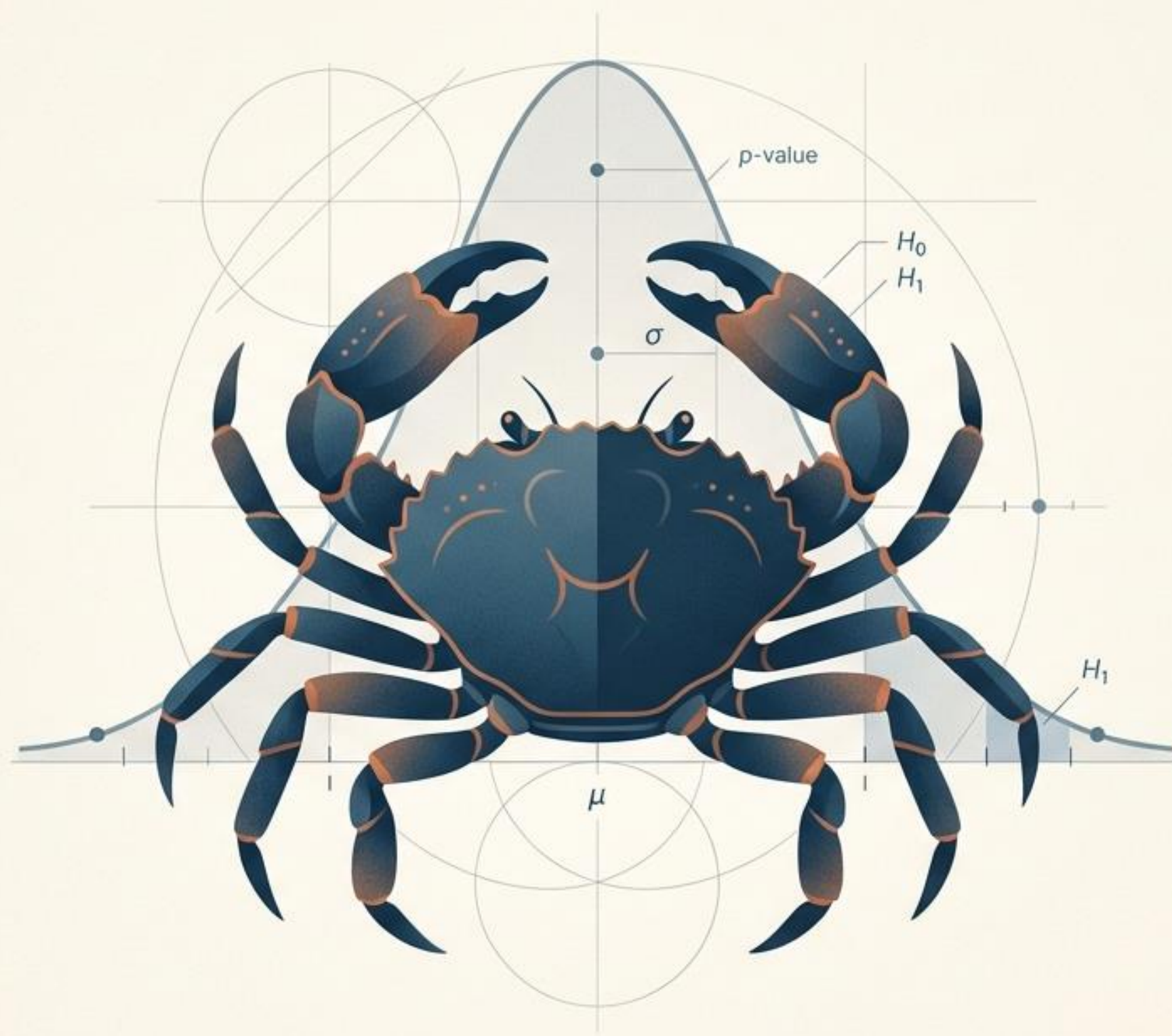


Décoder le Vivant par l'Inférence Statistique

Règles de décision, heuristiques visuelles et applications pratiques des tests de Student et de Wilcoxon en science des données biologiques.



L'Inaccessible Population

Nous ne pouvons jamais mesurer tous les individus. Ses vrais paramètres restent inconnus :

- Moyenne : μ
- Écart type : σ

Le défi de l'inférence :
Faire le pont entre
l'échantillon et la population
grâce aux lois du hasard.



Le Représentant Mesurable

Un échantillon aléatoire et indépendant.
Nos seuls outils tangibles :

- Moyenne calculée : $\bar{x} = \sum \frac{x_i}{n}$
- Écart type estimé : $s_x = \sqrt{\frac{\sum (x_i - \bar{x})^2}{n-1}}$

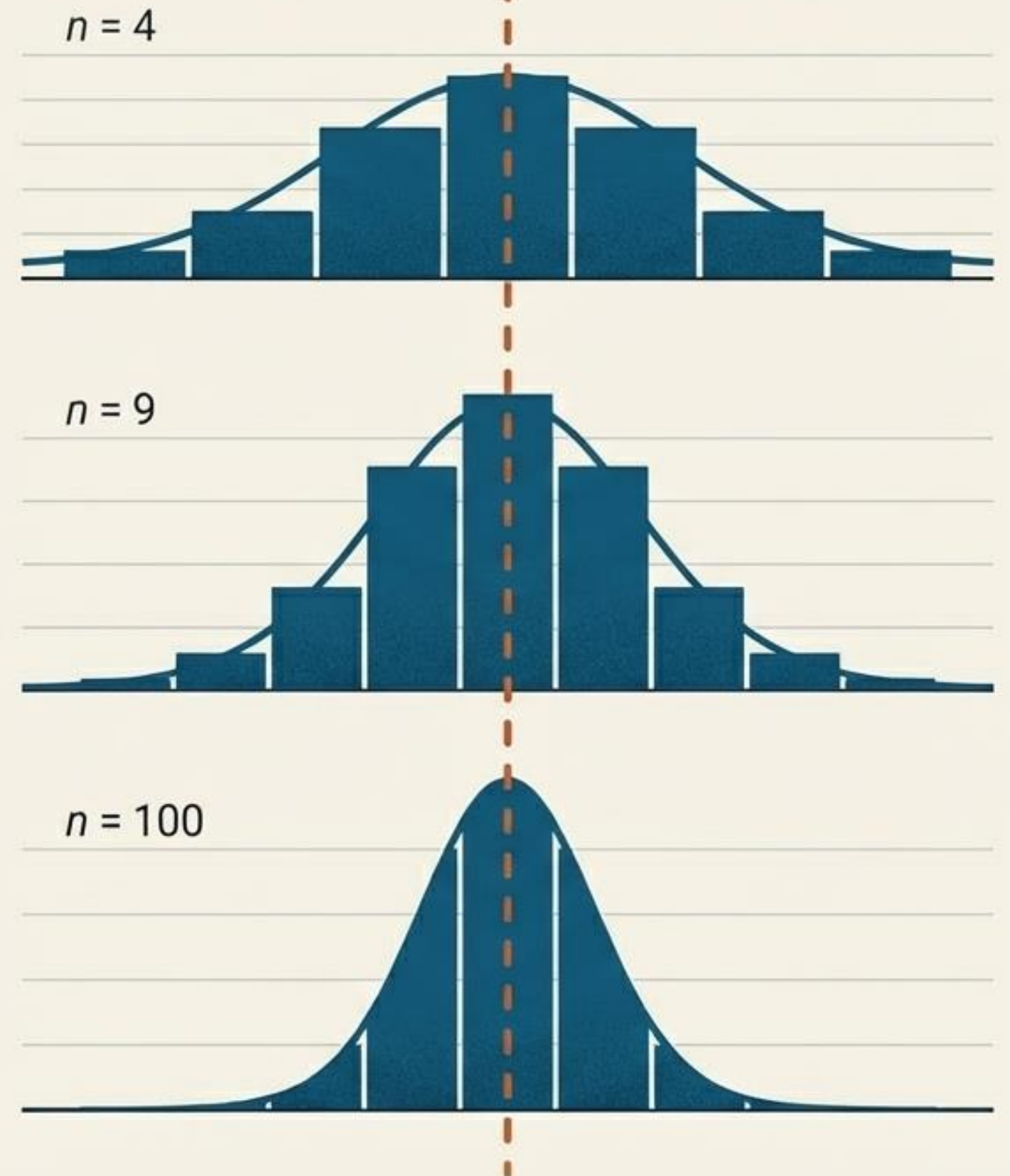
La Magie de la Distribution d'Échantillonnage

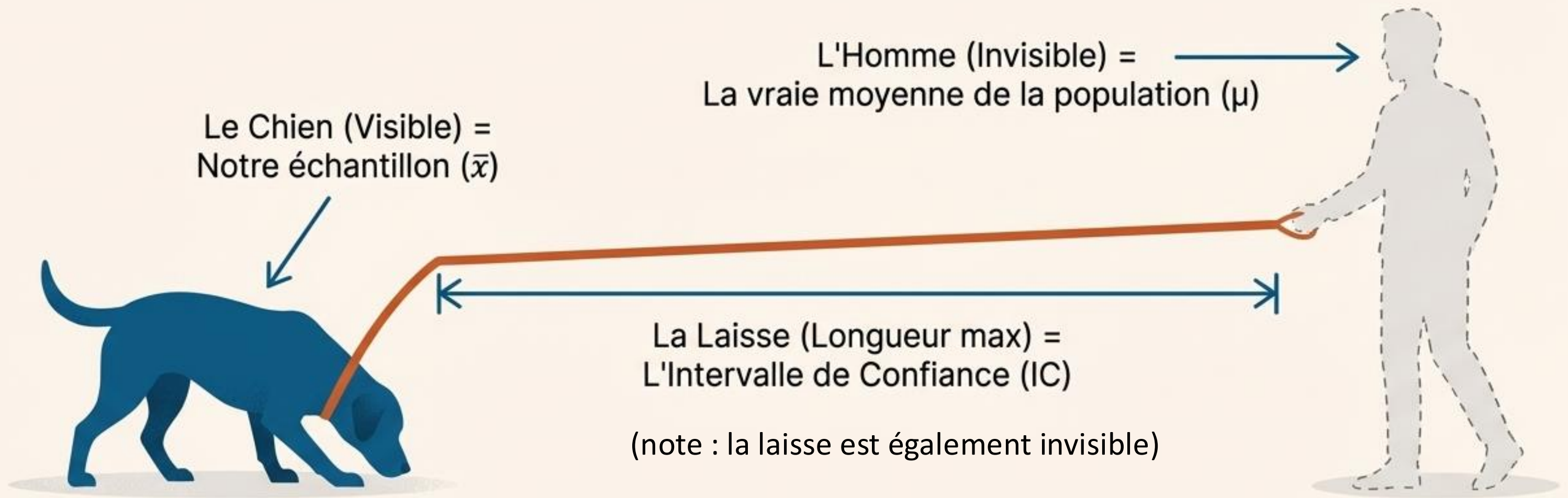
Si on répète un échantillonnage 10 000 fois, la moyenne des échantillons dessine une courbe symétrique. Mais estimer σ avec s_x déforme cette distribution.

La Solution de William Gosset

La distribution t de Student. Plus étalée aux extrémités qu'une loi normale, elle s'ajuste grâce à un paramètre crucial : les degrés de liberté ($ddl = n-1$).

Théorème Central Limite : Même si la population n'est pas normale, la distribution des moyennes tendra vers une normale si n est assez grand.





Calculer la longueur de la laisse

L'Intervalle de Confiance (IC) est défini par l'Erreur Standard ($SE_x = s_x / \sqrt{n}$).

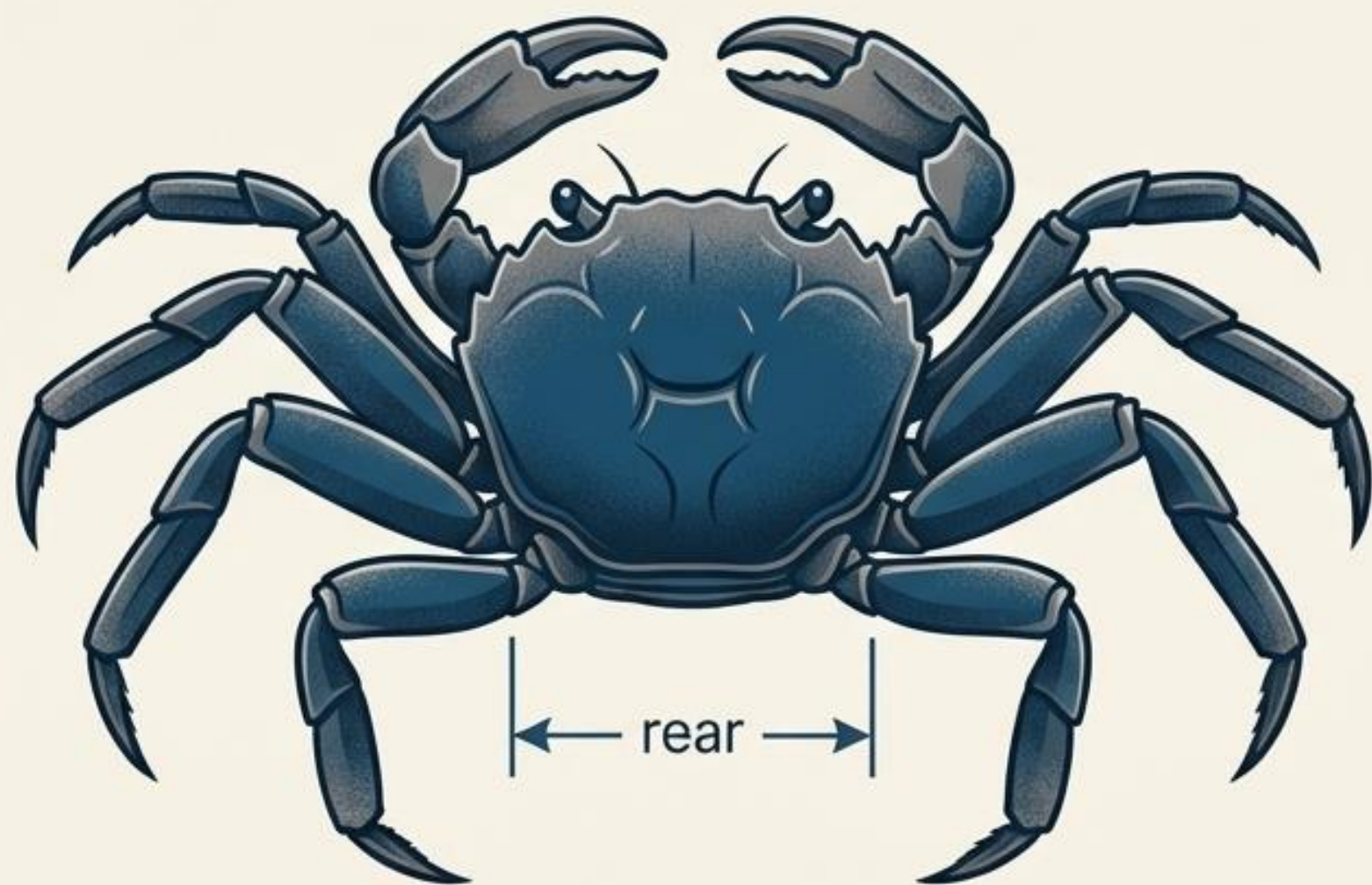
Formule : $IC \simeq \bar{x} \pm t_{\alpha/2}^{n-1} \cdot SE_x$

Le compromis du risque (α)

Fixer α à 5% signifie accepter de se tromper 1 fois sur 20 pour obtenir une précision utilisable.

Le Cas Pratique : Les crabes de Fremantle

Sur un échantillon de 200 crabes *L. variegatus* (100 mâles, 100 femelles), la carapace arrière (rear) des femelles semble plus large. Est-ce significatif ou dû au hasard ?



Le Test t de Student Indépendant (Bilatéral)

On teste l'hypothèse d'une différence de moyenne sans a priori sur la direction :

- H_0 (Hypothèse Nulle) : $\overline{\text{rear}_F} - \overline{\text{rear}_M} = 0$
(Aucune différence)
- H_1 (Alternative) : $\overline{\text{rear}_F} - \overline{\text{rear}_M} \neq 0$
(Différence réelle)

```
t.test(data = crabs, rear ~ sex, alternative = "two.sided",  
       conf.level = 0.95, var.equal = TRUE)
```

```
t = 4.2896, df = 198, p-value = 2.797e-05
```

```
95 percent confidence interval:  
 0.8087907  2.1852093
```

```
mean in group F: 13.487 | mean in group M: 11.990
```

t & df : Valeur observée (4.29) et degrés de liberté (200 - 2 = 198).

p-value : $\ll 0.05$ (α).
La probabilité d'obtenir ce résultat si H_0 était vraie est quasi-nulle. On rejette H_0 .

Intervalle de Confiance : [0.81, 2.19]. Il ne contient pas Zéro ! Confirmation du rejet de H_0 .

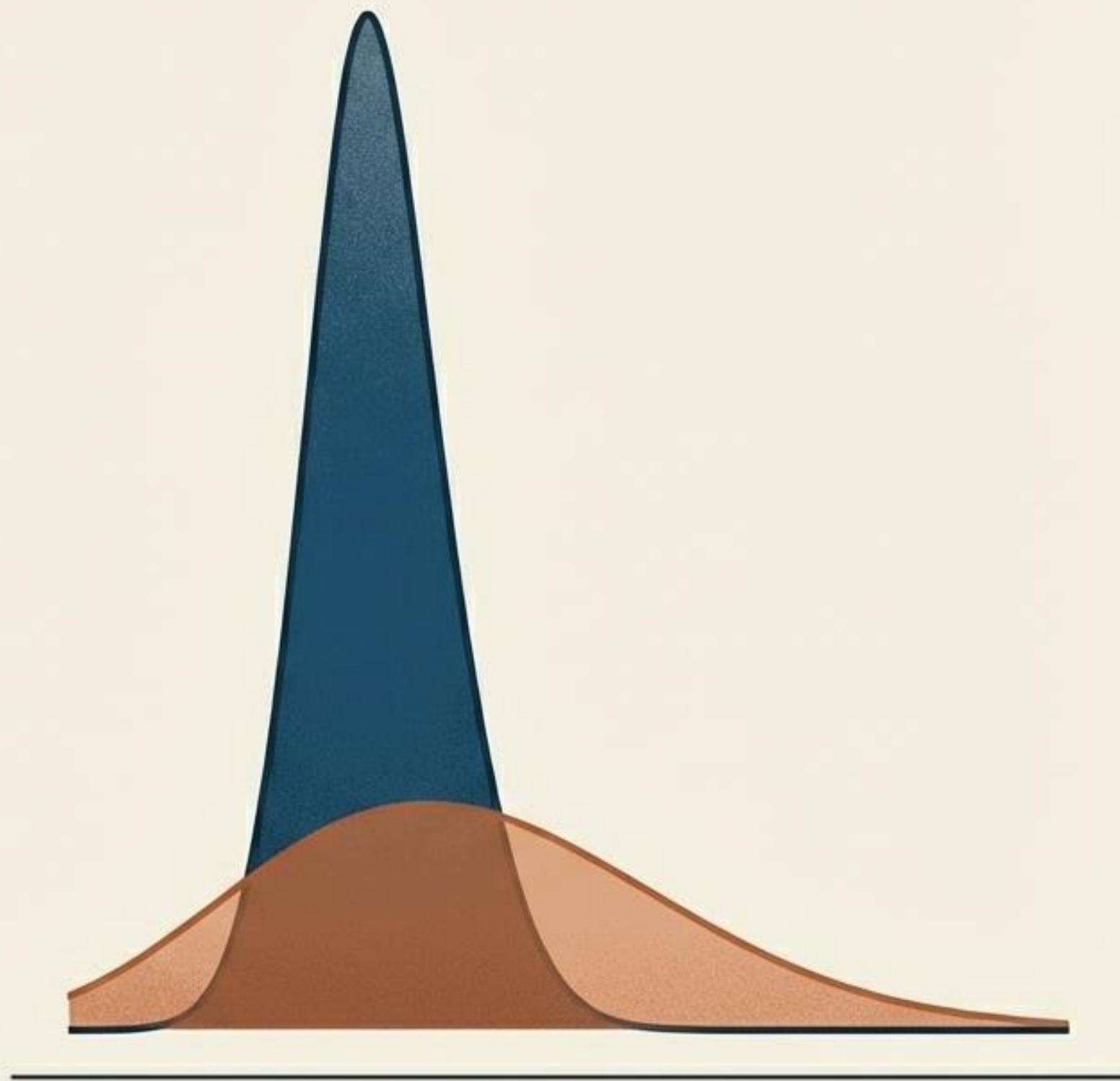
Que faire si les variances sont inégales ?

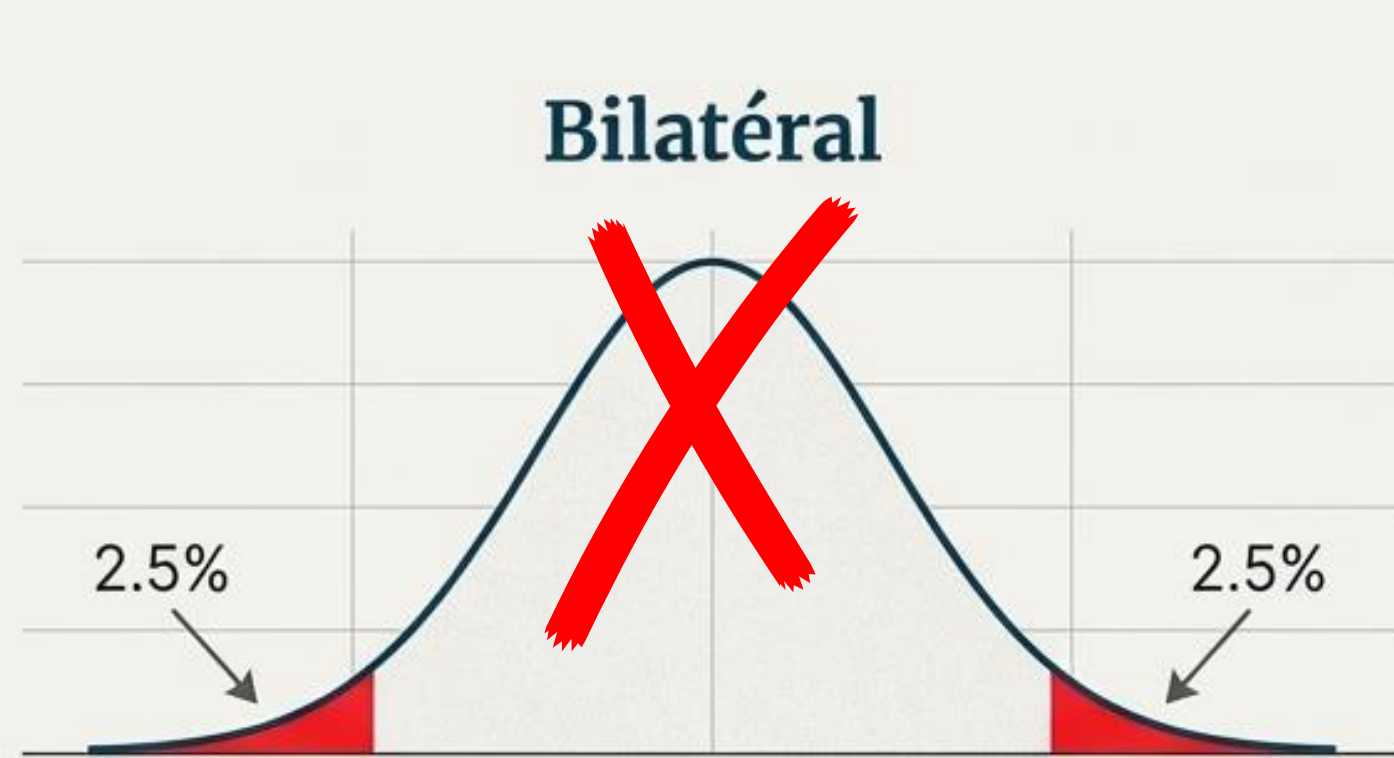
Le test classique exige une variance identique entre les deux groupes. Si la dispersion diffère fortement entre mâles et femelles, l'ajustement est immédiat.

La Solution : Le Test de Welch

Un simple changement de paramètre dans R :
`var.equal = FALSE`.

L'algorithme ajuste mathématiquement les degrés de liberté (ici, ddl passe de 198 à 187.76) pour compenser l'inégalité de variance sans perdre la robustesse du test.





Intégrer la Connaissance Biologique

La littérature indique que chez les Grapsidae, si dimorphisme il y a, la femelle est toujours plus large.

Nouvelle Hypothèse H_1 : $\text{rear}_F - \text{rear}_M > 0$

L'Impact Analytique

En utilisant alternative = "greater", toute la zone de rejet (5%) bascule à droite.

Résultat direct : La p-value est divisée par deux ($1.431e-05$ au lieu de $2.862e-05$). Le test devient plus sensible dans la direction attendue.

Le Design Optimal : Le Test t Apparié

Comparer des mesures répétées sur les mêmes individus permet d'éliminer la variabilité inter-individuelle. On ne teste plus la différence des moyennes, mais la moyenne des différences (med1 - med2).

Le Piège de l'Encodage

Un test apparié exige un tableau 'cas par variables' strict. Une ligne = Un individu. Si les identifiants se répètent sur plusieurs lignes, l'algorithme appliquera un test indépendant (totalement faux ici). Utilisez **paired = TRUE**.



extra	group	ID
-0.7	1	1
-1.2	1	2
-0.2	1	3
0.7	2	1
2.3	2	2
1.8	2	3



id	med1	med2
1	-0.7	0.7
2	-1.2	2.3
3	-0.2	1.8

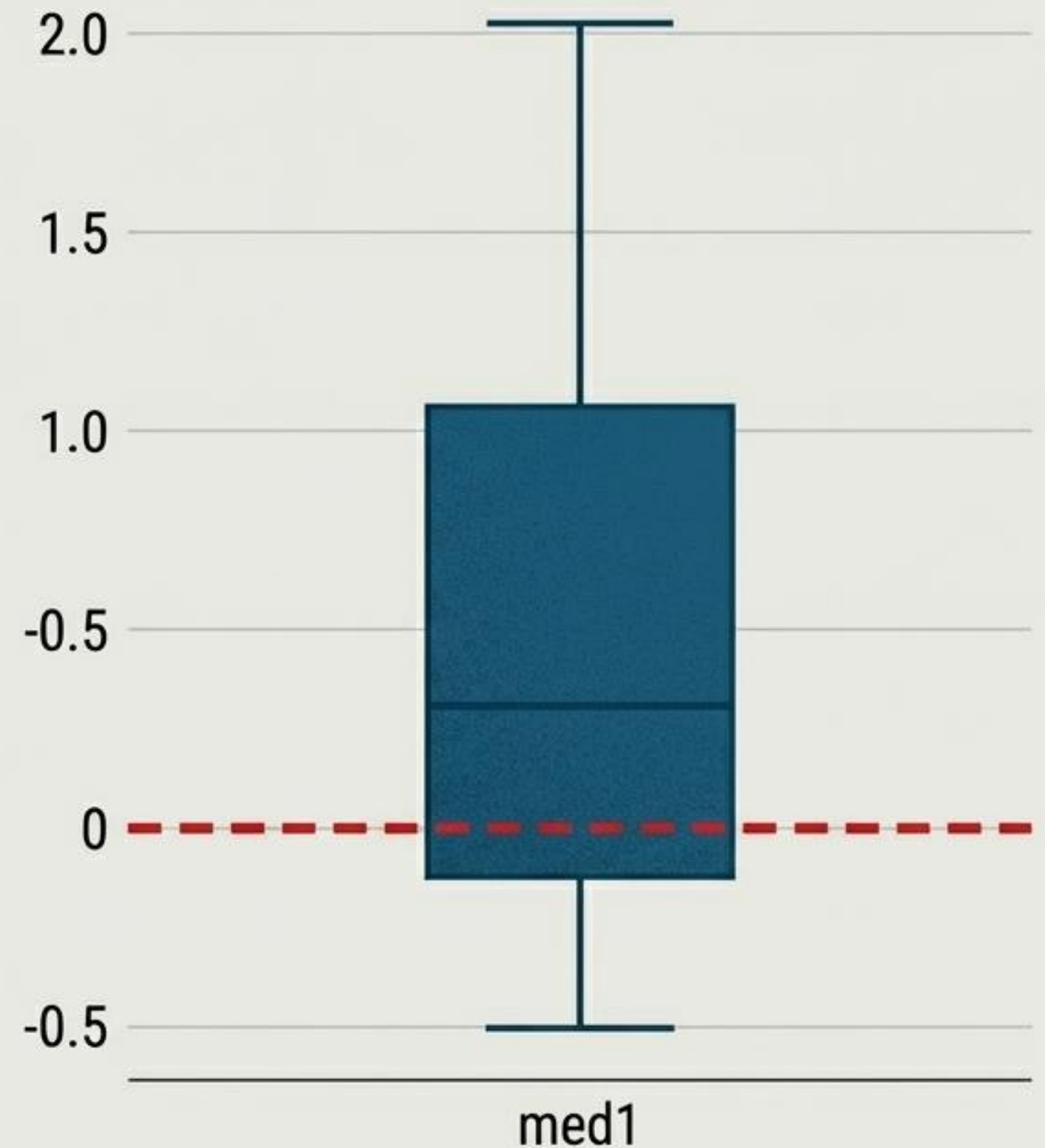
Le Test t Univarié : Comparaison à une Norme

Objectif : Comparer la moyenne d'une seule variable à une constante de référence (ex. : Le médicament 1 modifie-t-il la durée du sommeil par rapport à zéro heure supplémentaire ?).

- $H_0 : \text{med1} = 0$
- $H_1 : \text{med1} \neq 0$

Paramètre R : $\mu = 0$

Ici, la p-value = 0.22 (> 0.05) et l'IC [-0.53, 2.03] contient le zéro. On ne rejette pas H_0 .





Le Plan B : Le Test de Wilcoxon-Mann-Whitney

Le test t de Student s'effondre si l'échantillon est trop petit et que la distribution n'est pas normale.

L'Approche Non-Paramétrique

La fonction `wilcox.test()` ne fait aucune hypothèse sur la distribution (pas de μ , pas de σ). Au lieu de comparer les moyennes, le test se concentre sur le classement des données et compare les médianes. C'est le bouclier ultime contre les données asymétriques ou les tailles d'échantillons critiques.

Pourquoi préférer Student quand c'est possible ?

La réponse est la Puissance ($1 - \beta$).

- Erreur Type I (α) : Condamner un innocent.
- Erreur Type II (β) : Relâcher un coupable.

Le test t paramétrique est mathématiquement plus puissant. Il nécessite moins de réplicats que Wilcoxon pour détecter un effet biologique réel.

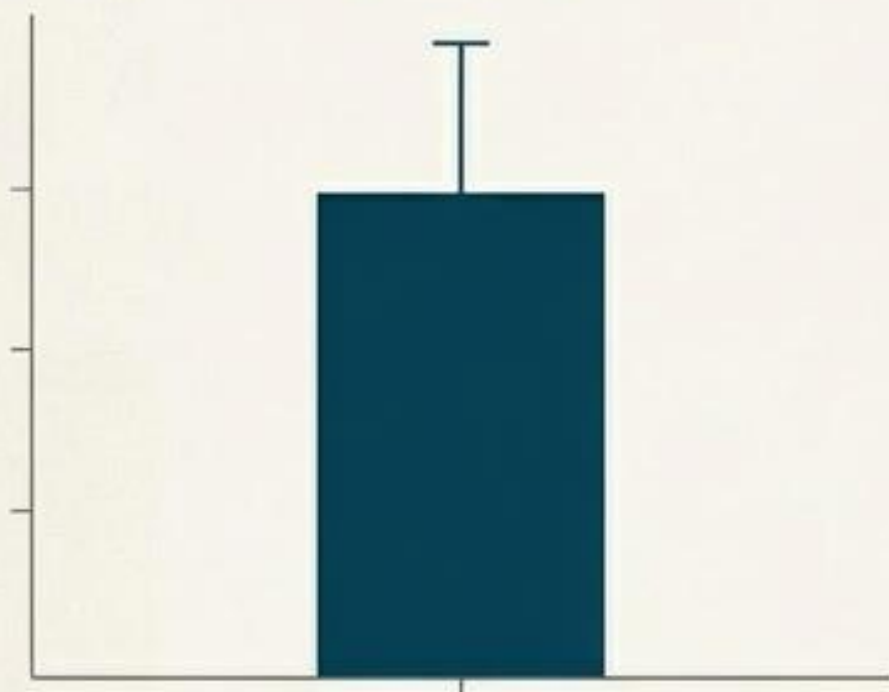
Règle d'or : Wilcoxon est une solution de repli, pas le choix par défaut.

Justice System of Statistics

	H_0 est Vraie	H_0 est Fausse
On Rejette H_0	Erreur Type I (α) Faux positif	Correct
On Ne Rejette Pas H_0	Correct	Erreur Type II (β) Faux négatif

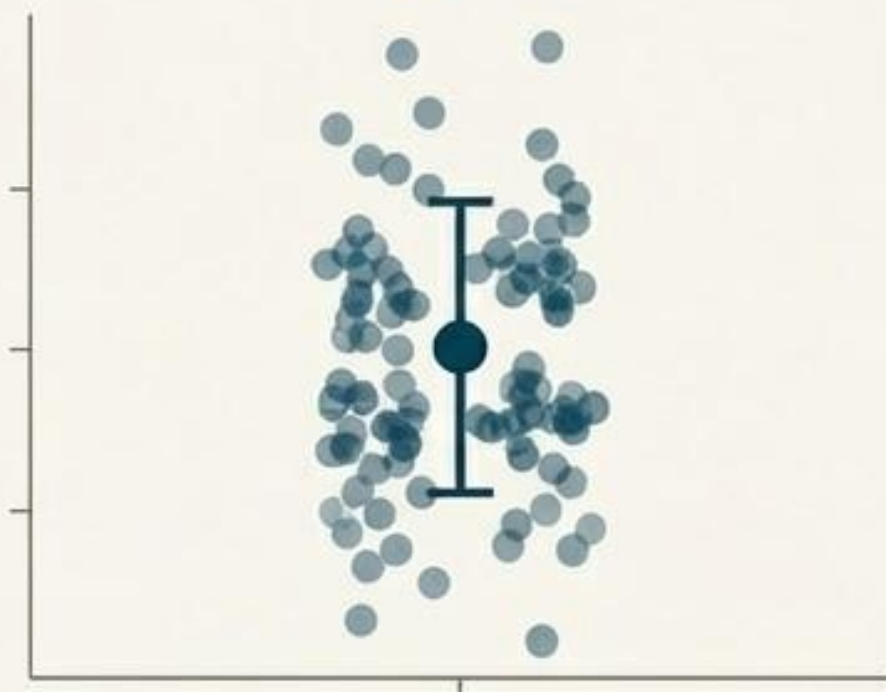
Communiquer la Vérité Visuellement

~~À Bannir~~
(La Brute)



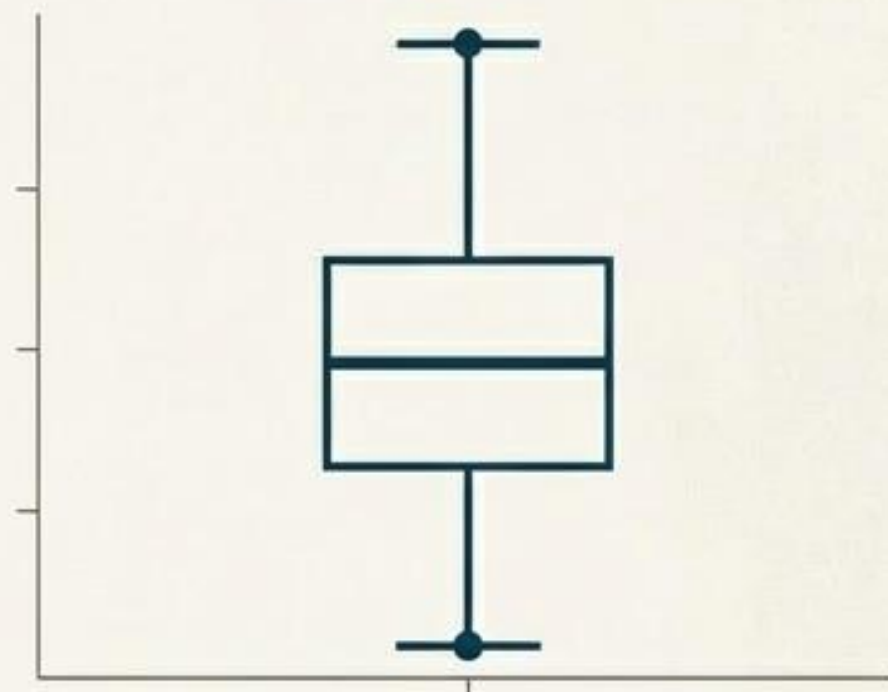
Cache la distribution réelle et le nombre d'observations.

Le Standard Student
(Le Bon)



Révèle chaque point de donnée, superposé à la moyenne et à l'IC calculé.

Le Standard Wilcoxon
(Le Juste)



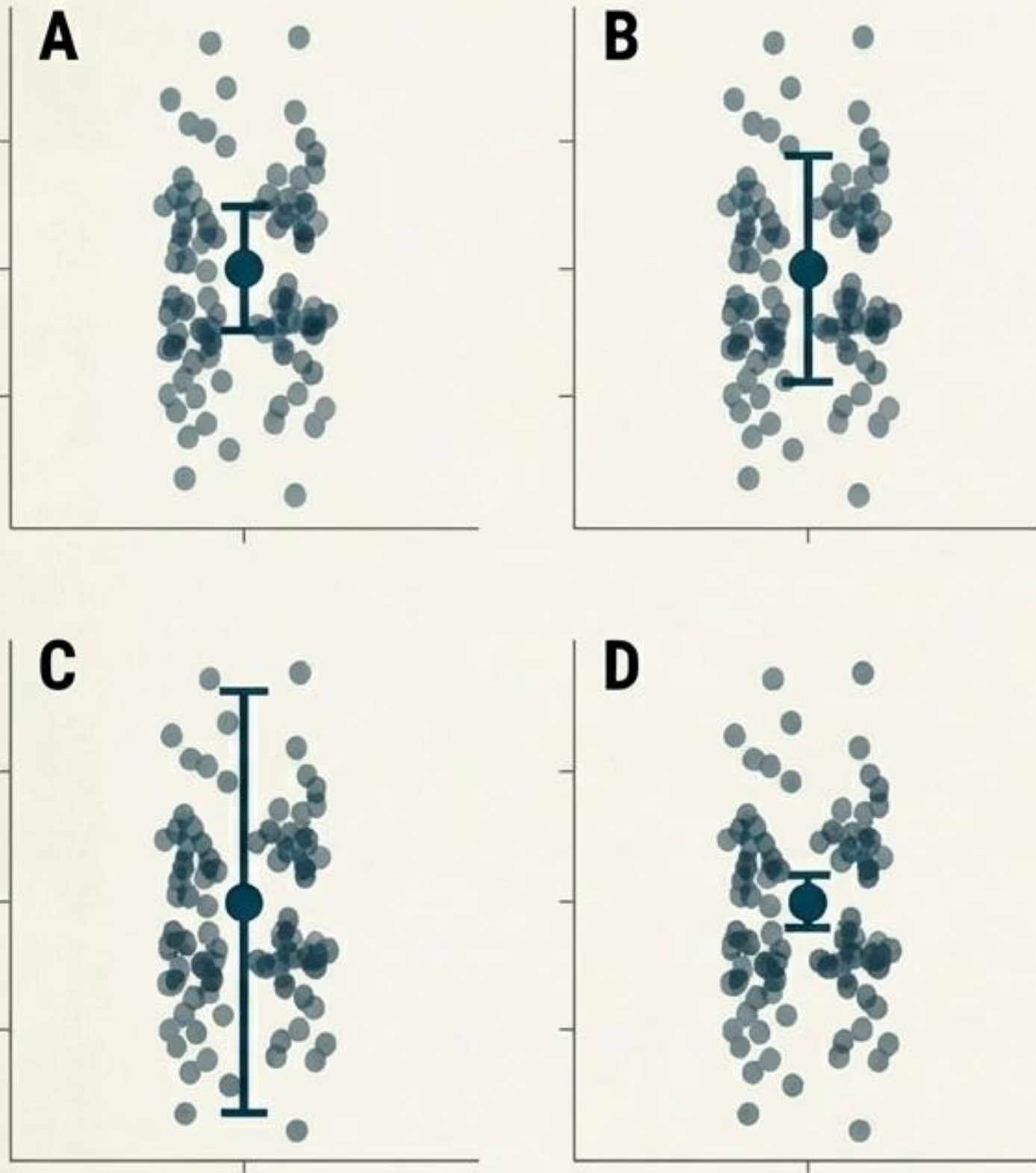
La boîte à moustaches décrit parfaitement les données non-paramétriques.

Le Piège des Barres d'Erreurs

Ces quatre graphiques montrent les mêmes données. Si les barres ne sont pas légendées, le graphique est illisible et trompeur.

- **A. IC 95%** : Idéal pour comparer visuellement la significativité.
- **B & C. Écart-Type (SD)** : Dispersion intrinsèque des observations, indépendamment de n .
- **D. Erreur Standard (SE)** : Paraît très précise, mais se réduit artificiellement à mesure que n grandit.

Règle absolue : Toujours préciser explicitement dans la légende la fonction mathématique derrière les moustaches.



L'Arbre de Décision de l'Inférence Biologique

Ce logigramme synthétise le cheminement analytique : de la structuration de l'expérience jusqu'à la représentation finale, en garantissant la rigueur mathématique à chaque intersection.

