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 2 : Statistique Descriptive

Objectifs :

- Citer les différentes mesures de position et identifier leurs pouvoirs et limites statistiques ;
- Citer les différentes mesures de dispersion et identifier leurs pouvoirs et limites ;
- Synthétiser les data en tableaux et en figures

A). Introduction

Les données sont classées en 2 catédories distenctes :

- Données métriques (quantitatives continues) sont analysées par utilisation des tests statistiques paramétriques comme student, fisher, Z., coefficient de corrélation…. Ces test sont plus robustes et exige des conditions pour etre plus satisfaisants.
- Données non métriques (qualitatives et quantitative discrètes) : à l'aide des tests non paramétrique, ces données sont analysées (khé deux, Mann-Whiteny test, Kruskall-Wallis test. Ces tests sont plus flexibles et facile à utiliser

L'objectif global de la SD est de résumer et synthétiser les données d'une expérience afin d'avoir une image globale plus claire et illustrative (un profil global). La qualité de cette image étant en relation étroite avec la quantité de ces données recueillis. Pour obtention de cette image de synthèse on doit:
- Organiser les data pour savoir à quelle fréquence les différents valeurs se produisent ;
- Condenser les informations en les réduisant à une taille gérable ;
- Obtenir une vue instantanée qui aide à la compréhension et à l'iterprétation.

B). Statégies de SD : on peut utiliser :

- ✓ **Tableaux (tables) :** pour présebter les caractéristiques des données ;
- ✓ **Figures (diagrams) :** pour illuster les motifs ;
- ✓ **Mesures numériques :** pour résumer les données.

La nature de la variable impose la figure et la mesure statisque les plus appropriées. Si la variable est de nature **qualitative (couleur de la robe chez l'espèce bovine)** ou quantitative **discrète(taille de la porté chez les ovins)**, on présente les données dans des tableaux ou dans des figures comme digramme en batton (bar chart) ou en secteur(pie chart , cammembert). La fréquence relative et cummulative sont pour ce cas les mesures les plus utilisées. On préfère le histogrammeet ou la boite à moustache pour la variable quantitative continu (hauteur au garrot en cm chez les caprins).

C). Représentations graphiques :

I. Cas d'une variable qualitative

1. En barres (bar chart):

Chaque catégorie de la variable est représentée, la largeur de chaque barre doit etre égale tendisque la hauteur variée selon le nombre ou le % des individus appartiennent à chaque catégorie (proportionnelle aux effectifs ou ou frequence relative).

2. En secteurs (Pie chart) : cerle divisé en segments qui représentent les différentes catégorie d'une variable qualitative. Generalement , les barres sont plus claires et interprétables que les secteurs.

II. Cas d'une variable quantitative

1. Diagramme en points (dot diagram/plot) : pour une série de donnée ou chaque observation est représenté par un point sur une line verticla ou horizontale standardisé selon l' unité de mesures de la variable . comme il est utile pour comparer graphiquement entre deux (ou plus)groupes.
2. Histogrammes : il synthétise les données. L'axe des abscisses représente les unités de mesures de la variable. La surface de chaque rectangle, qui sont continuées, est proportionnelle des effectifis des classes. Si l'intervalle des classes ont des largeurs identiques, la hauteur sera proportionnelle aux fréquences des classes. Il peut prendre une forme symétrique (l'image gauche est la mirroire de l'image groite). Comme il peut dévier à droite ; avoir une queue à droite (positivement déviée : skewed to the right) or vers la gauche.
   NB* Les données biologique sont fréquement deviées à droite.
3. Boite à moustache (box and whisker plot or box plot)
   Une boite avec des limites horizontales définissent la quantille supérieures et inferieures . elle reptésente l'amplitite interquartille et rnferme les 50% des observation centrales ? le median est marqué par une ligne horizontale à l'intérieur. Les moustaches sont verticaux dont si le supérieur représente les données au dela de 97,5 percentiles, l'inférieur représente celle en bas de 2,5
4. Nuage des points (Scatter diagram) : une méthode plus efficace pour répresenter la relation entre deux variables numérique ou ordinale

D).

Mesures numériques

Il est essentiel de complémenter la visualition des données par des mesures numériques appropriées qui les synthétisent et les résument. Utile de distinguer entre les mésures sur populations (paramètres) notés en grec et les mesures sur échantillon (estimateurs ou les statistiques) notés en roman

1. Mesures de position (average : tendence centrale)

1.1. La moyenne arithmétique

- Influencé par les outliers (une observation hautement inconsistante avec le radical des données) ;
- Une mesure de tendence centrale la plus approprié pour les données symétriquement distribuées.

1.2. **Mediane** : une valeur centrale partage la sérien de données en deux portions agales (50% en bas et 50% au dessus (50 percentile) :

- = moyenne arithémique pour les données symétriques ; (< :droite, > : gauche))
- Non affecté par les outliers.

1.3. **Moyenne géométrique** (logarithme x)

On utilise le log x pour la transformation les données pour les rendre symétrique.

- Petit que la moyenne arithmétique et proche à la mediane ;
- Plus utile pour traiter les données biologiques fréquement dévié à droit.

1.4. **Mode** : l'observation la plus courante dans une série de données.

En bref :

Nature de la variable	**Mesure de position appropriée**
Données nominales	Mode
Données ordinales	Mediane
Variable continue symétriquement distribué	Moyenne
Variable continue asymétriquement distribu	Mediane

2. Mesures de dispersion (spread)

2.1. Variance

La moyenne arithmétique des écart à la moyenne est une mesure de consisstance des données. Plus elle est grande plus les données sont hétérogènes (s^2 : estimateur de la variance de la population)

2.2. Ecart type (SD)

Plus les données sont étroitement regroupées plus SD est de faible valeur.

- Mesuré pour une variable quantitative continue,
- Très utilisés lors de la distribution gaussienne. Dans la cas pareil, Moyenne±2SD englobe les 95% des données. 4SD indique l'amplitude de la majorité des données.

2.3. **Rang (Amplitude)**

La différence ente le max et mim observation.

2.4. Rang interquartille

Amplitude des valeurs qui englobe 50% centrales des données (Q3-Q1)

2.5. Coefficient de variation (CV)

Mesure la variabilité dans la série des données en relation avec leur moyenne.(variabilité relative). Il est utilisé lors de la comparaison entre deux groupes avec deux moyennes non identiques ou avec des unités de mesures différentes.

NB : L'intervalle de référence décrit l'amplitude des observations définissent la population saine (healthy population) (2,5 percentiles_97,5 percentiles).

3. Mesures décrivent la forme de la distribution

Skewness $ß_1$:

- mesure de la symétrie (=0 : allure symétrique, >0 : allure vers la droite : la majorité des observations sont infériures à la moyenne, la mediane est plus grand que le mode, vis versa à gauche),
- $ß_1$ > 2 SD : déparature de la symétrie. Il faudra de tester la sigbnificativité.

Kurtosis (aplitissement)

Mesure décrit la distribution des observations autour de la moyenne.
=0 : distribution normale >0 : données sont homogènes, <0 : données hétérogènes