

# Estimation ponctuelle

Statistique mathématique  
M2 santé publique, université Paris-Sud

10 novembre 2017

1. Notion de statistique
2. Statistique libre
3. Exhaustivité
  - ▶ Exhaustivité
  - ▶ Caractérisation de l'exhaustivité
4. Exhaustivité minimale
  - ▶ Caractérisation de l'exhaustivité minimale
5. Complétude
  - ▶ Relation entre *libre* et *exhaustive*
  - ▶ Relation entre *Complète* et *exhaustive minimale*

# Modèle statistique et problème d'inférence

Rappelons que nous avons

- ▶ Collection de v.a (un vecteur aléatoire)  $X = (X_1, \dots, X_n)$
- ▶  $X \sim F_\theta \in \mathcal{F}$ , souvent on fera appel à  $f(x; \theta)$  au lieu de  $F_\theta$ .
- ▶  $\mathcal{F}$  une famille paramétrique  $\theta \in \Theta \subseteq \mathbb{R}^d$

# Modèle statistique et problème d'inférence

Rappelons que nous avons

- ▶ Collection de v.a (un vecteur aléatoire)  $X = (X_1, \dots, X_n)$
- ▶  $X \sim F_\theta \in \mathcal{F}$ , souvent on fera appel à  $f(x; \theta)$  au lieu de  $F_\theta$ .
- ▶  $\mathcal{F}$  une famille paramétrique  $\theta \in \Theta \subseteq \mathbb{R}^d$

## Le problème de l'estimation ponctuelle

- ▶ Supposons que  $F$  est complètement définie par son paramètre  $\theta$   
**inconnu**
- ▶ Soit  $(x_1, \dots, x_n)$  des réalisations de  $X \sim F_\theta$
- ▶ Estimer la valeur de  $\theta$  qui a **généré** les réalisations  $(x_1, \dots, x_n)$

# Modèle statistique et problème d'inférence

Rappelons que nous avons

- ▶ Collection de v.a (un vecteur aléatoire)  $X = (X_1, \dots, X_n)$
- ▶  $X \sim F_\theta \in \mathcal{F}$ , souvent on fera appel à  $f(x; \theta)$  au lieu de  $F_\theta$ .
- ▶  $\mathcal{F}$  une famille paramétrique  $\theta \in \Theta \subseteq \mathbb{R}^d$

## Le problème de l'estimation ponctuelle

- ▶ Supposons que  $F$  est complètement définie par son paramètre  $\theta$  **inconnu**
- ▶ Soit  $(x_1, \dots, x_n)$  des réalisations de  $X \sim F_\theta$
- ▶ Estimer la valeur de  $\theta$  qui a **généré** les réalisations  $(x_1, \dots, x_n)$

L'information qu'on possède : c'est  $(x_1, \dots, x_n)$  et  $\mathcal{F}$

- ▶ Ce qu'on peut construire n'est rien d'autre qu'une **fonction** des données  $g(x_1, \dots, x_n)$
- ▶ Nous allons étudier les propriétés de telles fonctions et la perte d'information qu'on subit (une fonction de  $(x_1, \dots, x_n)$  apporte au plus la même information que l'échantillon entier. Souvent on subit une perte d'information)

# Notion de statistique

## Définition d'une statistique

Soit  $X$  un échantillon (des v.a iid) issu de  $F_\theta$ . une **statistique** est une fonction (ou application) **mesurable**  $T$  qui envoie  $X$  dans  $\mathbb{R}^d$  et ne dépend pas de  $\theta$ .

- ▶ Intuitivement, toute fonction de l'échantillon est une statistique.
- ▶ Toute statistique est elle même une v.a avec sa propre loi.

# Notion de statistique

## Définition d'une statistique

Soit  $X$  un échantillon (des v.a iid) issu de  $F_\theta$ . une **statistique** est une fonction (ou application) **mesurable**  $T$  qui envoie  $X$  dans  $\mathbb{R}^d$  et ne dépend pas de  $\theta$ .

- ▶ Intuitivement, toute fonction de l'échantillon est une statistique.
- ▶ Toute statistique est elle même une v.a avec sa propre loi.

## Exemple

- ▶  $T(X) = n^{-1} \sum_{i=1}^n X_i$  est une statistique (rappelons que la taille de l'échantillon  $n$  est connue).
- ▶  $T(X) = (X_{(1)}, \dots, X_{(n)})$  où  $X_{(1)} \leq X_{(2)} \leq \dots \leq X_{(n)}$  sont les statistiques d'ordre de  $X$ . Puisque  $T$  dépend seulement des valeurs de  $X$ ,  $T$  est une statistique.
- ▶ Soit  $T(X) = c$ , où  $c$  est une constante. Alors  $T$  est une statistique.

# Statistique et information sur $\theta$

- ▶ Parmi les exemples précédents, certaines statistiques sont plus informatives que d'autres au vu de la vraie valeur de  $\theta$ .
- ▶ Une question naturelle : Quelles sont les *bonnes* statistiques et les *mauvaises* statistiques.

## Statistique libre

Une statistique  $T$  est dite **libre** (pour  $\theta$ ) si sa loi de probabilité ne dépend pas *fonctionnellement* de  $\theta$ .

$\rightsquigarrow$  Donc une statistique libre a la même loi  $\forall \theta \in \Theta$ .



# Statistique et information sur $\theta$

- ▶ Parmi les exemples précédents, certaines statistiques sont plus informatives que d'autres au vu de la vraie valeur de  $\theta$ .
- ▶ Une question naturelle : Quelles sont les *bonnes* statistiques et les *mauvaises* statistiques.

## Statistique libre

Une statistique  $T$  est dite **libre** (pour  $\theta$ ) si sa loi de probabilité ne dépend pas *fonctionnellement* de  $\theta$ .

$\rightsquigarrow$  Donc une statistique libre a la même loi  $\forall \theta \in \Theta$ .

## Exemple

Supposons que  $X_1, \dots, X_n \stackrel{\text{iid}}{\sim} \mathcal{N}(\mu, \sigma^2)$  (où  $\mu$  est inconnu et  $\sigma^2$  est connu). Soit  $T(X_1, \dots, X_n) = X_1 - X_2$ .

# Statistique et information sur $\theta$

- ▶ Parmi les exemples précédents, certaines statistiques sont plus informatives que d'autres au vu de la vraie valeur de  $\theta$ .
- ▶ Une question naturelle : Quelles sont les *bonnes* statistiques et les *mauvaises* statistiques.

## Statistique libre

Une statistique  $T$  est dite **libre** (pour  $\theta$ ) si sa loi de probabilité ne dépend pas *fonctionnellement* de  $\theta$ .

$\rightsquigarrow$  Donc une statistique libre a la même loi  $\forall \theta \in \Theta$ .

## Exemple

Supposons que  $X_1, \dots, X_n \stackrel{\text{iid}}{\sim} \mathcal{N}(\mu, \sigma^2)$  (où  $\mu$  est inconnu et  $\sigma^2$  est connu). Soit  $T(X_1, \dots, X_n) = X_1 - X_2$ . La loi de  $T$  est normale de moyenne 0 et de variance  $2\sigma^2$ . On déduit que  $T$  est une statistique libre pour le paramètre inconnu  $\mu$ . Si  $\mu$  et  $\sigma^2$  sont inconnus,  $T$  n'est pas libre pour le paramètre vectoriel  $\theta = (\mu, \sigma^2)$ .

# Statistique et information sur $\theta$

- ▶ Si  $T$  est libre pour  $\theta$  alors  $T$  ne contient pas d'information sur  $\theta$ .
- ▶ Pour contenir une information utile sur  $\theta$ , la loi de  $T$  doit dépendre explicitement de  $\theta$ .
- ▶ Intuitivement, la *quantité* d'information apportée par  $T$  sur  $\theta$  est proportionnelle à la **dépendance** de la loi de  $T$  de  $\theta$ .

# Statistique et information sur $\theta$

- ▶ Si  $T$  est libre pour  $\theta$  alors  $T$  ne contient pas d'information sur  $\theta$ .
- ▶ Pour contenir une information utile sur  $\theta$ , la loi de  $T$  doit dépendre explicitement de  $\theta$ .
- ▶ Intuitivement, la *quantité* d'information apportée par  $T$  sur  $\theta$  est proportionnelle à la **dépendance** de la loi de  $T$  de  $\theta$ .

## Exemple

Soit  $X_1, \dots, X_n \stackrel{\text{iid}}{\sim} \mathcal{U}[0, \theta]$ ,  $S = \min(X_1, \dots, X_n)$  et  $T = \max(X_1, \dots, X_n)$ .

- ▶  $f_S(x, \theta) = \frac{n}{\theta} \left(1 - \frac{x}{\theta}\right)^{n-1}$ , où  $0 \leq x \leq \theta$
- ▶  $f_T(x, \theta) = \frac{n}{\theta} \left(\frac{x}{\theta}\right)^{n-1}$ , où  $0 \leq x \leq \theta$ 
  - ▶ Aucune des deux statistiques  $S$  et  $T$  n'est libre pour  $\theta$ .
  - ▶ Quand  $n \rightarrow +\infty$   $f_S$  se concentre autour de 0.
  - ▶ Quand  $n \rightarrow +\infty$   $f_T$  se concentre autour de  $\theta$ .
- ▶ On déduit que  $T$  apporte plus d'information sur  $\theta$  que  $S$ .

# Statistique et information sur $\theta$

- ▶  $X = (X_1, \dots, X_n) \stackrel{\text{iid}}{\sim} F_\theta$  et  $T(X)$  une statistique.
- ▶ On définit les ensembles de niveaux de  $T$

$$A_t = \{x \in \mathbb{R}^n : T(x) = t\}.$$

(Ensemble des échantillons qui mènent à la même valeur  $t$  de  $T$ ).

- ▶  $T$  est constante quand on se restreint à  $A_t$ .
- ▶ Toutes les réalisations de  $X$  qui appartiennent au même ensemble de niveau sont équivalentes vis à vis de  $T$ .
- ▶ Toutes les inférences sont les mêmes à l'intérieur du même ensemble de niveau  $A_t$ .
- ▶ Regardons la loi de  $X$  dans l'ensemble de niveau  $A_t$  c'est à dire  $f_{X|T=t}$

# Statistique et information sur $\theta$

- (1) Si  $f_{X|T=t}$  dépend de  $\theta$  alors : **perte d'information**.
- (2) Si l'expression de  $f_{X|T=t}$  ne dépend pas de  $\theta$ 
  - ▶  $X$  ne contient pas d'information sur  $\theta$  dans l'ensemble  $A_t$ .
  - ▶ Autrement dit :  $X$  est libre pour  $\theta$  dans  $A_t$ .

# Statistique et information sur $\theta$

- (1) Si  $f_{X|T=t}$  dépend de  $\theta$  alors : **perte d'information**.
- (2) Si l'expression de  $f_{X|T=t}$  ne dépend pas de  $\theta$ 
  - ▶  $X$  ne contient pas d'information sur  $\theta$  dans l'ensemble  $A_t$ .
  - ▶ Autrement dit :  $X$  est libre pour  $\theta$  dans  $A_t$ .

## Interprétation de la deuxième situation

Si cela est vrai pour tout  $t \in \text{Image}(T)$  alors  $T(X)$  contient la même quantité d'information sur  $\theta$  que ce que peut contenir  $X$ .

- ▶ Il n'y pas de différence entre l'observation de  $X = (X_1, \dots, X_n)$  entier et  $T(X)$ .
- ▶ La connaissance de la valeur de  $X$  en plus de  $T(X)$  n'apporte aucune information supplémentaire sur  $X$ .

# Statistique et information sur $\theta$

- (1) Si  $f_{X|T=t}$  dépend de  $\theta$  alors : **perte d'information**.
- (2) Si l'expression de  $f_{X|T=t}$  ne dépend pas de  $\theta$ 
  - ▶  $X$  ne contient pas d'information sur  $\theta$  dans l'ensemble  $A_t$ .
  - ▶ Autrement dit :  $X$  est libre pour  $\theta$  dans  $A_t$ .

## Interprétation de la deuxième situation

Si cela est vrai pour tout  $t \in \text{Image}(T)$  alors  $T(X)$  contient la même quantité d'information sur  $\theta$  que ce que peut contenir  $X$ .

- ▶ Il n'y pas de différence entre l'observation de  $X = (X_1, \dots, X_n)$  entier et  $T(X)$ .
- ▶ La connaissance de la valeur de  $X$  en plus de  $T(X)$  n'apporte aucune information supplémentaire sur  $X$ .

## Statistique exhaustive

Une statistique  $T = T(X)$  est dite **exhaustive** pour le paramètre  $\theta$  si pour tout ensemble (**Borelien**)  $B$ , la probabilité  $\mathbb{P}[X \in B \mid T(X) = t]$  ne dépend pas de  $\theta$ .



# Statistique exhaustive

## Exemple

Soit  $X_1, \dots, X_n \stackrel{\text{iid}}{\sim} \text{Bernoulli}(\theta)$  et  $T(X) = \sum_{i=1}^n X_i$ . Soit  $x \in \{0, 1\}^n$ ,

$$\begin{aligned}\mathbb{P}[X = x \mid T = t] &= \frac{\mathbb{P}[X = x, T = t]}{\mathbb{P}[T = t]} \\ &= \frac{\mathbb{P}[X = x]}{\mathbb{P}[T = t]} \chi \left\{ \sum_{i=1}^n x_i = t \right\} \\ &= \frac{\theta^t (1 - \theta)^{n-t}}{C_n^t \theta^t (1 - \theta)^{n-t}} = \frac{1}{C_n^t}.\end{aligned}$$

- ▶  $T$  est exhaustive pour  $\theta$ .
- ▶ Les positions des 1 dans les  $n$  réalisations de Bernoulli importent peu : 0011101 vs 1000111 vs 1010101.

# Statistique exhaustive

- ▶ La définition précédente est difficile à vérifier notamment dans le cas continu.
- ▶ Cette définition ne permet pas d'identifier facilement les statistiques exhaustives.

# Statistique exhaustive

- ▶ La définition précédente est difficile à vérifier notamment dans le cas continu.
- ▶ Cette définition ne permet pas d'identifier facilement les statistiques exhaustives.

## Théorème de factorisation de Fisher-Neyman

Supposons que l'échantillon  $X = (X_1, \dots, X_n)$  a une densité jointe  $f(x; \theta), \theta \in \Theta$ . Une statistique  $T = T(X)$  est exhaustive pour  $\theta$  si et seulement si

$$f(x; \theta) = g(T(x), \theta) h(x).$$

# Statistique exhaustive

## Exemple : loi uniforme

Soit  $X_1, \dots, X_n \stackrel{\text{iid}}{\sim} \mathcal{U}[0, \theta]$  où  $f(x; \theta) = \frac{1}{\theta} \chi\{x \in [0, \theta]\}$ . Montrons que  $X_{(n)}$  est exhaustive pour  $\theta$ .

# Statistique exhaustive

## Exemple : loi uniforme

Soit  $X_1, \dots, X_n \stackrel{\text{iid}}{\sim} \mathcal{U}[0, \theta]$  où  $f(x; \theta) = \frac{1}{\theta} \chi\{x \in [0, \theta]\}$ . Montrons que  $X_{(n)}$  est exhaustive pour  $\theta$ .

$$\begin{aligned} f(x_1, \dots, x_n; \theta) &= \frac{1}{\theta^n} \chi\{x \in [0, \theta]^n\} \\ &= \frac{\chi\{\max [x_1, \dots, x_n] \leq \theta\} \chi\{\min [x_1, \dots, x_n] \geq 0\}}{\theta^n} \\ &= g\left(\max [x_1, \dots, x_n]; \theta\right) h(x_1, \dots, x_n) \end{aligned}$$

On déduit que la statistique  $T(X) = X_{(n)} = \max [x_1, \dots, x_n]$  est exhaustive pour  $\theta$ .

# Statistique exhaustive

## Exemple : famille exponentielle

Soit  $X_1, \dots, X_n \stackrel{\text{iid}}{\sim} f(x; \theta)$  où  $f(x; \theta)$  est une densité associée à une loi issue de la famille exponentielle avec un certain nombre  $k$  de paramètres.

$$f(x; \theta) = \exp \left[ \sum_{i=1}^k c_i(\theta) T_i(x) - d(\theta) + S(x) \right] \chi \{x \in A\}.$$

On choisit  $h(x) = \exp [S(x)] \chi \{x \in A\}$ , on déduit de théorème factorisation que la statistique

$$T = (T_1(X), \dots, T_k(X))$$

est exhaustive pour  $\theta$ .

# Statistique exhaustive

## Preuve du théorème de Neyman-Fisher - cas discret

Supposons que  $T$  est exhaustive. Ainsi

# Statistique exhaustive

## Preuve du théorème de Neyman-Fisher - cas discret

Supposons que  $T$  est exhaustive. Ainsi

$$\begin{aligned} f(x; \theta) &= \mathbb{P}[X = x] = \sum_t \mathbb{P}[X = x, T = t] \\ &= \mathbb{P}[X = x, T = T(x)] = \mathbb{P}[T = T(x)] \mathbb{P}[X = x \mid T = T(x)]. \end{aligned}$$

Comme  $T$  est exhaustive,  $\mathbb{P}[X = x \mid T = T(x)]$  est indépendante de  $\theta$  et donc  $f(x; \theta) = g(T(x); \theta)h(x)$ .



# Statistique exhaustive

## Preuve du théorème de Neyman-Fisher - cas discret

Supposons que  $T$  est exhaustive. Ainsi

$$\begin{aligned} f(x; \theta) &= \mathbb{P}[X = x] = \sum_t \mathbb{P}[X = x, T = t] \\ &= \mathbb{P}[X = x, T = T(x)] = \mathbb{P}[T = T(x)] \mathbb{P}[X = x \mid T = T(x)]. \end{aligned}$$

Comme  $T$  est exhaustive,  $\mathbb{P}[X = x \mid T = T(x)]$  est indépendante de  $\theta$  et donc  $f(x; \theta) = g(T(x); \theta) h(x)$ .

Maintenant, supposons que  $f(x; \theta) = g(T(x); \theta) h(x)$ . Alors si  $T(x) = t$ ,

$$\begin{aligned} \mathbb{P}[X = x \mid T = t] &= \frac{\mathbb{P}[X = x, T = t]}{\mathbb{P}[T = t]} = \frac{\mathbb{P}[X = x]}{\mathbb{P}[T = t]} \chi\{T(x) = t\} \\ &= \frac{g(T(x); \theta) h(x) \chi\{T(x) = t\}}{\sum_{y: T(y)=t} g(T(y); \theta) h(y)} = \frac{h(x) \chi\{T(x) = t\}}{\sum_{y: T(y)=t} h(y)}. \end{aligned}$$

# Statistique exhaustive minimale

- ▶ Quand une statistique exhaustive apporte-t-elle de l'information importante seulement ?
- ▶ Peut-on renoncer à de l'information ? laquelle ? combien ?

# Statistique exhaustive minimale

- ▶ Quand une statistique exhaustive apporte-t-elle de l'information importante seulement ?
- ▶ Peut-on renoncer à de l'information ? laquelle ? combien ?

## Définition d'une statistique exhaustive minimale

Une statistique  $T = T(X)$  est dite **exhaustive minimale** pour un paramètre  $\theta$  si pour toute autre statistique  $S$  exhaustive pour  $\theta$ , il existe une fonction  $g(\cdot)$  où

$$T(X) = g(S(X)).$$

# Statistique exhaustive minimale

- ▶ Quand une statistique exhaustive apporte-t-elle de l'information importante seulement ?
- ▶ Peut-on renoncer à de l'information ? laquelle ? combien ?

## Définition d'une statistique exhaustive minimale

Une statistique  $T = T(X)$  est dite **exhaustive minimale** pour un paramètre  $\theta$  si pour toute autre statistique  $S$  exhaustive pour  $\theta$ , il existe une fonction  $g(\cdot)$  où

$$T(X) = g(S(X)).$$

## Proposition

Si  $T$  et  $S$  sont des statistiques exhaustives minimales pour un paramètre  $\theta$ , alors il existe des fonctions injectives  $g$  et  $h$  telles que  $S = g(T)$  et  $T = h(S)$ .

# Exhaustive minimale : caractérisation

## Théorème

Soit  $X = (X_1, \dots, X_n)$  un échantillon de densité jointe (ou fonction de masse)  $f(x; \theta)$  et  $T = T(X)$  une statistique. Si

$\frac{f(x; \theta)}{f(y; \theta)}$  est indépendant de  $\theta \Leftrightarrow T(x) = T(y)$ , alors  $T$  est **exhaustive minimale** pour  $\theta$ .

# Exhaustive minimale : caractérisation

## Théorème

Soit  $X = (X_1, \dots, X_n)$  un échantillon de densité jointe (ou fonction de masse)  $f(x; \theta)$  et  $T = T(X)$  une statistique. Si

$\frac{f(x; \theta)}{f(y; \theta)}$  est indépendant de  $\theta \Leftrightarrow T(x) = T(y)$ , alors  $T$  est **exhaustive minimale** pour  $\theta$ .

## Preuve (conditions $\Rightarrow$ exhaustivité)

Soit  $\mathcal{T} = \{T(y) : y \in \mathbb{R}^n\}$  l'image de  $\mathbb{R}^n$  par  $T$  et soit  $A_t$  un ensemble de niveau de  $T$ . Pour tout  $t$ , on choisit un représentant  $y_t \in A_t$ . Notons que pour tout  $x$ ,  $y_{T(x)}$  est dans le même ensemble de niveau que  $x$ , donc  $f(x; \theta) / f(y_{T(x)}; \theta)$  ne dépend pas de  $\theta$  par hypothèse. On pose  $g(t, \theta) = f(y_t; \theta)$  et notons

$$f(x; \theta) = \frac{f(y_{T(x)}; \theta) f(y; \theta)}{f(y_{T(x)}; \theta)} = g(T(x); \theta) h(x),$$

on obtient ainsi l'exhaustivité par le théorème de factorisation.

# Exhaustive minimale : caractérisation

## Preuve (conditions $\Rightarrow$ minimalité)

Soit  $T'$  une autre statistique exhaustive. Par le théorème de factorisation :  $\exists g', h'$  telles que  $f(x; \theta) = g'(T'(x); \theta) h'(x)$ . Soit  $x$  et  $y$  tels que  $T(x) = T(y)$ . Alors

$$\frac{f(x; \theta)}{f(y; \theta)} = \frac{g'(T'(x); \theta) h'(x)}{g'(T'(y); \theta) h'(y)} = \frac{h'(x)}{h'(y)}.$$

Puisque ce rapport ne dépend pas de  $\theta$ , nous avons par hypothèse  $T(x) = T(y)$ . Ainsi  $T$  est une fonction de  $T'$ ; donc elle est minimale par un choix arbitraire de  $T'$ .

# Exhaustive minimale : exemple

## Retour au modèle de Bernoulli

Soit  $X_1, \dots, X_n \stackrel{\text{iid}}{\sim} \text{Bernoulli}(\theta)$ . Soit  $x$  et  $y \in \{0, 1\}^n$  deux réalisations possibles. Montrons que  $T(x) = \sum_{i=1}^n x_i$  est exhaustive minimale.



# Exhaustive minimale : exemple

## Retour au modèle de Bernoulli

Soit  $X_1, \dots, X_n \stackrel{\text{iid}}{\sim} \text{Bernoulli}(\theta)$ . Soit  $x$  et  $y \in \{0, 1\}^n$  deux réalisations possibles. Montrons que  $T(x) = \sum_{i=1}^n x_i$  est exhaustive minimale. Nous avons

$$\frac{f(x; \theta)}{f(y; \theta)} = \frac{\theta^{\sum x_i} (1 - \theta)^{n - \sum x_i}}{\theta^{\sum y_i} (1 - \theta)^{n - \sum y_i}}$$

qui est constant si et seulement si  $T(x) = \sum x_i = \sum y_i = T(y)$  et donc  $T$  est exhaustive minimale.

# Statistique complète

- ▶ Statistique libre  $\rightarrow$  ne contient pas d'information sur  $\theta$
- ▶ Statistique exhaustive minimale  $\rightarrow$  contient toute l'information pertinente et un **peu** d'information non-pertinente.
- ▶ Ces deux aspects sont-ils indépendants ?

# Statistique complète

- ▶ Statistique libre  $\rightarrow$  ne contient pas d'information sur  $\theta$
- ▶ Statistique exhaustive minimale  $\rightarrow$  contient toute l'information pertinente et un **peu** d'information non-pertinente.
- ▶ Ces deux aspects sont-ils indépendants ?

## Définition d'une statistique complète

Soit  $\{g(t; \theta) : \theta \in \Theta\}$  une famille de densités (ou fonctions de masse) pour  $T(X)$ . La statistique  $T$  est dite *complète* si pour toute fonction mesurable  $h$ , nous avons

$$\int h(t)g(t; \theta)dt = 0 \quad \forall \theta \in \Theta \implies \mathbb{P}[h(T) = 0] = 1 \quad \forall \theta \in \Theta$$

# Statistique complète

## Exemple

Soit  $X_1, \dots, X_n \stackrel{\text{iid}}{\sim} \text{Bernoulli}(\theta)$ ,  $\theta \in (0, 1)$ , et  $T = \sum_{i=1}^n X_i$ . Vérifions que  $T$  est complète.

# Statistique complète

## Exemple

Soit  $X_1, \dots, X_n \stackrel{\text{iid}}{\sim} \text{Bernoulli}(\theta)$ ,  $\theta \in (0, 1)$ , et  $T = \sum_{i=1}^n X_i$ . Vérifions que  $T$  est complète. Soit  $h$  une fonction arbitraire,

$$\mathbb{E}[h(T)] = \sum_{t=0}^n h(t) C_t^n \theta^t (1 - \theta)^{n-t} = (1 - \theta)^n \sum_{t=0}^n h(t) C_t^n \left( \frac{\theta}{1 - \theta} \right)^t$$

Puisque  $\theta \in (0, 1)$ , le rapport  $\theta / (1 - \theta)$  varie dans  $(0, \infty)$ . Ainsi, supposons que  $\mathbb{E}[h(T)] = 0$  pour tout  $\theta \in (0, 1)$ , nous avons

$$P(x) = \sum_{t=0}^n h(t) C_t^n x^t = 0 \quad x > 0,$$

i.e le polynôme  $P(x)$  est uniformément nul sur l'ensemble des nombres réels positifs. Donc les coefficients du polynôme sont nuls et donc  $h(t) = 0, t = 1, \dots, n$ . On déduit  $\mathbb{P}[h(T) = 0] = 1 \forall \theta \in (0, \infty)$ .

# Statistique complète

- La complétude est-elle pertinente pour la réduction des données ?

## Proposition

Si  $T$  est complète, alors  $h(T)$  est libre pour  $\theta$  si et seulement si  $h(T) = c$  presque sûrement.

## Preuve

Soit  $h(T)$  une statistique libre. Sa loi ne dépend pas de  $\theta$  .....

# Statistique complète

- ▶ La complétude est-elle pertinente pour la réduction des données ?

## Proposition

Si  $T$  est complète, alors  $h(T)$  est libre pour  $\theta$  si et seulement si  $h(T) = c$  presque sûrement.

## Preuve

Soit  $h(T)$  une statistique libre. Sa loi ne dépend pas de  $\theta$  ....

Ainsi  $\mathbb{E}[h(T)] = c$ , pour une certaine constante  $c$  et donc

$\mathbb{E}[h(T) - c] = 0$ . La complétude de  $T$  implique que  $\mathbb{P}[h(T) = c] = 1$ .

- ▶ Autrement dit : Il n'y a que les fonctions triviales (= constante) de  $T$  qui sont des statistiques libres.
- ▶ Une **statistique complète ne contient pas d'information libre**
- ▶ Une statistique exhaustive contient toute l'information pertinente alors qu'une statistique complète est épurée de toute information non-pertinente.

# Statistique complète

## Théorème de Basu

Une statistique exhaustive complète est **indépendante** de toute statistique libre.



# Statistique complète

## Théorème de Basu

Une statistique exhaustive complète est **indépendante** de toute statistique libre.

## Preuve

On va s'intéresser au cas discret seulement. On pose  $T$  une statistique exhaustive complète et  $S$  libre. Il suffit de montrer que

$$\mathbb{P}[S(X) = s \mid T(X) = t] = \mathbb{P}[S(X) = s]$$

On définit  $h(t) = \mathbb{P}[S(X) = s \mid T(X) = t] - \mathbb{P}[S(X) = s]$ .

On remarque que

- ▶  $\mathbb{P}[S(X) = s]$  ne dépend pas de  $\theta$  (libre)
- ▶  $\mathbb{P}[S(X) = s \mid T(X) = t] = \mathbb{P}[X \in \{x : S(x) = s\} \mid T = t]$  ne dépend pas de  $\theta$  (à cause de exhaustivité).

Et  $h$  ne dépend pas de  $\theta$ .

Ainsi pour tout  $\theta \in \Theta$ ,

preuve (suite)

$$\begin{aligned}\mathbb{E}h(T) &= \sum_t (\mathbb{P}[S(X) = s \mid T(X) = t] - \mathbb{P}[S(X) = s])\mathbb{P}[T(X) = t] \\ &= \sum_t \mathbb{P}[S(X) = s \mid T(X) = t]\mathbb{P}[T(X) = t] \\ &\quad - \mathbb{P}[S(X) = s] \sum_t \mathbb{P}[T(X) = t] \\ &= \mathbb{P}[S(X) = s] - \mathbb{P}[S(X) = s] = 0.\end{aligned}$$

Or  $T$  est complète, donc  $h(t) = 0$  pour tout  $t$ .

Le théorème de Basu est utile pour déduire l'indépendance entre deux statistiques.

- ▶ On n'en a pas besoin de calculer la loi jointe
- ▶ Besoin de montrer la complétude (souvent difficile)
- ▶ On va voir des modèles où il est facile de vérifier la complétude

# Complétude et exhaustivité minimale

## Théorème de Lehmann-Scheffé

Soit  $X$  de densité  $f(x; \theta)$ . Si  $T(X)$  est exhaustive et complète pour  $\theta$  alors  $T$  est exhaustive minimale.

# Complétude et exhaustivité minimale

## Théorème de Lehmann-Scheffé

Soit  $X$  de densité  $f(x; \theta)$ . Si  $T(X)$  est exhaustive et complète pour  $\theta$  alors  $T$  est exhaustive minimale.

## Théorème

Si une statistique exhaustive minimale existe, alors toute statistique complète est aussi exhaustive minimale.