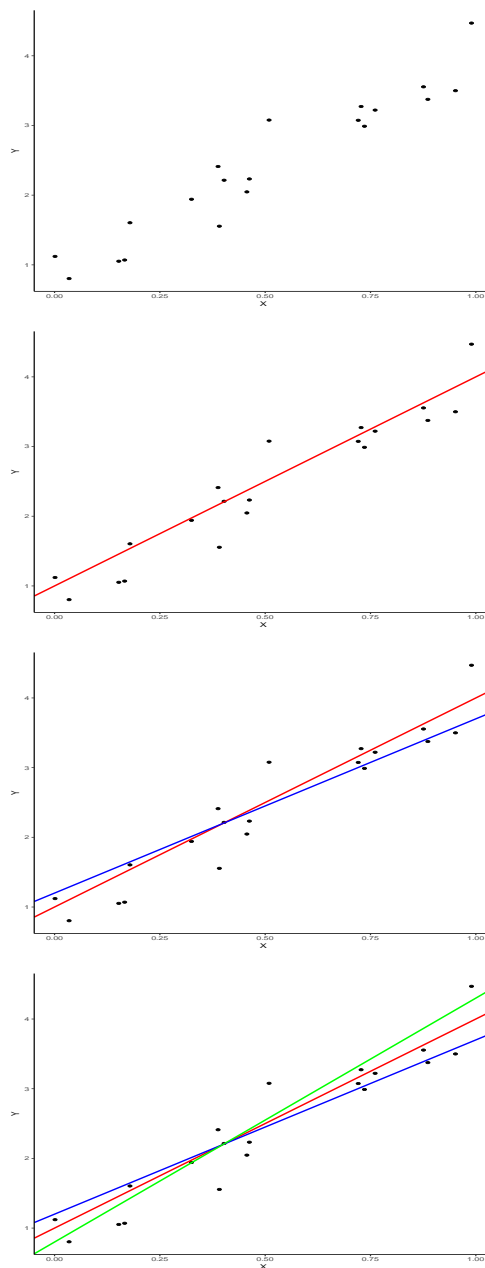
$$m(x) = \beta_0 + \beta_1 x, \quad \beta_0 \in \mathbb{R}, \beta_1 \in \mathbb{R}.$$

- Paramètres *inconnus* à estimer :  $\beta = (\beta_0, \beta_1) \in \mathbb{R}^2 \implies$  modèle *paramétrique*.

### Ajustement linéaire d'un nuage de points

#### Notations

- $n$  observations  $y_1, \dots, y_n$  de la *variable à expliquer* (maxO3).
- $n$  observations  $x_1, \dots, x_n$  de la *variable explicative* (T12).



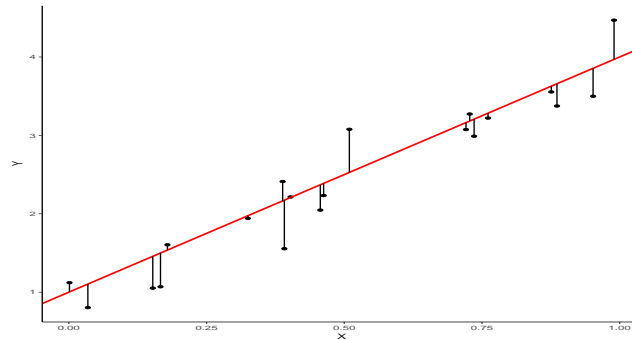


## Le problème

Trouver la droite qui ajuste *au mieux* le nuage de points.

- On cherche  $y = \beta_0 + \beta_1 x$  qui *ajuste au mieux le nuage des points*.
- Toutes les observations mesurées ne se trouvent **pas sur une droite** :

$$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i.$$



## Idée

Chercher à minimiser les **erreurs** ou les **bruits**  $\varepsilon_i$ .

## Le critère des moindres carrés

### Critère des MC

On cherche  $\beta = (\beta_0, \beta_1)$  qui minimise

$$\sum_{i=1}^n \varepsilon_i^2 = \sum_{i=1}^n (y_i - \beta_0 - \beta_1 x_i)^2.$$

## Solution

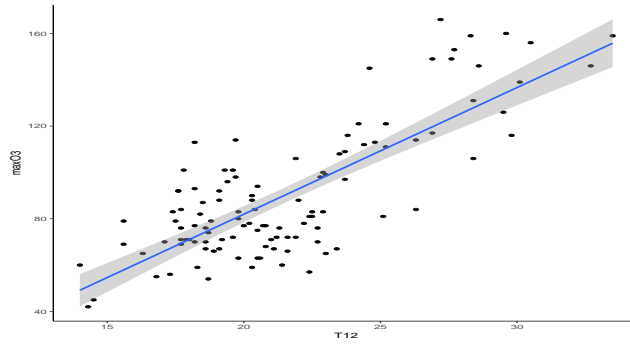
La solution est donnée par :

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x} \quad \text{et} \quad \hat{\beta}_1 = \frac{\sum_{i=1}^n (y_i - \bar{y})(x_i - \bar{x})}{\sum_{i=1}^n (x_i - \bar{x})^2}$$

à condition que tous les  $x_i$  ne soient pas égaux.

## Application à l'ozone

```
> modele.lin <- lm(maxO3~T12,data=ozone)
> modele.lin
Coefficients:
(Intercept)      T12
    -27.420      5.469
> ggplot(ozone)+aes(x=T12,y=maxO3)+geom_point()+theme_classic()+
  geom_smooth(method="lm")
```



## Les estimateurs des MCO

### Rappels

- Le *modèle*

$$Y_i = \beta_0 + \beta_1 x_i + \varepsilon_i, \quad i = 1, \dots, n,$$

où les  $\varepsilon_i$  sont i.i.d. de loi  $\mathcal{N}(0, \sigma^2)$ .

- Les *estimateurs des MCO* :

$$\hat{\beta}_0 = \bar{Y} - \hat{\beta}_1 \bar{x} \quad \text{et} \quad \hat{\beta}_1 = \frac{\sum_{i=1}^n (Y_i - \bar{Y})(x_i - \bar{x})}{\sum_{i=1}^n (x_i - \bar{x})^2}.$$

### Propriétés

- *Biais* :  $\mathbf{E}[\hat{\beta}_0] = \beta_0$  et  $\mathbf{E}[\hat{\beta}_1] = \beta_1$ .
- *Variance* :

$$\mathbf{V}(\hat{\beta}_0) = \sigma^2 \frac{\sum_{i=1}^n x_i^2}{n \sum_{i=1}^n (x_i - \bar{x})^2} \quad \text{et} \quad \mathbf{V}(\hat{\beta}_1) = \frac{\sigma^2}{\sum_{i=1}^n (x_i - \bar{x})^2}.$$

### Quelques remarques

- Les estimateurs des MCO sont *sans biais*.
- Sous des hypothèses peu contraignantes, on montre que leur *variance est en*  $1/n$ . On déduit

$$\mathcal{R}(\hat{\beta}_0) = O\left(\frac{1}{n}\right) \quad \text{et} \quad \mathcal{R}(\hat{\beta}_1) = O\left(\frac{1}{n}\right).$$

### Conclusion

Les estimateurs des MCO atteignent la *vitesse paramétrique* classique en  $1/n$ .

- On peut également obtenir la *loi* des estimateurs  $\hat{\beta}_0$  et  $\hat{\beta}_1$ .
- On déduit de cette loi des *intervalles de confiance* et des procédures de tests statistiques.

### IC et tests pour l'ozone

- *Intervalles de confiance* :

```
> confint(modele.lin)
              2.5 %      97.5 %
(Intercept) -45.321901 -9.517371
T12          4.651219  6.286151
```

- *Tests statistique* :

```
> summary(modele.lin)$coefficients
              Estimate Std. Error t value Pr(>|t|)
(Intercept) -27.419636   9.0334940  -3.03533 2.999431e-03
T12          5.468685   0.4124939  13.25761 1.512025e-24
```

## 2.2 Approche non paramétrique : l'estimateur à noyau

- En l'absence d'hypothèse paramétrique (forte), on regarde ce qui se passe au *voisinage du point* où on cherche à estimer la fonction de régression.
- Les méthodes non paramétriques consistent donc à définir des *voisinages* et à faire des *moyennes locales* à l'intérieur des voisinages :

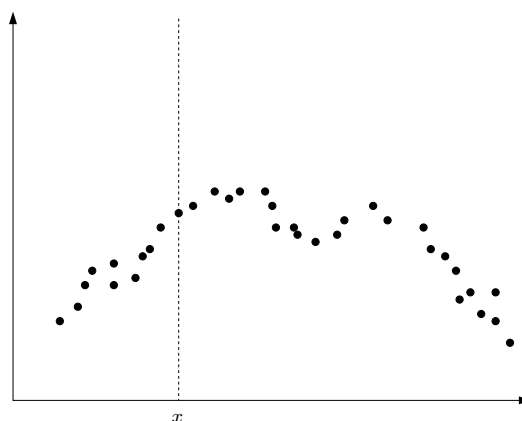
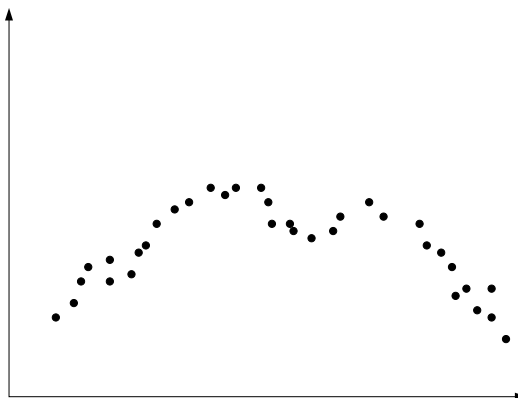
$$\hat{m}_n(x) = \sum_{i=1}^n W_{ni}(x) Y_i$$

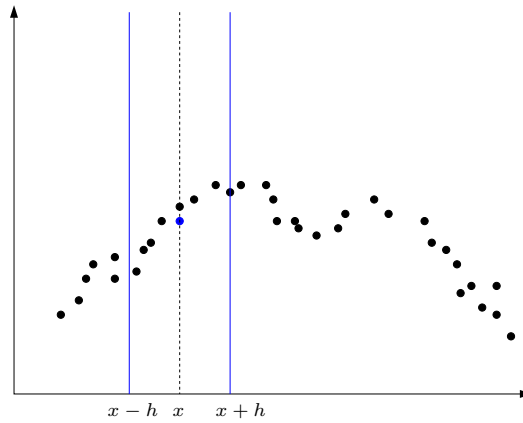
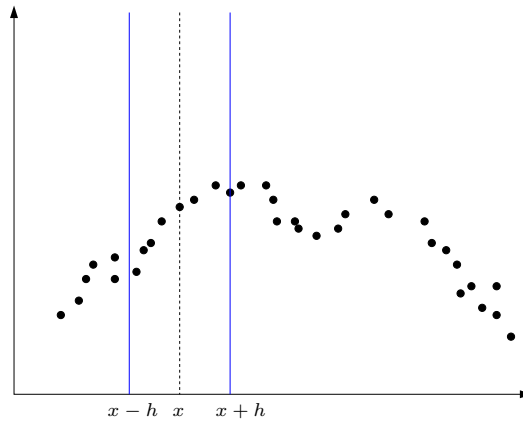
où  $W_{ni}(x)$  représente le **poids** à accorder à la  $i$ ème observation pour estimer  $m$  en  $x$ .

- Nous illustrons ce principe à travers l'estimateur de *Nadaraya Watson* [Nadaraya, 1964, Watson, 1964] (on aurait aussi pu faire l'algorithme des *plus proches voisins*).

### La méthode

- $(x_1, Y_1), \dots, (x_n, Y_n)$  i.i.d.
- *But* : estimer  $m$  tel que  $Y = m(x) + \varepsilon$ .





— L'estimateur s'écrit

$$\hat{m}_n(x) = \frac{\sum_{i=1}^n \mathbf{1}_{x-h \leq X_i \leq x+h} Y_i}{\sum_{i=1}^n \mathbf{1}_{x-h \leq X_i \leq x+h}} = \frac{\sum_{i=1}^n \mathbf{1}_{\left| \frac{X_i - x}{h} \right| \leq 1} Y_i}{\sum_{i=1}^n \mathbf{1}_{\left| \frac{X_i - x}{h} \right| \leq 1}}.$$

### Définition

Soit  $h > 0$  et  $K : \mathbb{R} \rightarrow \mathbb{R}^+$ . L'estimateur à noyau de *fenêtre*  $h$  et de *noyau*  $K$  est défini par

$$\hat{m}_n(x) = \frac{\sum_{i=1}^n K\left(\frac{X_i - x}{h}\right) Y_i}{\sum_{i=1}^n K\left(\frac{X_i - x}{h}\right)}.$$

### Noyau et fenêtre

— *Noyau usuel* :

1. **Uniforme** :  $K(x) = \mathbf{1}_{|x| \leq 1}$  ;
2. **Gaussien** :  $K(x) = \exp(-|x|^2)$  ;
3. **Epanechnikov** :  $K(x) = \frac{3}{4}(1 - x^2)\mathbf{1}_{|x| \leq 1}$ .

— Le choix de  $h$  est *crucial* pour la qualité de l'estimation :

1.  **$h$  grand** : estimateur « constant », variance faible, biais fort ;
2.  **$h$  petit** : « interpolation », variance forte, biais faible ;

### Un exemple

- On génère un échantillon  $(X_I, Y_I), i = 1, \dots, n = 200$  selon

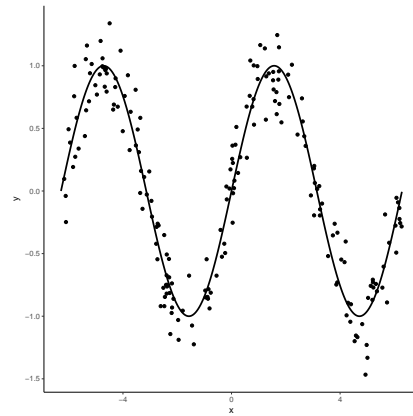
$$Y_i = \sin(X_i) + \varepsilon_i, \quad i = 1, \dots, n$$

avec  $X_i$  uniforme sur  $[-2\pi, 2\pi]$ ,  $\varepsilon_i$  de loi gaussienne  $\mathcal{N}(0, 0.2^2)$ .

```

> n <- 200; set.seed(1234)
> X <- runif(n,-2*pi,2*pi)
> set.seed(5678)
> eps <- rnorm(n,0,0.2)
> Y <- sin(X)+eps
> df <- data.frame(X=X,Y=Y)
> x <- seq(-2*pi,2*pi,by=0.01)
> df1 <- data.frame(x=x,y=sin(x))
> ggplot(df1)+aes(x=x,y=y)+
  geom_line(size=1)+
  geom_point(data=df,aes(x=X,y=Y))

```

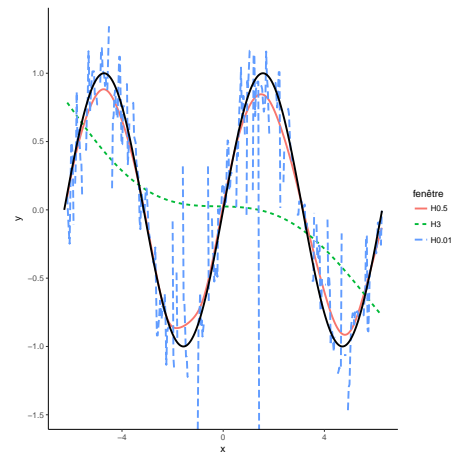


— La fonction *locpoly* du package **kernSmooth** permet de construire des estimateurs à noyau.

```

> h1 <- 0.5; h2 <- 3; h3 <- 0.01
> fx1 <- locpoly(X,Y,bandwidth=h1)
> fx2 <- locpoly(X,Y,bandwidth=h2)
> fx3 <- locpoly(X,Y,bandwidth=h3)
> df1 <- data.frame(x=x,y=sin(x))
> df2 <- data.frame(x=fx1$y,
  "H0.5"=fx1$y, "H3"=fx2$y,
  "H0.01"=fx3$y)
> df22 <- melt(df2,id.vars=1)
> names(df22)[2:3] <- c("fenêtre",
  "y")
> ggplot(df22)+aes(x=x,y=y)+
  geom_line(aes(color=fenêtre,
    lty=fenêtre))+geom_line
  (data=df1,aes(x=x,y=y),size=1)

```



## Propriétés des estimateurs

- Là encore, on peut quantifier le *compromis biais/variance*.
- On considère le noyau uniforme et on suppose que  $m$  est dérivable et que sa dérivée est Lipschitzienne :

$$|m'(x) - m'(y)| \leq L|x - y|, \quad \forall x, \forall y \in \mathbb{R}.$$

### Théorème

Sous les hypothèses ci-dessus, on a

$$|b(\hat{m}_n(x))| = O(h^2) \quad \text{et} \quad \mathbf{V}[\hat{m}_n(x)] = O\left(\frac{1}{nh}\right).$$

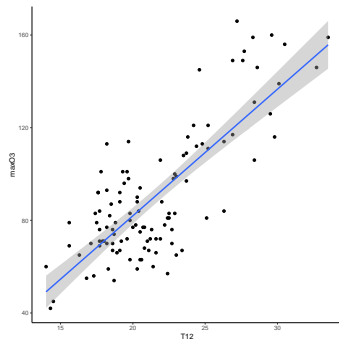
- Toutes les remarques faites pour l'estimateur à noyau de la densité sont valables pour l'estimateur de Nadaraya Watson.
- Le  $h$  optimal est de l'ordre de  $n^{-1/5}$ . Pour cette valeur de  $h$ , le *risque quadratique* est de l'ordre de  $n^{-4/5}$ .
- On obtient donc une vitesse de convergence *plus lente* que pour les estimateurs paramétriques.
- C'est le *prix à payer* pour un modèle *plus flexible*.

## Retour à l'ozone

### Paramétrique (linéaire)

```
> ggplot(ozone)+aes(x=T12,y=maxO3)+  
geom_point()+
```

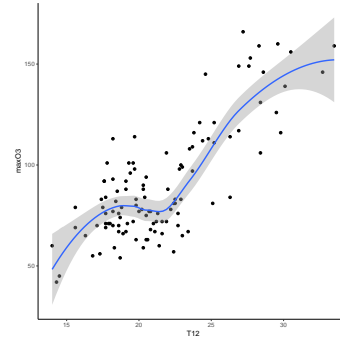
```
geom_smooth(method="lm",size=1)+  
theme_classic()
```



### Non paramétrique

```
> ggplot(ozone)+aes(x=T12,y=maxO3)+  
geom_point()+
```

```
geom_smooth(size=1)+  
theme_classic()
```



## 3 Bibliographie

### Références

#### Biblio5

- [Nadaraya, 1964] Nadaraya, E. A. (1964). On estimating regression. *Theory of Probability and its Applications*, 9.
- [Parzen, 1962] Parzen, E. (1962). On estimation of a probability density function and mode. *Ann. Math. Stat.*, 33 :1065–1076.
- [Watson, 1964] Watson, G. S. (1964). Smooth regression analysis. *Sankhya : The Indian Journal of Statistics, Series A*, 26 :359–372.