

Cours statistiques :

Chapitre 1 :

Définition 1 : La statistique appelée également la science des données est la science de la collecte, de l'analyse, de la présentation et de l'interprétation des données.

Session 1 : Éléments fondamentaux d'une étude statistique

Définition 1 : Une population est l'ensemble des éléments sur lesquels une étude se porte.

Une population peut être d'une taille finie ou infinie

Définition 2 : L'unité statistique, ou l'élément, est l'unité par laquelle les données sont recueillies

Unité stat peut être de diff nat : une personne, un pays, un objet, ...

La pop est le vrai élément d'intérêt

Définition 3 : Un échantillon est un sous ensemble de la pop.

Ns notons la taille d'échantillon N ou T selon l'application

Définition 4 : L'échantillonnage aléatoire est une procédure utilisée pour sélectionner un échantillon de N individus d'une pop de telle manière que chaque individu de la pop est choisi strictement au hasard, la sélection d'un individu n'influence pas la sélection d'aucun autre individu, chaque individu de la pop est également susceptible d'être choisi, et chaque échantillon possible d'une taille donnée, N , a la même chance d'être sélectionné.

Définition 5 : Une variable statistique, ou simplement variable, est une caractéristique des éléments à laquelle on s'intéresse.

Ns utilisons les lettres minuscules de l'alphabet latin pour les var statistiques : x , y , z etc

Définition 6 : Une modalité est une val prise par une var statistique

Définition 7 : Une observation est un ensemble de mesures obtenues par chaque élément d'un ensemble de données.

Définition 8 : Un paramètre est une mesure num qui décrit une caractéristique spécifique d'une pop. Une statistique est une mesure num qui décrit une caractéristique d'un échantillon.

Définition 9 : L'erreur d'échantillonnage est la Dif entre paramètres d'intérêt et son estimation.

Définition 10 : La statistique descriptive se concentre sur les procédures graphiques et numériques utilisées pr résumer et traiter les données. L'inférence stat se concentre sur l'utilisation des données pr faire des prévisions, des estimations et tester les théories pr rendre de meilleures décisions.

Session 2 : Classification des variables statistique

Variables catégoriques :

Définition 11 : Une variable catégorique, ou qualitative, ou un facteur, est une variable statistique dont les modalités sont des catégories.

Les var cat classifient les unités stat, telles que chaque individu est mis ds une seule catégorie.

Remarque 1 : Une variable indicatrice est une var qui prend seulement 2 dft val : 0 et 1. Une var catégorique avec L cat peut être ré-exprimée à l'aide de L var indicatrices.

Définition 12 : Une var num ou quantitative est une var statistique dont les modalités sont mesurables. Une var num qui peut prendre soit un nbr fini de val, soit un ensemble infini de val dénombrables telles que 0, 1, 2, ... est dite var discrètes. Une var qui peut prendre ses val ds un intervalle ou une suite d'intervalles est apl var continue.

Session 3 :

Données en coupe transversale et séries temporelle :

Définition 13 : Les données en coupe transversale sont les données liées aux dft individus collectés au mm moment. Les séries temporelles sont des données collectées sur pls périodes de temps dfts.

Table 1: Croissance du PIB par tête et population pour 10 pays en 2020

Pays	Croissance du PIB (%)	Population
Allemagne	-4.65	83160871
Australie	-1.28	25693267
Belgique	-6.13	11544241
Bésil	-4.56	212559409
Colombie	-8.04	50882884
Éthiopie	3.40	114963583
France	-8.03	67379908
Grèce	-8.84	10700556
Italie	-8.60	59449527
Slovénie	-4.87	2102419

Résumé :

- 1- Éléments fondamentaux d'une étude statistique
 - > Pop, échantillon, échantillonnage aléatoire
 - > Paramètre, estimation, erreur d'échantillonnage
 - > Stat descriptives et inférence statistique
- 2- Classification des var statistique
 - > Var catégorique
 - > Var num
- 3- Types des données
 - > Données en coupe transversales
 - > Série temporelles

Session 4 : Représentation des var catégoriques

Représentation des var catégoriques :

Définition 14 : La fréquence d'une cat est le nbr d'unités statistique présentant cette dernière.

Soit x_i , $i = 1, 2, \dots, N$, une var catégoriques avec les cat C_I .

$I = 1, 2, \dots, L$.

Nous notons la fréquence de la cat I , n_i .

La fréquence de la cat C_I est le nbr d'unités statistique i telle que x_i appartient à C_I

Définition 15 : Une distribution de fréquence est un résumé des données sous forme de tableau décrivant la fréquence des observations ds dfts cat définie par une var.

Définition 16 : La fréquence d'une cat correspond à la pop des observation appartenant à cette cat :

$$\text{Fréquence relative s'une cat} = \text{Fréquence d'une cat} / N$$

Nous notons la fréquence relative de la classe $C_1 = F_1$

Nous avons alors : $f_i = n_i/N$

n_i : fréquence

F_1 : fréquence relative

N : somme fréquence chaque classe

I : indice classe

L : nombre de classe

$$f_i = n_i/N$$

Proposition 1 : La somme des fréquences relative est égale a $1 = \sum_{i=1}^L f_i$.

Session 5 : Représentation des val numériques

Tableau de fréquences :

Définition 17 : La fréquence relative cumulée d'une classe est la somme des fréquences de toutes les classes dont les observations sont $<$ ou $=$ à la borne sup de cette classe

Nous notons la fréquence relative de la classe I , F_I .

Nous avons $F_I = \sum$

Rappel :

$x_i \quad i=1, N$

$C_l \quad l=1, L$

$M_l \quad l=1, L$

Tableau de fréquence : (Classe C_l) intervalle des var num

$(a_1, b_1) \quad M_1$

$(a_2, b_2) \quad M_2$

$(a_l, b_l) \quad M_l$

Total N

Fréquence relative (f_i). $F_I = M_l/N \quad l=1, L$

f_1

f_2

f_l

Somme = 1

$$F_s = \sum_{s=1}^l f_s \quad l=1, L$$

$$F_1 = 1 \sum_{s=1} f_s = f_1$$

Exemple 6 : Les fréquences ne sont pas directement présentables pour une variable continue ou pour une variable discrète présentant beaucoup de modalités. Dans ce cas, il faut regrouper les données.

Table 4 : PIB:hab (1000\$) des pays de l'OCDE

Exercice 6 : Présenter la distribution du PIB par habitant (1000\$) des pays de l'OCDE en 2020 (Tableau 4) sous la forme d'un tableau de fréquences en regroupant les données en 5 classes.

C_l	n_l	f_l	$f_l(x100)$	F_l	$F_l(x100)$
(5,31 ;27,31)	16	0,42	42	$F_1=f_1$	42
(27,31 ;49,31)	12	0,32	32	0,74	74
(49,31 ;71,31)	7	0,18	18	0,92	92
(71,31 ;93,31)	2	0,05	5	0,97	97
(115,31 ;137,31)	1	0,03	3	1	100
Total	38	1	100	1	100

$$L = 5 \quad N = 38$$

$$f_1 = n_1/N = 16/38 = 0,42$$

$$f_2 = n_2/N = 12/38 = 0,32$$

$$F_2 = f_1 + f_2 = 0,42 + 0,32 = 0,74$$

$$F_3 = f_1 + f_2 + f_3 = 0,74 + 0,18 = 0,92$$

Session 6 : Représentation des données relatives à 2 variables

Définition 18 : Un tableau de contingence ou une tabulation croisée, est un résumé sous forme de tableau des données relative à 2 variables stat.

Supposons que nous voulons étudier la distribution des fréquences ds un échantillon de 2 variable x_i et y_i , $i = 1, 2, \dots, N$

$$C_l = 1, 2, \dots, L. \quad G_k \quad k=1, \dots, K$$

Tableau de contingence

C_l/G_k	G_1	G_2	G_k	Total
C_1	n_{11}	n_{12}	n_{1k}	$n_{1.}$
C_2	n_{21}	n_{22}	n_{2k}	$n_{2.}$
C_l	n_{L1}	n_{L2}	n_{lk}	$n_{l.}$
Total	$n_{.1}$	$n_{.2}$	$n_{.k}$	$n_{..} = N$

$n_{..}$ = nbr d'observation ds l'échantillon

Définition 19 : Une fréquence partielle n_{lk} $l=1, L. \quad k = 1, K$ est le nbr d'unité stat i telle que i appartient à C_l , y appartient à G_k . Autrement dit n_{lk} est le nbr d'unité stat présentant à la fois la classe l de x et la classe k de y .

Définition 20 : Une fréquence marginale de la variable x notée $n_{l.}$, $l=1, \dots, L$ est le nbr d'unité stat i telle que x_i appartient à C_l .

$$n_{l.} = \sum_{k=1}^K n_{lk}$$

Définition 21 : Une fréquence marginale de la variable y notée $\mathbf{n}_{.k}$, $k=1,\dots,N$ est le nbr d'unité stat i telle que y_i appartient à C_k .

$$\mathbf{n}_{.k} = \sum_{l=1}^L n_{lk}$$

Proposition : $\sum_{l=1}^L n_{l.} = \sum_{k=1}^K \mathbf{n}_{.k} = N$.
 $= \sum_{l=1}^L \sum_{k=1}^K n_{lk} = N$