Statistiques et informatique

Enseignante : Nesrine TABCHOUCHE [1]

1. **Université Akli Mohand Oulhadj-Bouira Faculté des Sciences de la Nature et de la Vie et des Sciences de la Terre. Département de Sciences Agronomiques. Spécialité Technologie Agroalimentaire et Contrôle de Qualité. e-mail : tabchouche_nesrine@yahoo.fr**

Table des matières

Liste des tableaux

Chapitre 1

PROBABILITES - STATISTIQUES

Symbole	Signification
σ	L′écart-type de la population
σ^2	La variance de la population
θ	Valeur estimée d′un paramètre
μ	La moyenne de la population
α	Le coefficient de risque
$1-\alpha$	Le coefficient de confiance
$Z_{\frac{\alpha}{2}}$	La valeur critique (l'écart réduit)
s	Ecart-type biaisé d'un échantillon

TAB. 1.1: Symboles utilisés

1.1 Rappels de probabilités théoriques

1.2 Outils probabilistes de base

– Expérience aléatoire : expérience où le hasard intervient rendant le résultat imprévisible (Ex : lancer un dé)

ensemble de tous les résultats possibles = univers des possibles = Ω

(Ex : $\Omega = \{1, 2, ..., 6\}$)

Evénement : assertion relative au résultat d'une expérience, se réalise ou non

(Ex : obtenir un nbre pair)

=toute partie de Ω (Ex :$\{2, 4, 6\}$)

- Soit C un ensemble d'événements =ensemble de parties de Ω satisfaisant les propriétés suivantes (algèbre de Boole) :

$$\left.\begin{array}{c} \forall A \in \varphi : \text{ son contraire } \bar{A} \in \varphi(\bar{A} = \Omega \backslash A) \\ \forall A_1, A2, ..., A_n \in \varphi : \cup A_i \in \varphi \\ \Omega \in \varphi \end{array}\right\} \Longrightarrow \begin{array}{c}(\Omega, \varphi) \text{ est un espace} \\ \text{probabilisable}\end{array}$$

- Loi de probabilité P : $(\Omega, \varphi)\Pi[0, 1]$ telle que

$$\left.\begin{array}{c} P(\Omega) = 1 \text{ et } \forall A_1, A2, ..., A_n \in \varphi \text{ tels que } A_i \cap A_j = \emptyset(\ \forall i \neq j), \\ \text{on a } P(\cup A_i) = \sum P(A_i) \end{array}\right\} (\Omega, \varphi, P)\text{est}$$

un espace probabilisé

- Propriétés élémentaires
  - $P(\emptyset) = 0$
  - $P(\bar{A}) = 1 - P(A)$
  - $P(A) \leq P(B)$ si $A \subset B$
  - $P(A \cup B) = P(A) + P(B) - P(A \cap B)$
  - $P(\cup A_i) \leq \sum P(A_i),\ 0 \leq P(A_i) \leq 1$
- Si $P(A) = 0$ alors A est presque impossible. On écrit $A = \emptyset$

  Si $P(A) = 1$ alors A est presque sûr. On écrit $A = \Omega$.
- Assignement d'une probabilité à un événement :
  - vision classique (jeux de hasard) :

    Ω est un ensemble fini de cas possibles dont chaque singleton (événement élémentaire) a la même probabilité de se réaliser.

    (Ex : lancer un dé parfait => Ω est constitué de 6 éléments équiprobables) d'où :

$$P(A) = \frac{\text{Nbre de cas favorables}}{\text{Nbre de cas possibles}}$$

  - vision "fréquentiste" (loi des grands nombres) :

    répéter un grand nbre de fois N l'expérience et observer le nbre de fois que l'événement d'intérêt A se produit :

    d'où :

$$f(A) = \frac{\text{Nbre d'occurrences de } A}{\text{Nbre d'expériences } (N)} \text{ et } P(A) = \lim_{N \to \infty} f(A)$$

Example 1.2.1. *: jet de deux pièces de monnaie distinguables* $\Omega = \{(P,P);(P,F);(F,P);(F,F)\}$ *est équiprobable. Soit* $A =$ *"On obtient au moins une fois* $\{(P,P);(P,F);(F,P)\}$. $P(A) = \frac{3}{4}$.

1.2.1 Variables aléatoires et distributions

- Variable aléatoire : entitéprenant différentes valeurs ('variable'), chacune avec une certaine probabilité ('aléatoire')
  - nbre fini ou dénombrables de valeurs : variable discrète,
  - toute valeur dans un certain intervalle de $\Re$: variable continue
- Loi de probabilitéd'une variable aléatoire X :
  assignation des probabilités sur les différentes valeurs de X (discrète) ou sur des intervalles de valeurs de X(continue)
  - Pour une variable discrète : masses ponctuelles $P(X = x_i)$,
  - Pour une variable continue : densité de probabilité $P(a < X < b)$.

Moments d'une variable aléatoire X :

valeurs typiques :

- centrales : moyenne
- de dispersion : variance, écart-type(déviation standard)
- de forme de distribution : coefficient d'asymétrie('skewness'), d'aplatissement('kurtosis').

notion d'espérance mathématique : $E(X) =$ moyenne (= centre de masse) :

- var. discrète : $\mu = E(X) = \sum_i x_i P(X = x_i)$
- var. continue de densité $f(x) =: \mu = E(X) = \int_{\Re} x f(x) dx$ (n'existe pas tjrs)
- propriétés élémentaires :
  - $E(a) = a$
  - $E(aX) = aE(X)$
  - $E(X + Y) = E(X) + E(Y)$
- variance : $V(X) = \sigma^2 = E((X = E(X))^2)$ (moment centré d′ordre 2)
  $= E(X)^2 - \mu^2$
- écart-type : $\sigma = \sqrt{V(X)}$

1.3 Estimation statistique

Définition 1.3.1. *Soit X une variable aléatoire dont la densité de probabilité $f(x, \theta)$dépend d'un paramètre appartenant à $I \subset R$. A l'aide d'un échantillon issu de X, il s'agit de déterminer au mieux la vraie valeur θ_0 de θ. On pourra utiliser deux méthodes :*

- *estimation ponctuelle : on calcule une valeur vraisemblable $\hat{\theta}$ de θ_0.*
- *estimation par intervalle : on cherche un intervalle dans lequel θ_0 se trouve avec une probabilité élevée.*

1.4 L'estimation ponctuelle

Définition 1.4.1. *Estimer un paramètre, c'est en chercher une valeur approchée en se basant sur les résultats obtenus dans un échantillon. Lorsqu'un paramètre est estimé par un seul nombre, déduit des résultats de l'échantillon, ce nombre est appelé estimation ponctuelle du paramètre.*

L'estimation ponctuelle se fait à l'aide d'un estimateur, qui est une variable aléatoire d'échantillon. L'estimation est la valeur que prend la variable aléatoire dans l'échantillon observé.

1.4.1 Propriétiés des estimateurs ponctuels

Estimateur non biaisé.

Définition 1.4.2. *Un estimateur est sans biais si la moyenne de sa distribution d'échantillonnage est égale à la valeur θ du paramètre de la population à estimer, c'est-à-dire si $E(\hat{\theta}) = \theta$*

Si l'estimateur est biaisé, son biais est mesuré par l'écart suivant : BIAIS =$E(\hat{\theta}) - \theta$

Example 1.4.1. *Si $E(\bar{X}) = m$. Donc la moyenne d'échantillon $\bar{X}$ est un estimateur sans biais du paramètre m, moyenne de la population. En revanche, la médiane d'échantillon M_e est un estimateur biaisé lorsque la population échantillonnée est asymétrique.*

1.4.2 Estimateur efficace

Définition 1.4.3. *Un estimateur sans biais est efficace si sa variance est la plus faible parmi les variances des autres estimateurs sans biais. Ainsi, si* $\hat{\theta}_1$*et* $\hat{\theta}_2$ *sont deux estimateurs sans biais du paramètre* θ*, l'estimateur* $\hat{\theta}_1$ *est efficace si :*

$$V(\hat{\theta}_1) < V(\hat{\theta}_2) \quad et \quad E(\hat{\theta}_1) = E(\hat{\theta}_2) = \theta.$$

1.4.3 Estimateur convergent

Définition 1.4.4. *Un estimateur* $\hat{\theta}$ *est convergent si sa distribution tend à se concentrer autour de la valeur inconnue à estimer,* θ*, , à mesure que la taille d'échantillon augmente, c-à-d si* $\lim_{n\to\infty} V(\hat{\theta}) = 0.$

Remarque 1.4.1. *Un estimateur sans biais et convergent est dit absolument correct.*

1.5 Estimation par intervalle de confiance

Définition 1.5.1. *L'estimation par intervalle d'un paramètre inconnu* θ*consiste à calculer, à partir d'un estimateur choisi* $\hat{\theta}$*, un intervalle dans lequel il est vraisemblable que la valeur correspondante du paramètre s'y trouve .* ***L'intervalle de confiance*** *est défini par deux limites* LI *et* LS *aux quelles est associée une certaine probabilité, fixée à l'avance et aussi élevée qu'on le désire, de contenir la valeur vraie du paramètre. La probabilité associée à l'intervalle de confiance et exprimée en pourcentage est égale à S où S est le seuil de confiance ou niveau de* ***confiance de l'intervalle****, exprimé également en pourcentage.*

$$P(LI \leq \theta \leq LS) = S$$

- LI *: limite inférieure de l'intervalle de confiance.*
- LS *: limite supérieure de l'intervalle de confiance*
- S *: probabilité associée à l'intervalle d'encadrer la vraie valeur du paramètre.*

LI et LS sont appelées les limites de confiance de l'intervalle et sont des quantités qui tiennent compte des fluctuations d'échantillonnage, de l'estimateur $\hat{\theta}$ et du seuil de confiance S.

La quantité $1 - S$ est égale à la probabilité, exprimée en pourcentage, que l'intervalle n'encadre pas la vraie valeur du paramètre. On note $\alpha = 1 - S,\ \alpha$ s'appelle le risque ou le seuil de signification de l'intervalle.

Remarque 1.5.1. – *L'intervalle ainsi défini est un intervalle aléatoire puisqu'avant l'expérience, les limites de l'intervalle sont des variables aléatoires (elles sont fonctions des observations de l'échantillon).*

– *Le niveau de confiance est toujours associé à l'intervalle et non au paramètre inconnu θ. θ n'est pas une variable aléatoire : il est ou n'est pas dans l'intervalle $[LI, LS]$.*

– *Le niveau de confiance doit être choisi avant que ne s'effectue l'estimation par intervalle. Il arrive souvent que le chercheur non averti calcule plusieurs intervalles d'estimation à des niveaux de confiance différents et choisisse par la suite l'intervalle qui lui semble le plus approprié. Une telle approche constitue en réalité une interprétation inacceptable des données en ce qu'elle fait dire aux résultats échantillonnaux ce que l'on veut bien entendre.*

– *Il y a une infinité de solutions possibles pour déterminer l'intervalle $[LI, LS]$. On choisira de prendre des risques symétriques, c'est-à-dire de choisir LIet LS tels que :*

$$\begin{cases} P(\theta \leq LI) = \frac{1-S}{2} \\ P(\theta \geq LS) = \frac{1-S}{2}. \end{cases}$$

1.5.1 Estimation d'une moyenne par intervalle de confiance

On se propose d'estimer, par intervalle de confiance, la moyenne m d'un caractère mesurable d'une population. Il s'agit donc de calculer, à partir de la moyenne $\bar{x}$ (valeur prise par l'estimateur $\bar{X}$) de l'échantillon, un intervalle dans lequel il est vraisemblable que la vraie valeur de m s'y trouve.

Cet intervalle se définit d'après l'équation suivante : $P(A \leq m \leq B) = S$.

Les limites A et B de cet intervalle sont des quantités aléatoires et prendront, après avoir prélevé l'échantillon et calculé l'estimation $\bar{x}$, la forme suivante : $LI \leq m \leq LS$.

Nous allons déterminer LI et LS en utilisant la distribution d'échantillonnage de $\bar{X}$.

1.5.2 On dispose d'un grand échantillon $(n \geq 30)$ ou d'un petit échantillon $(n < 30)$ dont la distribution est normale d'écart-type σ

On veut estimer la moyenne μ de la population à l'aide d'un échantillon aléatoire (si la taille de l'échantillon est grande $n \geq 30$ la distribution d'échantillonnage de la moyenne est normale quelle que soit la distribution de la population).

$\bar{X} \rightarrow N(\mu_{\bar{X}}, \sigma_{\bar{X}})$ avec :

$\sigma_{\bar{X}} = \frac{\sigma}{\sqrt{n}}$, si le tirage est non exhaustif,

$\sigma_{\bar{X}} = \frac{\sigma}{\sqrt{n}}\sqrt{\frac{N-n}{N-1}}$, tirage est exhaustif.

On a : $\frac{\bar{X}-\mu_{\bar{X}}}{\sigma_{\bar{X}}}$suit une loi normale centrée réduite $N(1, 0)$.

On cherche un intervalle centré sur μ avec une probabilité égale à $P(\frac{\bar{X}-\mu_{\bar{X}}}{\sigma_{\bar{X}}} < Z_{\frac{\alpha}{2}}) = \alpha$.

On peut distinguer entre les cas suivants :

– σ connue : $\mu \in [\bar{x} - Z_{\frac{\alpha}{2}}\sigma_{\bar{X}}, \bar{x} + Z_{\frac{\alpha}{2}}\sigma_{\bar{X}}]$.

– σ inconnue $(n \geq 30)$:

$$\mu \in [\bar{x} - Z_{\frac{\alpha}{2}}\sigma_{\bar{X}}, \bar{x} + Z_{\frac{\alpha}{2}}\sigma_{\bar{X}},]$$

avec

$$\sigma_{\bar{X}} = \frac{s}{\sqrt{n-1}}.$$

– σ inconnue $(n < 30)$

$$\mu \in [\bar{x} - t_{\frac{\alpha}{2}}\sigma_{\bar{X}}, \bar{x} + t_{\frac{\alpha}{2}}\sigma_{\bar{X}},]$$

1.5.3 Estimation d'une variance par intervalle de confiance

On se propose d'estimer, par intervalle de confiance, la variance σ^2 d'un caractère mesurable d'une population. Il s'agit donc de déterminer, à partir de la variance de l'échantillon σ^2_{ech}, un intervalle dans lequel il est vraisemblable que la vraie valeur de σ^2 s'y trouve,

On cherche un intervalle $[A, B]$ vérifiant : $P(A \leq \sigma^2 \leq B) = S$.

Les limites de cet intervalle prendront, après avoir prélevé l'échantillon et calculé l'estimation les valeurs prises par les deux quantités aléatoires A et B, la forme suivante :

$$a \leq \sigma^2 \leq b.$$

Nous allons déterminer A et B en utilisant la distribution d'échantillonnage de la variance d'échantillon S

Bibliographie

[1] Dagnelie P., Statistique théorique et appliquée, 2ème edition, De Boeck Université, 2007.

[2] Morgenthaler S., Introduction à la statistique, 3ème édition, Presses polytechniques et universitaires romandes, 2007.

[3] Saporta G., Probabilités, analyse de données et statistique, 2ème édition, Technip, 2006.