

Def statistique : la stat (science des données) est la science de la collecte, de l'analyse, de la prez et de l'interprétation des données.

Objectifs :

- maitrise outils de l'analyse stat pr synthétiser l'info des données écos
- maitrise formalisation math pr dev capacités de trad d'un pb en formalisation math.

Livres :

- statistics for business and economics (amphi sieges verts)
- statistiques pour l'économie et la gestion (blanc/violet/bleu)

CHAPITRE 1 : INTRODUCTION, GRAPHES ET TABLEAUX

Section 1. Définitions fondamentales

1. Une **population** est l'E des éléments sur lesquels une étude se porte.
2. L'**unité statistique**, ou élément, est l'unité pour laquelle des données sont recueillies.
↳ Diff natures : prsn, pays, objet, bactérie...
3. Un **échantillon** est un sous-ensemble de la pop (noté N ou T)
4. L'**échantillonnage aléatoire (simple)** «est une procédure utilisée pr select un échantillon de N individus ds une pop de telle way que chaque indiv de la pop est choisi au hasard, aucune influence, chacun peut être choisi et chaque échantillon possible d'une taille donnée N a la mm chance d'être sélectionné.
5. Une **variable stat** est une caractéristique des éléments à laquelle on s'intéresse.
↳ Variables notées y ; x ou z
6. Une **modalité** est une valeur prise par une valeur stat
7. Une **observation** est un E de mesures obtenues pour chaque élément d'un E de données

Exercice 1. Pr connaître la fréquentat° des theatres de l'UB, supposons que 200 étudiants sont selected pr rep à un qst sur le nb de fois ou ils sont allés au theatre

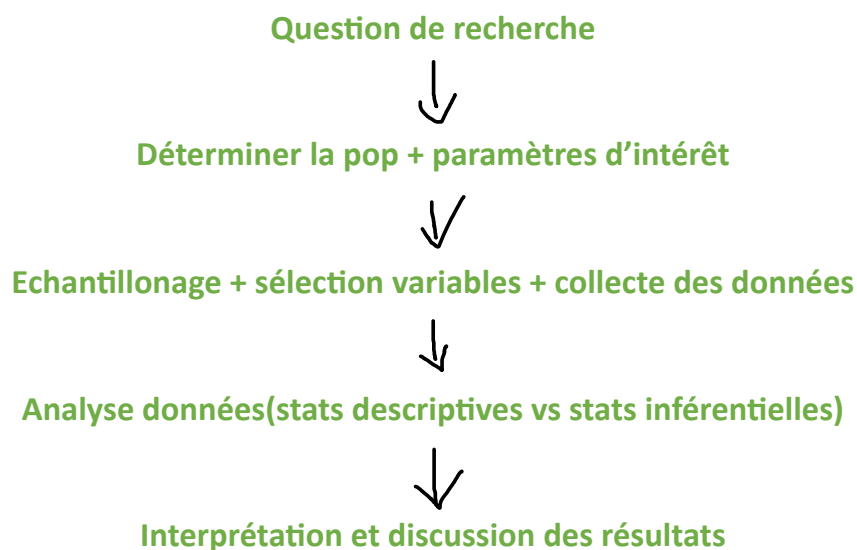
Population : les étudiants de l'UB

Echantillon : 200 étudiants
Unité stat : 1 étudiant
Variable : nb de fois au théâtre
Modalités : 2,3..... fois

Exercice 2 : Effet du lvl d'éducation sur les salaires des f mariées. 428 femmes
Pop : femmes mariées
Ech : 428 femmes
Unité stat : 1 femme mariée
Variables : niveau d'études // salaire
Modalités : diplôme, nb années d'études // dollars/euros...

8. Un paramètre est une mesure numérique qui décrit une caractéristique spécifique d'une pop. Une statistique/estimation est une mesure numérique qui décrit une caract d'un échantillon.
9. (L'erreur d'échantillonnage est la diff entre le paramètre d'intérêt et son estimation.)
10. La statistique descriptive focus sur les procédures graphiques et numériques used pr résumer et traiter les données. L'inférence statistique focus sur l'utilisation des données pr faire des prévisions/ estimations/ tester les théories pr prendre de meilleures décisions.

ETAPES D'UN PROJET DE RECHERCHE :



SECTION 2. Variables catégoriques

11. Une variable catégorique/qualitative/facteur est une variable stat dont les modalités sont des catégories (classes)

+classifient les unités stat tq chaque indiv est mis ds une seule catégorie

REMARQUE 1 : Une variable indicatrice est une variable catégorique qui prend seulement 2 diff valeurs : 0 et 1. Une variable cat avk L cats peut être ré-exprimée avk L variables indicatrices.

Exemple 1 :

Indiv	sexe	homme	femme
1	h	1	0

↳ v. indicatrices

Exemple 2 :

Pays	Groupe	CATEGORIQUE			
Algerie.	Riti				
Indonésie	rits				
Nepal	riti				
Roumanie.	Re				
Soudan	fr				

PAYS	RITI	RITS	RE	FR	INDICATIF
Algerie	1	0	0	0	
Indonesie	0	1	0	0	
Nepal	1	0	0	0	
Roumanie	0	0	1	0	
Soudan.	0	0	0	1	

12. Une variable numérique/quantitative est une variable stat dont les modalités sont mesurables. Elle est numérique discrète si elle peut prendre un nb fini/infini de valeurs dénombrables, tq 0,1,2...

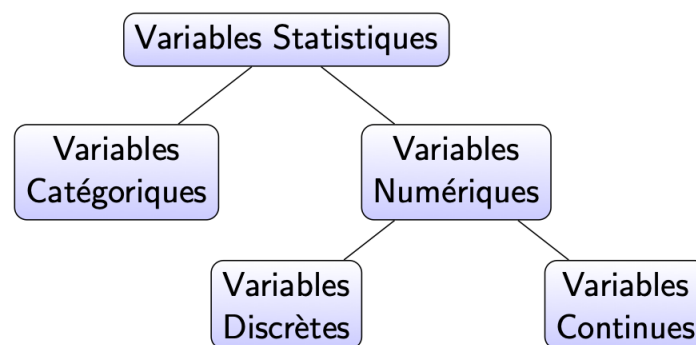
Une variable qui peut prendre ses valeurs nu mds un intervalle/ suite d'intervalles est appelée variable continue.

Exemple 3 : variables discrètes :

- Nb appels reçus en 5min ds un centre d'appel
- Nb voitures arrivant au péage

Exemple 4 : variables continues

- Temps écoulé entre les appels reçus
- Tx inflation en France



Section 3 : les types de données

13. Les données en coupe transversale sont les données liées aux diff indivs collectées au mm moment (ou approximativement au mm moment). Les séries temporelles sont des données collectées sur plusieurs périodes de temps diff.

Section 4. représentation des variables catégoriques

Déf 14. La fréquence d'une catégorie (ou classe) est le nb d'unités stats présentant cette dernière.

Soit x_i , $i=1,2,\dots,N$, une variable cat avec les cats C_l , $l=1,2,\dots,L$

Nous notons la fréquence de la cat l , n_l

La fréquence de la cat C_l est le nb d'unités stats i (tq x_i appartient à C_l)

Nous avons l'identité : $\sum_{I=1}^L n_I = N$

$$\sum_{i=1}^N x_i = x_1 + x_2 + \dots + x_N$$

Déf 15 : Une distribution de fréquence est un résumé des données sous forme de tableau décrivant la fréquence des ds diff catégories déf par une variable.

Les fréquences ne sont pas direct présentables pr une variable continue/discrete présentant bcp de modalités. On doit parfois regrouper les données.

Déf 16. La fréquence relative d'une catégorie correspond à la proportion des observations appartenant à cette catégorie :

$$\text{Fréquence relative d'une catégorie} = \frac{\text{fréquence d'une catégorie}}{N}$$

Nous notons la fréquence relative de la classe C_I , f_I

$$\text{Nous avons alors : } f_I = \frac{n_I}{N}$$

Proposition 1 : la somme des fréquences relatives est égale à 1 : $\sum_{I=1}^L f_I = 1$

Section 5 : rpz variables numériques

Tableaux de fréquence :

Table 5: Distribution des fréquences d'une variable numérique

Classes	Fréquence	Fréquence relative	Fréquence relative cumulée
$[a_1; b_1)$	n_1	f_1	F_1
$[a_2; b_2)$	n_2	f_2	F_2
$\vdots$	$\vdots$	$\vdots$	$\vdots$
$[a_I; b_I)$	n_I	f_I	F_I
$\vdots$	$\vdots$	$\vdots$	$\vdots$
$[a_{L-1}; b_{L-1})$	n_{L-1}	f_{L-1}	F_{L-1}
$[a_L; b_L)$	n_L	f_L	1.00
Total	N	1.00	

Les trois étapes nécessaires à la définition des classes d'une distribution de fréquence pr des données qtt :

- 1- Déter le nb de classes (L) en fonction du nb d'observations
- 2- Déter la largeur approximative de chaque classe : $d \approx \frac{\max - \min}{L}$
- 3- Déter les bornes a_l , b_l de chaque classe. Une méthode est de déf la borne inférieure de la 1^e classe comme min et calculer itérativement les autres bornes en using la largeur approximative trouvée ds l'étape précédente.
- 4- $A_1 = \min$, $b_l = a_l + d$, et $a_l = b_{l-1}$, $\forall l=2, \dots, L$

Déf 17. La fréquence relative cumulée d'une classe est la somme des fréquences de ttes les classes dont les observations sont inférieures ou égales à la borne supérieure de cette classe.

$$F_l = \sum_{s=1}^l f_s \quad l = 1$$

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

$$F_2 = \sum_{s=1}^2 f_s = f_1 + f_2$$

$$F_L = \sum_{s=1}^L f_s = f_1 + f_2 + \dots + f_L = 1$$

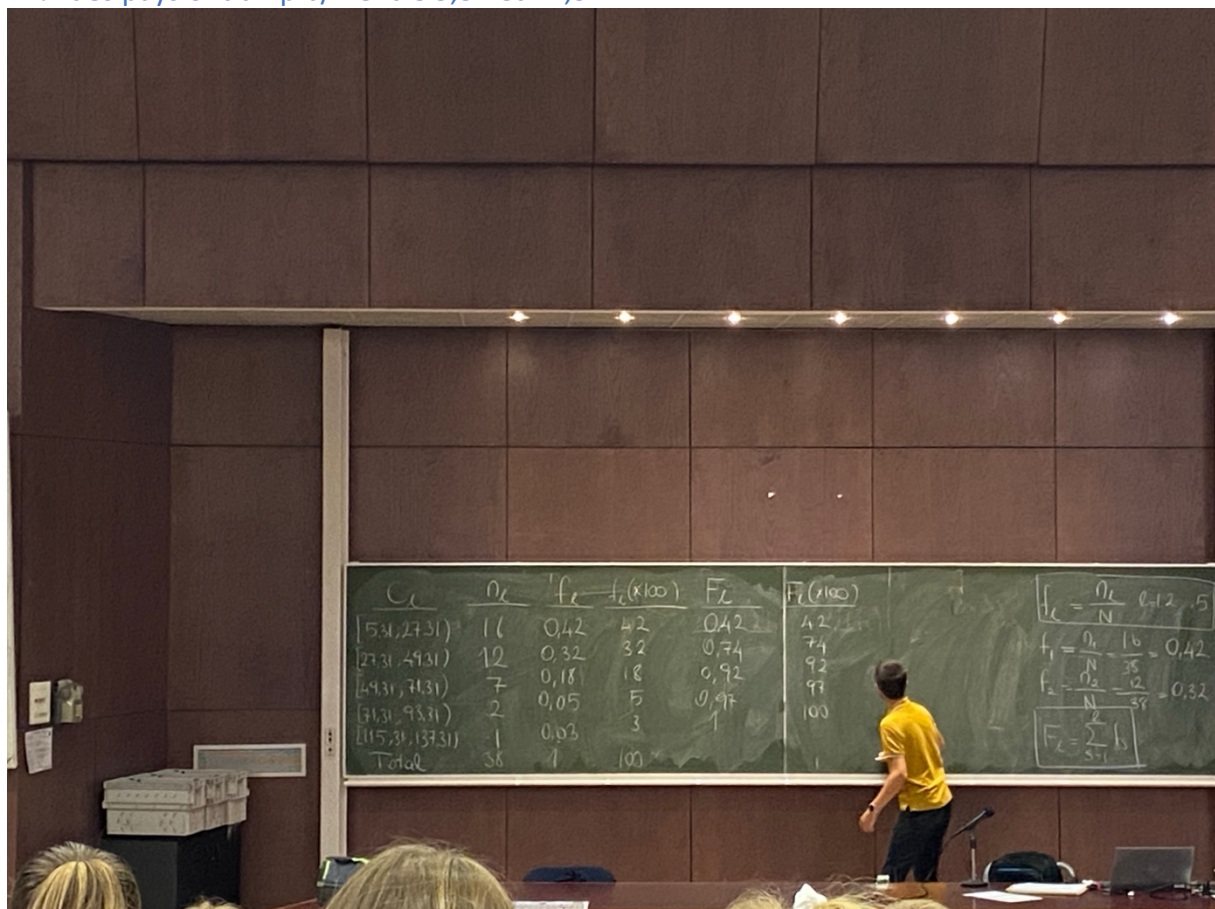
Ex : présenter la distrib du pib/h (1000\$) des pays de l'OCDE en 2020 sous la forme d'un tableau de fréquences

Nb observations dépend du nb de valeurs ds l'intervalle voulu de l'échantillon

$$CI (5,31 ; 27,31) \rightarrow n_l = 16 \rightarrow f_l = \frac{n_l}{N} = 0,42 \rightarrow f_l \times 100 = 42\% \rightarrow F_l = \sum_{s=1}^l f_s = 42$$

...

42% des pays ont un pib/h entre 5,31 et 27,31

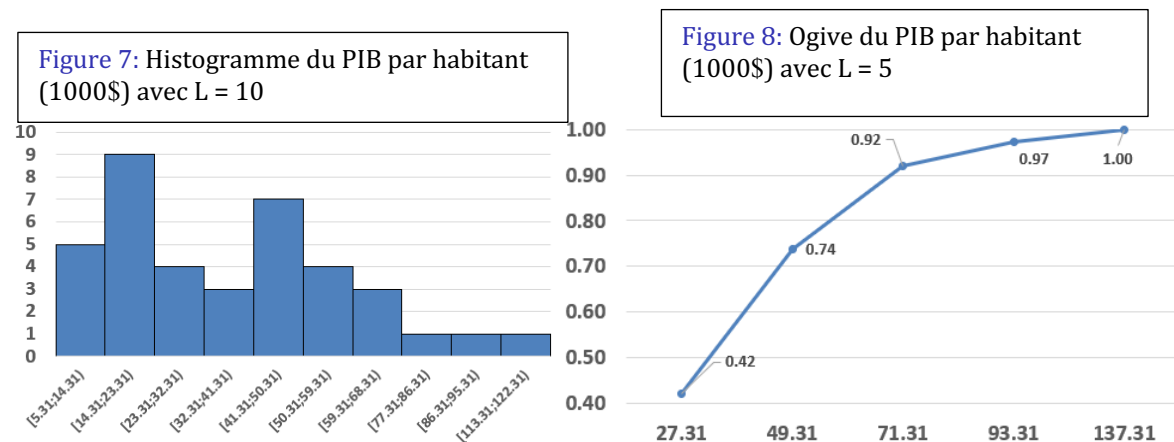


C_l	n_l	f_l	$f_l (\times 100)$	F_l	$F_l (100)$
$[5,31, 27,31)$	16	0,42	42	0,42	42
$[27,31, 49,31)$	12	0,32	32	0,74	74
$[49,31, 71,31)$	7	0,18	18	0,92	92
$[71,31, 93,31)$	2	0,05	5	0,97	97
$[93,31, 115,31)$	1	0,03	3	1	100
Total	38	1	100		

$f_l = \frac{n_l}{N} = \frac{16}{38} = 0,42$
 $f_l = \frac{n_l}{N} = \frac{12}{38} = 0,32$
 $F_l = \sum_{s=1}^l f_s$

Un histogramme est construit en plaçant les classes la variable considérée sur l'axe horizontal et la fréquence ou la fréquence (relative) sur l'axe vertical. La fréquence (relative) de chaque classe est représentée par un rectangle dont la base est déterminée par les limites de classes et dont la hauteur correspond à la fréquence (relative).

Un ogive est construit en plaçant les bornes supérieures des classes de la variable considérée sur l'axe horizontal et les fréquences relative cumulées sur l'axe vertical.



SECTION 6 : REPRÉSENTATION DES DONNÉES RELATIVES À DEUX VARIABLES

Definition 18 = Un tableau de contingence, ou une tabulation croisée, est un resume sous forme de tableau des donnees relatives a deux variables.

Les deux variables peuvent être catégoriques ou numériques.

Supposons que nous voulons étudier la distribution des fréquences dans un échantillon selon deux variables x_i et y_i , $i=1,2,\dots,N$.

Notons les catégories (ou classes) de x_i et y_i , C_l , $l = 1,2,\dots,L$ et G_k , $k = 1,2,\dots,K$, respectivement.

Table 6: La forme générale d'un tableau de contingence

$x_i \backslash y_j$	G_1	G_2	$\dots$	G_K	Total
C_1	n_{11}	n_{12}	$\dots$	n_{1K}	$n_{1.}$
C_2	n_{21}	n_{22}	$\dots$	n_{2K}	$n_{2.}$
$\vdots$	$\vdots$	$\vdots$	$\ddots$	$\vdots$	$\vdots$
C_L	n_{L1}	n_{L2}	$\dots$	n_{LK}	$n_{L.}$
Total	$n_{.1}$	$n_{.2}$	$\dots$	$n_{.K}$	N

Definition 19 = Une fréquence partielle, notée n_{lk} , $l = 1, 2, \dots, L$, $k = 1, 2, \dots, K$, est le nombre d'unités statistiques i telle que $x_i \in C_l$ et $y_i \in G_k$.

Autrement dit, n_{lk} est le nombre d'unités statistiques présentant à la fois la classe l de x et la classe k de y .

Definition 20 = Une fréquence marginale de la variable x , notée n_l , $l = 1, 2, \dots, L$, est le nombre d'unités statistiques i telle que $x_i \in C_l$. Cette définition est identique à la Définition

14. $f_{LK} = \frac{n_{LK}}{N}$ où n_{LK} l'effectif d'une case et N l'effectif total de l'échantillon (en bas à droite)

Exemple 7

Table 7: Fréquences de 217 pays classés selon la région et le niveau de revenu

	AME	AEP	EAC	MOA	AS	Total
FR	0	1	0	26	1	28
RE	20	14	38	9	0	81
RITI	5	14	4	25	6	54
RITS	19	9	16	9	1	54
Total	44	38	58	69	8	217

Fréquence conditionnelle : $\frac{\text{eff de chaque case}}{\text{eff total ligne}}$