Modélisation Statistique (MAP-STA1)

M1-Mathématiques Appliquées
1ère partie: Modélisation statistique
Cours 1: Introduction et rappels

Christine Keribin

[1]Laboratoire de Mathématiques d'Orsay
Université Paris-Saclay

2021-2022

Sommaire

Introduction

- On dispose d'un échantillon de 6 relevés du temps de trajet (en min) domicile/bureau d'un employé

```
x=(15,17,15,18,16,15)
```

$$\bar{x} = \frac{1}{6}\sum_i x_i = 16$$

1. Quelle est la durée moyenne d'un trajet sur l'année ?
2. Peut-on affirmer avec peu de risque que la durée moyenne d'un trajet est supérieure à 15 min ?
3. La durée d'un trajet a-t-elle augmenté d'une année sur l'autre ?
4. Y a-t-il des facteurs qui influencent la durée de trajet ?
5. Quelle sera la durée d'un trajet demain ?

Introduction

- Population : ensemble d'objets "équivalents" (individus, unités statistiques), sur lesquels on observe des caractéristiques (variables qualitatives ou quantitatives)
  - finie : recensement
  - infinie : sondage
- Étude de la variabilité
- Statistique exploratoire / statistique inférentielle

Statistique inférentielle

- Probabilité : étudier les propriétés d'une loi connue
- Statistique : à partir d'un ensemble d'observations d'une loi inconnue, inférer des propriétés de cette loi pour répondre à une question

  $\hookrightarrow$ résoudre un problème inverse
  - Modéliser
  - Estimer
  - Utiliser à des fins explicatives ou prédictives

Quelques objectifs de la statistique inférentielle

- estimation : valeur d'un paramètre d'intérêt, intervalle de confiance
- test : comparaison à une situation de référence, de deux échantillons, ...
- prédiction pour une nouvelle unité non encore observée
- classification dans un groupe

Problématiques :

- construction, comparaison, choix des procédures
- fiabilité (risque) de l'information obtenue ?

Objectifs du cours STA201

Être capable, en utilisant les bases de la statistique mathématique, de :

- définir une modélisation adaptée à un jeu de données
- construire des estimateurs (MV, MC)
  - étudier le risque, l'efficacité et l'asymptotique
- construire des tests et ICs (Wald, RV)
- travailler avec le modèle linéaire
- estimer un modèle statistique (linéaire) avec un logiciel (R) et interpréter les résultats obtenus
- prendre en compte le risque de toute décision statistique.

Le cours STA201 est un prérequis des cours

- STA202 (séries chronologiques), dans lequel seront développés les modèles d'observations non indépendantes
- STA203 (apprentissage statistique), dans lequel seront développés les modèles de régression non linéaires, logistique, ... et la prédiction ($\hookrightarrow$ machine learning)

Evaluation

Note = (PR+EX)/2

- EX : examen qui pourra comporter des questions théoriques et des questions pratiques d'interprétation de résultats.
- PR : projet à rendre par binôme

Chargés de TDs : Jérémie Capitao-Miniconi, Olivier Coudray, Guillermo Durand, Zacharie Naulet, Christine Keribin

documents, bibliographie, ... : moodle e-campus

Sommaire

Modélisation statistique

Modéliser l'expérience, c'est proposer une loi théorique pour l'échantillon $X = (X_1, \ldots, X_n)$.

- Modèle : $\mathcal{M} = (\mathcal{X}^n, \mathcal{A}^n, \mathcal{P})$
  - espace mesurable $(\mathcal{X}^n, \mathcal{A}^n)$
  - muni d'une famille de lois de proba $\mathcal{P} = (\mathbb{P}_\theta^n)_{\theta \in \Theta}$

  Quand il existe $d \in \mathbb{N}^*$ tel que $\Theta \subset \mathbb{R}^d$, le modèle est dit *paramétrique*
- Une observation est une variable aléatoire X à valeur dans $\mathcal{X}^n$ et dont la loi appartient à $(\mathbb{P}_\theta^n)_{\theta \in \Theta}$
- Un n-échantillon i.i.d. est une observation $X = (X_1, \ldots, X_n)$ de n variables aléatoires de même loi η_θ et indépendantes. Alors, $\mathbb{P}_\theta^n = (\eta_\theta)^{\otimes n}$
- Les données sont les réalisations (valeurs) $x_1, \ldots, x_n$ prises par l'échantillon $X_1, \ldots, X_n$

Exemples de modèles (simples)

- Etude de la moyenne (espérance) d'un temps de trajet : $\mathcal{M} = (\mathbb{R}^n, \mathcal{A}^n, \mathbb{P}_\theta^{\otimes n}, \theta \in \Theta)$

$$X_i \sim_{i.i.d.} \mathbb{P}(\mu, \sigma^2)$$

- Comparaison du rendement de maïs sous deux conditions de culture

$$X \sim \mathcal{N}^{\otimes n}(\mu_x, \sigma_x^2) \text{ et } Y \sim \mathcal{N}^{\otimes m}(\mu_y, \sigma_y^2) \text{ indépendants}$$

- Estimation d'une proportion par sondage $\mathcal{M} = (\{0,1\}^n, \mathcal{A}^n, \mathcal{B}_\pi^{\otimes n}, \pi \in [0;1])$

$$X_i \sim_{i.i.d.} \mathcal{B}(1, \pi)$$

Régression linéaire (simple)

$$Y_{i|X_i=x_i} = \alpha + \beta x_i + \varepsilon_i \text{ où } \varepsilon_{|X=x} \sim_{i.i.d.} \mathcal{L}(0, \sigma^2 I_n).$$

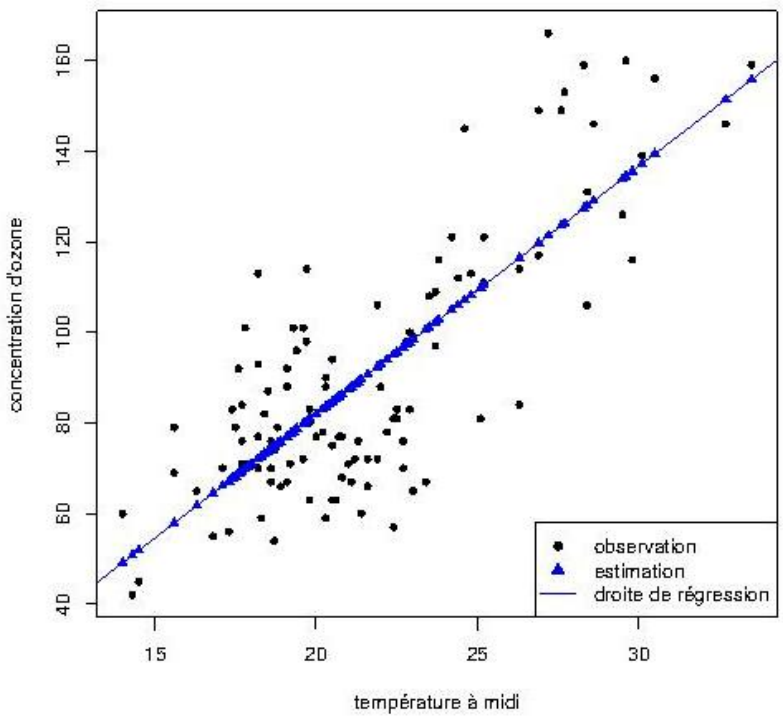


Figure – *source des données : Cornillon et Matzner-Løber, 2007*

Modèle paramétré identifiable

Définition

Un modèle à paramétrage dans Θ est identifiable si

$$\forall \theta, \theta' \in \Theta, \quad \mathbb{P}_\theta = \mathbb{P}_{\theta'} \Rightarrow \theta = \theta'$$

Les étapes de la démarche statistique

A partir des données d'un n-échantillon, déduire -ou inférer- certaines propriétés du modèle inconnu

1. Acquérir et préparer les données
2. Définir un modèle adapté à la situation observée.
3. Estimer les paramètres du modèle grâce aux observations.
4. Vérifier l'adéquation de l'estimation aux observations
5. Choisir entre des modèles
6. Utiliser le modèle à des fins de décision ou de prédiction.

Estimation ponctuelle

Un objet mathématique : l'estimateur

Soit $\theta \in \Theta \in \mathbb{R}$ le paramètre d'une loi $\mathbb{P}_\theta$ et soit $X = (X_1, \ldots, X_n)$ un n-échantillon issu de cette loi

Définition

Une statistique est une variable aléatoire T_n, fonction mesurable de l'échantillon et calculable à partir de l'échantillon

$$T_n = t(X_1, \ldots, X_n).$$

Un estimateur est une statistique utilisée pour estimer un paramètre ou une quantité d'intérêt $\nu(\theta)$

- Notation $T_n = \widehat{\nu}_n$ ou $T_n = \widehat{\nu}$.
- Exemples : estimateur empirique de l'espérance, de la variance

Performance en moyenne

Soit $\widehat{\nu}_n$ un estimateur de $\nu(\theta)$, fonction du paramètre d'une loi $\mathbb{P}_\theta$:

- On appelle biais de $\widehat{\nu}_n$ pour $\nu(\theta)$ la valeur

$$b_\theta(\widehat{\nu}_n) = \mathbb{E}_\theta(\widehat{\nu}_n) - \nu(\theta)$$

  Si $b_\theta(\widehat{\nu}_n) = 0$ pour tout $\theta \in \Theta$, T_n est sans biais pour $\nu(\theta)$

- On appelle variance de $\widehat{\nu}_n$ la valeur

$$\text{var}_\theta(\widehat{\nu}_n) = \mathbb{E}_\theta\left((\widehat{\nu}_n - \mathbb{E}_\theta(\widehat{\nu}_n))^2\right)$$

- On appelle risque quadratique de $\widehat{\nu}_n$ la valeur

$$R_\theta(\widehat{\nu}_n) = \mathbb{E}_\theta\left((\widehat{\nu}_n - \nu(\theta))^2\right)$$

Décomposition du risque quadratique

$$R_\theta(\widehat{\nu}_n) = \mathrm{var}_\theta(\widehat{\nu}_n) + (b_\theta(\widehat{\nu}_n))^2$$

Définition

Un estimateur δ_1 de $\nu(\theta)$ domine l'estimateur δ_2 si, pour tout $\theta \in \Theta$,

$$R_\theta(\delta_1) \leq R_\theta(\delta_2),$$

cette inégalité étant stricte pour au moins une valeur de θ.
Un estimateur est admissible s'il n'existe aucun estimateur le dominant.

- Soit $\theta_0 \in \Theta$. L'estimateur constant $\widehat{\nu}_n = \theta_0$ est-il admissible ?

Il n'existe en général pas d'estimateur dominant tous les autres $\hookrightarrow$ Recherche d'estimateurs UVMB

Performance asymptotique

Soit $\widehat{\nu}_n$ un estimateur de $\nu(\theta)$, défini à partir d'une observation de loi $\mathbb{P}_\theta$:

- $\widehat{\nu}_n$ est asymptotiquement sans biais pour $\nu(\theta)$:

$$\forall \theta \in \Theta, \lim_{n\to\infty} b_\theta(\widehat{\nu}_n) = 0$$

- $\widehat{\nu}_n$ est consistant : $\widehat{\nu}_n$ tend en probabilité vers $\nu(\theta)$ quand $n \to \infty$:

$$\forall \theta \in \Theta, \forall \epsilon, \lim_{n\to\infty} \mathbb{P}_\theta(|\widehat{\nu}_n - \nu(\theta)| > \epsilon) = 0$$

- $\widehat{\nu}$ est fortement consistant ssi

$$\forall \theta \in \Theta, \mathbb{P}_\theta(\lim_{n\to\infty} \widehat{\nu}_n = \nu(\theta)) = 1$$

- $\widehat{\nu}$ est consistant en moyenne quadratique :

$$\forall \theta \in \Theta, \lim_{n\to\infty} R_\theta(\widehat{\nu}_n) = 0$$

Vitesse de convergence

Définition

Si un estimateur $\widehat{\nu}_n$ *de* $\nu \in \mathbb{R}^p$ *de variance* $\text{var}(\widehat{\nu}_n) = V_n$ *a un comportement asymptotique tel que*

$$V_n^{-1/2}(\widehat{\nu}_n - \nu) \xrightarrow{\mathcal{L}} \mathcal{N}_p(0, Id_p)$$

on dit que l'estimateur est asymptotiquement normal. Si $nV_n \to V_0$ *où* $V_0 > 0$ *est finie, on dit que la vitesse de l'estimateur est en* $\sqrt{n}$

$\hookrightarrow$ Un estimateur est d'autant meilleur que sa vitesse de convergence est rapide et sa loi limite concentrée autour de 0.

$\hookrightarrow$ Il existe des estimateurs consistants non asymptotiquement normaux

L'approche test

L'approche test : exemple du test de Student

Choisir entre deux hypothèses (H_0) et (H_1), en calibrant le risque de première espèce α (5%, 10%,..) de choisir (H_1) à tort

1. Définir le modèle : $X \sim \mathcal{M}$ (par exemple, $\mathcal{N}^{\otimes n}(\mu, \sigma^2)$)
2. Définir les hypothèses
   - Hypothèse nulle (H_0) : $\mu = \mu_0 = 15$
   - Hypothèse alternative (H_1) : $\mu = \mu_1 > \mu_0$
3. Définir une statistique de test T, sa loi sous (H_0),

$$T = \sqrt{n}\frac{\bar{X} - \mu_0}{\widehat{\sigma}} \sim \mathcal{T}(n-1)$$

avec $\bar{X} = \sum X_i/n$ et $\widehat{\sigma}^2 = \sum(X_i - \bar{X})^2/(n-1)$

L'approche test (suite)

4. Définir la règle de décision en calibrant la région de rejet $\mathcal{R}$ suivant le niveau α choisi a priori :

$$\delta(T) = \mathbb{1}\{T \in \mathcal{R}\}, \text{ tel que}$$

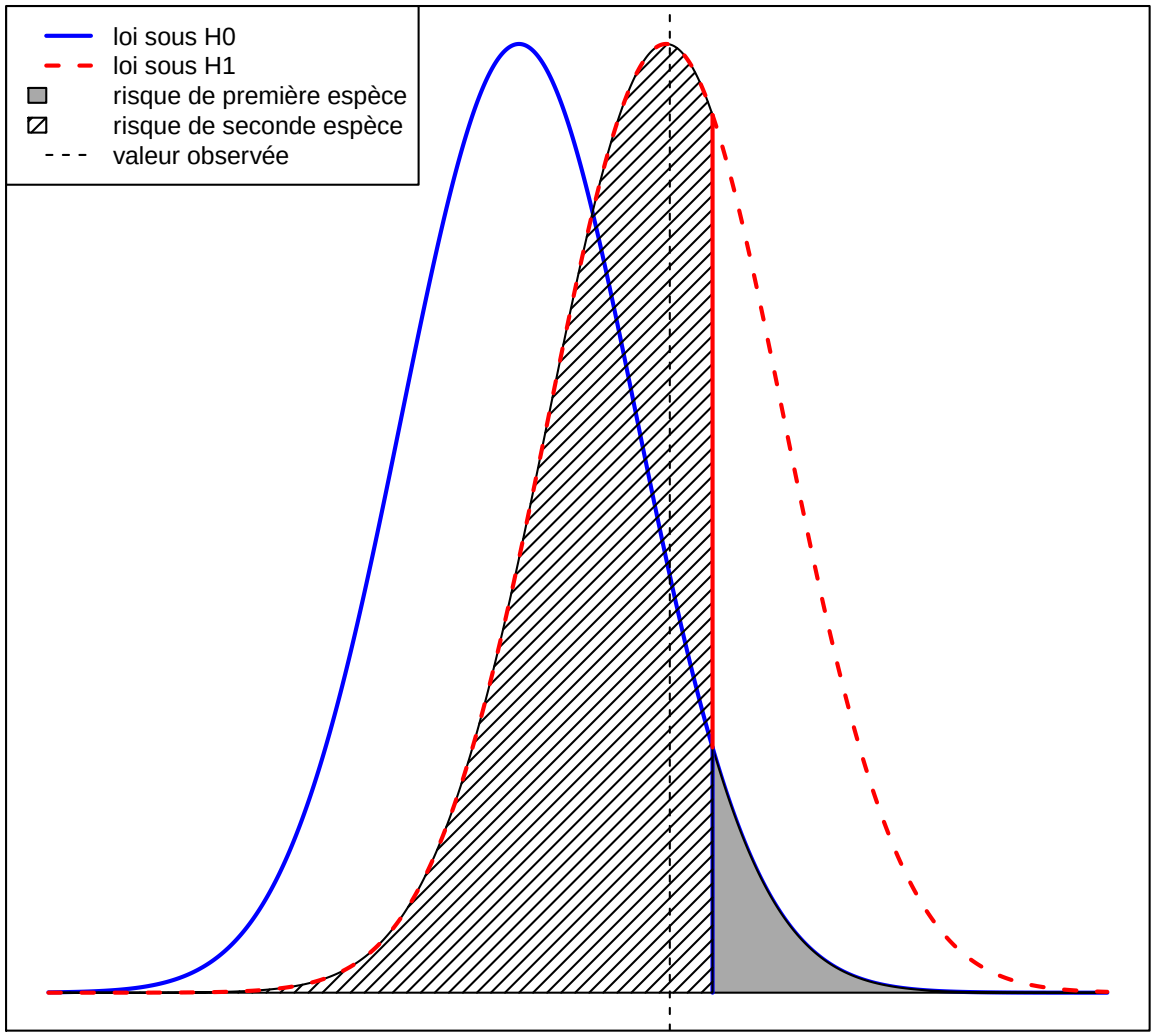
$$\alpha = \mathbb{P}_{(H_0)}\Big(\underbrace{T > q_{1-\alpha}^{\mathcal{T}(n-1)}}_{\mathcal{R}\text{: Région de rejet}} \Big)$$

5. Décider
   - si T est dans la région de rejet, on rejette H_0
   - sinon, on conserve (ou accepte) (H_0) faute de preuves suffisantes

A.N.

- Hypothèse nulle (H_0) : $\mu = \mu_0 = 15$
- Hypothèse alternative (H_1) : $\mu = \mu_1 > \mu_0$
- Risque $\alpha = 0.1$
- Données observées : $\bar{x} = 16$,
  $s^2 = \sum_{i=1}^{6}(x_i - \bar{x})^2/5 = 1.6$
- statistique de test observée : $t_{obs} = 1.88$
- quantiles : $q_{0.90}^{\mathcal{T}(5)} = 1.48$; $q_{0.95}^{\mathcal{T}(5)} = 2.02$
- décision ? risque ?

Les risques d'un test



Les risques d'un test

Soit $\mathcal{R}_\alpha$ la région de rejet du test de niveau α.

- Risque de 1ère espèce : probabilité de rejeter (H_0) à tort :
$$\theta_0 \in \Theta_0, \alpha(\theta_0) = \mathbb{P}_{\theta_0}(\{T \in \mathcal{R}\})$$
- Risque de 2nde espèce : probabilité de conserver (H_0) à tort
$$\theta_1 \in \Theta_1, \beta(\theta_1) = \mathbb{P}_{\theta_1}(\{T \notin \mathcal{R}\}).$$
- Puissance du test : probabilité de refuser (H_0) à raison :
$$\theta_1 \in \Theta_1, \pi(\theta_1) = \mathbb{P}_{\theta_1}(\{T \in \mathcal{R}\}) = 1 - \beta(\theta_1).$$
- le test est sans biais si $\pi > \alpha$; consistant si $\lim_{n\to\infty} \pi_n = 1$

Les risques d'un test

A l'issue d'un test, les quatre situations suivantes sont possibles

	Choix (H_0)	Choix (H_1)
(H_0) vraie	$1-\alpha$ bonne décision	$\alpha = \mathbb{P}(T \in \mathcal{R}\|(H_0))$ erreur première espèce mauvaise décision
(H_1) vraie	$\beta = \mathbb{P}(T \notin \mathcal{R}\|(H_1))$ erreur seconde espèce mauvaise décision	$\pi = 1-\beta$ puissance bonne décision

- Le rejet de (H_0) assure-t-il que (H_1) est vraie ?
- L'acceptation de (H_0) assure-t-elle que (H_0) est vraie ?
- Du point du vue du risque de la décision, est-il préférable de conserver (H_0) ?

p-value

- C'est le plus petit niveau qui fait rejeter (H_0) au vu des données
- C'est une variable aléatoire

$$\text{p-value} = \alpha(\omega) = \inf\{\alpha \in [0,1];\, T(\omega) \in \mathcal{R}_\alpha\}$$

Si $\{T_n > q_\alpha\}$ est de niveau α

$$\text{p-value} = \inf\{\alpha \in [0,1];\, T_n(\omega) > q_\alpha\} = \mathbb{P}_{\theta_0}\{T_n > T_n(\omega)\}.$$

Dans un test de niveau α, (H_0) est rejetée si $\alpha >$ p-value, conservée si $\alpha <$ p-value :

- si $0.05 >$ p-value > 0.01, le test est significatif,
- si $0.01 >$ p-value > 0.001, le test est très significatif,
- si $0.001 >$ p-value, le test est hautement significatif.

A.N.

- Hypothèse nulle (H_0) : $\mu = \mu_0 = 15$
- Hypothèse alternative (H_1) : $\mu = \mu_1 > \mu_0$
- Risque $\alpha = 0.1$
- Données observées : $\bar{x} = 16$,
  $s^2 = \sum_{i=1}^{6}(x_i - \bar{x})^2/5 = 1.6$
- statistique de test observée : $t_{obs} = 1.88$
- quantiles : $q_{0.90}^{\mathcal{T}(5)} = 1.48$; $q_{0.95}^{\mathcal{T}(5)} = 2.02$
- décision ? risque ?

p-value $= 0.059$

L'approche test : région de rejet bilatérale

1. Même modèle que prédémment
2. $(H_0) : \mu = \mu_0$ contre $(H_1) : \mu \neq \mu_0$
3. statistique de test :

$$T = \sqrt{n}\frac{\bar{X} - \mu_0}{\widehat{\sigma}} \sim_{(H_0)} \mathcal{T}(n-1)$$

4. Région de rejet

$$\alpha = \mathbb{P}_{(H_0)}\Big(\underbrace{|T| > q_{1-\alpha/2}^{\mathcal{T}(n-1)}}_{\mathcal{R}\text{: Région de rejet}} \Big)$$

Estimation par intervalle de confiance

L'approche Intervalle de confiance

Sous les mêmes hypothèses que précédemment, on peut aussi écrire, avec $q = q_{1-\alpha/2}^{\mathcal{T}(n-1)}$

$$1-\alpha = \mathbb{P}_{(H_0)}\Big(\bar{X} - q\frac{\widehat{\sigma}}{\sqrt{n}} < \mu < \bar{X} + q\frac{\widehat{\sigma}}{\sqrt{n}}\Big) = \mathbb{P}_{(H_0)}(IC \ni \mu)$$

Définition

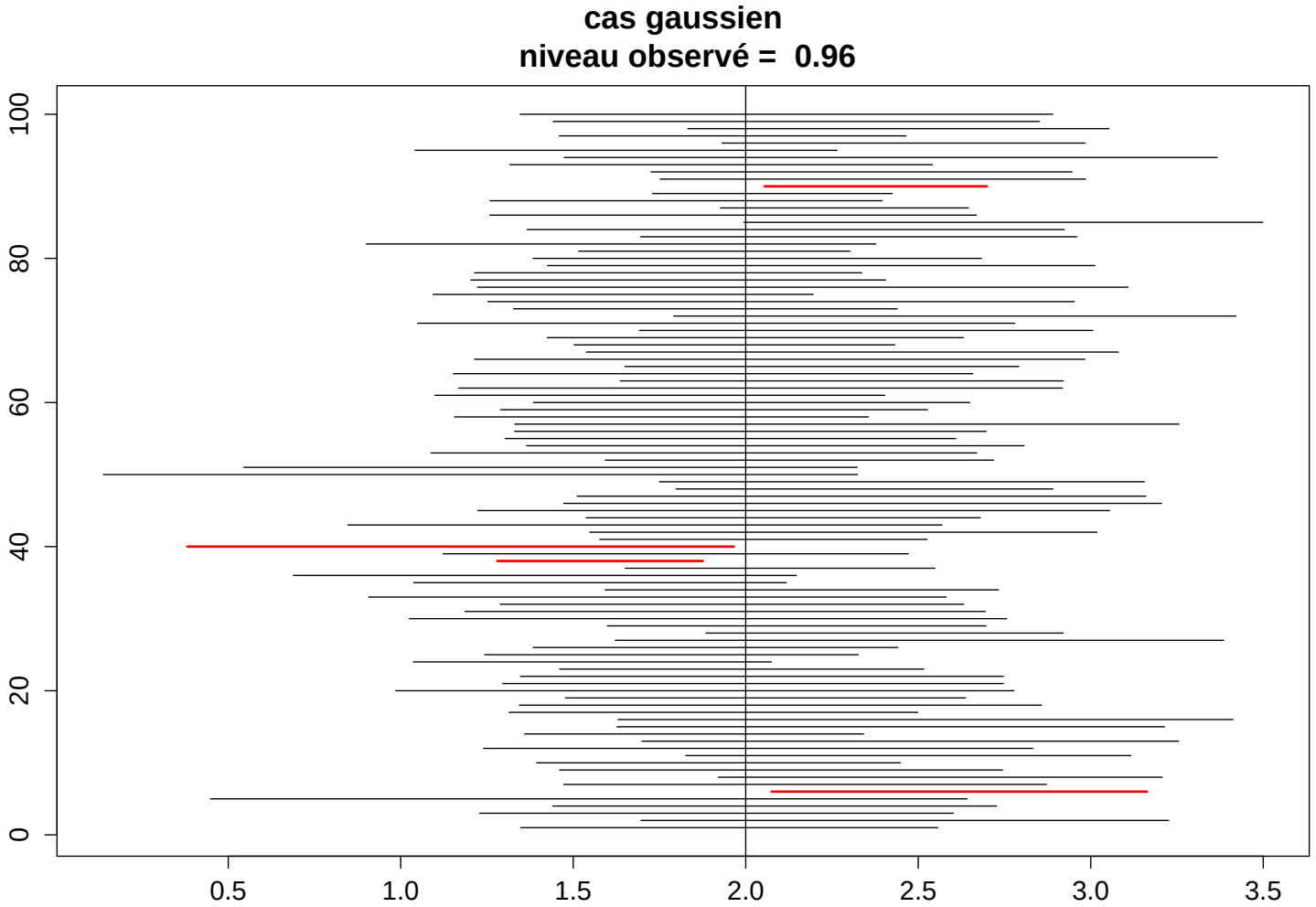
Soit $X = (X_1, \ldots, X_n)$ un n-échantillon de loi $\mathbb{P}_\theta$. Un intervalle de confiance de niveau $1-\alpha$ pour $\theta \in \mathbb{R}$ est un intervalle $IC = [\widehat{\theta}_{inf}(X), \widehat{\theta}_{sup}(X)]$ dont les bornes sont aléatoires, telles que, pour tout $\theta \in \Theta$

$$P_\theta(IC \ni \theta) \geq 1-\alpha.$$

où α est "petit".
Une réalisation $[\widehat{\theta}_{inf}(x), \widehat{\theta}_{sup}(x)]$ est obtenue à partir des données $x = (x_1, \ldots, x_n)$.

L'approche Intervalle de confiance (suite)

Fournir un intervalle (fourchette) permet de prendre en compte la fluctuation d'échantillonnage plutôt que de donner une valeur ponctuelle $\widehat{\theta}$



ICs et tests paramétriques classiques

- Test de Student de l'espérance d'une loi (ou test de la moyenne)
- Test de la variance d'une loi gaussienne
- Test de comparaison des espérances de deux lois gaussiennes
  - avec même variance
  - avec variance différents
  - cas des échantillons appariés
- Test de comparaison des variances de deux lois gaussiennes
- Test sur le paramètre de proportion d'une loi de Bernoulli
- Test de comparaison des paramètres de deux lois de Bernoulli

Remarque : tests non paramétriques : adéquation à une loi, indépendance