# INTERVALLE DE CONFIANCE DE LA MÉDIANE

Estimateur $\widetilde{X}$ de la médiane $m$

Objectif : encadrer $m$, c'est-à-dire :
proposer un intervalle
et savoir quelle est la probabilité
que cet intervalle contienne $m$

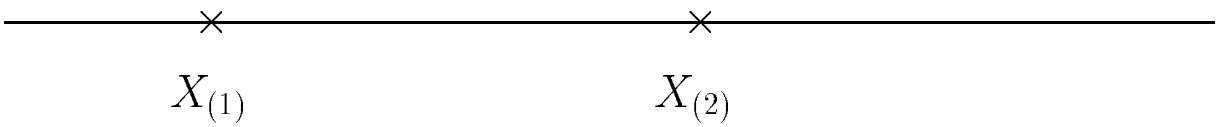
Quel intervalle ? Quelle probabilité ?

Étapes : 1) 2-échantillon

2) $n$-échantillon

**2-échantillon** : $X_1, X_2$

$$X_{(1)} = \min(X_1, X_2) \qquad X_{(2)} = \max(X_1, X_2)$$




On propose l'intervalle $X_{(1)} \leq m \leq X_{(2)}$.

Probabilité que cet intervalle contienne $m$ ?

Calcul de $\Pr\{X_{(1)} \leq m \leq X_{(2)}\}$

1) $\Pr\{m > X_1\} = \left(\frac{1}{2}\right)^2$

2) $\Pr\{m < X_2\} = \left(\frac{1}{2}\right)^2$

3) $\Pr\{X_1 \leq m \leq X_2\} = 1 - \Pr\{m < X_{(1)}\} - \Pr\{m > X_{(2)}\}$

$$= 1 - \tfrac{1}{4} - \tfrac{1}{4} = \tfrac{1}{2}$$

Conclusion : $[X_{(1)}\ ;\ X_{(2)}]$ contient $m$ une fois sur deux.

$n$**-échantillon** : $X_1, X_2, \ldots, X_n$

On propose $[X_{(1)}\ ;\ X_{(n)}]$.

Calcul de la probabilité que $m$ contienne cet intervalle :

$$\gamma = \Pr\{X_{(1)} \leq m \leq X_{(2)}\}$$

$$\gamma = 1 - \frac{1}{2^{n-1}} \qquad \text{erreur } = \frac{1}{2^{n-1}}$$

$$\gamma = \text{coefficient de confiance}$$

$$[X_{(1)}\ ;\ X^{(1)}],\ \textbf{ou}\ [X_{(2)}\ ;\ X^{(2)}],\ldots,\ \textbf{ou}\ [X_{(c)}\ ;\ X^{(c)}]\ \textbf{?}$$

À chaque intervalle est associé un coefficient de confiance $\gamma$.

On choisit $\gamma$ *a priori*, et on détermine un intervalle :

de coefficient de confiance au moins égal à $\gamma$,

à l'aide d'une **table**.

$$n = 9 \quad \gamma \geq 0,95 \longrightarrow [X_{(\ )}\ ;\ X^{(\ )}]$$

$$\gamma \geq 0,99 \longrightarrow [X_{(\ )}\ ;\ X^{(\ )}]$$

## Simulation de 1 000 tirages de 9-échantillons

Pour chaque tirage calcul de $[X_{(2)} \ ; \ X^{(2)}]$.

$\gamma = 0,960\,9$

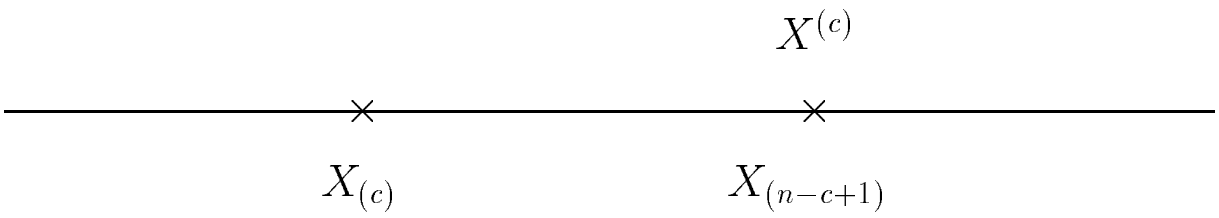
$m$ connue = 41,2

| Nombre de cas où $m \in [X_{(2)} \ ; \ X^{(2)}]$ | nombre de cas où $m \notin [X_{(2)} \ ; \ X^{(2)}]$ |
|---|---|
| 964 | 36 |

Conclusion : on s'est trompé dans un certain nombre de cas.

# CALCUL DE LA TABLE

Calcul de $\Pr\{X_{(c)} \leq m \leq X^{(c)}\}$.




$Y =$ nombre de valeurs plus petites que la médiane $m$

$$\{X_{(c)} \leq m \leq X^{(c)}\} \iff \{X_{(c)} \leq m \leq X_{(n-c+1)}\}$$

$$X_{(c)} \leq m \iff Y \geq c$$

$$m \leq X_{(n-c+1)} \iff Y \leq n - c + 1$$

Résultat : $\Pr\{X_{(c)} \leq m \leq X^{(c)}\} = \Pr\{c \leq Y \leq n - c + 1\}$

Calculable si l'on connaît la loi de $Y$.

| $Y$ = nombre de valeurs plus petites que $m$ | | $\mathcal{B}(n,p)$ |
|---|---|---|
| épreuve : | | épreuve : |
| à chaque tirage 2 issues | ↗ + petit<br>↘ + grand ou égal | 2 issues ↗ 1<br>↘ 0 |
| probabilité de {+ petit} | $= \frac{1}{2}$ | $\Pr\{1\} = p$ |
| $\Pr\{$+ grand ou égal$\}$ | $= 1 - \frac{1}{2} = \frac{1}{2}$ | $\Pr\{0\} = q = 1 - p$ |
| $n$-échantillon | | $n$ réalisations |
| variable aléatoire $Y =$ | nombre de réalisations de {+ petit} | variable aléatoire $Y$ = nombre de réalisations de {1} |
| $Y \sim \mathcal{B}(n, 1/2)$ | | $Y \sim \mathcal{B}(n,p)$ |

Donc on sait calculer

$Pr\{c \leq Y \leq n - c + 1\}$.

Erreur = probabilité que $[X_{(c)}\ ;\ X^{(c)}]$

ne contienne pas $m$

est notée $\alpha$

$$\boxed{\alpha = 1 - \gamma}$$

Symétrie de la distribution binomiale

$$\Longrightarrow \Pr\{m < X_{(c)}\} = \Pr\{m > X^{(c)}\}$$

notée $\alpha'$

$$\alpha = \Pr\{m < X_{(c)}\} + \Pr\{m > X^{(c)}\} = 2\alpha'$$

$$\boxed{\alpha = 2\alpha'}$$

La *loi* de $Y$ est *connue* :

c'est une loi binomiale,

calculable, tabulée.

$\Longrightarrow$ $Y$ est une variable *pivot*.

$Y$ ne peut pas être *calculée* à partir du $n$-échantillon

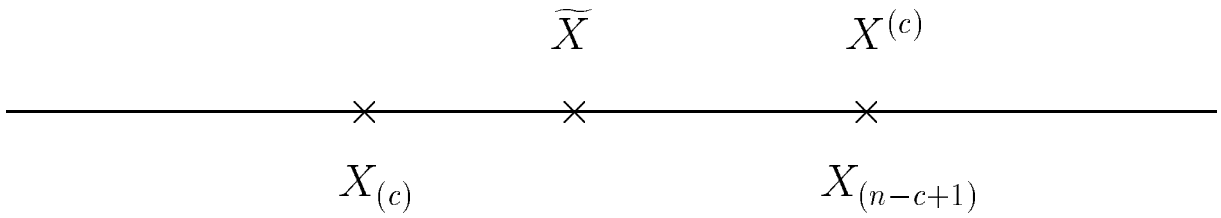
(on ne connaît pas le nombre de valeurs plus
petites que $m$, car on ne connaît pas $m$)

$\Longrightarrow$ $Y$ n'est pas une *statistique*.

# DÉFINITION D'UN INTERVALLE DE CONFIANCE DE LA MÉDIANE

Médiane de l'échantillon $\widetilde{X}$ : estimateur de $m$.

Valeurs de l'échantillon réordonnées :




On supprime $(c-1)$ valeurs à chaque extrémité :

$$\Longrightarrow \text{intervalle } [X_{(c)} \ ; \ X^{(c)}]$$

$$\text{niveau de confiance : } \gamma$$

$$\text{erreur : } \alpha = 2\alpha' = 1 - \gamma$$

| $n$ | min à max | $X_{(2)}$ à $X^{(2)}$ | $X_{(3)}$ à $X^{(3)}$ | $X_{(4)}$ à $X^{(4)}$ | $X_{(5)}$ à $X^{(5)}$ |
|---|---|---|---|---|---|
| 2 | 0,500 0 | | | | |
| 3 | 0,750 0 | | | | |
| 4 | 0,875 0 | 0,375 0 | | | |
| 5 | 0,937 5 | 0,625 0 | | | |
| 6 | 0,968 8 | 0,781 3 | 0,312 5 | | |
| 7 | 0,984 4 | 0,875 0 | 0,546 8 | | |
| 8 | 0,992 2 | 0,929 7 | 0,711 0 | 0,273 4 | |
| 9 | 0,996 1 | 0,960 9 | 0,820 3 | 0,492 2 | |
| 10 | 0,998 0 | 0,978 5 | 0,890 6 | 0,656 3 | 0,246 1 |
| 11 | 0,999 0 | 0,988 3 | 0,934 6 | 0,773 4 | 0,451 2 |
| 12 | 0,999 5 | 0,993 7 | 0,961 4 | 0,854 0 | 0,612 3 |
| 13 | 0,999 76 | 0,996 6 | 0,977 5 | 0,907 7 | 0,733 2 |
| 14 | 0,999 88 | 0,998 2 | 0,987 1 | 0,942 6 | 0,820 4 |
| 15 | 0,999 939 | 0,999 02 | 0,992 1 | 0,964 8 | 0,881 5 |
| 16 | 0,999 969 | 0,999 48 | 0,995 8 | 0,978 7 | 0,923 2 |
| 17 | 0,999 985 | 0,999 73 | 0,997 7 | 0,987 3 | 0,951 0 |
| 18 | 0,999 992 3 | 0,999 86 | 0,998 7 | 0,992 5 | 0,969 1 |
| 19 | 0,999 996 2 | 0,999 924 | 0,999 27 | 0,995 6 | 0,980 8 |
| 20 | 0,999 998 1 | 0,999 960 | 0,999 60 | 0,997 4 | 0,988 2 |
| 25 | 0,999 999 934 | 0,999 998 3 | 0,999 978 6 | 0,999 83 | 0,999 0 |
| $c =$ | 1 | 2 | 3 | 4 | 5 |

$c$ dépend de $\gamma$, et de $n$.

# GRANDS ÉCHANTILLONS

Si $n$ est grand :

$$\mathcal{B}(n,p) \approx \mathcal{N}(np, npq)$$

donc

$$\mathcal{B}(n,1/2) \approx \mathcal{N}(n/2, n/4)$$

$$c = \frac{n+1}{2} - z\frac{\sqrt{n}}{2}$$

($z$ écart gaussien)

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| 20 | 23,5 | 24 | 25 | 32,5 | 33 | 35 | 38 |
| 39 | 40 | 41 | 46,5 | 48 | 49 | 50 | 51 |
| 51,5 | 52 | 59 | 62 | 63,5 | 68 | 68,5 | 69 |
| 79 | 79,5 | 80 | 87 | 90 | 97,5 | 99 | 100 |
| 102 | 102,5 | 104 | 105 | 175 | 250 | | |

38-échantillon : revenus des familles en kF.

Intervalle unilatéral : $]-\infty\ ;\ X^{(c')}]$

On choisit $\gamma \geq 0,90$.

Trouver $c'$ (table 3) :

$$\begin{aligned} 1-\gamma &= \\ \alpha' &= \\ c' &= \\ \gamma &= \end{aligned}$$

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| 20 | 23,5 | 24 | 25 | 32,5 | 33 | 35 | 38 |
| 39 | 40 | 41 | 46,5 | 48 | 49 | 50 | 51 |
| 51,5 | 52 | 59 | 62 | 63,5 | 68 | 68,5 | 69 |
| 79 | 79,5 | 80 | 87 | 90 | 97,5 | 99 | 100 |
| 102 | 102,5 | 104 | 105 | 175 | 250 | | |

38-échantillon : revenus des familles en kF.

Intervalle unilatéral : $c' = 15$ $\gamma' = 0.928$

Intervalle bilatéral : $c =$ $\gamma =$

$$\Pr\{\,]-\infty\ ;\ X^{(c')}]\ni m\}\approx\Pr\{\,[X_{(c)}\ ;\ X^{(c)}]\ni m\}$$

Les intervalles observés sur le 38-échantillon sont :

$]-\infty\ ;\ X^{(15)}]=]-\infty\ ;\ 69]$ et $[X_{(\ )}\ ;\ X^{(\ )}]=[49\ ;\ 79]$

## DISTRIBUTION ASYMÉTRIQUE

## DISTRIBUTION SYMÉTRIQUE

la moyenne et la médiane sont confondues

$\Downarrow$

construction d'un intervalle symétrique
autour du paramètre considéré, fonction de :

– la taille de l'échantillon,

– la dispersion des données de l'échantillon,

– le coefficient de confiance $\gamma$ choisi.

# INTERVALLE DE CONFIANCE DE LA MÉDIANE SOUS L'HYPOTHESE D'UNE DISTRIBUTION SYMÉTRIQUE

- Soient $n$ variables indépendantes $X_1, X_2, \ldots, X_n$
- Hypothèse : la distribution des variables $X_i$ est symétrique.
- Construction des variables $Z_{ij} = \frac{X_i + X_j}{2}$
- Tableau des $Z_{ij}$ :

| | $X_1$ | $X_2$ | $X_3$ | $\ldots$ | $X_n$ |
|---|---|---|---|---|---|
| $X_1$ | $X_1$ | | | | |
| $X_2$ | $\frac{X_1+X_2}{2}$ | $X_2$ | | | |
| $X_3$ | $\frac{X_1+X_3}{2}$ | $\frac{X_2+X_3}{2}$ | $X_3$ | | |
| $\vdots$ | $\vdots$ | $\vdots$ | $\vdots$ | $\ddots$ | |
| $X_n$ | $\frac{X_1+X_n}{2}$ | $\frac{X_2+X_n}{2}$ | $\frac{X_3+X_n}{2}$ | $\ldots$ | $X_n$ |

**Recherche graphique** de la médiane des $Z_{ij}$, pour une population entière complètement symétrique de 7 individus.

• Propriété :

Si les $X_i$ sont indépendantes et à distribution symétrique, alors la loi des $Z_{ij}$ a la même médiane que celle des $X_i$.

**Application à un 9-échantillon**

tiré d'une population dont la distribution est symétrique.

## Construction d'un intervalle de confiance $[Z_{(c)} \; ; \; Z^{(c)}]$

Cet intervalle est fonction :

- de l'effectif du $n$-échantillon, et
- du coefficient de confiance $\gamma$.

Il tient compte de la dispersion effective des données.

• Remarque : les variables $Z_{ij}$ ne sont pas indépendantes !!!

$$\downarrow$$

$Y_Z$ = nombre de $Z_{ij} \leq m$ n'est pas une variable binomiale.

$$\downarrow$$

Utilisation d'une table spécifique
différente de la précédente
pour déterminer la valeur de $c$.

# INTERVALLE DE CONFIANCE DE LA MOYENNE SOUS L'HYPOTHÈSE D'UNE DISTRIBUTION GAUSSIENNE MÉTHODE DE STUDENT

• On construit un intervalle symétrique autour de la moyenne empirique, $\overline{X}$, de l'échantillon :

$$\overline{X} - \text{marge} \leq \mu \leq \overline{X} + \text{marge}$$

• La *marge* sera calculée en tenant compte à nouveau :

- de la taille de l'échantillon $n$,

- de la dispersion des données de l'échantillon $\widehat{\sigma}^2$,

- et du coefficient de confiance choisi $\gamma$.

- Cas des **grands échantillons** :

$$\Pr\left\{\overline{X} - z\frac{\widehat{\sigma}}{\sqrt{n}} \leq \mu \leq \overline{X} + z\frac{\widehat{\sigma}}{\sqrt{n}}\right\} = 1 - \alpha$$

- Cas des **petits échantillons** :

$$X_1, X_2, \ldots, X_n \sim \mathcal{N}(\mu, \sigma^2)$$

$$\overline{X} = \frac{1}{n}\sum_{i=1}^{n} X_i \sim \mathcal{N}(\mu, \frac{\sigma^2}{n})$$

$$\frac{\overline{X} - \mu}{\sigma/\sqrt{n}} \sim \mathcal{N}(0, 1)$$

• Cas des **petits échantillons** (suite)

Soit $z$ l'écart gaussien associé au risque $\alpha$ :

$$\Pr\left\{-z \leq \frac{\overline{X}-\mu}{\sigma/\sqrt{n}} \leq z\right\} = 1-\alpha$$

$$\Pr\left\{\overline{X} - z\frac{\sigma}{\sqrt{n}} \leq \mu \leq \overline{X} + z\frac{\sigma}{\sqrt{n}}\right\} = 1-\alpha$$

Le paramètre $\sigma^2$ n'est pas connu *a priori*, on l'estime à partir des données expérimentales.

La variable $T = \frac{\overline{X}-\mu}{\widehat{\sigma}/\sqrt{n}}$ ne suit pas une loi normale,
mais une loi de Student à $(n-1)$ degrés de liberté : $t_{n-1}$.

On remplace donc l'écart gaussien $z$ par l'écart de Student $t$ :

$$\Pr\left\{\overline{X} - t\frac{\widehat{\sigma}}{\sqrt{n}} \leq \mu \leq \overline{X} + t\frac{\widehat{\sigma}}{\sqrt{n}}\right\} = 1-\alpha$$

## BILAN SUR LES INTERVALLES DE CONFIANCE

- IDENTIFIER :

  - une population,

  - une variable d'étude,

  - un paramètre de la loi de cette variable.

- PRÉCISER :

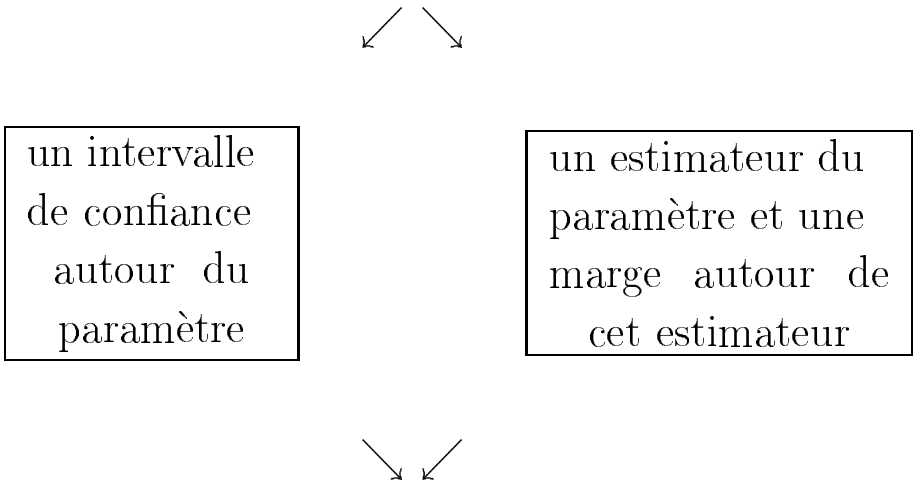
  - la distribution *a priori* de la variable.

- DÉFINIR :

  - un niveau de confiance $\gamma$.

## BILAN (suite) SUR LES INTERVALLES DE CONFIANCE

- Calculer, pour un échantillon de taille $n$ :




- Remarque :

Les intervalles prennent en compte la distribution des $X_i$.