ETUDE DE LA DISTRIBUTION D'UNE VARIABLE

G. TIXIER

IFFA MERIEUX, 254 rue Marcel Mérieux - 69007 LYON

RESUME

A l'occasion de l'étude des paramètres de la distribution d'une variable, nous insistons sur :

- *la distinction entre écart-type et erreur standard qui caractérisent respectivement la dispersion des individus et la précision sur leur moyenne.*
- *la correspondance entre les deux modes d'expression des paramètres d'une distribution log-normale (arithmétique et logarithmique).*

L'application des statistiques aux études épidémiologiques a facilité grandement l'interprétation des différentes enquêtes effectuées dans ce domaine et permis d'améliorer ainsi les connaissances sur l'évolution et les prévisions d'évolution des maladies contagieuses. Toutefois, dans cette discipline comme dans d'autres, il importe que les auteurs parlent le même langage. Or, il apparaît que des confusions s'établissent :

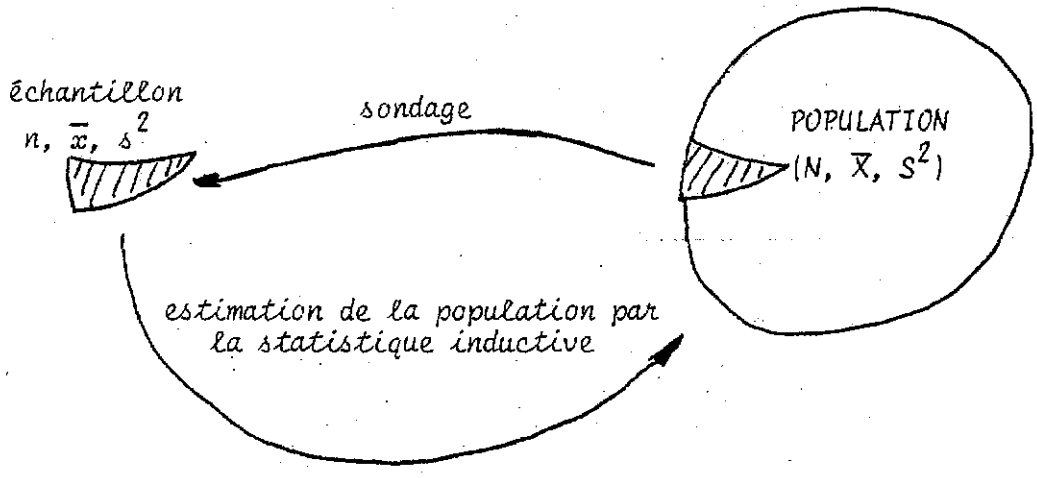
- d'une part sur la signification et l'utilisation d'un certain nombre de données à travers les termes qui les désignent, en particulier "l'écart type" et "l'erreur standard" dont les traductions "franglaises" ne simplifient pas la distinction.
- d'autre part, à cause de l'utilisation de modes d'expression différents, arithmétique et logarithmique, au sein d'une même étude.

C'est pourquoi, il nous a semblé intéressant d'apporter quelques éclaircissements à ce sujet, accompagnés d'exemples significatifs pour éviter des erreurs d'interprétation aux biologistes utilisateurs de la méthode statistique.

Il importe pour commencer de faire un bref rappel de statistique descriptive et inductive avant d'aborder :

- d'une part, le choix à faire entre écart type et erreur standard,
- d'autre part, le cas des transformations de variables.

Quand on doit étudier la distribution d'une caractéristique propre à un groupe de sujets, le plus souvent ce groupe ne doit être considéré que comme un échantillon que l'on espère représentatif d'une population trop grande pour être appréhendée dans sa totalité. L'étude de l'échantillon est du domaine de la statistique descriptive, mais c'est grâce à la statistique inductive que l'on acquiert des renseignements sur la population inconnue à partir de l'échantillon.



Pour favoriser cette distinction dans le développement qui suit, les symboles des paramètres relatifs à l'échantillon seront écrits en caractères minuscules et les caractères majuscules seront réservés à ce qui à trait à la population.

I - RAPPEL DE STATISTIQUE DESCRIPTIVE

Il s'agit de préciser le contenu d'un certain nombre de termes d'usage courant désignant différents paramètres utilisés en statistique.

1. VARIABLE (en anglais: variable)

C'est la caractéristique observée dans l'échantillon.

Trois types de variables sont à distinguer :

- *les variables qualitatives* qui ne sont pas susceptibles de mesure mais de classement (sexe, survie ou non d'un animal...),
- *les variables quantitatives discontinues* qui sont mesurables mais ne peuvent prendre que des valeurs entières (nombre de particules, d'animaux protégés...).
  Le diagramme en bâton est la représentation graphique de la distribution adaptée à ce type de variable, il représente l'effectif des individus pour chaque valeur.

- *les variables quantitatives continues* qui peuvent prendre toutes les valeurs numériques entières ou non dans leur intervalle de variation. La représentation graphique de la distribution de ces variables est l'histogramme représentant l'effectif des individus pour chaque intervalle de classe. L'histogramme cumulatif des effectifs est obtenu en traçant une courbe à partir de la somme, pour une classe donnée, des effectifs de toutes les classes antérieures, jusqu'à la classe considérée comprise.

2. LE MODE (en anglais:mode)

Ce dernier désigne la valeur ou la classe la plus fréquemment observée. S'il y a un seul mode, la distribution de la variable est dite "unimodale" ; dans le cas contraire, elle est dite "plurimodale" (bimodale s'il y a deux modes).

3. LA MEDIANE (en anglais:median)

Il s'agit de la valeur de l'observation qui partage la distribution en deux : 50 % des cas de part et d'autre de sa valeur. Elle ne provient pas d'un calcul, mais d'un classement, elle désigne un rang. La médiane a l'avantage d'être très peu influencée par un résultat isolé, très différent des autres

4. L'ETENDUE (en anglais:range)

C'est l'écart entre la plus petite et la plus grande des valeurs observées. C'est une notion de dispersion.

5. LA FONCTION DE REPARTITION (en anglais:distribution function)

Dans la grande majorité des cas, les distributions que l'on rencontre en biologie peuvent être assimilées à trois lois de probabilité :

- la loi de Laplace-Gauss ou loi normale,
- la loi log-Laplacienne ou loi log normale,
- la loi de Poisson.

Nous ne retiendrons ici que les deux premières formes de distribution.

a) La distribution de Laplace-Gauss ou normale (en anglais:normal distribution) se traduit graphiquement par une courbe en cloche et symétrique. Une distribution est considérée comme gaussienne lorsque en représentant les fréquences relatives cumulées en fonction des valeurs de la variable dans son étendue sur un papier graphique où sont disposées l'échelle gaussienne en ordonnée et l'échelle millimétrée en abscisse, on obtient une série de points alignés de façon franche.

Il faut alors confirmer le caractère normal de cette distribution ; c'est le but des tests de normalité dont le plus simple est l'ajustement de ces points par le tracé graphique de la droite de Henry.

b) La forme de distribution log-Laplacienne ou log normale n'est pas symétrique à cause de son allure "étirée" vers les grandes valeurs.

Le fait qu'au sein d'une population la variance à l'intérieur d'un groupe de sujets est proportionnelle au carré de la moyenne est une caractéristique de la distribution log-normale. La transformation log des valeurs observées ramène celle-ci à une distribution normale et en même temps égalise les variances. On satisfait ainsi aux conditions requises pour effectuer des tests paramètriques.

6. LA MOYENNE (en anglais:mean)

C'est un paramètre de position.

Selon que l'on s'adresse à une distribution normale ou log normale, on calculera une moyenne arithmétique ou géométrique. Le terme "moyenne", désigne, sauf indication contraire, la moyenne arithmétique.

La moyenne arithmétique (en anglais:arithmetic mean) est le quotient de la somme des observations par leur nombre :

$$\bar{x} = \frac{n1\ x1 + n2\ x2 + \ldots\ n4\ x4}{n} \quad \text{où} \quad \bar{x} = \frac{\Sigma\ ni\ xi}{\Sigma ni}$$

où ni : est l'effectif de la valeur particulière xi, ou de la classe de centre xi

n : est l'effectif total.

Elle s'applique bien à une distribution normale mais l'information qu'elle fournit est peu significative si la distribution est trop dispersée et/ou trop dissymétrique (il paraît qu'un statisticien ne sachant pas nager s'est noyé dans une rivière d'une profondeur moyenne de un mètre !...).

La moyenne géométrique (en anglais:geometric mean) est aux distributions log-normales ce que la moyenne arithmétique est aux distributions normales. C'est par définition, dans le cas de n observations strictement positives ou nulles, la racine nième de leur produit. De manière plus pratique, c'est l'anti-log de la moyenne arithmétique des valeurs transformées en log.

Exemple : calculer la moyenne géométrique de 2, 4, 4, 8 et 16 :

$$\bar{x}g = \sqrt[5]{2 \times 4 \times 4 \times 8 \times 16}$$

ou de manière plus pratique :

$$\bar{x}\log = \frac{\log_{10}2 + 2\ \log_{10}4 + \log_{10}8 + \log_{10}16}{5} = \frac{3,6}{5} = 0,72$$

0,72 est la moyenne des valeurs log,

$10^{0,72} = 5,28$ est la moyenne géométrique.

La moyenne géométrique a pour but de prendre en compte la dissymétrie ; son calcul a pour conséquence la réduction des fortes valeurs.

7. LA VARIANCE (en anglais:variance)

La variance est un paramètre de dispersion. C'est le quotient de la somme des carrés des écarts à la moyenne, par l'effectif -1 de l'échantillon.

$$s^2 = \frac{\sum_{i=1}^{i=n} (xi - \bar{x})^2}{n - 1}$$

En pratique, la variance se calcule plus aisément ainsi :

$$s^2 = \frac{T2 - T_1^2/n}{n - 1}$$

Avec T1 : somme des valeurs observées,
T2 : somme des carrés des valeurs observées.

8. L'ECART TYPE (en anglais:standard deviation)

La variance étant un carré, n'est pas interprétable directement. Par contre, l'écart type qui est la racine carrée de la variance, est du même ordre d'unité que la moyenne et constitue un paramètre de dispersion.

$$et. = \sqrt{s^2}$$

L'écart type (comme la variance) n'a pas de relation de proportionnalité avec l'effectif de l'échantillon.

On utilise également comme critère de dispersion le coefficient de variation: $et/\bar{x}$, qui s'exprime en pourcentage.

II - STATISTIQUE INDUCTIVE OU LA NOTION DE POPULATION ET D'ECHANTILLON

La connaissance des principaux paramètres de la distribution d'une variable au sein d'un *échantillon* nous permet ensuite, par la technique de la statistique inductive, d'acquérir des renseignements sur la *population inconnue*. La moyenne et la variance des valeurs d'un échantillon ne sont que des estimations des vrais paramètres de la population inconnue ; aussi, toute extrapolation de l'échantillon à l'ensemble de la population sera associée à une probabilité. Dans ces conditions, on démontre que :

- dans l'ignorance où l'on est de la vraie valeur de la moyenne de la population : $\overline{X}$, on montre que la meilleure estimation que l'on puisse faire de $\overline{X}$ est $\bar{x}$; on note $\overline{X} \neq \bar{x}$.

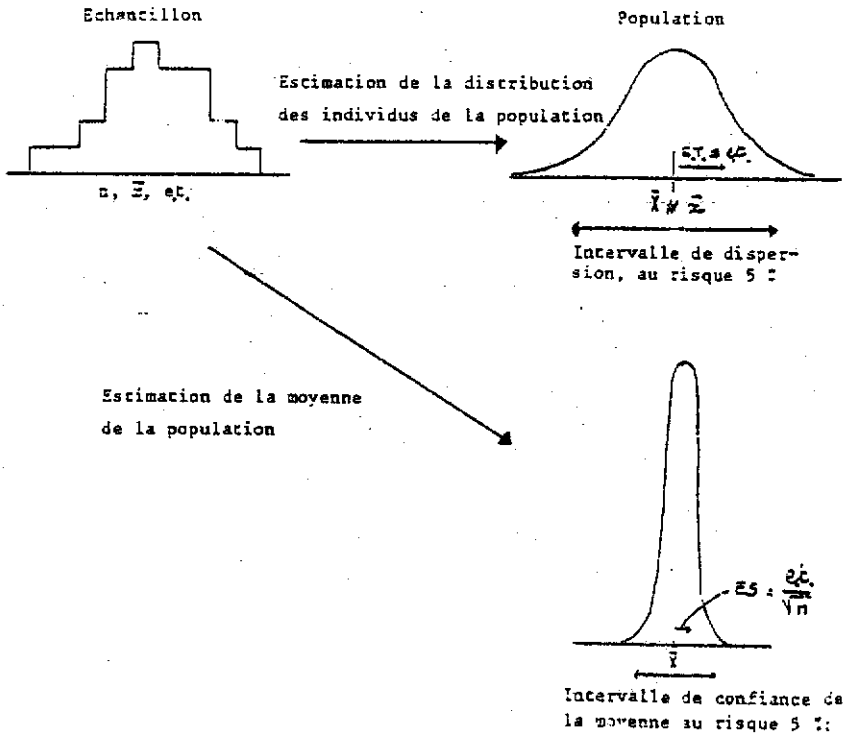
  Il en est de même pour la variance et l'écart-type de la population ; ces deux paramètres sont estimés à partir de l'échantillon :

  $S^2 \neq s^2$ et $ET. \neq et.$

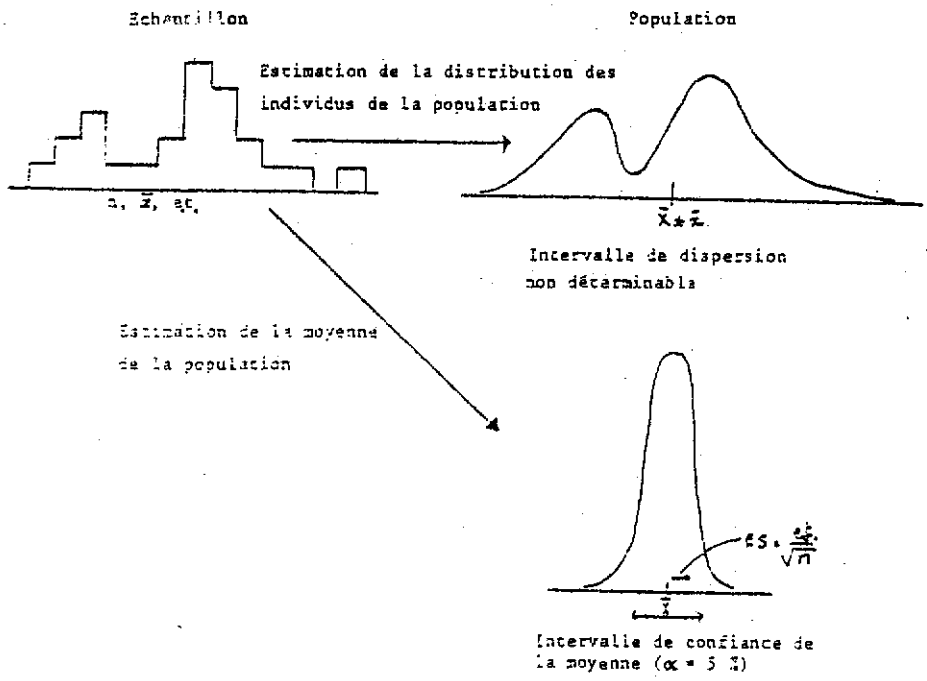
DE L'ECHANTILLON A LA POPULATION

DEUX EXEMPLES DE STATISTIQUE INDUCTIVE

A. L'HYPOTHESE DE NORMALITE DE LA DISTRIBUTION DE L'ECHANTILLON EST ACCEPTEE



B. L'HYPOTHESE DE NORMALITE DE LA DISTRIBUTION DE L'ECHANTILLON EST REFUSEE



- si la distribution de la population est assimilable à une loi normale (directement ou après transformation) on peut déterminer un intervalle de dispersion statistique des individus au sein de la population, égal à $\bar{x} \pm t.et$. La valeur de t est donnée par une table de t de Student en fonction du risque α et de la taille de l'échantillon.

  Pour un risque de 5 % et un échantillon au moins égal à 20, t vaut 2 ; c'est-à-dire qu'un point appartenant à la population d'où est tiré l'échantillon a 95 % de chances de se situer à l'intérieur de cet intervalle et le risque de se situer à l'extérieur est de 5 %. Avec cette même taille d'échantillon, cette probabilité est de 68 % (environ 2/3) avec un intervalle de ± 1et autour de la moyenne.

Nous avons dit que la moyenne vraie de la population est estimée par celle de l'échantillon mais avec une certaine incertitude. Cette dernière correspond à la distribution des moyennes de plusieurs échantillons représentatifs de la population. *Cette distribution des moyennes des échantillons suit une loi normale et ce, même si la distribution des individus suit une toute autre loi.*

L'erreur standard (en anglais: standard error) est l'écart type de la distribution des moyennes de chacun des échantillons représentatifs de la population.

Elle peut se calculer à partir d'un échantillon en divisant l'écart-type de ses individus par la racine carrée de son effectif.

$$ES = \sqrt{\frac{s^2}{n}} \quad \text{ou} \quad \frac{et.}{\sqrt{n}}$$

Ainsi, l'erreur standard caractérise *l'imprécision* avec laquelle on connaît la vraie moyenne de la population inconnue d'où est extrait l'échantillon

$\bar{x} \pm t.ES$ détermine l'intervalle de confiance de la moyenne au risque α bilatéral choisi par la valeur du t.

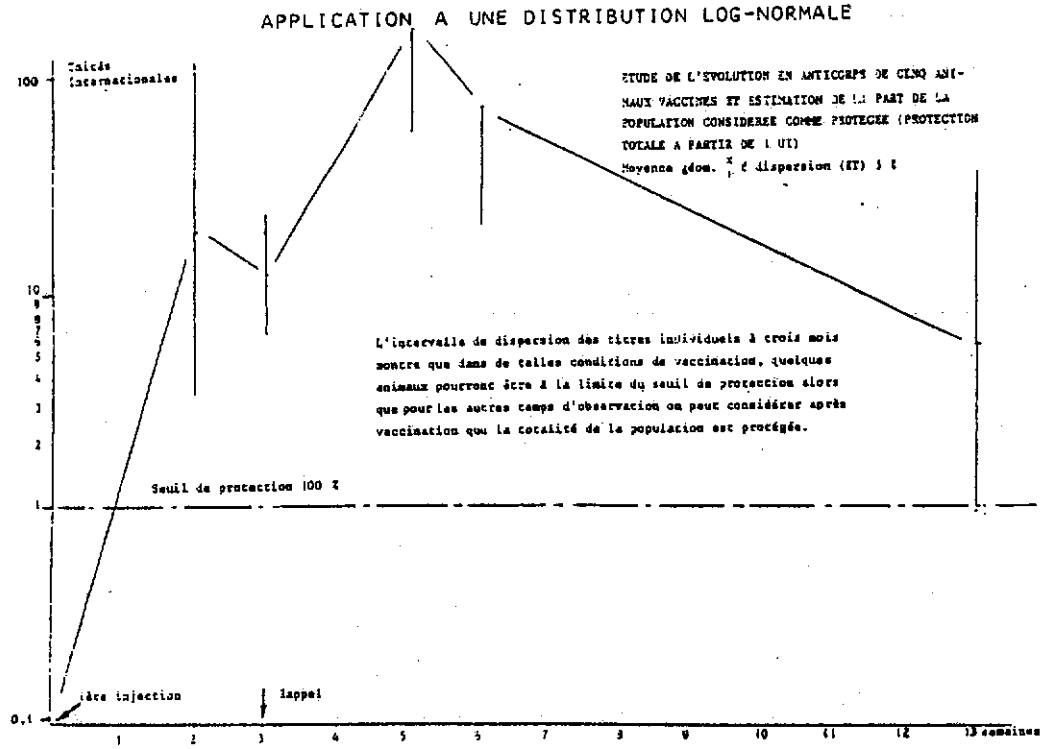
La vraie moyenne se situe à l'intérieur de cet intervalle avec une probabilité de $1-\alpha$.

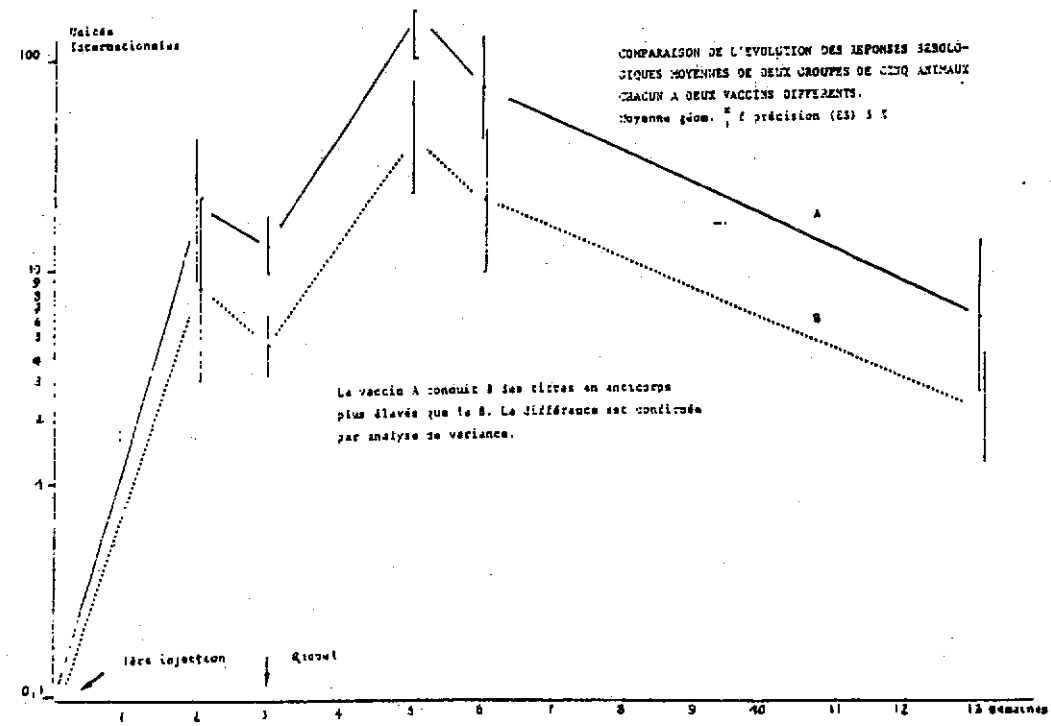
L'intervalle de confiance (ou fourchettes encadrant la vraie valeur) *est d'autant plus réduit* que la dispersion des points est faible et, (par opposition à l'écart type) *que la taille de l'échantillon sur lequel est basée l'estimation est élevée.*

III - CHOIX ENTRE ECART-TYPE ET ERREUR STANDARD

Une des raisons de la confusion que l'on rencontre entre ces deux paramètres provient des termes utilisés dans la littérature tant française qu'anglaise. Ce que nous appelons ici écart type ou écart type des individus se nomme standard deviation en anglais. Quant à notre erreur standard, on la trouve également sous le nom d'écart type de la moyenne et en anglais : standard error.

CHOIX ENTRE L'ECART TYPE ET L'ERREUR STANDARD
APPLICATION A UNE DISTRIBUTION LOG-NORMALE





Il est capital de bien comprendre la distinction que nous avons faite entre la distribution des individus et celle des moyennes pour choisir à bon escient le paramètre (écart type ou erreur standard) qui accompagnera la moyenne, cette dernière étant figurée dans un tableau de résultats ou sur un graphique.

Rappelons que :

- la distribution des individus est estimée par l'écart type, paramètre de dispersion, indépendant de l'effectif.

- l'intervalle de précision de la moyenne vraie de la population est déterminé à partir d'un échantillon grâce à l'erreur standard. Cette fourchette autour de la moyenne est d'autant plus étroite que l'effectif de l'échantillon est élevé.

Le choix entre écart type et erreur standard dépend de l'intention de l'auteur.
Un exemple illustrera ceci : étude des titres en anticorps à un temps donné, chez plusieurs groupes d'animaux soumis respectivement à des vaccins différents. On vérifie auparavant que le seul facteur différenciant les groupes est bien le vaccin, autrement dit qu'il n'y a pas de biais.

Si l'on veut comparer les résultats d'un groupe par rapport à une norme ou à une valeur seuil garantissant 100 % de protection par exemple, on s'intéressera tout d'abord à la position de la moyenne des titres par rapport à cette valeur. Mais cela ne suffit pas et l'on cherchera à savoir si tous les animaux de la population ou une partie seulement atteignent cette norme ou ce seuil ; la réponse sera apportée par l'intervalle de dispersion, basé sur l'écart type.

Autrement dit, on s'intéressera autant à la position de la moyenne des titres qu'à la dispersion de ceux-ci à l'intérieur du groupe. Une bonne vaccination sera celle qui induit une moyenne de titres, élevée certes, mais aussi la plus grande homogénéité des réponses (écart type faible).

Par contre, si la comparaison des réponses sérologiques aux différents vaccins est le but de l'expérience, on s'intéressera préférentiellement à l'imprécision avec laquelle on connaît la moyenne des titres de chaque groupe. Plus elle est réduite, plus l'expérience sera puissante*, c'est-à-dire plus l'aptitude à différencier deux moyennes sera grande. Dans ce cas, chaque moyenne sera accompagnée de son intervalle de précision, basé sur l'erreur standard.

*La puissance d'une expérience caractérise son aptitude à différencier les groupes que l'on compare. Ce pouvoir séparateur peut être comparé à celui d'un microscope où 2 points sont confondus avec une certaine puissance de grossissement et sont visibles séparément avec un appareil de plus forte puissance. Mais le prix du microscope sera alors plus élevé.

L'imprécision sur la moyenne est un paramètre peu utilisé dans la littérature. On rencontre plus souvent l'intervalle de dispersion ou l'écart type qui conviennent très bien si l'on veut mettre en évidence l'étendue des valeurs individuelles autour d'un point moyen, mais sûrement pas lors de comparaison de plusieurs moyennes où, dans ce cas, c'est l'incertitude sur la position réelle du point moyen donnée par l'erreur standard qui nous intéresse. Mais, quel que soit le choix effectué, il est impératif de le signifier dans la légende des tableaux ou des graphiques ainsi que de préciser l'effectif des individus à l'origine.

IV - LES TRANSFORMATIONS DE VARIABLES

L'autre confusion que l'on peut rencontrer dans la littérature provient de la nécessité où l'on se trouve d'avoir à passer de l'expression arithmétique à l'expression logarithmique pour effectuer des tests statistiques (comparaison de moyennes, régressions...) sur des variables dont la distribution est log normale. Le même problème se pose inversement, pour revenir à la forme arithmétique, pour une meilleure compréhension des résultats.

Prenons par exemple cinq animaux dont on titre les anticorps, exprimés en Unités Internationales. Sachant que la distribution des titres en anticorps des animaux est de la forme log normale pour des valeurs observées de 2, 4,4, 8 et 16 UI.

La moyenne des valeurs log sera alors égale à :

$$\frac{0,3 + 0,6 + 0,6 + 0,9 + 1,2}{5} = 0,72 \text{ log UI}$$

la variance log, à 0,1178;
l'écart type, à 0,34 log UI, $(\sqrt{0,1178})$
l'erreur standard, à 0,15 log UI $(\sqrt{0,1178/5})$

Il est fort possible d'exprimer ces résultats en log UI et ainsi utiliser des termes statistiques bien connus.

Toutefois, si le lecteur à l'habitude de raisonner en Unités Internationales, il risque d'être dérouté devant la notion de log UI, aussi est-il parfois nécessaire de revenir aux valeurs arithmétiques en utilisant pour cela les trois paramètres suivants :

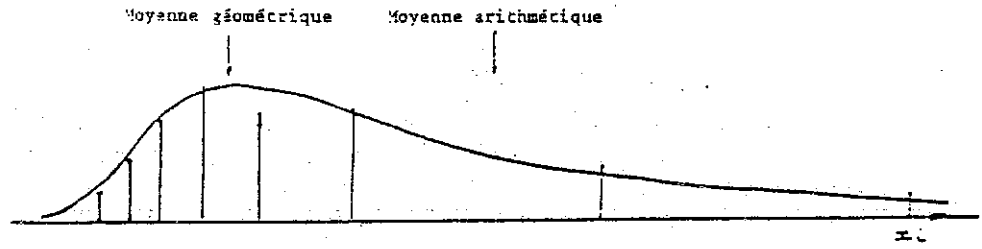
- *la moyenne géométrique* : $10^{\overline{x}\ \log}$

  Ici égale à $10^{0,72} = 5,28$ UI.

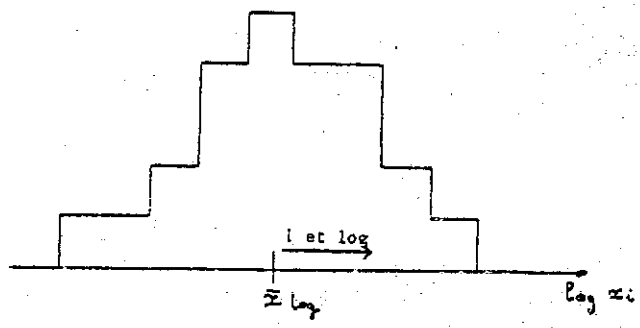
- *le facteur de dispersion* au risque 5 % bilatéral (Fd 5 %) qui est égal à $10^{t.et.}$ où t est la valeur du t de student pour n - 1 degrés de liberté et et l'écart type de la distribution des valeurs log. C'est un facteur *multiplicatif* qui permet de déterminer *l'intervalle de dispersion* à 5 % par la formule : moyenne géométrique $\overset{\times}{\div}$ Fd 5 % ou autrement dit Fd 5 % fois en plus et en moins de la moyenne géométrique.

APPLICATION A UNE DISTRIBUTION LOG-NORMALE

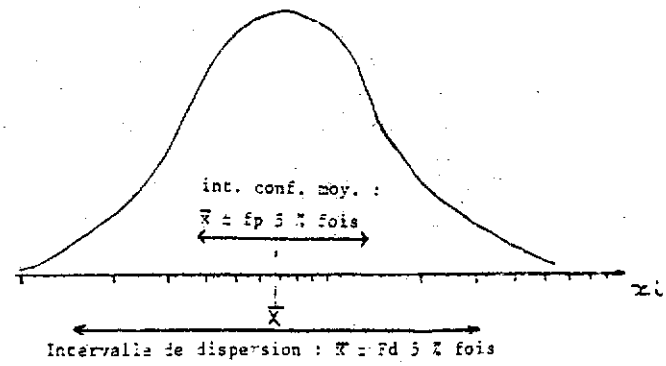
1. REPRESENTATION DE L'HISTOGRAMME DE L'ECHANTILLON ET DE LA DISTRIBUTION DE LA POPULATION SUR PAPIER MILLIMETRE



2. REPRESENTATION DE L'HISTOGRAMME DE L'ECHANTILLON SUR PAPIER MILLIMETRE ET APRES TRANSFORMATION LOG DES VALEURS



3. REPRESENTATION DE LA DISTRIBUTION ESTIMEE DE LA POPULATION ET INTERVALLES DE DISPERSION DES VALEURS ET DE CONFIANCE DE LA MOYENNE, AU RISQUE 5 %.
EXPRESSION EN VALEUR ARITHMETIQUE D'ORIGINE GRACE A L'EMPLOI DE PAPIER A ECHELLE LOG.



Dans notre exemple, l'écart type log est égal à $\sqrt{0,1178} = 0,34$ exprimé en log UI, le degré de liberté est égal à 4 et le t de Student correspondant est 2,8 au risque 5 %.

Fd 5 % = $10^{2,8\ 0,34} = 8,9$

et l'intervalle de dispersion qui en découle vaut :

$5,23 \overset{\times}{\div} 8,9$ soit 0,6 — 47 UI.

- *le facteur de précision de la moyenne* au risque 5 % bilatéral (Fp 5 %) est égal à $10^{t.ES}$, où ES est l'erreur standard, log.

Dans notre exemple, l'erreur standard log est égale à

$$\sqrt{\frac{0,1178}{5}} = 0,15 \text{ UI}$$

Fp 5 % = $10^{2,8.0,15} = 2,6$

L'intervalle de précision correspondant est égal à :

$5,28 \overset{\times}{\div} 2,6$ soit 2,0 — 14 UI.

La méconnaissance ou la non-utilisation des notions de facteur de dispersion et facteur de précision contribue à créer des erreurs ou des confusions, comme le calcul d'une moyenne arithmétique des valeurs dont la distribution est log-normale ou l'expression d'une moyenne géométrique associée à un écart type log. Ces paramètres ayant été développés ici, nous insistons sur l'importance d'*exprimer* l'ensemble des résultats soit dans le mode log, soit dans l'arithmétique pour assurer une *cohérence indispensable à la compréhension.*

Il est certain que tout ce qui vient d'être développé peut paraître banal aux yeux d'un statisticien confirmé. Mais parce que la statistique est devenue maintenant un outil de travail fréquemment utilisé dans de nombreuses disciplines en Biologie, dont l'Epidémiologie, il est important de souligner les pièges tendus par les mots ou par des applications à des cas particuliers. Nous souhaitons que cette contribution apporte une meilleure compréhension pour une meilleure utilisation de la statistique en Biologie.

BIBLIOGRAPHIE

AFNOR - Recueil des normes françaises de la statistique. Tome 1 - Vocabulaire, estimation et tests statistiques. AFNOR, Paris, 1970.

GREMY F. et SALMON D. - Bases statistiques pour la recherche médicale et biologique. Dunod, Paris, 1969.

MARTIN J. - Notions de base en Mathématiques et Statistiques à l'usage des Biologistes, Médecins et Pharmaciens. Gauthier-Villars, Paris, 1967.

SCHWARTZ D. - Méthodes statistiques à l'usage des médecins et des biologistes. Flammarion, Paris, 1970.

Nous remercions le Docteur Vétérinaire CAMAND pour son aide à la rédaction de ce document.

°°°