105 L'intervalle de confiance

105.1 Cadre conceptuel

105.1.1 Imprécision et incertitude

> *L'objectif de la plupart des observations scientifiques est de permettre d'établir des règles générales à partir d'observations particulières. Lorsqu'un traitement est lancé sur le marché, c'est parce que les observations réalisées sur un échantillon de sujets permettent d'affirmer avec suffisamment de confiance que le traitement aura un effet déterminé sur n'importe quel individu de la population. Ce concept est celui d'inférence. D'une manière générale, les méthodes d'inférence statistique permettent de faire certaines affirmations concernant les caractéristiques d'une population à partir d'observations réalisées dans un échantillon.*

De par la nature aléatoire des variables étudiées, l'inférence est entachée d'incertitude, et les conclusions tirées au sujet de la population peuvent être fausses. Le but des statistiques est d'indiquer la confiance que l'on peut avoir dans les résultats.

> *Imaginez que vous ayez l'intention d'acheter une voiture d'occasion. Quel vendeur trouverez -vous le plus crédible : le vendeur qui vous assure que cette voiture est comme neuve et qu'il est impossible qu'elle tombe en panne : vous roulerez 100.000 km sans problème » ou le vendeur qui vous dirait : « je vends une bonne centaine de voitures de ce type chaque année, et je peux dire que vous avez 95% de chances de rouler entre 20.000 et 40.000 km sans panne majeure, de plus de 250 euros». Assurément, je choisirais la seconde affirmation car elle contient les trois éléments fondamentaux d'une affirmation plausible basée sur un phénomène aléatoire.*

Trois éléments sont essentiels dans une estimation basée sur l'inférence : la fourchette, la confiance, la signification.

La fourchette[1] délimite l'imprécision d'une estimation (« entre 20.000 et 40.000 km »)

la confiance[2] est la probabilité (1 -α) qu'une observation particulière appartienne à l'intervalle ; inversement le risque[3] d'erreur est la probabilité (α) qu'une observation particulière n'appartienne pas à l'intervalle

[1] *Fourchette : Écart entre deux nombres, à l'intérieur duquel on fait une appréciation.*

[2] *Confiance : Sentiment de sécurité de celui qui se fie à une affirmation.*

« vous avez 95% de chances » (1 -α=95%), sous -entendu : cette voiture -là pourrait aussi bien tomber en panne au prochain tournant, mais cela n'arrive que rarement (α=5%)

la signification[4] est une appréciation de l'importance de la différence entre deux situations

« sans panne majeure » : la variable est ici le nombre de kilomètres parcourus avant d'être en panne, mais à partir de quelle gravité d'avarie peut -on dire qu'un véhicule est en panne ? Une réparation de moins de 250 euros est considérée comme mineure.

Dans le domaine biomédical, on trouvera maints exemples de différences « statistiquement significatives[5] » mais sans grande « signification » : repousse « significative » (de seulement 0.3%) des cheveux après le traitement X, diminution « significative » (de seulement 1%) du poids après le régime Y…

105.1.2 L'intervalle de confiance

Une fourchette, évaluée à partir de l'échantillon, situe la valeur d'un paramètre de la population dans un certain intervalle, avec une certaine confiance.

Le calcul de l'intervalle de confiance formalise l'observation déjà décrite qualitativement en statistique descriptive : « dans une distribution symétrique, les observations ont environ 95 chances sur cent de se trouver dans un intervalle compris entre Mx ± 2σ »

$P(Mx - 2\sigma \leq X \leq Mx + 2\sigma) \approx 0{,}95$

Suivant les circonstances,
0,95 sera généralisée par la valeur 1 -α (0,95,0,99,0,999…)
en fonction des distributions d'échantillonnage concernées, la valeur approximative ± 2 devra être remplacée par ± $Z_{1-\alpha/2}$, ± $t_{n-1\,;\,-\alpha/2}$ … qui sont

---

[3] Risque : Danger plus ou moins probable auquel on est exposé en prenant une décision.

[4] Signification : pertinence d'une différence entre deux valeurs d'une variable dans le contexte considéré.

[5] [Statistiquement] Significatif : caractère non aléatoire d'une différence entre deux valeurs d'une variable

relativement proches de 2 pour 1 -α = 0,95 (toutefois ± $t_{n\ -1\ ;\ -\alpha/2}$ peut être sensiblement plus grand que 2 si n est petit).
la valeur σ devra être remplacée par la valeur ad hoc[6] en fonction des distributions d'échantillonnage concernées.

A partir de 3 prélèvements de 10 épis de maïs dans un champ de 10 ha un agronome pourrait déterminer par un intervalle de confiance, avec une certitude de 95%, que le rendement de la récolte se situera entre 19 et 21 tonnes à l'hectare[7].

105.2 Intervalle de confiance de la moyenne

105.2.1 Principe

Plaçons -nous dans la situation la plus courante où se trouve l'expérimentateur. Après son expérience, il dispose des statistiques Mx et S^2 calculées sur base d'un échantillon de valeurs indépendantes, prélevé dans une population de paramètres inconnus. Son objectif est de préciser le mieux possible ces paramètres.

Sur un échantillon de 5 dosages de protéines dans le lait récolté au hasard dans le stock d'une laiterie, on obtient les statistiques Mx = 30 g/l et S^2 = 20 (g/l)^2 La variable mesurée est supposée normale : X v.a. N(μ,σ^2). Que peut -on dire au sujet de la moyenne réelle de la teneur en protéines du stock de la laiterie[8]?

A priori, l'expérimentateur doit savoir que le paramètre μ ne vaut pas exactement 30g/l. Il peut cependant affirmer que :

μ = 30 ± une certaine imprécision (ε)

La distribution d'échantillonnage de Mx est une variable normale de moyenne μ, prenant théoriquement des valeurs comprises entre ± ∞ .

Limiter l'intervalle à ± ε implique d'accepter que certaines valeurs de Mx sortent de l'intervalle. Ce risque d'erreur est α (petite incertitude) Si α = 0, ε=± ∞ (imprécision infinie)

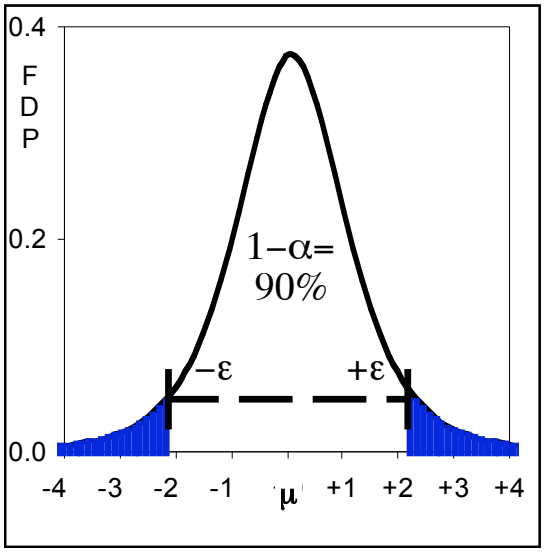
---

[6] *ad hoc : Qui convient à la situation*

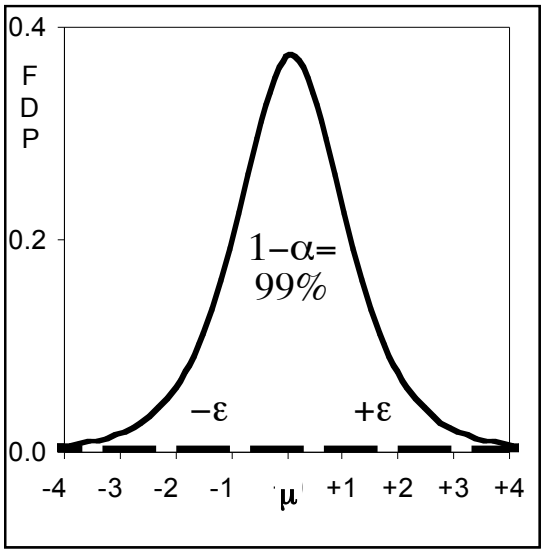
[7] *Hectare : unité de mesure d'aire ou de superficie (symb. ha) valant 10^4 mètres carrés.*

[8] *Laiterie : Usine où le lait est traité pour sa consommation et pour la production de produits dérivés (crème, beurre, fromage, yaourts)*

Plus l'incertitude est grande, plus l'imprécision est petite (fig. A), et réciproquement (fig. B). On ne peut donc fixer ε que pour une probabilité 1 -α donnée.



A



B

Figure 105 -1 Distribution d'échantillonnage de Mx, centrée sur μ. P(Mx = μ± ε) ≥ (1 - α) . Relation entre l'incertitude (A : α = 10%, B : α = 1%) et l'imprécision de la fourchette (±ε)

105.2.2 Technique

La détermination de ε dépend de la distribution d'échantillonnage. Pour déterminer l'intervalle à la moyenne (v.a. Normale suivant le théorème central limite), lorsque la variance de la population est inconnue, la variable à considérer est un t de Student.

$$t_{n-1;1-\alpha/2} = \frac{Mx_{1-\alpha/2} - \mu}{S/\sqrt{n}} = \frac{\varepsilon}{S/\sqrt{n}}$$

$$t_{n-1;\alpha/2} = \frac{Mx_{\alpha/2} - \mu}{S/\sqrt{n}} = \frac{-\varepsilon}{S/\sqrt{n}}$$

$$\pm\varepsilon = \pm t_{n-1;\alpha/2} S/\sqrt{n}$$ **Équation 105-1**

Dans notre exemple n = 5 : *t* a 4 degrés de liberté (d.l.)

La borne supérieure se trouve dans les tables : $t_{4\ ;0,975} = 2,78$

$$t_{4;0.975} = \frac{Mx - \mu}{S/\sqrt{n}} = 2.78$$

$$\frac{30 - \mu}{\sqrt{20/5}} = 2.78$$

$$\varepsilon = 2.78\sqrt{20/5} = 5.6$$

$$\mu_{0.975} = 30 + 5.6 = 35.6$$

De la même façon, on trouvera la borne inférieure en utilisant $t_{4\ ;0,025} = -t_{4\ ;0,975} = -2,78$, ce qui donne $\mu_{0,025} = 30 - 5,6 = 24,4$.

Revoyons graphiquement la signification des valeurs obtenues :

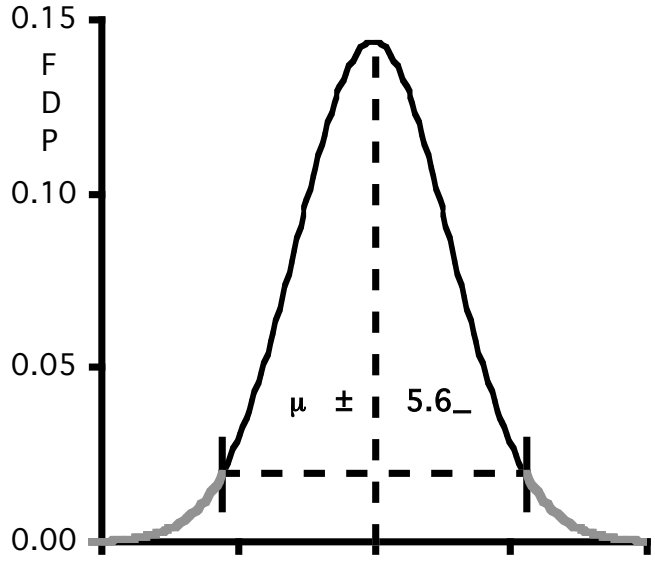


Figure 105 -2 Distribution d'échantillonnage de MX, basée sur une confiance de 95%.

La distribution d'échantillonnage de Mx ne donne pas de valeurs en abscisse car μ est inconnu : ε = 5,6.

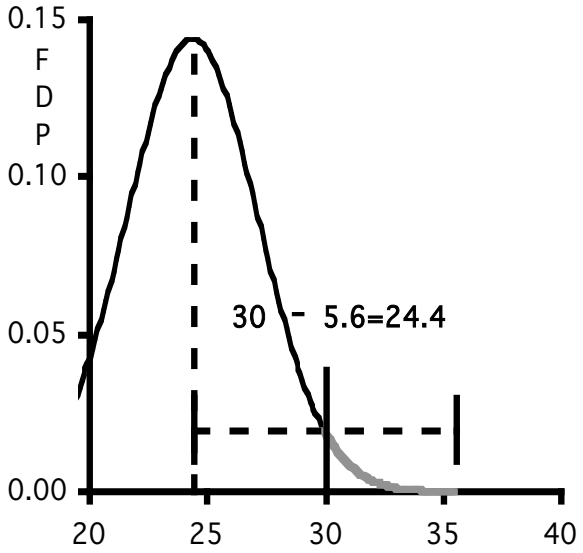


Figure 105 -3 Distribution d'échantillonnage de Mx centrée sur une moyenne hypothétique de 24.4.

Imaginons que la moyenne observée (30) soit la moyenne la plus grande que l'on puisse considérer : elle se situerait à la limite supérieure de la zone de probabilité 95% (au-delà de cette limite, on entre dans la zone improbable, non considérée). Dans ce cas, la valeur la plus petite que l'on puisse considérer pour μ est 30 -5,6 = 24,4

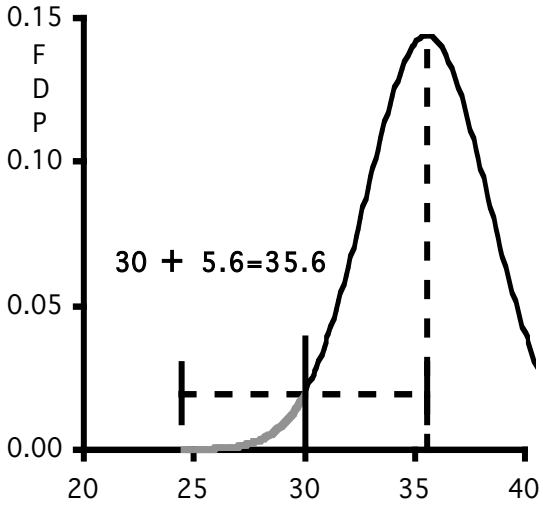


Figure 105 -4 Distribution d'échantillonnage de Mx centrée sur une moyenne hypothétique de 35,6.

Imaginons que la moyenne observée (30) soit la moyenne la plus petite que l'on puisse considérer : elle se situerait à la limite inférieure de la zone de probabilité 95% (au delà de cette limite, on entre dans la zone improbable, non considérée). Dans ce cas, la valeur la plus grande que l'on puisse considérer pour μ est 30 +5,6 = 35,6

105.2.3 Exemple

L'intervalle de confiance peut être nettement plus grand que celui que l'on estimerait intuitivement,

> *Un médecin obtient pour une patiente une mesure de glucose de 10,17 g/l sang et trouve cette valeur élevée. Il fait refaire une analyse indépendante et obtient 8,70 g/l. Ce résultat vient -il conforter le premier ? S'il ne dispose d'aucune autre valeur de variance que celle obtenue à partir de ces deux mesures, que peut -il dire de la moyenne réelle?*

$$\varepsilon = t_{1;0.975} \, S/\sqrt{n} = 12.71 \times 0.74 = 9.34$$

$$\mu = Mx \pm \varepsilon = 9.44 \pm 9.34$$

> *La moyenne réelle a 95% de chances de se situer entre 0,1 et 18,8 g/l (! !)*

A partir de quelques observations, un intervalle de confiance peut toutefois donner une estimation relativement précise d'une population tout à fait inconnue :

> *Sur 5 cultures dans un nouveau fermenteur[9] un chimiste obtient un rendement de 78, 74, 75, 72, 73%. Que peut -il dire du rendement moyen réel du fermenteur ?*

Mx = 74,75,Sx = 2,5
$1-\alpha$ = 95%
Tables : $t_{4\,;0,975}$ = 2,78

$$\varepsilon = t_{4;0.975} \, S/\sqrt{n} = 2.78 \times 1.12 = 3.10$$

$$\mu = Mx \pm \varepsilon = 74.75 \pm 3.10$$

[9] *Fermenteur : cuve pouvant atteindre une capacité de plusieurs milliers de litres et permettant la fermentation de micro-organismes dans un milieu parfaitement contrôlé. Utilisé en brasserie et dans de nombreuses applications bio -technologiques.*

105.3 Autres intervalles de confiance

105.3.1 Moyenne de variance connue

$\mu = Mx \pm \varepsilon$

Lorsque la variance σ^2 est connue, σ^2 doit remplacer S^2 et la distribution de Z peut être prise comme référence au lieu de la distribution de *t*.

L'expérimentateur a tout intérêt à le faire, car pour 1 -α donné, $Z_{1-\alpha/2} < t_{1-\alpha/2}$: la fourchette obtenue en référence à la distribution de Z sera la plus petite, et son estimation plus précise.

Exemple

> *Un médecin obtient pour une patiente une mesure de glucose de 10,17 g/l sang et trouve cette valeur élevée. Il fait refaire une analyse indépendante et obtient 8,70 g/l. Les tables d'analyses cliniques prévoient pour cette mesure un écart -type de 1g/l. Que peut -il dire de la valeur réelle ?*

$$\varepsilon = Z_{0.975} \sigma/\sqrt{n} = 1.96 \times 0.70 = 1.37$$

$$\mu = Mx \pm \varepsilon = 9.44 \pm 1.37$$

> *La moyenne réelle a 95% de chance de se situer entre 8,06 et 10,81 g/l.*

105.3.2 Proportion

$$\pi = P \pm \varepsilon$$

Une proportion estimée est un nombre de réalisations X sur une série d'observations n.
X répond donc à la définition d'une v.a. Bi(n ; π) de variance S_X^2 = nπ(1 -π)

L'intervalle de confiance d'une proportion est basé sur la variance de P = X/n qui estime π. Si l'on divise (multiplie) une variable par une constante, on divise (multiplie) la variance par le carré de cette constante, ce qui donne :

$$S^2_{X/n} = \frac{nP(1-P)}{n^2} = \frac{P(1-P)}{n}$$

Et donc pour 1 -α 95% :

$$\varepsilon \approx 2\sqrt{\frac{P(1-P)}{n}}$$

Une estimation plus précise de ε est basée sur une approximation de la Binomiale par une v.a. Normale

$$\varepsilon \approx z_{1-\alpha/2}\sqrt{\frac{P(1-P)}{n}}$$ Équation 105-2

Cette approximation est valable si X ≥ 5 et n ≥ X +5

Exemple : Dans le cas d'une contamination d'un cheptel bovin par la bactérie Brucella abortus un vétérinaire observe indépendamment 53 avortements pour 134 vaches gestantes[10]. Quel risque peut -il prédire dans la population (ici, le cheptel)?

Proportion : 53/134 = 0,40
Confiance 95%
Tables : $z_{1-\alpha/2} = 1.96$

$$\varepsilon \approx 1.96\sqrt{\frac{0.4 \times 0.6}{134}} = 0.083$$

La proportion réelle a 95 chances sur 100 d'être comprise entre 0,317 et 0,483.

[10] *En gestation : état d'une femelle vivipare, entre nidation et mise bas, chez les espèces qui nourrissent l'embryon, puis le fœtus, par voie placentaire.*

105.3.3 Dénombrement

$\mu = x \pm \varepsilon$

Le comptage d'observations indépendantes par unité de surface, de volume ou de temps suit une distribution de Poisson sous les conditions énoncées plus haut. Si X Po(μ), VAR(X) = μ.

$\varepsilon \approx 2\sqrt{x}$ pour 1 -α = 95%

Une estimation plus précise de ε est basée sur une approximation de la Poisson par une v.a. Normale. Cette approximation est valable si X ≥ 5

$$\varepsilon \approx Z_{1-\alpha/2}\sqrt{x}$$ **Équation 105-3**

Exemple

Un géologue mesure une radioactivité de 150 dpm dans un prélèvement de roche. Que peut -il dire, environ, de la valeur réelle de la radioactivité ?

$\varepsilon \approx 2\sqrt{150} \approx 25$

La radioactivité réelle a environ 95 chances sur 100 d'être comprise entre 125 et 175 dpm[11].

Plus précisément, en se servant des tables, $\varepsilon \approx 1.96\sqrt{150} \approx 24$.

[11] *Dpm = nombre de désintégrations par minute d'un isotope radioactif. Mesure de coups par minute mesurés dans un liquide à scintillation, corrigée par les dilutions et la mesure d'un témoin non radioactif.*

105.3.4 Pente de la droite de régression

$$\beta 1 = B1 \pm \varepsilon$$

Les conditions d'inférence sur la droite de régression sont strictes :

(1) la variable Y dépend de la variable X et non l'inverse, x est connu sans imprécision.

Si X et Y sont des v.a. aléatoires, le modèle adéquat est la droite des moindres rectangles, qui ne se prête pas à l'inférence.

(2) les observations sont indépendantes

Si plusieurs valeurs de Y sont obtenues pour une même valeur de X, la régression doit être étudiée dans le cadre d'une analyse de la variance.

(3) chaque valeur x_i correspond une population de valeurs de Y de distribution normale, de même variance

Si la variance de Y se modifie en fonction de X une transformation (souvent logarithmique) doit être envisagée.

(4) les moyennes μ_i de ces populations sont situées sur une droite de régression dans le domaine de X étudié.

Si les moyennes des populations ne sont pas colinéaires, le modèle, plus complexe, est celui d'une régression non linéaire.

Deux sources de variabilité interviennent dans la variance de la pente de la droite de régression B_1 : SCER et SCEx.

1.SCER représente la dispersion des points autour de la droite : plus cette dispersion est faible, moins la pente est variable.

$$SCER = \sum_{i=1}^{n} (yo_i - ym_i)^2 =$$

$$\sum_{i=1}^{n} (yo_i - (Bo - B_1 x_i))^2$$

En l'absence de réplicats, cette somme de carrés d'écart a n -2 degrés de liberté, car 2 paramètres estimés (B_o et B_1) interviennent dans son calcul.

CMR = SCER/n -2

Intuitivement : par deux points passe toujours exactement une droite. Il faut donc au moins 3 points pour obtenir une estimation de la variance de B_1.

2.SCEX représente la dispersion des valeurs de X : plus large est le domaine de X exploré (plus SCEx est élevée), moins la pente est variable.

$$SCEx = \sum_{i=1}^{n} (xi - Mx)^2$$

Il ressort de ceci que sous les conditions énoncées l'écart -type de B_1 est donné par l'expression :

$$\sqrt{\frac{CMR}{SCEx}}$$

et pour 1 -α = 95%

$$\varepsilon \approx 2\sqrt{\frac{CMR}{SCEx}}$$

Une estimation plus précise de ε est basée sur une distribution de t de Student (ce qui est indispensable si n est petit).

$$\varepsilon = t_{n-2;1-\alpha/2}\sqrt{\frac{CMR}{SCEx}} \qquad \textbf{Équation 105-4}$$

Exemple

Un démographe estime la croissance de la population pensionnée d'un quartier, année par année sur 5 ans : 26, 32, 40, 44, 55% : SCEx = 10, SCER = 9
Sur base d'un modèle linéaire, il estime une croissance de 7% par an.

Tables : t 3 ;0.975 = 3.18

$$\varepsilon = 3.18\sqrt{\frac{9/3}{10}} = 1.74$$

L'accroissement réel annuel est donc compris entre 5,25 et 8,74%