Année universitaire : 19/20
Promotion : M2 Production et nutrition animale
Module : Analyse des données statistique en sciences animales
Cours préparés par : M. HICHER Azzeddine-UMKB-Algérie
****

Cours 1 : Notions de base en Statistique

Objectifs cibles : -

- Apprendre certaines vocabulaires de la statistique ;
- Savoir les causes de la variabilité de mesure en science biologique ;
- Pouvoir catégoriser les différents types de variables ;
- Maitriser la représentation graphique des variables.

1. Statistiques ou Statistiques

StatistiqueS : (latin « status » état) – Ensemble cohérent de données numériques relatives à un groupe d'individus. Autrement dit, elles désignent des grandeurs (généralement numériques) que l'on calcule, ou que l'on est capable de calculer ,sur un ensemble de données observées. Exemple
– Statistiques démographiques – Statistiques du chômage – Statistiques de santé »
Etat de santé de la population » Activité : Statistiques d'activité hospitalière

A contrario, au singulier le

StatistiquE: est le nom de la science qui étudie ces grandeurs et propose des outils pour les concevoir. Elle désigne à la fois un ensemble de données observées et les méthodes de recueil, de traitement et d'analyse de celles-ci. En d'autre terme, c'est un Ensemble des méthodes qui permettent de rassembler et d'analyser les données numériques.

* Méthodes de mesures, d'échantillonnage, de présentation des résultats, de modélisation ….

* Calcul du Paramètre tel que moyenne... calculé à partir d'un ensemble de données.

2. Statistique et probabilité

La théorie des probabilités joue un rôle important en statistique car elle permet de modéliser certains phénomènes aléatoires, c'est-à-dire des expériences dont le résultat ne peut pas être prévu avec une totale certitude (généralement la déduction à partir d'un échantillon un jugement sur toute une population).

** L'intuition nous amène à penser que certains phénomènes obéissent à certaines lois probabilistes ;

** les valeurs observées d'un phénomène biologique seront concentrées autour d'une certaine «valeur moyenne». On considère assez souvent que ces valeurs observées se distribuent selon un certain modèle, une certaine loi, par exemple la loi normale. Cette hypothèse peut être confortée par un test d'ajustement. La statistique permet de confronter les modèles probabilistes avec la réalité observée afin de les valider ou de les invalider

À partir de 1843, la statistique désigne l'ensemble de techniques d'interprétation mathématique appliquées à des phénomènes pour lesquels une étude exhaustive de tous les facteurs est impossible, à cause de leur grand nombre ou de leur complexité. Les statistiques s'appuient sur les probabilités et sur la loi des grands nombres. La statistique vise à décrire, à résumer et à interpréter des phénomènes dont le caractère essentiel est la variabilité. Elle fournit de la manière la plus rigoureuse possible des éléments d'appréciation utiles à l'explication ou à la prévision de ces phénomènes.

3. Les données Biologiques

Les données numériques d'un phénomène biologique sont souvent très dispersées (variabilité dans ma mesure+ variabilité intra individu+ inter individu). Du fait de la variabilité, on est dans le domaine de l'incertain. Cette science de l'incertain, c'est le défi qu'a relevé la statistique en s'appuyant sur le concept de probabilité. On a besoin d'évaluer la signification réelle des ces valeurs et de les représenter d'une manière à ce qu'elles communiquent facilement leurs significations les unes pour les autres. C'est-à-dire, on a besoin de savoir si :

** ces valeurs se diffèrent l'une de l'autre sous effet du facteur étudié (traitement) ;

** ou cette variation est de nature biologique

4. La démarche générale d'une étude statistique (biométrique)

Toute étude statistique peut être décomposée en deux phases au moins : le recueil ou la collecte des données statistiques, et leur analyse ou leur interprétation.

4.1. Le recueil des données

Le recueil des données peut être réalisé soit par la simple observation des phénomènes (enquête), soit par l'expérimentation, c'est-à-dire en provoquant volontairement l'apparition de certains phénomènes contrôlés.

Exemple : le rôle de quelques substances (N, P, K) dans la production de biomasse chez les végétaux.

Lorsque les données sont très nombreuses, ou particulièrement difficiles à obtenir, il sera nécessaire pour la mise en œuvre rationnelle du recueil de définir des méthodes appropriées de collecte. Il s'agira de plans d'échantillonnage ou de plans d'expérience dont la mise en œuvre sera fonction du type de problème que l'on est amené à résoudre. Exemple : la numération des mammifères d'une aire protégée : inventaire et recensement.

4.2. 2. L'analyse et l'interprétation des données

L'analyse statistique se subdivise en deux étapes

- **La statistique déductive ou descriptive** : elle a pour but de résumer et de présenter les données observées sous la forme la plus accessible (simplification et réduction des données, à la fois visuelle et conceptuelle). Autrement dit, nous utilisons ces techniques pour réduire un ensemble de données à des proportions gérables, en résumant les tendances et tendance, afin de représenter clairement les résultats. À partir de ces procédures, nous pouvons produire des diagrammes, des tableaux et des descripteurs numériques.

Les descripteurs numériques incluent des mesures indiquant le centre de l'ensemble de données, telles que la moyenne ou la moyenne arithmétique, et des mesures de la dispersion ou de la dispersion des données, telles que la variance ou l'amplitude.

- **L'analyse inductive ou inférence statistique** est l'ensemble des méthodes permettant de formuler en termes probabilistes un jugement sur une population, à partir des résultats observés sur un échantillon extrait au hasard de cette population. Les méthodes statistiques les plus classiques sont celles de l'estimation (estimation par domaine de confiance). Un paramètre, tel que la moyenne ou la proportion, décrit une caractéristique particulière de la distribution d'une variable dans l'ensemble de la population. Habituellement, l'estimation est suivie d'une procédure appelée test d'hypothèse, un autre aspect de la statistique inférentielle qui étudie une théorie particulière des données. Les tests d'hypothèse permettent de tirer des conclusions relatives à la population à partir des informations contenues dans un échantillon.

5. La variation de la biométrie

Il bien connu en biologie que si on répète la mesure sur un phénomène quelconque. L'obtention d'une mesure identique c'est rarement possible. Un part de cette variabilité de mesure est du à la variation biologique de l'individu sujette de cette mesure (génétique, environnement, sexe, âge …). La sélection des individus aussi homogène que possible permet de réduire ce genre de variation biologique. L'autre part qui influence la valeur observé est de nature technique (erreur technique)

qui représente l'écart entre la valeur réelle et celle observée. L'erreur technique peut etre d'origine humaine (expérimentateur) ou instrumental.

6. types de variables

La variable est une caractéristique qui peut varier d'un individu à l'autre, d'un groupe à l'autre, d'un moment à l'autre. Certaines variables sont **qualitatives**, s'exprimant par l'appartenance à une catégorie : par exemple,

D'autres variables sont **quantitatives** (ou : **numériques**). Par exemple la taille, le poids, le volume, la durée de vie sont des variables quantitatives. Une variable quantitative est qualifiée de **discrète** dans le cas où l'on observe un nombre fini ou infini dénombrable de valeurs. Si on note n_i le nombre d'occurrences de x_i dans toute la population, et N la taille de la population, alors la fréquence correspondante est $f_i = n_i / N$.

Diagrammes en bâtons (ou à barres) : l'effectif ou la fréquence correspondant à chaque valeur du caractère est représenté par la longueur d'un segment ou d'un rectangle de largeur constante. La représentation de plusieurs séries de données sur un même graphique peut se faire en empilant les barres.

Diagramme circulaire : chaque valeur ou classe est représentée par un secteur angulaire d'un disque dont l'angle (et donc la surface) est proportionnel à sa fréquence.

Un caractère est dit **continu** lorsque les valeurs qu'il prend constituent un intervalle de R. Dans ce cas, il est fréquent de diviser la population en classes selon les intervalles de valeurs prises par le caractère. Ce procédé est parfois appelé **discrétisation** de la variable. Dans ce cas, on regroupe les valeurs observées en k classes; et on note pour chaque classe $[e_{i-1}, e_i[$ l'effectif n_i et la fréquence f_i , ainsi que les fréquences cumulées $F_i = \sum_{j=1}^{i} f_i$. On peut alors remarquer que Fi est la proportion d'individus pour lesquels $X < ei$.

On peut représenter cette série de données par un **histogramme** : chaque classe est représentée par un rectangle dont l'aire est proportionnelle à l'effectif.