

INTERVALLE DE CONFIANCE

On a un estimateur d'un paramètre

Objectif : encadrer la vraie valeur du paramètre, c'est-à-dire : proposer un intervalle et savoir quelle est la probabilité que cet intervalle contienne la vraie valeur

Quel intervalle ? Quelle probabilité ?

Etapas : - méthode pour déterminer un IC autour de l'estimateur du paramètre recherché

- introduire différents postulats sur la distribution des X_i pour réduire l'IC

POSTULATS SUR LA DISTRIBUTION DES X_i POUR DÉTERMINER L'INTERVALLE DE CONFIANCE

Par définition, ces postulats doivent être faits *a priori*

1) Aucun postulat

→ méthode du signe

2) Postulat 1 : distribution symétrique

méthode de Wilcoxon

3) Postulat 2 : distribution gaussienne

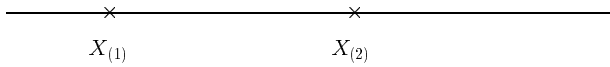
(postulat plus fort)

méthode de Student

Pour les 3 méthodes, on suppose que la distribution est continue.

2-échantillon : X_1, X_2

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



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

Probabilité que cet intervalle contienne m ?

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

$$\begin{aligned} 1) \Pr\{m < X_{(1)}\} &= \Pr\{m < X_1 \text{ et } m < X_2\} \\ &= \Pr\{m < X_1\} \times \Pr\{m < X_2\} \\ &= \frac{1}{2} \times \frac{1}{2} = \frac{1}{4} \end{aligned}$$

$$2) \Pr\{m > X_{(2)}\} = \frac{1}{4}$$

$$\begin{aligned} 3) \Pr\{X_{(1)} \leq m \leq X_{(2)}\} &= 1 - \Pr\{m < X_{(1)}\} - \Pr\{m > X_{(2)}\} \\ &= 1 - \frac{1}{4} - \frac{1}{4} = \frac{1}{2} \end{aligned}$$

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

n -échantillon : X_1, X_2, \dots, X_n

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

Calcul de la probabilité que cet intervalle contienne m :

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

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

$\gamma =$ coefficient de confiance

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

À chaque intervalle est associé un coefficient de confiance γ .

On choisit γ *a priori*, et on détermine un intervalle :

de coefficient de confiance au moins égal à γ ,

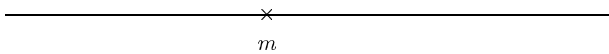
à l'aide d'une **table**.

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

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

CALCUL DE LA TABLE

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

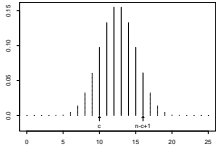


Y = nombre d'observations plus petites que la médiane m

- $\{X_{(c)} \leq m \leq X^{(c)}\} \iff ?$
- si aucun X_i ne coïncide avec m ,
 $X_{(c)} \leq m \iff Y \geq c$
- $m \leq X^{(c)} \iff Y \leq n - c$

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

Calculable si l'on connaît la loi de Y .

$Y =$ nombre de valeurs plus petites que m	$\mathcal{B}(n, p)$
épreuve : à chaque tirage 2 issues <div style="display: flex; align-items: center; margin-left: 40px;"> ↗ + petit ↘ + grand ou égal </div>	épreuve : 2 issues <div style="display: flex; align-items: center; margin-left: 40px;"> ↗ 1 ↘ 0 </div>
probabilité de {+ petit} = $\frac{1}{2}$ Pr{+ grand ou égal} = $1 - \frac{1}{2} = \frac{1}{2}$	Pr{1} = p 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 $Pr\{c \leq Y \leq n - c\}$. $= \sum_{i=c}^{n-c} C_n^i \left(\frac{1}{2}\right)^i \left(\frac{1}{2}\right)^{n-i}$	

Erreur = probabilité que $[X_{(c)} ; X^{(c)}]$ ne contienne pas m est notée α <div style="border: 1px solid black; padding: 5px; width: fit-content; margin: 10px auto;"> $\alpha = 1 - \gamma$ </div> Symétrie de la distribution binomiale $\implies Pr\{m < X_{(c)}\} = Pr\{m > X^{(c)}\}$ notée α' $\alpha = Pr\{m < X_{(c)}\} + Pr\{m > X^{(c)}\} = 2\alpha'$ <div style="border: 1px solid black; padding: 5px; width: fit-content; margin: 10px auto;"> $\alpha = 2\alpha'$ </div>
--

Y dépend du paramètre m

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)

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

La *loi* de Y est *connue* :

c'est une loi binomiale,

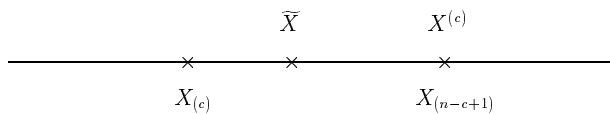
calculable, tabulée.

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

DÉFINITION D'UN INTERVALLE DE CONFIANCE DE LA MÉDIANE MÉTHODE DU SIGNE

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

Valeurs de l'échantillon réordonnées :



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

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

niveau de confiance : γ

erreur : $\alpha = 2\alpha' = 1 - \gamma$

IC de la médiane avec les statistiques d'ordre $X_{(c)}$ et $X^{(c)}$

$c =$	1	2	3	4	5
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 dépend de γ , et de n .

GRANDS ÉCHANTILLONS

Si n est grand :

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

donc :

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

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

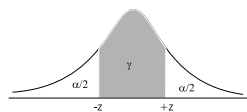
avec z tel que :

$$Pr(Z \geq z) = \alpha' = \frac{\alpha}{2}$$

$$Pr(Z < z) = 1 - \frac{\alpha}{2}$$

Z suit $\mathcal{N}(0, 1)$

à rechercher dans une table de la loi normale



**DE L'IC BILATÉRAL À L'IC UNILATÉRAL
IC BILATÉRAL**

Production laitière de 50 bêtes (litre de lait/an)

1: 2161	11: 3912	21: 4177	31: 4590	41: 5073
2: 2624	12: 3921	22: 4192	32: 4601	42: 5091
3: 2665	13: 3955	23: 4202	33: 4727	43: 5183
4: 2821	14: 3979	24: 4204	34: 4772	44: 5344
5: 2874	15: 4031	25: 4230	35: 4780	45: 5371
6: 3381	16: 4078	26: 4377	36: 4783	46: 5665
7: 3463	17: 4092	27: 4441	37: 4862	47: 5672
8: 3643	18: 4101	28: 4494	38: 4896	48: 5682
9: 3738	19: 4155	29: 4521	39: 4927	49: 5823
10: 3818	20: 4159	30: 4551	40: 4981	50: 5848

On veut estimer (par intervalle) la production laitière médiane.

$$\Pr\{[X_{(c)}, X^{(c)}] \ni m\} = \gamma \simeq 0,95$$

- $\gamma =$
- $\alpha =$
- $c =$

$$IC = [X_{()}, X^{()}] = [\quad , \quad]$$

**DE L'IC BILATÉRAL À L'IC UNILATÉRAL
IC UNILATÉRAL**

50-échantillon : production laitière (litre de lait/an) :

1: 2161	11: 3912	21: 4177	31: 4590	41: 5073
2: 2624	12: 3921	22: 4192	32: 4601	42: 5091
3: 2665	13: 3955	23: 4202	33: 4727	43: 5183
4: 2821	14: 3979	24: 4204	34: 4772	44: 5344
5: 2874	15: 4031	25: 4230	35: 4780	45: 5371
6: 3381	16: 4078	26: 4377	36: 4783	46: 5665
7: 3463	17: 4092	27: 4441	37: 4862	47: 5672
8: 3643	18: 4101	28: 4494	38: 4896	48: 5682
9: 3738	19: 4155	29: 4521	39: 4927	49: 5823
10: 3818	20: 4159	30: 4551	40: 4981	50: 5848

Est-ce que la production médiane est supérieure à la médiane régionale $R = 4190$ litres/an ?

Intervalle bilatéral : $c = 18 \quad \gamma' = 0.967$
 Intervalle unilatéral : $c' = \quad \gamma' =$

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

Les intervalles calculés sur le 50-échantillon sont :

$$[X_{()}; +\infty[= [\quad ; +\infty[\quad [X_{(18)}; X^{(18)}] = [4101; 4727]$$

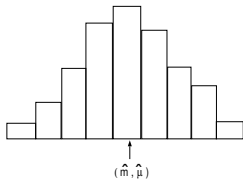
POSTULATS SUR LA DISTRIBUTION DES X_i POUR DÉTERMINER L'INTERVALLE DE CONFIANCE

Par définition, ces postulats doivent être faits *a priori*

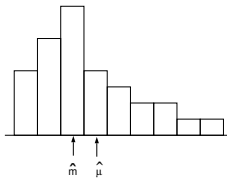
- 1) Aucun postulat
méthode du signe
- 2) Postulat 1 : distribution symétrique
→ méthode de Wilcoxon
- 3) Postulat 2 : distribution gaussienne
(postulat plus fort)
méthode de Student

Pour les 3 méthodes, on suppose que la distribution est continue.

DISTRIBUTION SYMÉTRIQUE



DISTRIBUTION ASYMÉTRIQUE



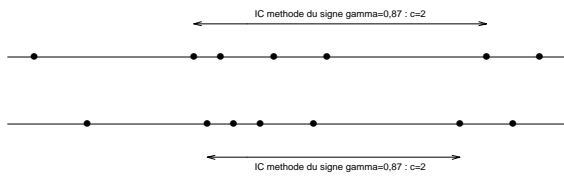
la moyenne et la médiane sont confondues

⇓

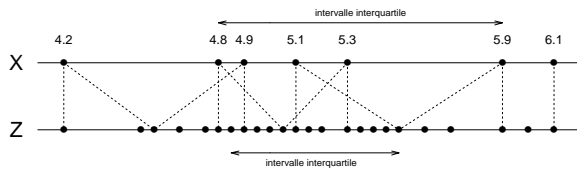
construction d'un intervalle symétrique
autour du paramètre considéré, fonction de :

- la taille de l'échantillon,
- la dispersion des données de l'échantillon,
- le coefficient de confiance γ choisi.

Variation de l'IC de la médiane en fonction de la dispersion des observations



Les moyennes 2 à 2 sont moins dispersées autour de m .



IC de la médiane par la méthode de Wilcoxon

- $X_1, \dots, X_n \rightarrow Z_{ij} = \frac{X_i + X_j}{2}, i \leq j$
- distribution des X_i symétrique \Rightarrow les X_i et les Z_{ij} ont même médiane
- IC pour la médiane : $[Z_{(c)}, Z^{(c)}]$
- calcul de c en fonction de γ et n :

Z_{ij} non indépendantes

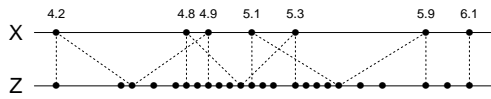
↓

$Y_Z =$ nombre de $Z_{ij} \leq m$ ne suit pas une binomiale

↓

utilisation d'une table spécifique différente de la précédente pour calculer le coefficient de confiance ou déterminer l'IC

calcul de l'IC de Wilcoxon



POSTULATS SUR LA DISTRIBUTION DES X_i POUR DÉTERMINER L'INTERVALLE DE CONFIANCE

Par définition, ces postulats doivent être faits *a priori*

1) Aucun postulat

méthode du signe

2) Postulat 1 : distribution symétrique

méthode de Wilcoxon

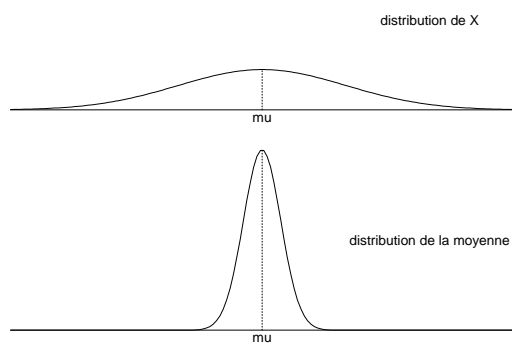
3) Postulat 2 : distribution gaussienne

(postulat plus fort)

→ méthode de Student

Pour les 3 méthodes, on suppose que la distribution est continue.

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



FPSTAT 2 – La décision statistique. 2. Intervalles de confiance. 21

Institut National de la Recherche Agronomique

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, \bar{X} , de l'échantillon :

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

- La *marge* sera calculée en tenant compte à nouveau :
 - de la taille de l'échantillon n ,
 - de la dispersion des données de l'échantillon $\hat{\sigma}^2$,
 - et du coefficient de confiance choisi γ .

FPSTAT 2 – La décision statistique. 2. Intervalles de confiance. 22

Institut National de la Recherche Agronomique

- Soient n variables X_i :

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

la moyenne empirique est :

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

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

- Soit z l'écart gaussien associé au risque α :

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

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

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

- Cas des **petits échantillons** ($n < 20$)

La variable $T = \frac{\bar{X} - \mu}{\hat{\sigma}/\sqrt{n}}$ 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 :

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

- Cas des **grands échantillons** ($n > 20-25$) :

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

BILAN SUR LES INTERVALLES DE CONFIANCE

- IDENTIFIER :

- une population,
- une variable d'étude,
- un paramètre de la loi de cette variable.

- PRÉCISER :

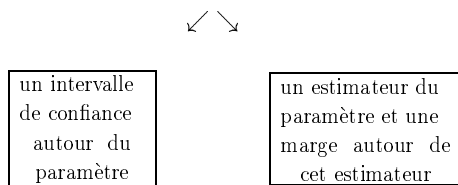
- la distribution *a priori* de la variable.

- DÉFINIR :

- un niveau de confiance γ .

BILAN (suite) SUR LES INTERVALLES DE CONFIANCE

- Calculer, pour un échantillon de taille n :



ces intervalles sont toujours aléatoires !!!

- Remarque :

Les intervalles prennent en compte la distribution des X_i .