Statistique

L. Rouvière

laurent.rouviere@univ-rennes2.fr

Septembre 2019

Première partie I

La modélisation statistique

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.

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.

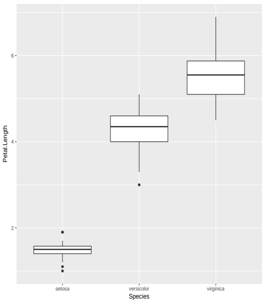
```
> data(iris)
> head(iris)
  Sepal.Length Sepal.Width Petal.Length Petal.Width Species
1          5.1         3.5          1.4         0.2  setosa
2          4.9         3.0          1.4         0.2  setosa
3          4.7         3.2          1.3         0.2  setosa
4          4.6         3.1          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
```

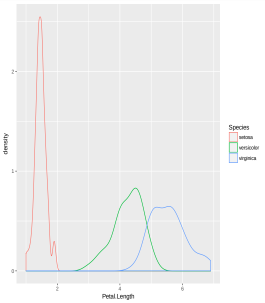
```
> summary(iris)
  Sepal.Length    Sepal.Width     Petal.Length    Petal.Width          Species
 Min.   :4.300   Min.   :2.000   Min.   :1.000   Min.   :0.100   setosa    :50
 1st Qu.:5.100   1st Qu.:2.800   1st Qu.:1.600   1st Qu.:0.300   versicolor:50
 Median :5.800   Median :3.000   Median :4.350   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
```

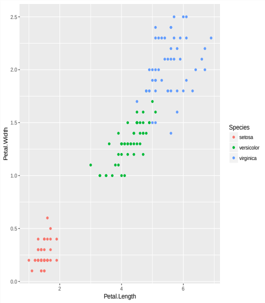
Statistiques descriptives

- 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 =

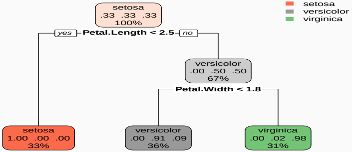
Modélisation

- 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 connaitre (prédire, estimer...) l'espèce de chacun.

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 connaitre (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 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 succés d'une étude statistique.

Dans ce cours

- On va s'intéresser à la phase de modélisation mathématique d'un problème.

- Chacune de ces étapes est primordiale pour le succés 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 $x_1, \ldots, x_n$.

- Chacune de ces étapes est primordiale pour le succés 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 $x_1, \ldots, x_n$.
- Les phases d'interprétation et de visualisation des résultats seront abordées plus tard.

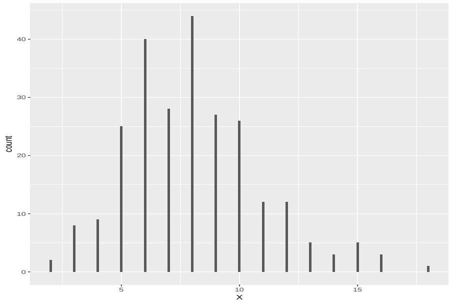
Nombre de voitures à un feu rouge

- Afin de mieux gérer la circulation, on s'intéresse au nombre de voitures à un feu rouge sur un créneau donné.

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 ?

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.

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.

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.

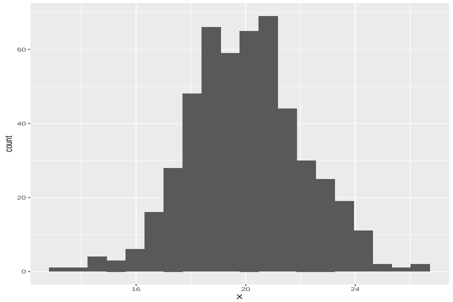
Durée d'un trajet

- 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 ?

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 ?

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.

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.

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.

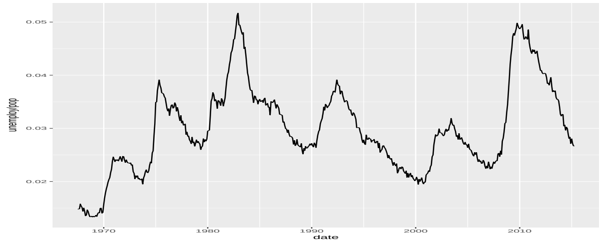
Séries temporelles

- On s'intéresse au taux de chomage d'une population entre deux dates t_0 et t_1. On souhaite prédire le taux de chomage futur.

Séries temporelles

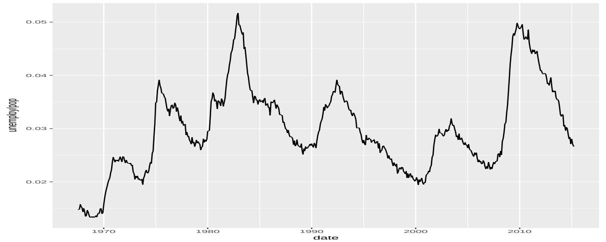
- On s'intéresse au taux de chomage d'une population entre deux dates t_0 et t_1. On souhaite prédire le taux de chomage futur.
- Expérience : on mesure le taux de chomage entre les deux dates

```
 > 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 chomage en 2012 ?

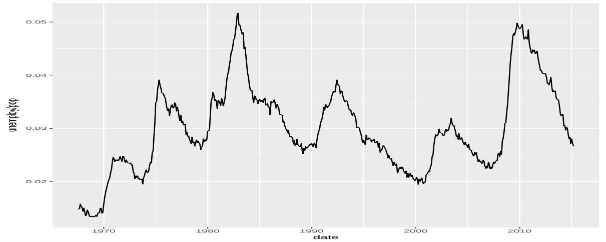


Question

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

Quantité d'intérêt

- Il serait intéressant d'avoir de l'information sur la loi de probabilité du taux de chomage à l'instant t sachant le taux de chomage avant t.

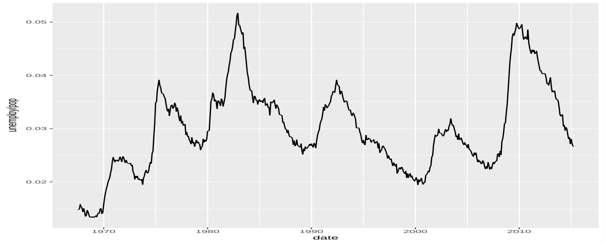


Question

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

Quantité d'intérêt

- Il serait intéressant d'avoir de l'information sur la loi de probabilité du taux de chomage à l'instant t sachant le taux de chomage avant t.
- On dispose juste de mesures, cette loi est donc inconnue.



Question

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

Quantité d'intérêt

- Il serait intéressant d'avoir de l'information sur la loi de probabilité du taux de chomage à l'instant t sachant le taux de chomage 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

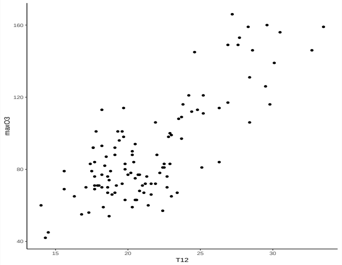
- On s'intéresse à la prévision de la concentration en ozone dans l'air.

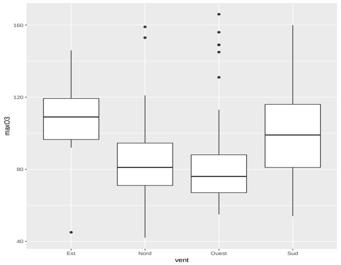
Prévision ozone

- 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.

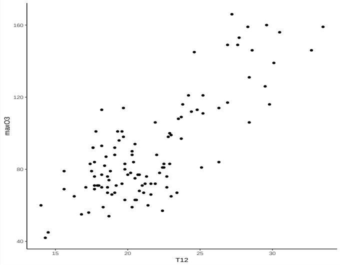
```
> head(ozone)
         maxO3   T9  T12  T15 Ne9 Ne12 Ne15     Vx9    Vx12    Vx15 maxO3v  vent pluie
20010601    87 15.6 18.5 18.4   4    4    8  0.6946 -1.7101 -0.6946     84  Nord   Sec
20010602    82 17.0 18.4 17.7   5    5    7 -4.3301 -4.0000 -3.0000     87  Nord   Sec
20010603    92 15.3 17.6 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
```

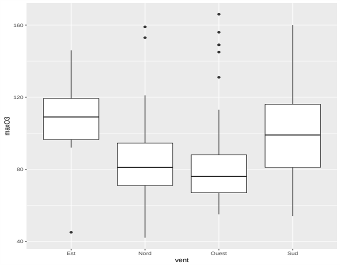
```
> ggplot(ozone)+aes(x=T12,y=maxO3)+geom_point()
> ggplot(ozone)+aes(x=vent,y=maxO3)+geom_boxplot()
```





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

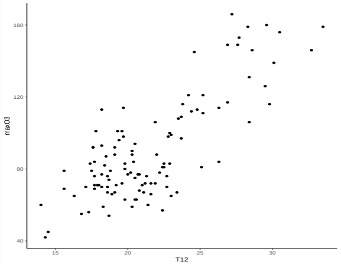


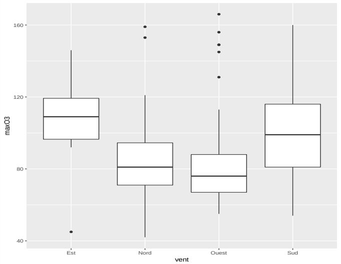


Question

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

```
> 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

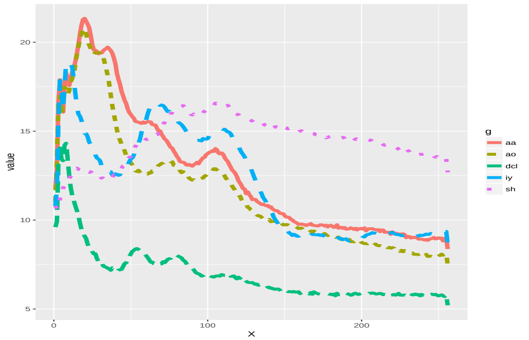
- On souhaite développer une procédure automatique permettant de reconnaitre un son.

Reconnaissance de la voix

- On souhaite développer une procédure automatique permettant de reconnaitre 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.

Reconnaissance de la voix

- On souhaite développer une procédure automatique permettant de reconnaitre 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 ?

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).

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)

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...).

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.

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

Modèle statistique

- Définition avec des mots : vision simplifiée de la réalité.

- 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.

- 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 ?

- On suppose que des données ont été collectées.

- On suppose que des données ont été collectées.
- Ces données sont le résultat d'une expérience répétée n fois.

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

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

Exemple des durées de trajet

- Données :

```
20.87 22.12 20.90 21.33 17.73
```

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

Exemple des durées de trajet

- Données :

```
20.87 22.12 20.90 21.33 17.73
```

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

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 ?

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.

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.

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.

Variable aléatoire

Un outil spécifique

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

Définition

Soit $(\Omega, \mathcal{A})$ un espace de probabilité et $(\Omega', \mathcal{A}')$ un espace mesurable. Une variable aléatoire à valeurs dans un ensemble F est une application

$$X : (\Omega, \mathcal{A}) \to (\Omega', \mathcal{A}')$$

telle que

$$\forall A \in \mathcal{A}', X^{-1}(A) \in \mathcal{A}.$$

Variable aléatoire

Un outil spécifique

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

Définition

Soit $(\Omega, \mathcal{A})$ un espace de probabilité et $(\Omega', \mathcal{A}')$ un espace mesurable. Une variable aléatoire à valeurs dans un ensemble F est une application

$$X : (\Omega, \mathcal{A}) \to (\Omega', \mathcal{A}')$$

telle que

$$\forall A \in \mathcal{A}', X^{-1}(A) \in \mathcal{A}.$$

- 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

- $x_1, \ldots, x_n$ représentent le résultat de l'expérience. On suppose que $x_i \in \mathbb{R}, i = 1, \ldots, n$.

V.a. et modélisation

- $x_1, \ldots, x_n$ représentent le résultat de l'expérience. On suppose que $x_i \in \mathbb{R}, i = 1, \ldots, n$.
- Pour prendre en compte l'aléa de l'expérience, on va considérer des variables aléatoires réelles (v.a.r.). C'est-à-dire que l'espace d'arrivée est $\mathbb{R}$ muni de sa tribu $\mathcal{B}(\mathbb{R})$.

- $x_1, \dots, x_n$ 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.). C'est-à-dire que l'espace d'arrivée est $\mathbb{R}$ muni de sa tribu $\mathcal{B}(\mathbb{R})$.

Lien observation/v.a.r.

Les x_i sont dés 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. $X_1, \ldots, X_n$ et des éléments $\omega_1, \ldots, \omega_n$ tels que

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

- 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 ω_i ?

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

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

Question

Que représentent les ω_i ?

Réponse

- ω_i représente les conditions expérimentales associées à la i^{e} mesure, c'est-a-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

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

Exemple : durée de trajet

- $x_1 = 20.87, x_2 = 22.12, x_3 = 20.90, x_4 = 21.33, x_5 = 17.73, \ldots$
- $X_1, \ldots, X_n$ 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^{e} temps de trajet.

Exemple : durée de trajet

- $x_1 = 20.87, x_2 = 22.12, x_3 = 20.90, x_4 = 21.33, x_5 = 17.73, \ldots$
- $X_1, \ldots, X_n$ 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^{e} temps de trajet.
- L'ensemble Ω 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...).

Exemple : durée de trajet

- $x_1 = 20.87, x_2 = 22.12, x_3 = 20.90, x_4 = 21.33, x_5 = 17.73, \ldots$
- $X_1, \ldots, X_n$ 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^{e} temps de trajet.
- L'ensemble Ω 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...).
- ω_i correspondant à ce qui s'est produit sur le i^{e} trajet.

Exemple : durée de trajet

- $x_1 = 20.87, x_2 = 22.12, x_3 = 20.90, x_4 = 21.33, x_5 = 17.73, \ldots$
- $X_1, \ldots, X_n$ 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^{e} temps de trajet.
- L'ensemble Ω 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...).
- ω_i correspondant à ce qui s'est produit sur le i^{e} trajet.
- Par exemple ω_1 représente tout ce qui s'est passé sur le trajet permettant d'expliquer qu'on a mis 20.87 minutes.

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^{e} temps de trajet.
- L'ensemble Ω 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...).
- ω_i correspondant à ce qui s'est produit sur le i^{e} trajet.
- Par exemple ω_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 Ω et les $\omega_i, i = 1, \dots, n$.

Récapitulatif

- n observations $x_1, \ldots, x_n$ telles que $x_i \in \mathcal{H}$.
- Les n valeurs observées $x_1, \ldots, x_n$ sont des réalisations de variables aléatoires $X_1, \ldots, X_n$ à valeurs dans $(\mathcal{H}, \mathcal{A})$.

Récapitulatif

- n observations $x_1, \ldots, x_n$ telles que $x_i \in \mathcal{H}$.
- Les n valeurs observées $x_1, \ldots, x_n$ sont des réalisations de variables aléatoires $X_1, \ldots, X_n$ à valeurs dans $(\mathcal{H}, \mathcal{A})$.

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.

Récapitulatif

- n observations $x_1, \ldots, x_n$ telles que $x_i \in \mathcal{H}$.
- Les n valeurs observées $x_1, \ldots, x_n$ sont des réalisations de variables aléatoires $X_1, \ldots, X_n$ à valeurs dans $(\mathcal{H}, \mathcal{A})$.

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. $X_1, \ldots, X_n$ 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

Etant donnée $\mathbf{P}$ une probabilité sur $(\Omega, \mathcal{A})$ et X une variable aléatoire réelle définie sur Ω et à valeurs dans $(\Omega', \mathcal{A}')$, on appelle loi de probabilité de X la mesure $\mathbf{P}_X$ définie par

$$\mathbf{P}_X(A) = \mathbf{P}(X^{-1}(A)) = \mathbf{P}(X \in A) = \mathbf{P}(\{\omega \in \Omega : X(\omega) \in A\}) \quad \forall A \in \mathcal{A}'.$$

Loi de probabilité

Etant donnée $\mathbf{P}$ une probabilité sur $(\Omega, \mathcal{A})$ et X une variable aléatoire réelle définie sur Ω et à valeurs dans $(\Omega', \mathcal{A}')$, on appelle loi de probabilité de X la mesure $\mathbf{P}_X$ définie par

$$\mathbf{P}_X(A) = \mathbf{P}(X^{-1}(A)) = \mathbf{P}(X \in A) = \mathbf{P}(\{\omega \in \Omega : X(\omega) \in A\}) \quad \forall A \in \mathcal{A}'.$$

Intérêt

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

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.

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 $X : (\Omega, \mathcal{A}, \mathbf{P}) \to (\mathbb{R}, \mathcal{B}(\mathbb{R}))$ une v.a.r. On appelle fonction de répartition de X la fonction $F_X : \mathbb{R} \to [0, 1]$ définie par

$$F(x) = \mathbf{P}_X \left( ] - \infty, x] \right) = \mathbf{P}(X \leq x).$$

Propriété

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

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

Propriété

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

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

Propriété

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

Densité

- Il est possible de présenter la notion de densité dans un cadre général pour les v.a.r. Pour simplifier, nous allons dissocier les cas discrets et (absolument) continues.

Densité

- Il est possible de présenter la notion de densité dans un cadre général pour les v.a.r. Pour simplifier, nous allons dissocier les cas discrets et (absolument) continues.

Définition

- La loi de probabilité $\mathbf{P}_X$ d'une v.a.r X est absolument continue si elle est absolument continue par rapport à la mesure de Lebesgue.

Densité

- Il est possible de présenter la notion de densité dans un cadre général pour les v.a.r. Pour simplifier, nous allons dissocier les cas discrets et (absolument) continues.

Définition

- La loi de probabilité $\mathbf{P}_X$ d'une v.a.r X est absolument continue si elle est absolument continue par rapport à la mesure de Lebesgue.
- Radon-Nikodym : si $\mathbf{P}_X$ est absolument continue, alors il existe une fonction f_X positive et intégrable telle que $\forall x \in \mathbb{R}$

$$F_X(x) = \int_{]-\infty,x]} \mathrm{d}\mathbf{P}_X = \int_{]-\infty,x]} f_X(t)\,\mathrm{d}\lambda(t).$$

Densité

- Il est possible de présenter la notion de densité dans un cadre général pour les v.a.r. Pour simplifier, nous allons dissocier les cas discrets et (absolument) continues.

Définition

- La loi de probabilité $\mathbf{P}_X$ d'une v.a.r X est absolument continue si elle est absolument continue par rapport à la mesure de Lebesgue.
- Radon-Nikodym : si $\mathbf{P}_X$ est absolument continue, alors il existe une fonction f_X positive et intégrable telle que $\forall x \in \mathbb{R}$

$$F_X(x) = \int_{]-\infty,x]} \mathrm{d}\mathbf{P}_X = \int_{]-\infty,x]} f_X(t)\,\mathrm{d}\lambda(t).$$

- Cette fonction f_X est appelée densité de la loi P_X (ou par abus densité de X).

Propriété

Soit X une v.a.r absolument continue de fonction de répartition F_X et de densité f_X. Alors

1. F_X est $\lambda - \text{p.p.}$ dérivable avec $F_X' = f_X \ \lambda - \text{p.p.}$;
2. $\int_{\mathbb{R}} f_X(t)\,\mathrm{d}\lambda(t) = 1$ et $\mathbf{P}(X \in [a,b]) = \int_{[a,b]} f_X(t)\,\mathrm{d}\lambda(t)$;
3. f_X est unique à une λ-équivalence près. f_X caractérise donc la loi de X à une λ-équivalence près ;
4. Si F_X admet en tout point de $\mathbb{R}$ une dérivée continue, alors la loi de X est absolument continue. La réciproque est fausse : on peut juste dire que si $\mathbf{P}_X$ est absolument continue, alors $F(x)$ est continue.

Propriété

Soit X une v.a.r absolument continue de fonction de répartition F_X et de densité f_X. Alors

1. F_X est $\lambda - \mathrm{p.p.}$ dérivable avec $F'_X = f_X \ \lambda - \mathrm{p.p.}$;
2. $\int_{\mathbb{R}} f_X(t) \,\mathrm{d}\lambda(t) = 1$ et $\mathbf{P}(X \in [a, b]) = \int_{[a,b]} f_X(t) \,\mathrm{d}\lambda(t)$;
3. f_X est unique à une λ-équivalence près. f_X caractérise donc la loi de X à une λ-équivalence près ;
4. Si F_X admet en tout point de $\mathbb{R}$ une dérivée continue, alors la loi de X est absolument continue. La réciproque est fausse : on peut juste dire que si $\mathbf{P}_X$ est absolument continue, alors $F(x)$ est continue.

Propriété

Toute fonction réelle positive, d'une variable réelle f, intégrable au sens de Lebesgue et telle que $\int_{\mathbb{R}} f(t) \,\mathrm{d}\lambda(t) = 1$ est la densité d'une loi de probabilité.

Cas discret

Définition

- On dit qu'une v.a.r X est discrète si son support $\mathcal{S}_X$ est fini ou dénombrable.

Cas discret

Définition

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

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

Cas discret

Définition

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

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

- Exemples : Bernoulli, binomiale, Poisson...

Cas discret

Définition

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

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

- Exemples : Bernoulli, binomiale, Poisson...

Propriété

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

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 $\mathbf{P}$:

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

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 $\mathbf{P}$:

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

Interprétation

- L'espérance revient à intégrer les valeurs de la v.a.r. X pour chaque évènement ω pondéré par la mesure de probabilité $\mathbf{P}$.

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 $\mathbf{P}$:

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

Interprétation

- L'espérance revient à intégrer les valeurs de la v.a.r. X pour chaque évènement ω pondéré par la mesure de probabilité $\mathbf{P}$.
- D'où l'interprétation de valeur moyenne prise par X.

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 $\mathbf{P}$:

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

Interprétation

- L'espérance revient à intégrer les valeurs de la v.a.r. X pour chaque évènement ω pondéré par la mesure de probabilité $\mathbf{P}$.
- D'où l'interprétation de valeur moyenne prise par X.

- Problème : l'espérance dépend de Ω que l'on ne peut généralement pas caractériser !

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 $\mathbf{P}$:

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

Interprétation

- L'espérance revient à intégrer les valeurs de la v.a.r. X pour chaque évènement ω pondéré par la mesure de probabilité $\mathbf{P}$.
- D'où l'interprétation de valeur moyenne prise par X.

- Problème : l'espérance dépend de Ω que l'on ne peut généralement pas caractériser !
- Le théorème de transfert permet de pallier à cette difficulté.

Théorème de transfert pour l'espérance

Soit X une v.a.r. de loi $\mathbf{P}_X$. Alors

$$\mathbf{E}[X] = \int X \,\mathrm{d}\mathbf{P} = \int_\Omega X(\omega)\,\mathrm{d}\mathbf{P}(\omega) = \int_{\mathbb{R}} x \,\mathrm{d}\mathbf{P}_X.$$

Plus généralement, si h est une fonction réelle telle que $\mathbf{E}[h(X)]$ existe, on a

$$\mathbf{E}[h(X)] = \int_{\mathbb{R}} h(x)\,\mathrm{d}\mathbf{P}_X.$$

Théorème de transfert pour l'espérance

Soit X une v.a.r. de loi $\mathbf{P}_X$. Alors

$$\mathbf{E}[X] = \int X \,\mathrm{d}\mathbf{P} = \int_\Omega X(\omega)\,\mathrm{d}\mathbf{P}(\omega) = \int_\mathbb{R} x\,\mathrm{d}\mathbf{P}_X.$$

Plus généralement, si h est une fonction réelle telle que $\mathbf{E}[h(X)]$ existe, on a

$$\mathbf{E}[h(X)] = \int_\mathbb{R} h(x)\,\mathrm{d}\mathbf{P}_X.$$

Corollaire

- Cas discret :

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

- Cas continu :

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

Variance

Définition

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

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

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

Variance

Définition

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

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

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

Interprétion

- La variance est un réel positif.
- Elle mesure l'écart entre les valeurs prises par X et l'espérance (moyenne) de X

Variance

Définition

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

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

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

Interprétion

- La variance est un réel positif.
- Elle mesure l'écart entre les valeurs prises par X et l'espérance (moyenne) de $X \Longrightarrow$ 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 φ une fonction réelle convexe sur $]a, b[$
$$\varphi(\mathbf{E}[X]) \leq \mathbf{E}[\varphi(X)].$$

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 φ 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 $a > 0$

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

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 $a > 0$

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

Définition

Soit $X_1, \dots, X_n$ n v.a.r de loi $\mathbf{P}_{X_1}, \dots, \mathbf{P}_{X_n}$

- $X = (X_1, \dots, X_n)$ est un vecteur aléatoire de loi (conjointe)

$$\begin{aligned} \mathbf{P}_X : (\mathbb{R}^n, \mathcal{B}(\mathbb{R}^n)) &\to [0,1] \\ B &\mapsto \mathbf{P}_X(B) = \mathbf{P}(X^{-1}(B)) \end{aligned}$$

- Les loi $\mathbf{P}_{X_i}, i = 1, \dots, n$ sont appelées lois marginales de X.

Définition

Soit $X_1, \ldots, X_n$ n v.a.r de loi $\mathbf{P}_{X_1}, \ldots, \mathbf{P}_{X_n}$

- $X = (X_1, \ldots, X_n)$ est un vecteur aléatoire de loi (conjointe)

$$\begin{aligned} \mathbf{P}_X : (\mathbb{R}^n, \mathcal{B}(\mathbb{R}^n)) &\to [0,1] \\ B &\mapsto \mathbf{P}_X(B) = \mathbf{P}(X^{-1}(B)) \end{aligned}$$

- Les loi $\mathbf{P}_{X_i}, i = 1, \ldots, n$ sont appelées lois marginales de X.

Fonction de répartition

- On appelle fonction de répartition (conjointe) de X la fonction

$$\begin{aligned} F_X : \mathbb{R}^n &\to [0,1] \\ (x_1, \ldots, x_n) &\mapsto \mathbf{P}_X(]-\infty, x_1] \times \ldots \times]-\infty, x_n]) = \mathbf{P}(X_1 \leq x_1, \ldots, X_n \leq x_n) \end{aligned}$$

Définition

Soit $X_1, \ldots, X_n$ n v.a.r de loi $\mathbf{P}_{X_1}, \ldots, \mathbf{P}_{X_n}$

- $X = (X_1, \ldots, X_n)$ est un vecteur aléatoire de loi (conjointe)

$$\begin{aligned} \mathbf{P}_X : (\mathbb{R}^n, \mathcal{B}(\mathbb{R}^n)) &\to [0,1] \\ B &\mapsto \mathbf{P}_X(B) = \mathbf{P}(X^{-1}(B)) \end{aligned}$$

- Les loi $\mathbf{P}_{X_i}, i = 1, \ldots, n$ sont appelées lois marginales de X.

Fonction de répartition

- On appelle fonction de répartition (conjointe) de X la fonction

$$\begin{aligned} F_X : \mathbb{R}^n &\to [0,1] \\ (x_1, \ldots, x_n) &\mapsto \mathbf{P}_X(]-\infty, x_1] \times \ldots \times]-\infty, x_n]) = \mathbf{P}(X_1 \leq x_1, \ldots, X_n \leq x_n \end{aligned}$$

- La fonction de répartition conjointe caractérise la loi conjointe.

Définition

Soit $X_1, \ldots, X_n$ n v.a.r de loi $\mathbf{P}_{X_1}, \ldots, \mathbf{P}_{X_n}$

- $X = (X_1, \ldots, X_n)$ est un vecteur aléatoire de loi (conjointe)

$$\begin{aligned} \mathbf{P}_X : (\mathbb{R}^n, \mathcal{B}(\mathbb{R}^n)) &\to [0,1] \\ B &\mapsto \mathbf{P}_X(B) = \mathbf{P}(X^{-1}(B)) \end{aligned}$$

- Les loi $\mathbf{P}_{X_i}, i = 1, \ldots, n$ sont appelées lois marginales de X.

Fonction de répartition

- On appelle fonction de répartition (conjointe) de X la fonction

$$\begin{aligned} F_X : \mathbb{R}^n &\to [0,1] \\ (x_1, \ldots, x_n) &\mapsto \mathbf{P}_X(]-\infty, x_1] \times \ldots \times]-\infty, x_n]) = \mathbf{P}(X_1 \leq x_1, \ldots, X_n \leq x \end{aligned}$$

- La fonction de répartition conjointe caractérise la loi conjointe.
- Pour $i = 1, \ldots, n$, on a $F_{X_i}(x_i) = \lim_{x_j \to \infty, j \neq i} F_X(x_1, \ldots, x_n)$.

Densité

Radon-Nikodym

- La loi conjointe d'un couple aléatoire $X = (X_1, X_2)$ est dite absolument continue si et seulement si il existe une fonction f_X positive et intégrable telle que pour tout borélien B de $\mathbb{R}^2$,

$$\mathbf{P}_X(B) = \int\int_B f_X(x_1, x_2)\,\mathrm{d}\lambda(x_1)\,\mathrm{d}\lambda(x_2).$$

- f_X est appelée densité de X.

Densité

Radon-Nikodym

- La loi conjointe d'un couple aléatoire $X = (X_1, X_2)$ est dite absolument continue si et seulement si il existe une fonction f_X positive et intégrable telle que pour tout borélien B de $\mathbb{R}^2$,

$$\mathbf{P}_X(B) = \int\int_B f_X(x_1, x_2)\,\mathrm{d}\lambda(x_1)\,\mathrm{d}\lambda(x_2).$$

- f_X est appelée densité de X.

Propriété

Toute fonction f_X de $\mathbb{R}^2$ dans $\mathbb{R}$, positive, intégrable et telle que

$$\int_{\mathbb{R}^2} f_X(x_1, x_2)\,\mathrm{d}\lambda(x_1)\,\mathrm{d}\lambda(x_2) = 1$$

est la densité d'une loi absolument continue sur $(\mathbb{R}^2, \mathcal{B}(\mathbb{R}^2))$.

Densités marginales

Si le couple $X = (X_1, X_2)$ admet une loi conjointe absolument continue, alors ses lois marginales sont absolument continues et sa densité détermine les densités des lois marginales (appelées densités marginales) :

$$f_{X_1}(x_1) = \int_{\mathbb{R}} f_X(x_1, x_2)\,\mathrm{d}\lambda(x_2) \quad \text{et} \quad f_{X_2}(x_2) = \int_{\mathbb{R}} f_X(x_1, x_2)\,\mathrm{d}\lambda(x_1).$$

Densités marginales

Si le couple $X = (X_1, X_2)$ admet une loi conjointe absolument continue, alors ses lois marginales sont absolument continues et sa densité détermine les densités des lois marginales (appelées densités marginales) :

$$f_{X_1}(x_1) = \int_{\mathbb{R}} f_X(x_1, x_2)\,\mathrm{d}\lambda(x_2) \quad \text{et} \quad f_{X_2}(x_2) = \int_{\mathbb{R}} f_X(x_1, x_2)\,\mathrm{d}\lambda(x_1).$$

Covariance

- Covariance entre X_1 et X_2 :

$$\mathbf{cov}(X_1, X_2) = \mathbf{E}[X_1 X_2] - \mathbf{E}[X_1]\mathbf{E}[X_2].$$

- Matrice de variance covariance d'un vecteur $X = (X_1, \ldots, X_n)$: matrice $n \times n$ de terme général

$$\sigma_{ij}^2 = \mathbf{cov}(X_i, X_j).$$

Indépendance

Définition

n v.a.r son indépendantes si pour toute suite finie $X_{i_1}, \ldots, X_{i_k}$ extraite de $(X_1, \ldots, X_n)$ et pour toute suite finie de boréliens $B_{i_1}, \ldots, B_{i_k}$, on a

$$\mathbf{P}(X_{i_1} \in B_{i_1}, \ldots, X_{i_k} \in B_{i_k}) = \mathbf{P}(X_{i_1} \in B_{i_1}) \ldots \mathbf{P}(X_{i_k} \in B_{i_k}).$$

Indépendance

Définition

n v.a.r son indépendantes si pour toute suite finie $X_{i_1}, \ldots, X_{i_k}$ extraite de $(X_1, \ldots, X_n)$ et pour toute suite finie de boréliens $B_{i_1}, \ldots, B_{i_k}$, on a

$$\mathbf{P}(X_{i_1} \in B_{i_1}, \ldots, X_{i_k} \in B_{i_k}) = \mathbf{P}(X_{i_1} \in B_{i_1}) \ldots \mathbf{P}(X_{i_k} \in B_{i_k}).$$

Remarque

- L'indépendance n'est généralement pas simple à montrer en utilisant la définition.
- On préfère souvent utiliser les outils (fonction de répartition, densités...) qui caractérisent les lois de probabilité pour l'étudier.

Propriété

Soit $X = (X_1, \ldots, X_n)$ un vecteur aléatoire. Les v.a.r $X_i, i = 1, \ldots, n$ sont indépendantes si et seulement si

$$F_X(x_1, \ldots, x_n) = F_{X_1}(x_1) \ldots F_{X_n}(x_n)$$

ou

$$f_X(x_1, \ldots, x_n) = f_{X_1}(x_1) \ldots f_{X_n}(x_n)$$

(si les densités existent).

Propriété

Soit $X = (X_1, \dots, X_n)$ un vecteur aléatoire. Les v.a.r $X_i, i = 1, \dots, n$ sont indépendantes si et seulement si

$$F_X(x_1, \dots, x_n) = F_{X_1}(x_1) \dots F_{X_n}(x_n)$$

ou

$$f_X(x_1, \dots, x_n) = f_{X_1}(x_1) \dots f_{X_n}(x_n)$$

(si les densités existent).

Propriété

Soit X_1 et X_2 2 v.a.r indépendantes. Alors

1. $\mathbf{E}[X_1 X_2] = \mathbf{E}[X_1]\mathbf{E}[X_2]$ et donc $\mathbf{cov}(X_1, X_2) = 0$
2. $\mathbf{V}[X_1 + X_2] = \mathbf{V}[X_1] + \mathbf{V}[X_2]$.

Propriété

Soit $X = (X_1, \dots, X_n)$ un vecteur aléatoire. Les v.a.r $X_i, i = 1, \dots, n$ sont indépendantes si et seulement si

$$F_X(x_1, \dots, x_n) = F_{X_1}(x_1) \dots F_{X_n}(x_n)$$

ou

$$f_X(x_1, \dots, x_n) = f_{X_1}(x_1) \dots f_{X_n}(x_n)$$

(si les densités existent).

Propriété

Soit X_1 et X_2 2 v.a.r indépendantes. Alors

1. $\mathbf{E}[X_1X_2] = \mathbf{E}[X_1]\mathbf{E}[X_2]$ et donc $\mathbf{cov}(X_1, X_2) = 0$
2. $\mathbf{V}[X_1 + X_2] = \mathbf{V}[X_1] + \mathbf{V}[X_2]$.

- Attention : les réciproques sont fausses !

Références i

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

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

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

Deuxième partie II

Théorie de l'estimation

Rappels

- n observations $x_1, \ldots, x_n$.
- Ces observations sont des réalisations de variables aléatoires $X_1, \ldots, X_n$

Rappels

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

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

Rappels

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

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

Hypothèse

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

Rappels

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

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

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 $X_1, \ldots, X_n$.

- Poser un modèle revient à supposer que la loi de probabilité inconnue **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 $(\mathcal{H}, \mathcal{A})$.

- Poser un modèle revient à supposer que la loi de probabilité inconnue **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.

- Poser un modèle revient à supposer que la loi de probabilité inconnue **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 **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.

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.

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 $p \in [0, 1]$.

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 $X_1, \ldots, X_n$ 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 $\{0, 1\}$.

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 $\{0, 1\}$.
- $\mathcal{P} = \{$lois de Bernoulli de paramètre $p \in [0, 1]\} = \{B(p) : p \in [0, 1]\}$.

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 $\{0, 1\}$.
- $\mathcal{P} = \{$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^{e} patient au traitement suit une loi de Bernoulli de paramètre inconnu $p \in [0, 1]$.

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 $\{0, 1\}$.
- $\mathcal{P} = \{$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^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 $X_1, \ldots, X_n$.

Autres exemples

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

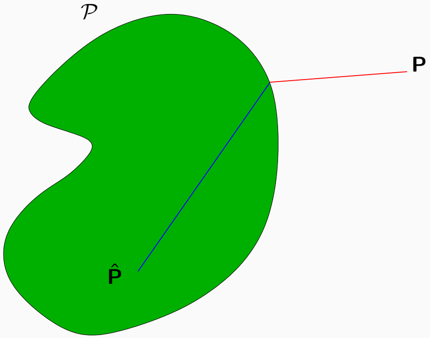
Autres exemples

- 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	$\mathbb{R}$	$\mathcal{B}(\mathbb{R})$	$\{\mathcal{N}(\mu,\sigma^2), \mu \in \mathbb{R}, \sigma \in \mathbb{R}^+\}$

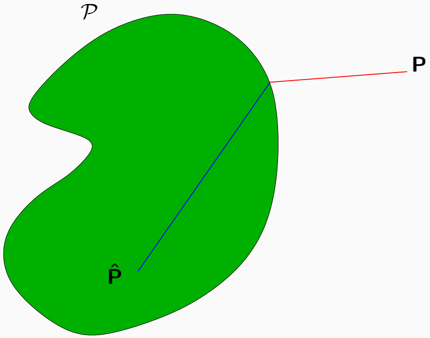
2 types d'erreur

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



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.

- Ces deux termes évoluent généralement en sens inverse.

Exemple des durées de trajet

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

- Ces deux termes évoluent généralement en sens inverse.

Exemple des durées de trajet

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

- Ces deux termes évoluent généralement en sens inverse.

Exemple des durées de trajet

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

- 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 $\mathcal{P} = \{\mathsf{P}_\theta, \theta \in \Theta\}$ où $\Theta \in \mathbb{R}^d$ alors on parle de modèle paramétrique et Θ est l'espace des paramètres.

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 Θ 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.

Paramétrique versus non paramétrique

Définition

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

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 Θ 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 (μ, σ^2) ou f à partir de l'échantillon $X_1, \ldots, X_n$.

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 $x_i \in \mathbb{R}^p$.

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 $m : \mathbb{R}^p \to \mathbb{R}$ vérifiant

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

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

Le problème de régression

- Données : $(x_1, y_1), \ldots, (x_n, y_n)$. 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 $(X_1, Y_1), \ldots, (X_n, Y_n)$ telles qu'il existe une fonction inconnue $m : \mathbb{R}^p \to \mathbb{R}$ vérifiant

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

où les ε_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 $(X_1, Y_1), \ldots, (X_n, Y_n)$.

Régression paramétrique vs non paramétrique

Modèle linéaire (paramétrique)

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

Régression paramétrique vs non paramétrique

Modèle linéaire (paramétrique)

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

Régression paramétrique vs non paramétrique

Modèle linéaire (paramétrique)

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

Un modèle non paramétrique

- On suppose que $m : \mathbb{R}^p \to \mathbb{R}$ est une fonction continue.

Régression paramétrique vs non paramétrique

Modèle linéaire (paramétrique)

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

Un modèle non paramétrique

- On suppose que $m : \mathbb{R}^p \to \mathbb{R}$ est une fonction continue.
- Le problème est d'estimer m à l'aide de $(X_1, Y_1), \ldots, (X_n, Y_n)$.

Régression paramétrique vs non paramétrique

Modèle linéaire (paramétrique)

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

Un modèle non paramétrique

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

Objectifs

Estimer...

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

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

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 $X_1, \ldots, X_n$.
- 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 Θ de dimension finie (typiquement $\mathbb{R}^p$).
- Choisir une loi reviendra donc à choisir un paramètre $\hat{\theta}$ à partir de l'échantillon $X_1, \ldots, X_n$.

- Les modèles que nous allons considérer auront pour espace d'observations un ensemble dénombrable Ω 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 $\mathcal{M} = (\mathcal{H}, \mathcal{P} = \{\mathbf{P}_\theta, \theta \in \Theta\})$.

- Les modèles que nous allons considérer auront pour espace d'observations un ensemble dénombrable Ω 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 $X_1, \ldots, X_n$ de n variables aléatoires indépendantes et de même loi $\mathbf{P}_\theta$, pour $\theta \in \Theta$.

- Les modèles que nous allons considérer auront pour espace d'observations un ensemble dénombrable Ω 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 $X_1, \ldots, X_n$ 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.

- Les modèles que nous allons considérer auront pour espace d'observations un ensemble dénombrable Ω 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 $X_1, \ldots, X_n$ 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.

- Les modèles que nous allons considérer auront pour espace d'observations un ensemble dénombrable Ω 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 $X_1, \ldots, X_n$ 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) $x_1, \ldots, x_n$ qui sont les résultats de n expériences aléatoires indépendantes.

La démarche statistique

1. On récolte n observations (n valeurs) $x_1, \ldots, x_n$ 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 $X_1, \ldots, X_n$ et de même loi P_θ. Ce qui nous amène à définir le modèle $\mathcal{M} = (\mathcal{H}, \{\mathsf{P}_\theta\}, \theta \in \Theta\})$.

La démarche statistique

1. On récolte n observations (n valeurs) $x_1, \ldots, x_n$ 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 $X_1, \ldots, X_n$ et de même loi P_θ. Ce qui nous amène à définir le modèle $\mathcal{M} = (\mathcal{H}, \{\mathsf{P}_\theta\}, \theta \in \Theta\})$.

3. **Estimation :** chercher dans le modèle une loi $\mathsf{P}_{\hat{\theta}}$ qui soit la plus proche possible de $\mathsf{P}_\theta \Longrightarrow$ chercher un **estimateur** $\hat{\theta}$ de θ.

Estimateurs

Définitions

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

Estimateurs

Définitions

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

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.

Estimateurs

Définitions

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

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.

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 $\hat{\theta}(X_1, \ldots, X_n)$ vis à vis de critères à définir.

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 $\hat{\theta}(X_1, \ldots, X_n)$ vis à vis de critères à définir.
- Une fois cette fonction trouvée, il faut donner une réponse (qui ne doit pas être abstraite !)...

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 $\hat{\theta}(X_1, \ldots, X_n)$ 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 $\hat{\theta}(x_1, \ldots, x_n)$.

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 $\hat{\theta}(X_1, \ldots, X_n)$ 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 $\hat{\theta}(x_1, \ldots, x_n)$.

Abus de notation

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

- on écrit $\hat{\theta}$ pour l'estimateur $\hat{\theta}(X_1, \ldots, X_n)$;
- on écrit $\hat{\theta}$ pour l'estimation $\hat{\theta}(x_1, \ldots, x_n)$;

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 $\hat{\theta}(X_1, \ldots, X_n)$ 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 $\hat{\theta}(x_1, \ldots, x_n)$.

Abus de notation

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

- on écrit $\hat{\theta}$ pour l'estimateur $\hat{\theta}(X_1, \ldots, X_n)$;
- on écrit $\hat{\theta}$ pour l'estimation $\hat{\theta}(x_1, \ldots, x_n)$;
- 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

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).

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.

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 :

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

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 :

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

- Estimation :

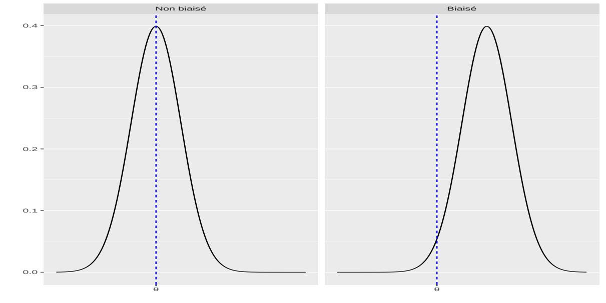
$$\hat{p} = \hat{p}(x_1, \ldots, x_n) = \frac{1}{n}\sum_{i=1}^{n} x_i = \frac{3}{8}.$$

- $X_1, \ldots, X_n$ i.i.d de loi $\mathbf{P}_\theta$ avec $\theta \in \Theta$ inconnu.
- On cherche un estimateur $\hat{\theta} = \hat{\theta}(X_1, \ldots, X_n)$.

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

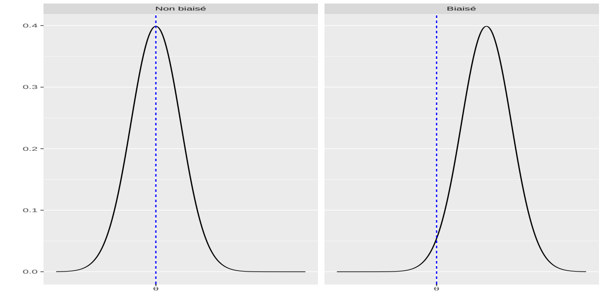
Espérance d'une estimateur

- On représente ci-dessous les lois de probabilité de 2 estimateurs de θ.



Espérance d'une estimateur

- On représente ci-dessous les lois de probabilité de 2 estimateurs de θ.



Commentaires

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

Biais d'un estimateur

- On désigne par E_θ l'espérance sous la loi P_θ :

$$\mathsf{E}_\theta(\hat{\theta}) = \mathsf{E}_\theta(\hat{\theta}(X_1, \ldots, X_n)) = \int_{\mathcal{H}^n} \hat{\theta}(x)\mathsf{P}_\theta \,\mathrm{d}x$$

où $x = (x_1, \ldots, x_n)$.

Biais d'un estimateur

- On désigne par $\mathbf{E}_\theta$ l'espérance sous la loi $\mathbf{P}_\theta$:

$$\mathbf{E}_\theta(\hat{\theta}) = \mathbf{E}_\theta(\hat{\theta}(X_1, \ldots, X_n)) = \int_{\mathcal{H}^n} \hat{\theta}(x) \mathbf{P}_\theta \, \mathrm{d}x$$

où $x = (x_1, \ldots, x_n)$.

Définition

Soit $\hat{\theta}$ un estimateur d'ordre 1.

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

Biais d'un estimateur

- On désigne par $\mathbf{E}_\theta$ l'espérance sous la loi $\mathbf{P}_\theta$:

$$\mathbf{E}_\theta(\hat{\theta}) = \mathbf{E}_\theta(\hat{\theta}(X_1, \dots, X_n)) = \int_{\mathcal{H}^n} \hat{\theta}(x)\mathbf{P}_\theta \,\mathrm{d}x$$

  où $x = (x_1, \dots, x_n)$.

Définition

Soit $\hat{\theta}$ un estimateur d'ordre 1.

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

Exemple 1

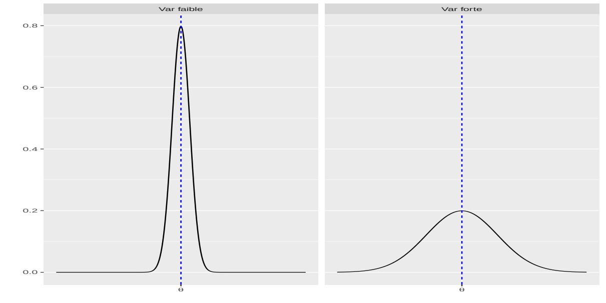
Les estimateurs $\hat{p}_1$ et $\hat{p}_2$ sont sans biais.

Variance d'un estimateur

- Mesurer le biais n'est pas suffisant

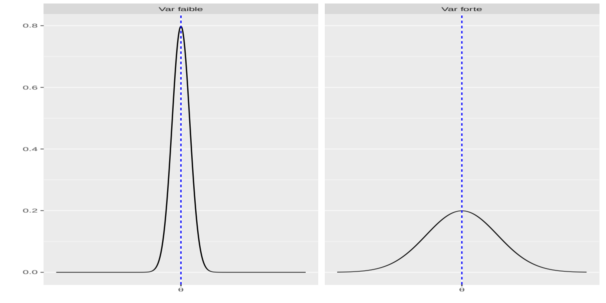
Variance d'un estimateur

- Mesurer le biais n'est pas suffisant, il faut également mesurer la dispersion des estimateurs.



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 : $\Longrightarrow \mathbf{V}[\hat{\theta}_1] \leq \mathbf{V}[\hat{\theta}_2]$.

Risque quadratique

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

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 $\theta \in \mathbb{R}$:

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

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

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

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 $\theta \in \mathbb{R}$:
$$\mathcal{R}(\theta, \hat{\theta}) = \mathbf{E}_\theta(\hat{\theta} - \theta)^2$$
2. Soit $\hat{\theta}'$ un autre estimateur d'ordre 2. On dit que $\hat{\theta}$ est préférable à $\hat{\theta}'$ si
$$\mathcal{R}(\theta, \hat{\theta}) \leq \mathcal{R}(\theta, \hat{\theta}') \ \forall \theta \in \Theta.$$

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

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

Estimateur VUMSB

Propriété décomposition biais variance

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

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

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 :

Estimateur VUMSB

Propriété décomposition biais variance

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

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

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 :

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

Estimateur VUMSB

Propriété décomposition biais variance

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

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

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 :

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

Exemple

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

- $X_1, \ldots, X_n$ 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 $\hat{\theta}$.

- $X_1, \dots, X_n$ 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 ?

- $X_1, \dots, X_n$ 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.

- C'est une approche intuitive qui repose sur le fait que pour de nombreux modèles les moments empiriques doivent être proches des moments théoriques.

- 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 :

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

- 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 :

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

Définition

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

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

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 $\mathcal{E}(\lambda)$	$\hat{\lambda}_m = 1/\bar{X}_n$

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 $\mathcal{E}(\lambda)$	$\hat{\lambda}_m = 1/\bar{X}_n$

Remarque

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

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 $X_1, \dots, X_n$.

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 $L(x_1, \dots, x_n; p) = \mathbf{P}_p(X_1 = x_1, \dots, X_n = x_n)$ peut être vue comme une mesure de la probabilité d'observer les données observées.

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

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 $L(x_1, \dots, x_n; p) = \mathbf{P}_p(X_1 = x_1, \dots, X_n = x_n)$ 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 θ pour la réalisation $(x_1, \ldots, x_n)$ est l'application $L : \mathcal{H}^n \times \Theta$ définie par

$$L(x_1, \ldots, x_n; \theta) = \mathsf{P}_\theta^{\otimes n}(\{x_1, \ldots, x_n\}) = \prod_{i=1}^{n} \mathsf{P}_\theta(\{x_i\}).$$

Vraisemblance

Cas discret

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

$$L(x_1, \ldots, x_n; \theta) = \mathrm{P}_\theta^{\otimes n}(\{x_1, \ldots, x_n\}) = \prod_{i=1}^{n} \mathrm{P}_\theta(\{x_i\}).$$

Cas absolument continu

Soit $f(., \theta)$ la densité associé à P_θ. La **vraisemblance** du paramètre θ pour la réalisation $(x_1, \ldots, x_n)$ est l'application $L : \mathcal{H}^n \times \Theta$ définie par

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

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 $\forall (x_1, \ldots, x_n) \in \mathcal{H}^n$

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

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

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

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

- $X_1, \dots, X_n$ i.i.d de loi $\mathbf{P}_\theta$ avec θ 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 θ, 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 P_θ.
- On désigne par $L_1(.;\theta)$ la vraisemblance associée.

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

Définition

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

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

Interprétation

L'information de Fisher peut s'interpréter comme :

- la quantité d'information apportée par l'observation X pour estimer le paramètre inconnu.

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 θ :
  - $I(\theta)$ grand : il sera "facile" d'identifier quel paramètre est le meilleur.
  - $I(\theta)$ petit : l'identification sera plus difficile.

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 θ :
  - $I(\theta)$ 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

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

- 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

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

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 $X_1, \ldots, X_n$ de loi $\mathbf{P}_\theta$.
- On désigne par $L_1(.;\theta)$ la vraisemblance associée.

Définition

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

$$
\begin{aligned}
I_n : \Theta &\to \mathbb{R}^+ \\
\theta &\mapsto \mathbf{E}_\theta\left[\left(\frac{\partial}{\partial\theta}\log(L(X_1, \ldots, X_n, \theta))\right)^2\right]
\end{aligned}
$$

Propriété d'additivité

L'information de Fisher est additive :

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

Propriété d'additivité

L'information de Fisher est additive :

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

Modèle de Bernoulli

- $X_1, \ldots, X_n$ 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 θ de biais $b(\theta) = \mathbf{E}_\theta[\hat{\theta}] - \theta$. Alors sous certaines hypothèses de régularité (voir [Guyader, 2017]), on a

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

Cramér-Rao

Proposition

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

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

Corollaire : Inégalité de Cramér-Rao

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

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

- 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.

- 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

- $X_1, \ldots, X_n$ 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 $\frac{p(1-p)}{n}$.

- 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

- $X_1, \ldots, X_n$ 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 $\hat{p} = \bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$.

- 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

- $X_1, \ldots, X_n$ 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

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

- 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

- $X_1, \ldots, X_n$ 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.

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 $\mathcal{B}(p)$:

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

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 $\mathcal{P}(\lambda)$:

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

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].

Références i

Cadre, B. and Vial, C. (2012).
Statistique mathématique, cours et exercices corrigés.
Ellipses.

Guyader, A. (2017).
Statistique mathématique.
Polycopié de cours, http ://www.lsta.upmc.fr/guyader/index.html.

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

Troisième partie III

Convergences stochastiques

Motivations

- $X_1, \ldots, X_n$ i.i.d. de loi P_θ avec θ inconnu dans Θ.
- Un estimateur : une fonction $\hat{\theta}(X_1, \ldots, X_n)$.

Motivations

- $X_1, \ldots, X_n$ i.i.d. de loi P_θ avec θ inconnu dans Θ.
- Un estimateur : une fonction $\hat{\theta}(X_1, \ldots, X_n)$.
- 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 θ.

Motivations

- $X_1, \ldots, X_n$ i.i.d. de loi $\mathbf{P}_\theta$ avec θ inconnu dans Θ.
- Un estimateur : une fonction $\hat{\theta}(X_1, \ldots, X_n)$.
- 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 θ.

Conséquence

- Plus on a d'information, plus on doit être précis.

Motivations

- $X_1, \ldots, X_n$ i.i.d. de loi $\mathbf{P}_\theta$ avec θ inconnu dans Θ.
- Un estimateur : une fonction $\hat{\theta}(X_1, \ldots, X_n)$.
- 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 θ.

Conséquence

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

Motivations

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

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 θ.
- On a donc envie de traduire cela par $\lim_{n\to\infty} \hat{\theta}(X_1, \dots, X_n) = \theta$.

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 : $\lim_{n\to\infty} X_n = X$.

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 : $\lim_{n\to\infty} X_n = X$.

Première idée

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

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 : $\lim_{n\to\infty} X_n = X$.

Première idée

- Une variable aléatoire réelle est une fonction qui va de Ω 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 $X(\omega)$:

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

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

- 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^{e} jet, 0 sinon.

- 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^{e} jet, 0 sinon. $X_i \sim \mathcal{B}(1/2)$.

- 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^{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

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

- 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 ie 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

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

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

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

- 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^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

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

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

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

Exemple du pile ou face (retour)

- Il est facile de voir que l'évènement ω_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

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

Exemple du pile ou face (retour)

- Il est facile de voir que l'évènement ω_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

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

- 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 ω tels que la suite numérique $(X_n(\omega))_n$ ne converge pas vers $X(\omega)$ est négligeable (c'est-à-dire vérifie $\mathsf{P}(N) = 0$). On note

$$\lim_{n\to\infty} X_n = X \quad \text{p.s.} \qquad \text{ou} \qquad X_n \overset{p.s.}{\to} X.$$

Définition

On dit que $(X_n)_n$ converge presque sûrement vers une variable aléatoire X si l'ensemble N des ω 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

$$\lim_{n\to\infty} X_n = X \quad \text{p.s.} \qquad \text{ou} \qquad X_n \overset{p.s.}{\to} X.$$

Remarque

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

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

ou encore

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

Proposition : opérations sur la cv ps

1. Si $X_n \overset{p.s.}{\to} X$ et si $\varphi : \mathbb{R} \to \mathbb{R}$ est une fonction continue sur $\mathbb{R}$ alors $\varphi(X_n) \overset{p.s.}{\to} \varphi(X)$.

Proposition : opérations sur la cv ps

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

Proposition : opérations sur la cv ps

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

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 $\varepsilon > 0$,

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

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 $\varepsilon > 0$,

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

Lemme de Borel-Cantelli

Si pour tout $\varepsilon > 0$,

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

alors $X_n \overset{p.s.}{\to} X$.

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

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

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
$$\frac{1}{n^2}\sum_{i=1}^{n} X_i \overset{p.s.}{\to} 0\ ?$$
- On a d'après B.T.
$$\mathbf{P}\left(\left|\frac{1}{n^2}\sum_{i=1}^{n} X_i\right| > \varepsilon\right) \leq \frac{1}{n^3\varepsilon^2}.$$

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

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

- 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

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

Définition

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

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

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

Définition

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

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

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

Exemple

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

Définition

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

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

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-Tchebytchev, on a

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

Définition

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

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

On note $X_n \overset{\mathbf{P}}{\to} X$.

Exemple

- Soit $X_1, \ldots, X_n, n \geq 1$ 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-Tchebytchev, 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 $\bar{X}_n \overset{\mathbf{P}}{\to} 0$.

Exemple

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

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

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,

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

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

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

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

- 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.

- 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 $\varphi : \mathbb{R} \to \mathbb{R}$ 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 $\mathbf{P}(Y = 0) = 0$.

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

Théorème

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

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

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

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

On note $X_n \overset{L_p}{\to} X$.

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

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

On note $X_n \overset{L_p}{\to} X$.

- Les cas les plus importants sont $p = 1$ (convergence en moyenne) et $p = 2$ (convergence en moyenne quadratique).

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

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

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

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

Convergence dans L_2

- Il est facile de voir que

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

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

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

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

$$X_n \overset{L_2}{\to} a \Longleftrightarrow \begin{cases} \lim_{n\to\infty} \mathbf{E}[X_n] = a \\ \lim_{n\to\infty} \mathbf{V}[X_n] = 0 \end{cases}$$

Application en statistique

Si $\hat{\theta}_n \overset{L_2}{\to} \theta$ 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

$$X_n \xrightarrow{L_2} X \quad \Longrightarrow \quad X_n \xrightarrow{L_1} X.$$

- 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

$$X_n \xrightarrow{L_2} X \quad \Longrightarrow \quad X_n \xrightarrow{L_1} X.$$

Théorème

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

- 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

$$X_n \xrightarrow{L_2} X \quad \Longrightarrow \quad X_n \xrightarrow{L_1} X.$$

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

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

- 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.

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

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

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

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

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.

(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 $\varepsilon > 0$

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

(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

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

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

- 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

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

- On a également pour $x \neq 0$ $\lim_{n\to\infty} F_{X_n}(x) = F_x(x)$.

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
$$\lim_{n\to\infty} \mathbf{P}(X_n \in [a, b]) = \mathbf{P}(X \in [a, b]).$$
- On a également pour $x \neq 0$ $\lim_{n\to\infty} F_{X_n}(x) = F_x(x)$.
- 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 $\lim_{n\to\infty} F_{X_n}(x) = F(x)$. On note $X_n \stackrel{\mathcal{L}}{\to} X$.

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 $\lim_{n\to\infty} F_{X_n}(x) = F(x)$. On note $X_n \overset{\mathcal{L}}{\to} X$.

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,

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

- Comme F_X est discontinue en 0, on conclut que $X_n \overset{\mathcal{L}}{\to} X$.

Attention

Remarque

- Les opérations conservées par les cv en probabilités et presque sure ne le sont pas forcément par la convergence en loi !
- Par exemple, $X_n \xrightarrow{\mathcal{L}} X$ n'implique pas
  - $\mathbf{P}(X_n \in A) \to \mathbf{P}(X \in A), \ \forall A$ (déjà vu) ;

Attention

Remarque

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

Attention

Remarque

- Les opérations conservées par les cv en probabilités et presque sure ne le sont pas forcément par la convergence en loi !
- Par exemple, $X_n \xrightarrow{\mathcal{L}} X$ n'implique pas
  - $\mathbf{P}(X_n \in A) \to \mathbf{P}(X \in A), \ \forall A$ (déjà vu) ;
  - $\mathbf{E}[X_n] \to \mathbf{E}[X]$. 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 $X_n = (-1)^n X$.

- Très souvent utilisées pour montrer des convergences en loi.

Définition

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

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

Fonctions caractéristiques

- Très souvent utilisées pour montrer des convergences en loi.

Définition

On appelle fonction caractéristique de X la fonction $\varphi_X : \mathbb{R} \to \mathbb{C}$ 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 π_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

$$\varphi_X(t) = \int e^{itx} f_X(x) \,\mathrm{d}x.$$

Exemple

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,σ^2)	$e^{im}e^{-\sigma^2 t^2/2}$

Proposition

1. φ_X est définie et continue pour tout nombre réel t ;
2. φ_X est bornée et $\forall t \; |\varphi_X(t)| \leq 1$;
3. $\forall (a, b) \in \mathbb{R}^2, \varphi_{aX+b}(t) = e^{ibt}\varphi_X(at)$;

Proposition

1. φ_X est définie et continue pour tout nombre réel t ;
2. φ_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 φ_X est une fonction réelle paire ;
5. φ_X caractérise la loi de X.

Proposition

1. φ_X est définie et continue pour tout nombre réel t ;
2. φ_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 φ_X est une fonction réelle paire ;
5. φ_X caractérise la loi de X.

Proposition

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

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

Proposition

1. φ_X est définie et continue pour tout nombre réel t ;
2. φ_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 φ_X est une fonction réelle paire ;
5. φ_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 $n \in \mathbb{N}^\star$ tel que $\mathbf{E}[|X|^n] < \infty$, alors

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

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

   lorsque $t \to 0$.

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. $X_n \overset{\mathcal{L}}{\to} X$;

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. $X_n \overset{\mathcal{L}}{\to} X$;
2. Pour toute fonction $f : \mathbb{R} \to \mathbb{R}$ continue bornée, on a $\lim_{n\to\infty} \mathbf{E}[f(X_n)] = \mathbf{E}[f(X)]$.

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

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

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

Binomiale vers Poisson

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

Binomiale vers Poisson

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

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

Binomiale vers Poisson

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

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

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

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 $np_n \to \lambda$ lorsque $n \to \infty$. On a lorsque $n \to \infty$ (faire un DL)

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

2. On déduit $X_n \xrightarrow{\mathcal{L}} X$ avec X qui suit une loi de Poisson de paramètre λ. 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 λ_n avec $\lambda_n \to \infty$ lorsque $n \to \infty$.
- 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.

- 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

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

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
$$\forall j \in E, \quad \lim_{n \to \infty} \mathbf{P}(X_n = j) = \mathbf{P}(X = j).$$
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 $\lim_{n\to\infty} f_n(x) = f(x)$, alors $X_n \xrightarrow{\mathcal{L}} X$.

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

- 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 $g : \mathbb{R} \to \mathbb{R}$ est continue en tout point de $\mathbb{R}$ alors $g(X_n) \xrightarrow{\mathcal{L}} g(X)$.

- 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 $g : \mathbb{R} \to \mathbb{R}$ 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 $X_n \xrightarrow{\mathcal{L}} X$.

Relation entre les convergences

Théorème

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

- Réciproque fausse : il suffit de prendre $X \sim \mathcal{N}(0,1)$ et $X_n = (-1)^n X$.

Relation entre les convergences

Théorème

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

- Réciproque fausse : 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

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

Relation entre les convergences

Théorème

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

- Réciproque fausse : 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
$$X_n \xrightarrow{\mathcal{L}} a \Longleftrightarrow X_n \xrightarrow{\mathbf{P}} a.$$
- On peut résumer les relations entre les différents modes de convergence par le diagramme suivant :

$$\begin{tikzcd}
L_p \arrow[r] & \text{Proba} \arrow[r] & \text{Loi} \\
& \text{p.s.} \arrow[u] &
\end{tikzcd}$$

Présentation

- $X_1, \ldots, X_n$ i.i.d. admettant une espérance $\mu = \mathbf{E}[X_1]$.

- $X_1, \ldots, X_n$ 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 μ.

Présentation

- $X_1, \ldots, X_n$ 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 μ.
- Les lois des grands nombres et le théorème central limite permettent de préciser rigoureusement ce rapprochement.

Un exemple

- Soit $X_1, \ldots, X_n$ n v.a.r. indépendantes de loi Bernoulli de paramètre p.
- Question : est-ce que $\bar{X}_n$ converge en probabilité vers p ?

Un exemple

- Soit $X_1, \ldots, X_n$ 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 $\forall \varepsilon > 0$

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

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

Un exemple

- Soit $X_1, \ldots, X_n$ 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 $\forall \varepsilon > 0$

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

- 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.

Un exemple

- Soit $X_1, \ldots, X_n$ 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 $\forall \varepsilon > 0$

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

- 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

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

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

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

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

- 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

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

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

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

Application

- $X_1, \ldots, X_n$ i.i.d de loi $\mathcal{E}(\lambda)$ avec $\lambda > 0$ (inconnu).

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

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

Application

- $X_1, \ldots, X_n$ i.i.d de loi $\mathcal{E}(\lambda)$ avec $\lambda > 0$ (inconnu).
- LFGN $\Longrightarrow$ $\bar{X}_n \stackrel{p.s.}{\to} 1/\lambda$.

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

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

Application

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

Méthode de Monte-Carlo

- Soit $f :]0,1[\to \mathbb{R}$ intégrable. On cherche à approcher $I = \int_0^1 f(x)\,\mathrm{d}x$.

Méthode de Monte-Carlo

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

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

Méthode de Monte-Carlo

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

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

- 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

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

Méthode de Monte-Carlo

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

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

- 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

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

Algorithme de Monte-Carlo

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

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

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.

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

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

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

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

- Les hypothèses sont faibles : on demande juste des v.a.r i.i.d. qui admettent une variance.

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

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

- 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 $\mathcal{N}(\mu, \sigma^2/n)$.

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

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

- 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

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

Eléments 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 φ 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

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

et

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

- On déduit
$$\varphi(u) = 1 - \frac{\sigma^2 u^2}{2} + \mathrm{o}(u^2)$$
et
$$\varphi_{Y_n}(t) = \exp\left(n \log(1 - t^2/2n + \mathrm{o}(1/n))\right).$$

- Par conséquent
$$\lim_{n\to\infty} \varphi_{Y_n}(t) = \exp(-t^2/2)$$
et $t \mapsto \exp(-t^2/2)$ est la fonction caractéristique de la loi $\mathcal{N}(0,1)$.

- On déduit
$$\varphi(u) = 1 - \frac{\sigma^2 u^2}{2} + \mathrm{o}(u^2)$$
et
$$\varphi_{Y_n}(t) = \exp\left(n \log(1 - t^2/2n + \mathrm{o}(1/n))\right).$$

- Par conséquent
$$\lim_{n\to\infty} \varphi_{Y_n}(t) = \exp(-t^2/2)$$
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 $p \in]0,1[$.

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

$$\bar{X}_n \overset{p.s.}{\to} p \quad \text{quand } n \to \infty$$

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

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

et d'après le théorème central limite

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

Illustration

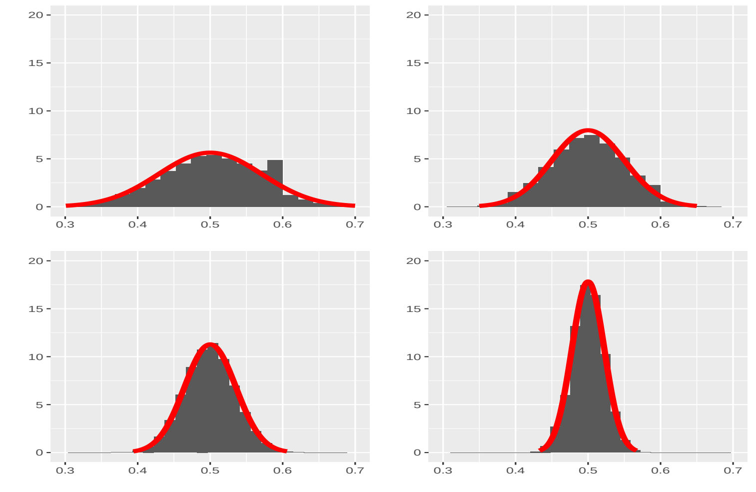


Figure 1 – Approximation TCL pour le modèle de Bernoulli $\mathcal{B}(1/2)$ avec $n = 50, 100, 200, 500$.

Slutsky

- Par continuité, on a

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

et donc

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

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

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 $\lambda > 0$.

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

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

et d'après le théorème central limite

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

Problème

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

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

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

et d'après le théorème central limite

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

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.

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

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

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

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

Application : loi exponentielle

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

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

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

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

  ou encore

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

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

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

Références i

Foata, D. and Fuchs, A. (2003).
***Calcul des probabilités*.**
Dunod, 2e edition.

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

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

Quatrième partie IV

Critères de performance asymptotiques, intervalles de confiance et estimation multivariée

Rappel

- $X_1, \ldots, X_n$ i.i.d de loi P_θ avec $\theta \in \Theta$ univarié.
- $\hat{\theta} = \hat{\theta}(X_1, \ldots, X_n) = \hat{\theta}_n$ un estimateur de θ.

Rappel

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

Rappel

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

Dans cette partie

- Critères de performance asymptotiques ;

Rappel

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

Dans cette partie

- Critères de performance asymptotiques ;
- Estimation par intervalles ;

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 θ.
- Critère de performance pour $\hat{\theta}_n$: biais, variance, risque quadratique, VUMSB...

Dans cette partie

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

Pourquoi ?

Postulat

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

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.

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.

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 θ.

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

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

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

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

Définition

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

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

où σ_θ est positif.

Outils consistance

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

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.

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

Motivations

- Donner une seule valeur pour estimer un paramètre peut se révéler trop ambitieux.

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.

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 P_θ avec $\theta \in \Theta$ inconnu.

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 θ tout intervalle de la forme $[A_n, B_n]$, où A_n et B_n sont des fonctions telles que :

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

Intervalle de confiance

- $X_1, \ldots, X_n$ 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 θ 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 $\lim_{n\to\infty} \mathbf{P}(\theta \in [A_n, B_n]) = 1 - \alpha$, on dit que $[A_n, B_n]$ est un intervalle de confiance asymptotique pour θ au niveau $1 - \alpha$.

Intervalle de confiance

- $X_1, \ldots, X_n$ 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 θ 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 $\lim_{n\to\infty} \mathbf{P}(\theta \in [A_n, B_n]) = 1 - \alpha$, on dit que $[A_n, B_n]$ est un intervalle de confiance asymptotique pour θ au niveau $1 - \alpha$.

Remarque importante

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

Construction d'un IC

- Inégalité de Bienaymé-Tchebychev (intervalle de confiance par excés) :

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

Construction d'un IC

- Inégalité de Bienaymé-Tchebychev (intervalle de confiance par excés) :

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

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

Construction d'un IC

- Inégalité de Bienaymé-Tchebychev (intervalle de confiance par excés) :

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

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

Méthode

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

Construction d'IC

- Un intervalle de confiance pour un paramètre inconnu θ se construit généralement à partir d'un estimateur de θ dont on connait la loi.

Construction d'IC

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

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

Construction d'IC

- Un intervalle de confiance pour un paramètre inconnu θ se construit généralement à partir d'un estimateur de θ dont on connait 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 α est petit, plus l'intervalle aura un grande amplitude.

Exemple

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

Exemple

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

Construction de l'IC

- Estimateur : $\hat{\mu} = \bar{X}_n$.

Exemple

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

Construction de l'IC

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

Exemple

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

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

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

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 μ.

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

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

Quantiles

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

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

Définition

Plus généralement, le quantile d'ordre α d'une variable aléatoire X est défini par le réel q_α vérifiant

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

Quantiles

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

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

Définition

Plus généralement, le quantile d'ordre α d'une variable aléatoire X est défini par le réel q_α vérifiant

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

- Les quantiles sont généralement renvoyés par les logiciels statistique :

```
> 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
```

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 μ :

```
> 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

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

Intervalle de confiance pour une proportion

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

Construction de l'IC

- Estimateur : $\hat{p}_n = \bar{X}_n$.

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 :

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

Intervalle de confiance pour une proportion

- $X_1, \ldots, X_n$ 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

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

Première version de l'IC

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

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

Première version de l'IC

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

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

- Problème : l'IC dépend de p qui est inconnu !

Première version de l'IC

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

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

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

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

Première version de l'IC

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

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

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

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

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 $\mathcal{B}(p)$.

- $n = 500$ 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
```

- $n = 500$ 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 μ est donné par

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

Loi normale (cas réel)

- $X_1, \ldots, X_n$ i.i.d de loi $\mathcal{N}(\mu, \sigma^2)$.
- On a vu qu'un IC pour μ 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, σ est inconnu !
- L'intervalle de confiance n'est donc pas calculable.

Idée

1. Estimer σ^2 par

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

Idée

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

2. Et considérer l'IC :
$$\left[\hat{\mu} - q_{1-\alpha/2}\frac{\widehat{\sigma}}{\sqrt{n}}, \hat{\mu} + q_{1-\alpha/2}\frac{\widehat{\sigma}}{\sqrt{n}}\right]. \tag{2}$$

Idée

1. Estimer σ^2 par

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

2. Et considérer l'IC :

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

Problème

- On a bien

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

Idée

1. Estimer σ^2 par

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

2. Et considérer l'IC :

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

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

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

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

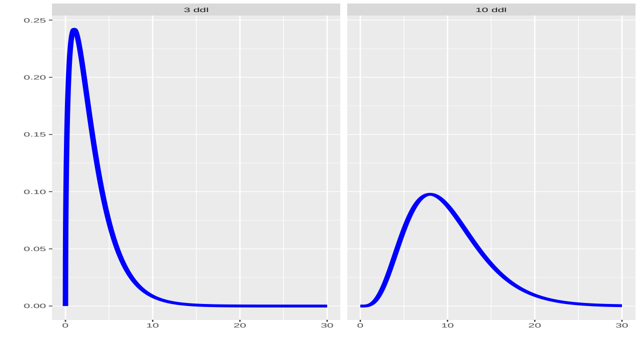
- $\mathbf{E}[X] = \mu$ et $\mathbf{V}[X] = \sigma^2$.
- Si $X \sim N(\mu, \sigma^2)$ alors

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

Loi du χ^2

Définition

- Soit $X_1, \ldots, X_n$ n variables aléatoires réelles indépendantes de loi $\mathcal{N}(0,1)$. La variable $Y = X_1^2 + \ldots + X_n^2$ 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 $\mathcal{T}(n)$.

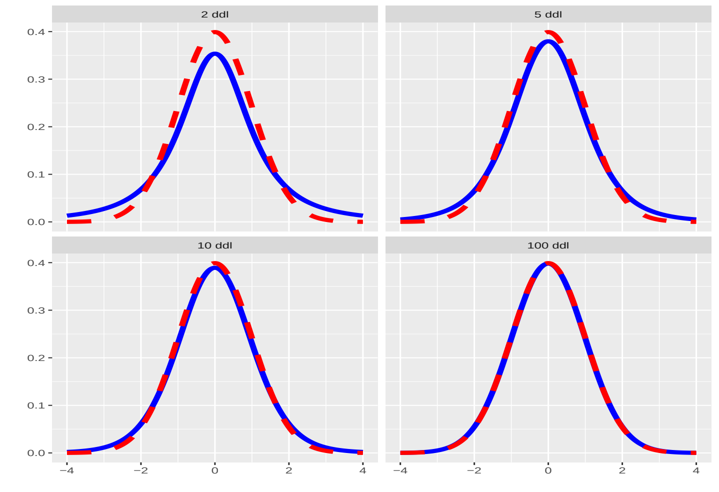
Loi de Student

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 $\mathbf{V}[T] = n/(n-2)$.

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 $\mathcal{N}(0,1)$.



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
$$F = \frac{X/m}{Y/m}$$
suit une loi de Fisher à m et n degrés de liberté. On note $\mathcal{F}(m, n)$.

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

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

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

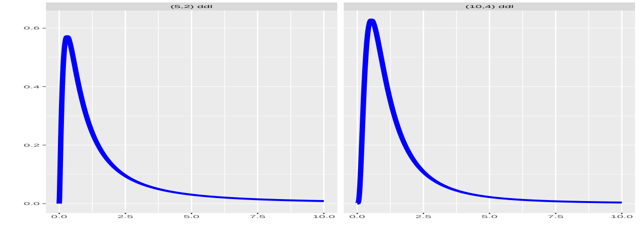


Figure 2 – Densités $\mathcal{F}(5, 2)$ et $\mathcal{F}(10, 4)$

Théorème de Cochran

- $X_1, \ldots, X_n$ 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. $(n-1)\frac{S^2}{\sigma^2} \sim \chi^2(n-1)$.

Théorème de Cochran

- $X_1, \ldots, X_n$ 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.

Théorème de Cochran

- $X_1, \ldots, X_n$ 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
$$\sqrt{n}\frac{\bar{X}_n - \mu}{S} \sim \mathcal{T}(n-1).$$

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 μ

On déduit du résultat précédent qu'un IC de niveau $1-\alpha$ pour μ 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 la loi gaussienne

IC pour μ

On déduit du résultat précédent qu'un IC de niveau $1-\alpha$ pour μ 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 σ^2

Un IC de niveau $1-\alpha$ pour σ^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 μ

- $n = 50$ 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
```

Exemple : modèle Gaussien - IC pour μ

- $n = 50$ 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 μ :

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

Exemple : modèle Gaussien - IC pour μ

- $n = 50$ 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 μ :

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

- Estimation de σ^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
```

- 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 σ^2

- On obtient l'IC pour σ^2 à l'aide de la formule

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

Exemple : modèle gaussien - IC pour σ^2

- On obtient l'IC pour σ^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
```

Jusqu'à présent

- $X_1, \ldots, X_n$ i.i.d de loi P_θ avec $\theta \in \mathbb{R}$.
- La loi P_θ dépend donc d'un seul paramètre (à estimer).

Jusqu'à présent

- $X_1, \ldots, X_n$ 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.

Jusqu'à présent

- $X_1, \ldots, X_n$ 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.
- $X_1, \ldots, X_n$ i.i.d de loi $\mathbf{P}_\theta$ avec $\theta = (\theta_1, \theta_2)$ inconnu dans $\mathbb{R}^2$.

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 θ à valeurs dans $\mathbb{R}^2$.

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 θ à valeurs dans $\mathbb{R}^2$.

Exemple : le modèle gaussien

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

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

- 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 :

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

Exemple : le modèle gaussien

- $\theta = (\mu, \sigma^2)$ et $\hat{\theta} = (\bar{X}_n, S^2)$.

Exemple : le modèle gaussien

- $\theta = (\mu, \sigma^2)$ et $\hat{\theta} = (\bar{X}_n, S^2)$.
- On a $b(\hat{\theta}) = (0, 0)$.

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

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

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

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

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".

Consistance

Définition

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

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

Consistance

Définition

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

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

- La valeur absolue est juste remplacée par la norme euclidienne.

Consistance

Définition

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

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

- 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 $\hat{\theta}_2 \xrightarrow{\mathbf{P}} \theta_2$.

Consistance

Définition

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

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

- 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 $v_n \to \infty$. On dit que $\hat{\theta} = (\hat{\theta}_1, \hat{\theta}_2)$ est asymptotiquent normal, de vitesse v_n si

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

où Σ_θ est une matrice symétrique 2×2 définie positive.

Normalité asymptotique

Définition

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

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

où Σ_θ est une matrice symétrique 2×2 définie positive.

- La loi limite est une loi gaussienne multivariée.

Normalité asymptotique

Définition

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

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

où Σ_θ est une matrice symétrique 2×2 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.

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 Σ est la matrice (2×2) de variance covariance de X.

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 Σ est la matrice (2×2) 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 (2×2) Σ, alors

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

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 (2×2) Σ, 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 $h : \mathbb{R}^d \to \mathbb{R}^m$ admet des dérivées partielles au point θ, alors

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

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

- $X_1, \ldots, X_n$ 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 θ 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

- $X_1, \ldots, X_n$ i.i.d de loi P_θ avec $\theta = (\theta_1, \theta_2) \in \mathbb{R}^2$.

Retour au cas multivarié

- $X_1, \ldots, X_n$ 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 θ pour une observation x.

Retour au cas multivarié

- $X_1, \ldots, X_n$ 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 θ 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 θ est la matrice de dimension 2×2 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 $1 \leq i, j \leq 2$.

Matrice d'information de Fisher

Définition

La matrice d'information de Fisher (si elle existe) au point θ est la matrice de dimension 2×2 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

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

Borne de Cramer Rao

- $X_1, \ldots, X_n$ 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 θ, on a

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

Borne de Cramer Rao

- $X_1, \ldots, X_n$ 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 θ, 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 :

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

Borne de Cramer Rao

- $X_1, \ldots, X_n$ 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 θ, 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

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

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 θ est

- consistant ;
- asymptotiquement normal :

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

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 θ 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 V

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.

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.

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

L'estimation de densité.

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

Le problème

Estimer f.

L'estimation de densité.

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

Le problème

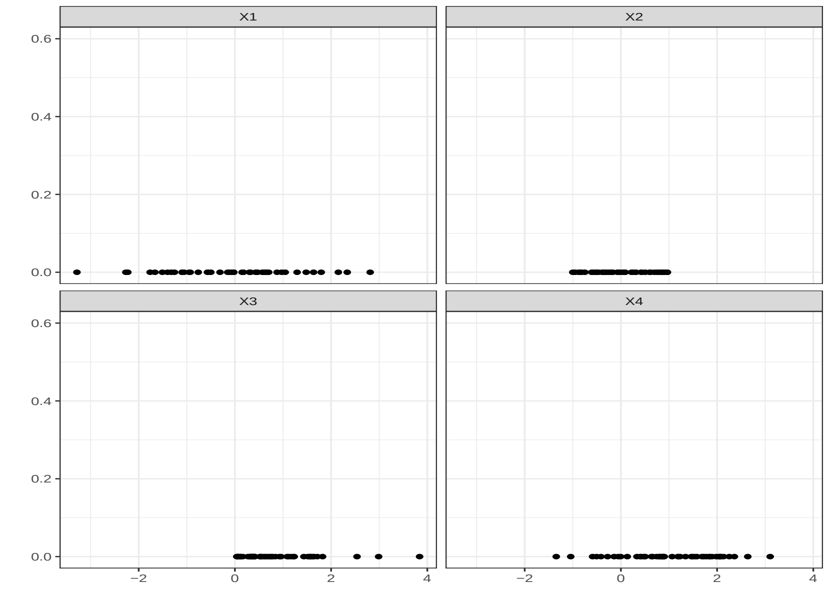
Estimer f.

Performance d'un estimateur

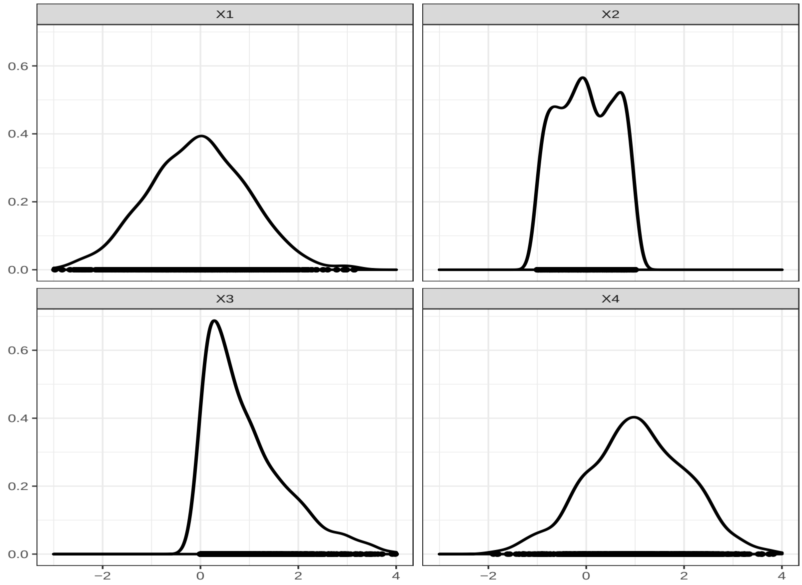
On mesurera la performance d'un estimateur $\hat{f}(.) = \hat{f}(., X_1, \ldots, X_n)$ par son risque quadratique ponctuel :

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

Exemple



Exemple



Le problème de la régression

- Données : $(x_1, y_1), \ldots, (x_n, y_n)$. On veut expliquer les sorties $y_i \in \mathbb{R}$ par les entrées $x_i \in \mathbb{R}^p$.

Le problème de la régression

- Données : $(x_1, y_1), \ldots, (x_n, y_n)$. 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. $(X_1, Y_1), \ldots, (X_n, Y_n)$ i.i.d. telles qu'il existe une fonction inconnue $m : \mathbb{R}^p \to \mathbb{R}$ vérifiant

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

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

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 $m : \mathbb{R}^p \to \mathbb{R}$ vérifiant

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

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

Le problème

Estimer m.

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 $m : \mathbb{R}^p \to \mathbb{R}$ vérifiant

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

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

Le problème

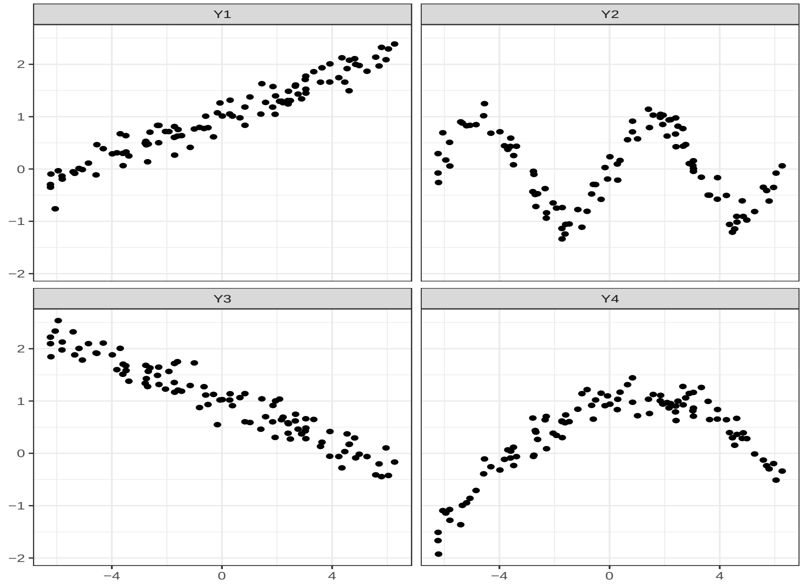
Estimer m.

Performance d'un estimateur

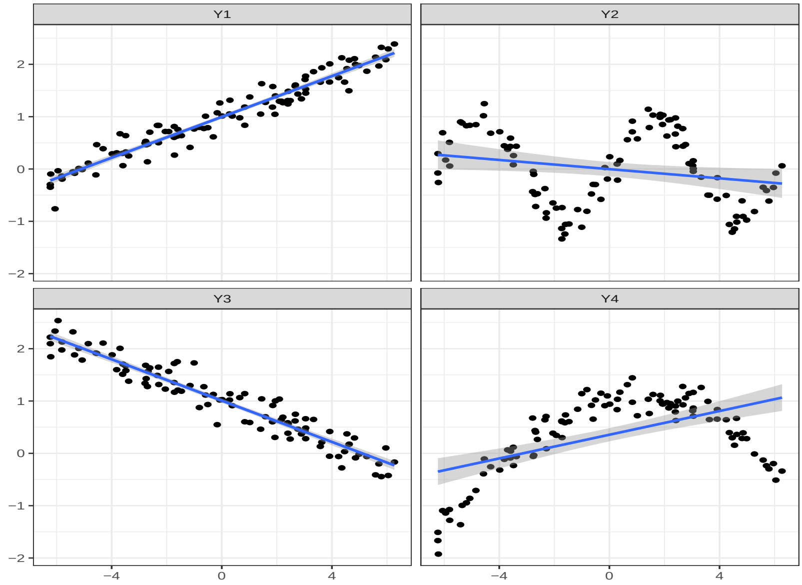
On mesurera la performance d'un estimateur $\hat{m}(.) = \hat{m}(., X_1, \dots, X_n)$ par son risque quadratique ponctuel :

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

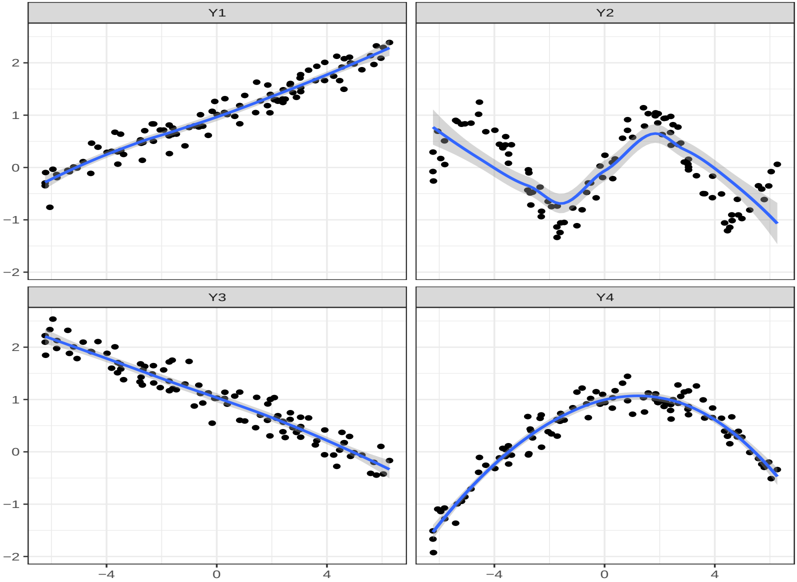
Exemple



Exemple



Exemple



- Dans les deux cas, le problème est d'estimer une fonction.

- 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 $\mathcal{F}$.

- 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.

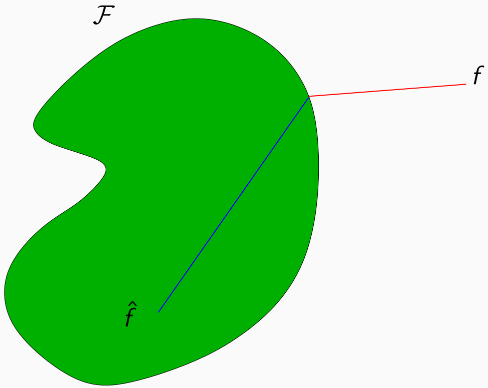
- Dans les deux cas, le problème est d'estimer une fonction.
- Poser un modèle revient à supposer que cette fonction appartient à un certain espace $\mathcal{F}$.

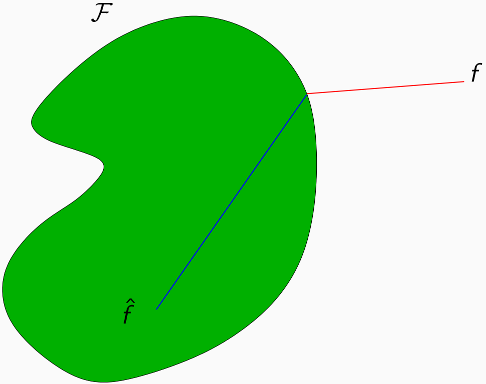
Définition

- 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.


$\mathcal{F}$
f
$\hat{f}$



- 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.

- $X_1, \ldots, X_n$ i.i.d. de densité f inconnue.
- On suppose que $f \in \mathcal{F} = \{f_\theta, \theta \in \Theta\}$ avec Θ 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 μ et σ^2.

- On peut estimer ces paramètres par maximum de vraisemblance :

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

- On peut estimer ces paramètres par maximum de vraisemblance :

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

- On montre "facilement" que

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

- On peut estimer ces paramètres par maximum de vraisemblance :

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

- On montre "facilement" que

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

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

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

- On peut estimer ces paramètres par maximum de vraisemblance :

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

- On montre "facilement" que

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

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

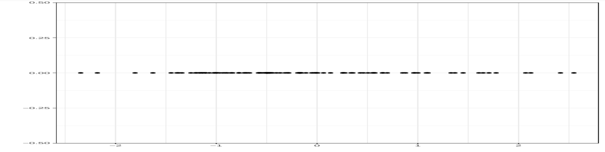
$$\mathbf{E}[\|\widehat{\theta} - \theta\|^2] = O\left(\frac{1}{n}\right).$$

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()
```

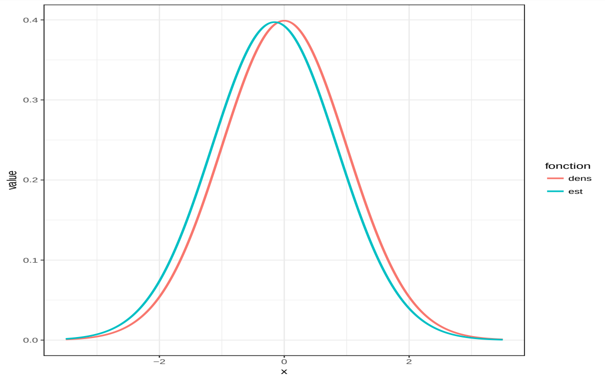


- On estime μ et σ^2 :

```
> 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()
```



Des moyennes locales

- En l'absence d'hypothèse paramétrique forte, on se base sur ce qui se passe au voisinage de x pour estimer $f(x)$.

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.

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

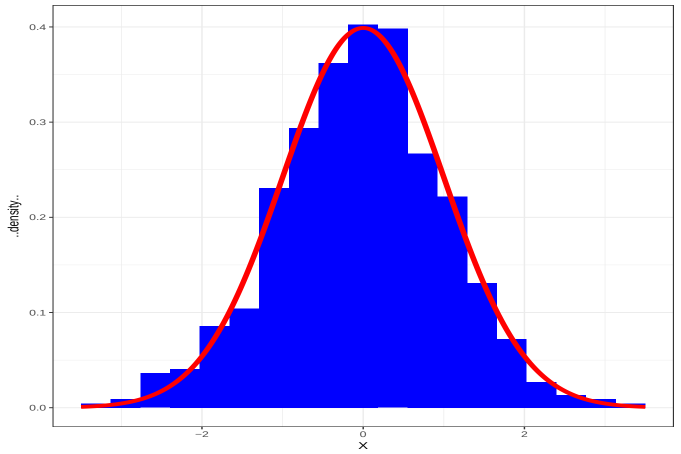
- $\mathcal{P} = \{I_1, \ldots, I_K\}$ une partition de $\mathbb{R}$ en K intervalles.
- L'histogramme est défini par

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

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.

Estimateurs à noyau

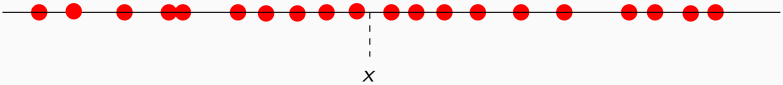
- L'histogramme n'est pas continu.
- L'estimateur à noyau permet de pallier à ce problème en ne fixant pas de partition.

Estimateurs à noyau

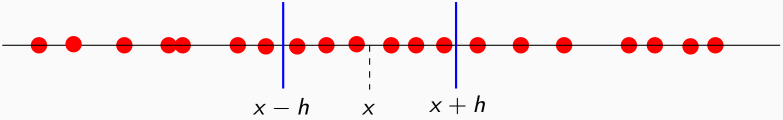
- L'histogramme n'est pas continu.
- L'estimateur à noyau permet de pallier à ce problème en ne fixant pas de partition.
- L'idée est d'utiliser une fenêtre glissante.

- $n = 20$ observations.

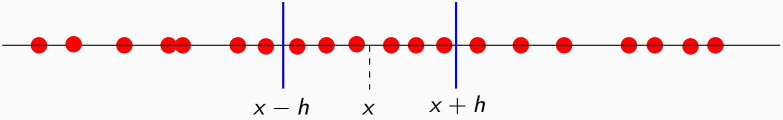
- $n = 20$ observations.
- On veut estimer la densité en x.



- $n = 20$ observations.
- On veut estimer la densité en x.
- On considère une fenêtre $[x - h, x + h]$.



- $n = 20$ observations.
- On veut estimer la densité en x.



- 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 $K : \mathbb{R} \to \mathbb{R}$ intégrable et tel que $\int K(u)\,\mathrm{d}u = 1$, l'estimateur à noyau de la densité est défini par

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

Estimateur à noyau de la densité

Définition [Parzen, 1962]

Etant donné $h > 0$ et $K : \mathbb{R} \to \mathbb{R}$ intégrable et tel que $\int K(u)\,\mathrm{d}u = 1$, 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

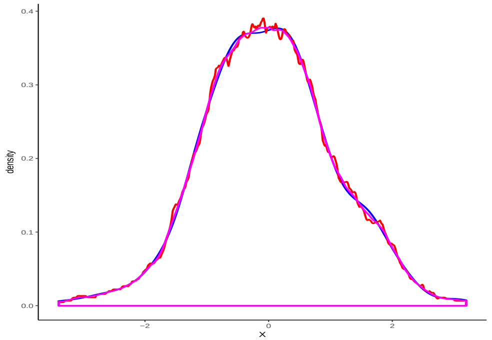
L'utilisateur doit choisir deux paramètres : un réel positif h et un noyau K

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

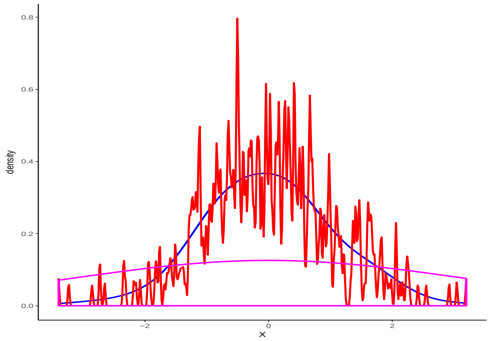
```
> X <- rnorm(500)
> df <- data.frame(X)
> ggplot(df)+aes(X)+geom_density(kernel=c("gaussian"),color="blue",size=1)+
   geom_density(kernel=c("rectangular"),color="red",size=1)+
   geom_density(kernel=c("epanechnikov"),color="black",size=1)+theme_classic()
```



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 $\Longrightarrow$ beaucoup d'observations dans les fenêtres $\Longrightarrow$ densités proches $\forall x$ $\Longrightarrow$ biais fort, variance faible.

Choix de h

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

Choix de h

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

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 $\int K(u)\,\mathrm{d}u = 1$, $\int uK(u)\,\mathrm{d}u = 0$ et $\int K(u)^2\,\mathrm{d}u < +\infty$.

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

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

Contrôle de la variance

Théorème

On suppose que :

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

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 $\int u^2 K(u)\,\mathrm{d}u < +\infty$.

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

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

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 $\int u^2 K(u) \, \mathrm{d}u < +\infty$.

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.

Risque quadratique

Corollaire (convergence L_2)

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

Risque quadratique

Corollaire (convergence L_2)

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

Corollaire (choix de h)

Le $h^\star$ qui minimise l'erreur quadratique vérifie

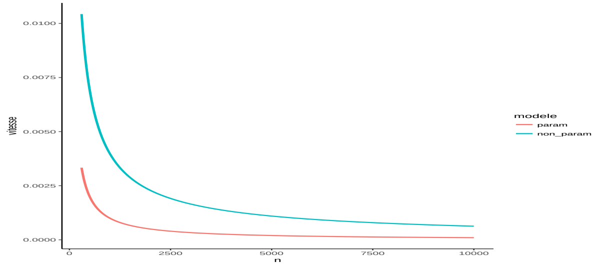
$$h^\star = Cn^{-\frac{1}{5}}.$$

Pour cette valeur de h, on a

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

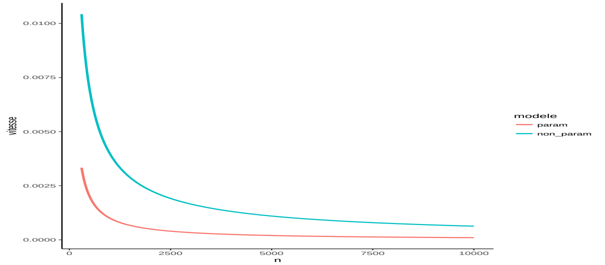
Remarque importante

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



Remarque importante

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



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

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

- La théorie nous dit que le h optimal est

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

- Ce résultat n'est quasiment d'aucune utilité pratique.

- La théorie nous dit que le h optimal est
$$h^\star = Cn^{-\frac{1}{5}}.$$
- 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.

Présentation du modèle

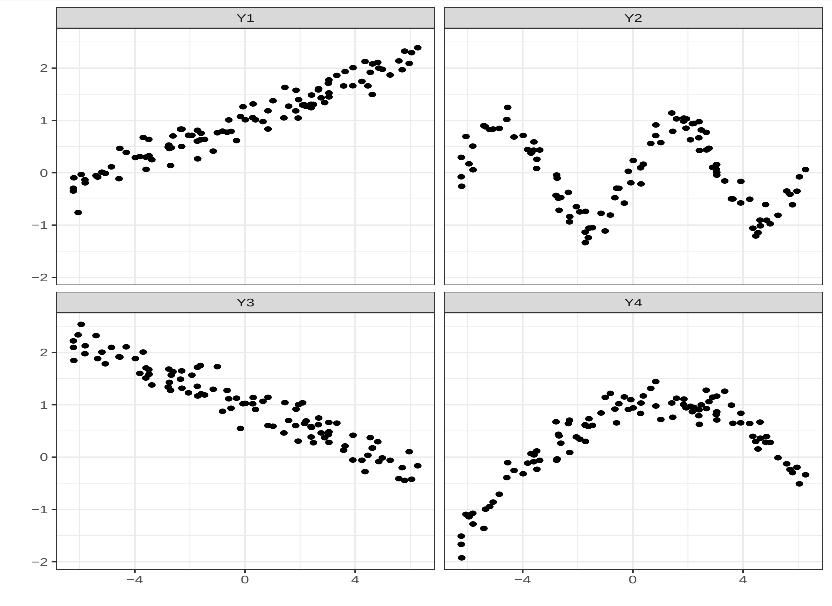
- Les données : $(x_1, y_1), \dots, (x_n, y_n)$ où $y_i \in \mathbb{R}$ et $x_i \in \mathbb{R}$ (pour simplifier).
- L'échantillon $(x_1, Y_1) \dots, (x_n, Y_n)$ i.i.d. (on suppose que les x_i sont déterministes).
- Le problème : expliquer les sorties Y_i par les entrées X_i.

- Les données : $(x_1, y_1), \ldots, (x_n, y_n)$ où $y_i \in \mathbb{R}$ et $x_i \in \mathbb{R}$ (pour simplifier).
- L'échantillon $(x_1, Y_1) \ldots, (x_n, Y_n)$ 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 $m : \mathbb{R} \to \mathbb{R}$ telle que

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

  où les termes d'erreurs ε_i sont i.i.d. de loi $\mathcal{N}(0, \sigma^2)$.
- Le problème statistique : estimer m.

Exemples



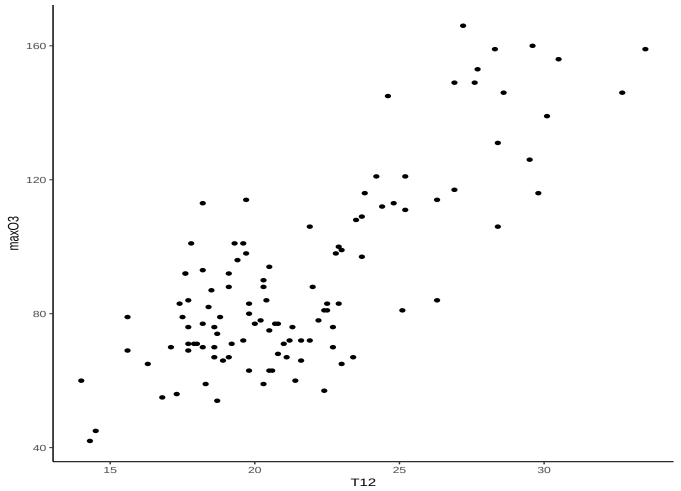
Un exemple concret

- 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()
```



Le modèle linéaire

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

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

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 : $\beta = (\beta_0, \beta_1) \in \mathbb{R}^2$

Le modèle linéaire

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

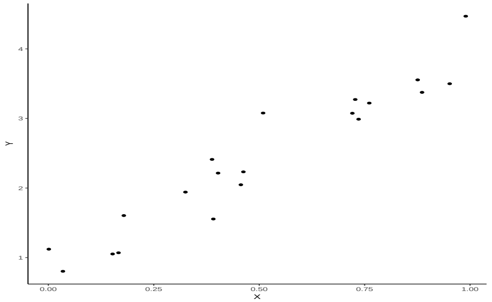
$$m(x) = \beta_0 + \beta_1 x, \quad \beta_0 \in \mathbb{R}, \beta_1 \in \mathbb{R}.$$

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

Ajustement linéaire d'un nuage de points

Notations

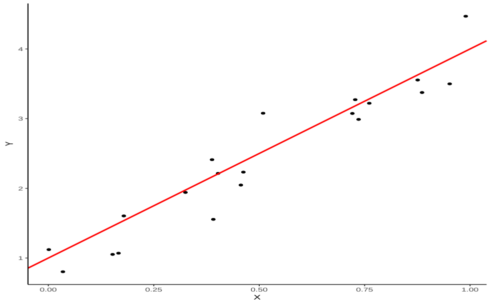
- n observations $y_1, \ldots, y_n$ de la variable à expliquer (maxO3).
- n observations $x_1, \ldots, x_n$ de la variable explicative (T12).



Ajustement linéaire d'un nuage de points

Notations

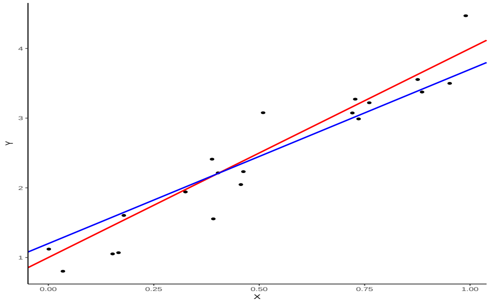
- n observations $y_1, \dots, y_n$ de la variable à expliquer (maxO3).
- n observations $x_1, \dots, x_n$ de la variable explicative (T12).



Ajustement linéaire d'un nuage de points

Notations

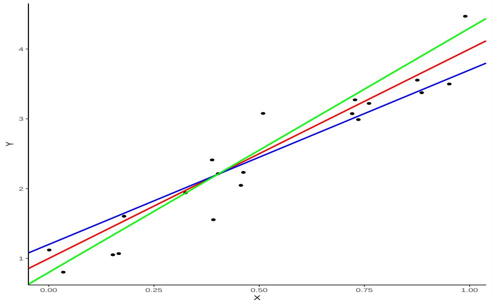
- n observations $y_1, \ldots, y_n$ de la variable à expliquer (maxO3).
- n observations $x_1, \ldots, x_n$ de la variable explicative (T12).



Ajustement linéaire d'un nuage de points

Notations

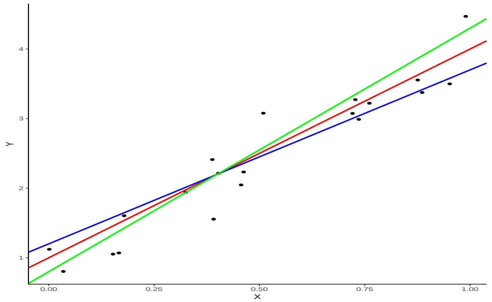
- n observations $y_1, \dots, y_n$ de la variable à expliquer (maxO3).
- n observations $x_1, \dots, x_n$ de la variable explicative (T12).



Ajustement linéaire d'un nuage de points

Notations

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



Le problème

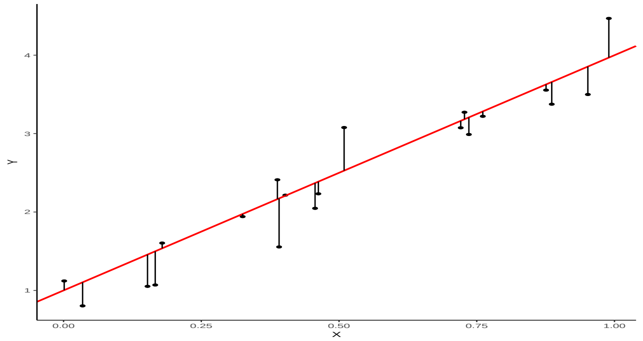
Trouver la droite qui ajuste au mieux le nuage de points.

- 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 :

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

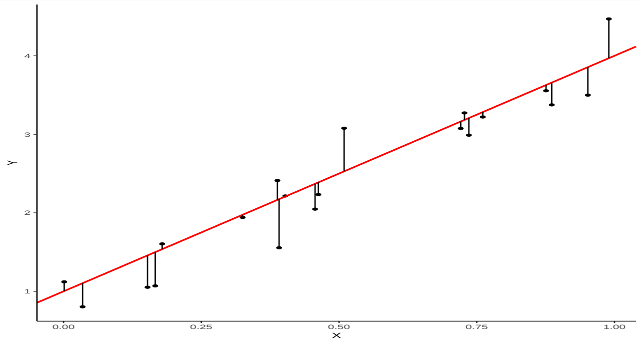
- On cherche $y = \beta_0 + \beta_1 x$ qui ajuste au mieux le nuage des points.
- Toutes les observations mesurées ne se trouvent pas sur une droite :

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



- 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 ε_i.

Le critère des moindres carrés

Critère des MC

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

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

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

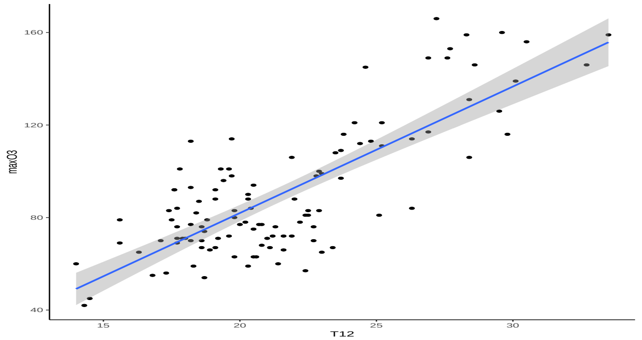
La solution est donnée par :

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

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 ε_i sont i.i.d. de loi $\mathcal{N}(0, \sigma^2)$.
- Les estimateurs des MCO :
$$\hat{\beta}_0 = \bar{Y} - \hat{\beta}_1 \bar{x} \quad \text{et} \quad \hat{\beta}_1 = \frac{\sum_{i=1}^n (Y_i - \bar{Y})(x_i - \bar{x})}{\sum_{i=1}^n (x_i - \bar{x})^2}.$$

Les estimateurs des MCO

Rappels

- Le modèle

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

où les ε_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

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

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 $1/n$.

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
```

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
```

- 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.

- 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 :

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

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

- 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 :

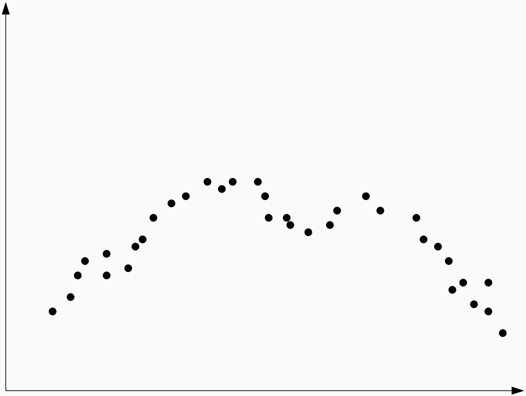
$$\widehat{m}_n(x) = \sum_{i=1}^{n} W_{ni}(x) Y_i$$

  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).

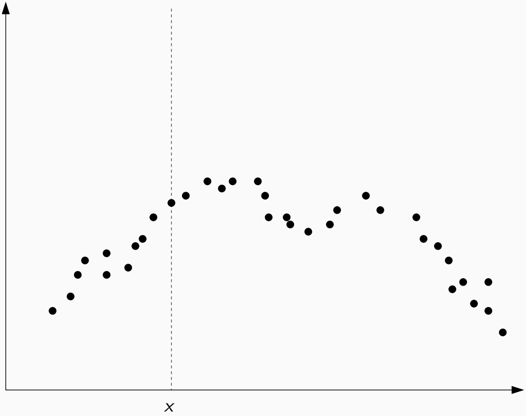
La méthode

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



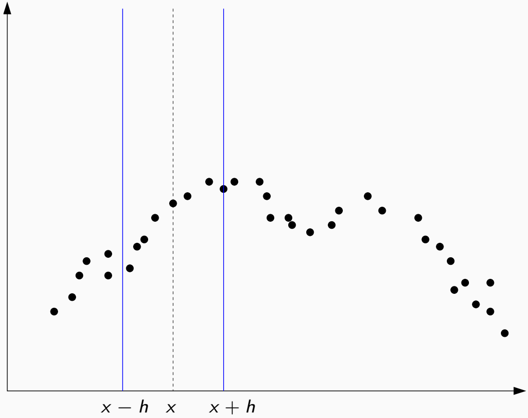
La méthode

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



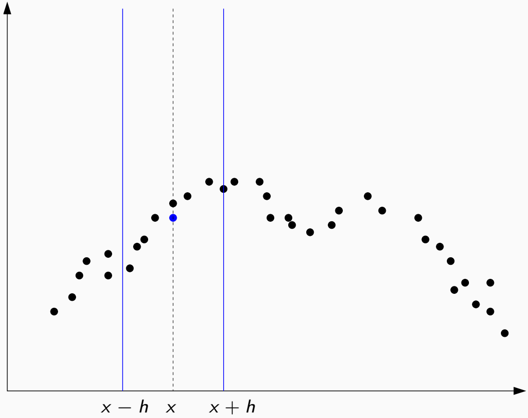
La méthode

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



La méthode

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



- L'estimateur s'écrit

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

- 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 $K : \mathbb{R} \to \mathbb{R}^+$. 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 : $K(x) = \frac{3}{4}(1 - x^2)\mathbf{1}_{|x| \leq 1}$.

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 ;

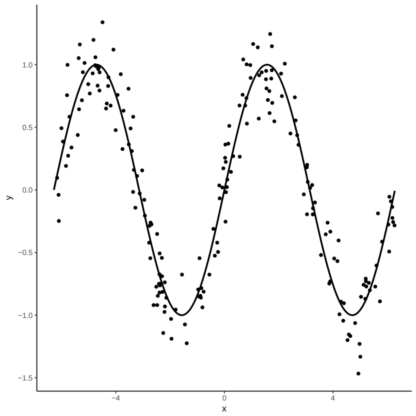
Un exemple

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

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

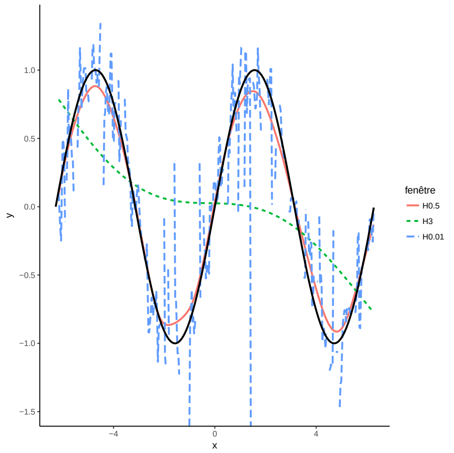
avec X_i uniforme sur $[-2\pi, 2\pi]$, ε_i de loi gaussienne $\mathcal{N}(0, 0.2^2)$.

```
> 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$x,
     "H0.5"=fx1$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.

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.

- 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 $n^{-4/5}$.

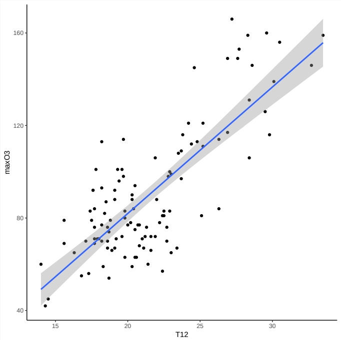
- Toutes les remarques faites pour l'estimateur à noyau de la densité sont valables pour l'estimateur de Nadaraya Watson.
- Le h optimal est de l'ordre de $n^{-1/5}$. Pour cette valeur de h, le risque quadratique est de l'ordre de $n^{-4/5}$.
- On obtient donc une vitesse de convergence plus lente que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

- 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

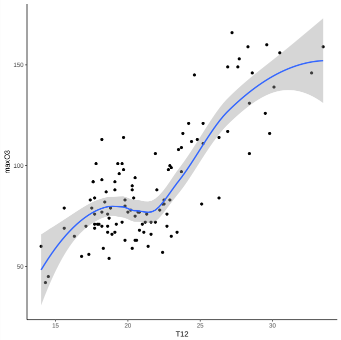
Paramétrique (linéaire)

```
> 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()
```



Références i

Nadaraya, E. A. (1964).
On estimating regression.
Theory of Probability and its Applications, 9.

Parzen, E. (1962).
On estimation of a probability density function and mode.
Ann. Math. Stat., 33 :1065–1076.

Watson, G. S. (1964).
Smooth regression analysis.
Sankhya : The Indian Journal of Statistics, Series A, 26 :359–372.