Probabilités et statistiques

Licence 1-semestre 1

Allegret Audrey

Maître de Conférences - Université de Toulon, LEAD

Responsable de la Licence-Directrice des études

Coordonnées

- Adresse e-mail :audrey.sallenave@univ-tln.fr
- Page personnelle: https://sites.google.com/view/audreyallegret
- Deux contrôles continus en TD
- Partiel de 1h30h-2h

Plan

Variables statistiques

Notions de base

La statistique est un ensemble de méthodes et d'outils permettant d'analyser des données.

Variables statistiques

Notions de base

La statistique est un ensemble de méthodes et d'outils permettant d'analyser des données.
Elle définit des techniques à la fois pour collecter les données, les arranger, les présenter, les résumer et les analyser.

Variables statistiques

Notions de base

La statistique est un ensemble de méthodes et d'outils permettant d'analyser des données.
Elle définit des techniques à la fois pour collecter les données, les arranger, les présenter, les résumer et les analyser.
Le terme provient du latin **statisticum** qui signifie "ce qui a rapport à l'État". Les premières enquêtes statistiques datent du XVIII siècle.

Variables statistiques

Notions de base

La statistique est un ensemble de méthodes et d'outils permettant d'analyser des données.
Elle définit des techniques à la fois pour collecter les données, les arranger, les présenter, les résumer et les analyser.
Le terme provient du latin **statisticum** qui signifie "ce qui a rapport à l'État". Les premières enquêtes statistiques datent du XVIII siècle.
On distingue deux grandes branches dans la statistique :

Notions de base

- les **statistiques descriptives** ont pour but d'obtenir une vue synthétique de données. Il s'agit de dégager et de résumer l'essentiel de l'information contenue dans les données. (La statistique descriptive fait l'objet de la première année de licence).

Notions de base

- les **statistiques descriptives** ont pour but d'obtenir une vue synthétique de données. Il s'agit de dégager et de résumer l'essentiel de l'information contenue dans les données. (La statistique descriptive fait l'objet de la première année de licence).
- les **statistiques inférentielles** ont pour objectif d'utiliser les données collectées afin de tester des hypothèses, de rechercher des modèles ou de faire des prévisions. (La statistique descriptive fait l'objet de la deuxième année de licence)

Vocabulaire

La statistique utilise une terminologie rigoureuse qu'il faut assimiler afin de savoir avec précision quels sont les objets et les concepts qu'on manipule.

Vocabulaire

La statistique utilise une terminologie rigoureuse qu'il faut assimiler afin de savoir avec précision quels sont les objets et les concepts qu'on manipule.
Tout d'abord, une enquête statistique se déroule toujours dans une **population**.

Vocabulaire

La statistique utilise une terminologie rigoureuse qu'il faut assimiler afin de savoir avec précision quels sont les objets et les concepts qu'on manipule.
Tout d'abord, une enquête statistique se déroule toujours dans une **population**.
C'est un ensemble de référence dont les éléments sont désignés comme **individus**. Ces individus peuvent être des personnes aussi bien que des entités.

Vocabulaire

Les études statistiques consistent à observer chez ces individus des **caractères** ou **variables statistiques**. Ceux-ci peuvent être de nature très variée.

Vocabulaire

Les études statistiques consistent à observer chez ces individus des **caractères** ou **variables statistiques**. Ceux-ci peuvent être de nature très variée.
L'ensemble des valeurs des caractères relevés chez un "individu" porte le nom d'observation. Les observations sont rassemblées dans des bases de données appelées fréquemment des **jeux de données**.

Vocabulaire

Les études statistiques consistent à observer chez ces individus des **caractères** ou **variables statistiques**. Ceux-ci peuvent être de nature très variée.
L'ensemble des valeurs des caractères relevés chez un "individu" porte le nom d'observation. Les observations sont rassemblées dans des bases de données appelées fréquemment des **jeux de données**.
Les valeurs des caractères portent le nom de **modalités**. Les modalités doivent être choisies de telle sorte que tout individu puisse se voir attribuer une unique valeur.

Vocabulaire

Les études statistiques consistent à observer chez ces individus des **caractères** ou **variables statistiques**. Ceux-ci peuvent être de nature très variée.
L'ensemble des valeurs des caractères relevés chez un "individu" porte le nom d'observation. Les observations sont rassemblées dans des bases de données appelées fréquemment des **jeux de données**.
Les valeurs des caractères portent le nom de **modalités**. Les modalités doivent être choisies de telle sorte que tout individu puisse se voir attribuer une unique valeur.
Les modalités sont comme des catégories. Elles constituent une partition des valeurs possibles.

Vocabulaire

On distingue deux types de caractères ou variables :

- les **caractères qualitatifs** : ce sont ceux qu'on ne peut pas représenter par une mesure.

Vocabulaire

On distingue deux types de caractères ou variables :

- les **caractères qualitatifs** : ce sont ceux qu'on ne peut pas représenter par une mesure.
  Par exemple : couleur des yeux, sexe, situation familiale, mention au baccalauréat, catégorie socio-professionnelle etc...

Vocabulaire

On distingue deux types de caractères ou variables :

- les **caractères qualitatifs** : ce sont ceux qu'on ne peut pas représenter par une mesure.
  Par exemple : couleur des yeux, sexe, situation familiale, mention au baccalauréat, catégorie socio-professionnelle etc...
- les **caractères quantitatifs** : ce sont ceux qu'on peut mesurer et représenter numériquement. Parmi eux on distingue :

Vocabulaire

On distingue deux types de caractères ou variables :

- les **caractères qualitatifs** : ce sont ceux qu'on ne peut pas représenter par une mesure.
  Par exemple : couleur des yeux, sexe, situation familiale, mention au baccalauréat, catégorie socio-professionnelle etc...
- les **caractères quantitatifs** : ce sont ceux qu'on peut mesurer et représenter numériquement. Parmi eux on distingue :
  - les **caractères quantitatifs discrets** : leur valeur est en général un nombre entier ou appartient à un ensemble fini de valeurs. Par exemple : nombre d'enfants, nombre de pièces etc..

Vocabulaire

On distingue deux types de caractères ou variables :

- les **caractères qualitatifs** : ce sont ceux qu'on ne peut pas représenter par une mesure.
  Par exemple : couleur des yeux, sexe, situation familiale, mention au baccalauréat, catégorie socio-professionnelle etc...
- les **caractères quantitatifs** : ce sont ceux qu'on peut mesurer et représenter numériquement. Parmi eux on distingue :
  - les **caractères quantitatifs discrets** : leur valeur est en général un nombre entier ou appartient à un ensemble fini de valeurs. Par exemple : nombre d'enfants, nombre de pièces etc..
  - les **caractères quantitatifs continus** : leur valeur est en général un nombre réel pris dans un certain intervalle. Par exemple : taille, poids, température, chiffre d'affaire, montant imposable etc...

Vocabulaire

Les variables qualitatives peuvent être **nominales** ou **ordinales**. Dan le premier cas, les modalités ne peuvent être ordonnées, contrairement au cas des variables ordinales.

Vocabulaire

Les variables qualitatives peuvent être **nominales** ou **ordinales**. Dan le premier cas, les modalités ne peuvent être ordonnées, contrairement au cas des variables ordinales.

Exemple

Un exemple usuel de variable nominale est par exemple le sexe (modalités : féminin, masculin) ou encore l'état-civil (modalités : célibataire, marié, divorcé, pacsé ou veuf). Des variables comme le niveau d'études (modalités : sans diplôme, primaire, secondaire et universitaire) sont des variables ordinales.

Vocabulaire

Les valeurs d'une variable quantitative continue sont fréquemment regroupées en classes ou en intervalles contigus.

Vocabulaire

Les valeurs d'une variable quantitative continue sont fréquemment regroupées en classes ou en intervalles contigus.
Leur domaine de définition est partitionné en intervalles de la forme $[e_i; e_{i+1}[$.

Vocabulaire

Les valeurs d'une variable quantitative continue sont fréquemment regroupées en classes ou en intervalles contigus.
Leur domaine de définition est partitionné en intervalles de la forme $[e_i; e_{i+1}[$.
Les intervalles peuvent être ouverts à gauche et fermés à droite, ou le contraire. Il faut s'assurer qu'ils sont disjoints et que leur réunion recouvre toutes les valeurs possibles.

Vocabulaire

Les valeurs d'une variable quantitative continue sont fréquemment regroupées en classes ou en intervalles contigus.
Leur domaine de définition est partitionné en intervalles de la forme $[e_i; e_{i+1}[$.
Les intervalles peuvent être ouverts à gauche et fermés à droite, ou le contraire. Il faut s'assurer qu'ils sont disjoints et que leur réunion recouvre toutes les valeurs possibles.
Dans les calculs, il arrive qu'on veuille représenter un intervalle par une valeur numérique. On utilise souvent pour cela le **centre de l'intervalle** :

Vocabulaire

$$c_i = \frac{e_i + e_{i+1}}{2}$$

Vocabulaire

$$c_i = \frac{e_i + e_{i+1}}{2}$$

La taille de l'intervalle s'appelle **l'amplitude** :

$$a_i = e_{i+1} - e_i$$

Exemple

On a relevé les poids suivants (en kg) parmi 100 individus.

Exemple

On a relevé les poids suivants (en kg) parmi 100 individus.

Exemple

On a relevé les poids suivants (en kg) parmi 100 individus.

64	85	79	84	68	74	94	75	64	65
72	74	78	69	67	64	70	63	69	82
62	64	71	74	77	73	77	76	82	82
86	48	50	69	76	59	70	61	55	77
73	81	76	56	63	84	63	57	76	86
62	70	69	66	63	90	72	73	73	76
75	70	68	66	74	66	52	66	81	57
77	79	55	69	78	60	85	70	67	64
76	78	65	81	69	76	72	71	74	58
67	76	74	78	79	69	92	64	73	65

Exemple

1. Déterminer les valeurs extrêmes (min et max).
2. Répartir les données en classes d'amplitude 10 en partant de 45 kg.
3. Préciser les centres des classes.

Correction

1. Déterminer les valeurs extrêmes (min et max).

Correction

1. Déterminer les valeurs extrêmes (min et max).
   Le poids minimal est 48 kg et le poids maximal est 94 kg.
2. Répartir les données en classes d'amplitude 10 en partant de 45 kg.

Correction

1. Déterminer les valeurs extrêmes (min et max).
   Le poids minimal est 48 kg et le poids maximal est 94 kg.
2. Répartir les données en classes d'amplitude 10 en partant de 45 kg.

[45 ; 55[	[55 ; 65[	[65 ;75[	[75 ;85 [	[85 ; 95[
3	21	40	29	7

Correction

1. Déterminer les valeurs extrêmes (min et max).
   Le poids minimal est 48 kg et le poids maximal est 94 kg.
2. Répartir les données en classes d'amplitude 10 en partant de 45 kg.

[45 ; 55[	[55 ; 65[	[65 ;75[	[75 ;85 [	[85 ; 95[
3	21	40	29	7

Remarque : si on avait choisi des intervalles ouverts à gauche, on aurait obtenu des résultats différents :

Correction

1. Déterminer les valeurs extrêmes (min et max).
   Le poids minimal est 48 kg et le poids maximal est 94 kg.
2. Répartir les données en classes d'amplitude 10 en partant de 45 kg.

[45 ; 55[	[55 ; 65[	[65 ;75[	[75 ;85 [	[85 ; 95[
3	21	40	29	7

Remarque : si on avait choisi des intervalles ouverts à gauche, on aurait obtenu des résultats différents :

]45 ; 55]	]55 ; 65]	]65 ;75]	]75 ;85]	]85 ; 95]
5	22	39	29	5

Correction

1. Déterminer les valeurs extrêmes (min et max).
   Le poids minimal est 48 kg et le poids maximal est 94 kg.
2. Répartir les données en classes d'amplitude 10 en partant de 45 kg.

[45 ; 55[	[55 ; 65[	[65 ;75[	[75 ;85 [	[85 ; 95[
3	21	40	29	7

   Remarque : si on avait choisi des intervalles ouverts à gauche, on aurait obtenu des résultats différents :

]45 ; 55]	]55 ; 65]	]65 ;75]	]75 ;85]	]85 ; 95]
5	22	39	29	5

3. Préciser les centres des classes.

Correction

1. Déterminer les valeurs extrêmes (min et max).
   Le poids minimal est 48 kg et le poids maximal est 94 kg.
2. Répartir les données en classes d'amplitude 10 en partant de 45 kg.

[45 ; 55[	[55 ; 65[	[65 ;75[	[75 ;85 [	[85 ; 95[
3	21	40	29	7

   Remarque : si on avait choisi des intervalles ouverts à gauche, on aurait obtenu des résultats différents :

]45 ; 55]	]55 ; 65]	]65 ;75]	]75 ;85]	]85 ; 95]
5	22	39	29	5

3. Préciser les centres des classes.

50	60	70	80	90

Vocabulaire

La distinction n'est pas toujours très rigoureuse entre variables discrètes et continues. Il arrive que des variables exprimées en nombres entiers soient quand même considérées comme variables continues. C'est le cas par exemple de l'âge.

Table des observations

Ce sont des tableaux qui représentent l'intégralité des observations collectées auprès des individus constituant la population ou un échantillon extrait.

Table des observations

Ce sont des tableaux qui représentent l'intégralité des observations collectées auprès des individus constituant la population ou un échantillon extrait.
Les tables d'observations (dites, en anglais, dataframes) peuvent avoir des colonnes de nature différente : variable qualitative, quantitative, etc... Ce ne sont donc pas des matrices.

Table des observations

Ce sont des tableaux qui représentent l'intégralité des observations collectées auprès des individus constituant la population ou un échantillon extrait.
Les tables d'observations (dites, en anglais, dataframes) peuvent avoir des colonnes de nature différente : variable qualitative, quantitative, etc... Ce ne sont donc pas des matrices.
On présente ici trois types de tableau (à une variable, à deux variables et à trois variables).

- Tableau à une variable X :

Table des observations

Observations	X
Obs1	x_1
Obs2	x_2
$\vdots$	$\vdots$
ObsN	x_N

Table des observations

- Tableau à deux variables X et Y :

Table des observations

- Tableau à deux variables X et Y :

Observations	X	Y
Obs1	x_1	y_1
Obs2	x_2	y_2
Obs3	x_3	y_3
$\vdots$	$\vdots$	$\vdots$
ObsN	x_N	y_N

Table des observations

- Tableau à trois variables X, Y et Z :

Table des observations

- Tableau à trois variables X, Y et Z :

Observations	X	Y	Z
Obs1	x_1	y_1	z_1
Obs2	x_2	y_2	z_2
Obs3	x_3	y_3	z_3
$\vdots$	$\vdots$	$\vdots$	$\vdots$
ObsN	x_N	y_N	z_N

Tables d'effectifs

Pour chaque modalité v_i d'une variable qualitative, chaque valeur v_i d'une variable quantitative discrète ou chaque classe modale c_i d'une variable quantitative continue, on note le nombre n_i d'individus présentant cette modalité ou appartenant à cette classe : n_i est l'effectif de la modalité ou la classe modale.

Tables d'effectifs

Pour chaque modalité v_i d'une variable qualitative, chaque valeur v_i d'une variable quantitative discrète ou chaque classe modale c_i d'une variable quantitative continue, on note le nombre n_i d'individus présentant cette modalité ou appartenant à cette classe : n_i est l'effectif de la modalité ou la classe modale.
Obtient ainsi une table d'effectifs de la forme (dans le cas discret) :

Tables d'effectifs

Pour chaque modalité v_i d'une variable qualitative, chaque valeur v_i d'une variable quantitative discrète ou chaque classe modale c_i d'une variable quantitative continue, on note le nombre n_i d'individus présentant cette modalité ou appartenant à cette classe : n_i est l'effectif de la modalité ou la classe modale.
Obtient ainsi une table d'effectifs de la forme (dans le cas discret) :

Valeurs	v_1	v_2	v_3	...	v_k
Effectifs	n_1	n_2	n_3	...	n_k

Tables d'effectifs

Pour chaque modalité v_i d'une variable qualitative, chaque valeur v_i d'une variable quantitative discrète ou chaque classe modale c_i d'une variable quantitative continue, on note le nombre n_i d'individus présentant cette modalité ou appartenant à cette classe : n_i est l'effectif de la modalité ou la classe modale.

Obtient ainsi une table d'effectifs de la forme (dans le cas discret) :

Valeurs	v_1	v_2	v_3	$\dots$	v_k
Effectifs	n_1	n_2	n_3	$\dots$	n_k

Où, dans le cas continu

Valeurs	$[e_1; e_2[$	$[e_2; e_3[$	$[e_3; e_4[$	$[\dots[$	$[e_k; e_{k+1}[$
Effectifs	n_1	n_2	n_3	$\dots$	n_k

Tables d'effectifs

Pour chaque modalité v_i d'une variable qualitative, chaque valeur v_i d'une variable quantitative discrète ou chaque classe modale c_i d'une variable quantitative continue, on note le nombre n_i d'individus présentant cette modalité ou appartenant à cette classe : n_i est l'effectif de la modalité ou la classe modale.

Obtient ainsi une table d'effectifs de la forme (dans le cas discret) :

Valeurs	v_1	v_2	v_3	$\dots$	v_k
Effectifs	n_1	n_2	n_3	$\dots$	n_k

Où, dans le cas continu

Valeurs	$[e_1; e_2[$	$[e_2; e_3[$	$[e_3; e_4[$	$[\dots[$	$[e_k; e_{k+1}[$
Effectifs	n_1	n_2	n_3	$\dots$	n_k

Où, k est le nombre de modalités ou de classes.

Tables d'effectifs

Le nombre total des observations est noté N :

Tables d'effectifs

Le nombre total des observations est noté N :

$$N = n_1 + n_2 + n_3 + \ldots + n_k = \sum_{i=1}^{k} n_i$$

L'ensemble des couples (v_i, n_i) constitue une distribution statistique.

Exemple

Le tableau suivant est issu du recensement de population de 2011 et dénombre les logements déclarés comme résidences principales en fonction du nombre de pièces :

Exemple

Le tableau suivant est issu du recensement de population de 2011 et dénombre les logements déclarés comme résidences principales en fonction du nombre de pièces :

Nombre de pièces	Effectifs
1	1571 903
2	3417233
3	5723944
4	6914989
5	5315838
6	4403719

Exercice

Reprendre la table des régions et des départements : Table des régions et départements de France métropolitaine Dresser une table d'effectifs pour chaque région.

Exercice

Reprendre la table des régions et des départements : Table des régions et départements de France métropolitaine Dresser une table d'effectifs pour chaque région.

Alsace	Aquitaine	Auvergne	B-N	Bourgogne	Bretagne
2	5	4	3	4	4
Centre	C-Ardenne	Corse	Franche-Comté	H-Nor.	I-L-D
6	4	2	4	2	8
L-Roussillon	Limousin	Lorraine	Midi-P	N-P-D-C	P-D-L-L
5	3	4	8	2	5
Picardie	P-C	PACA	Rhônes-Alpes		
3	4	6	8		

Table de fréquences

On appelle fréquence (ou proportion) le rapport entre l'effectif d'une modalité ou d'une classe et l'effectif total :

Table de fréquences

On appelle fréquence (ou proportion) le rapport entre l'effectif d'une modalité ou d'une classe et l'effectif total :

$$f_i = \frac{n_i}{N}$$

A partir de là, on peut définir des tables de fréquences, dans le cas discret comme dans le cas continu.

Table de fréquences

On appelle fréquence (ou proportion) le rapport entre l'effectif d'une modalité ou d'une classe et l'effectif total :

$$f_i = \frac{n_i}{N}$$

A partir de là, on peut définir des tables de fréquences, dans le cas discret comme dans le cas continu.

Valeurs	v_1	v_2	v_3	$\ldots$	v_k
Fréquences	f_1	f_2	f_3	$\ldots$	f_k

Table de fréquences

On appelle fréquence (ou proportion) le rapport entre l'effectif d'une modalité ou d'une classe et l'effectif total :

$$f_i = \frac{n_i}{N}$$

A partir de là, on peut définir des tables de fréquences, dans le cas discret comme dans le cas continu.

Valeurs	v_1	v_2	v_3	$\ldots$	v_k
Fréquences	f_1	f_2	f_3	$\ldots$	f_k

Où, dans le cas continu :

Valeurs	$[e_1; e_2[$	$[e_2; e_3[$	$[e_3; e_4[$	$[\ldots[$	$[e_k; e_{k+1}[$
Fréquences	f_1	f_2	f_3	$\ldots$	f_k

Table de fréquences

On appelle fréquence (ou proportion) le rapport entre l'effectif d'une modalité ou d'une classe et l'effectif total :

$$f_i = \frac{n_i}{N}$$

A partir de là, on peut définir des tables de fréquences, dans le cas discret comme dans le cas continu.

Valeurs	v_1	v_2	v_3	$\dots$	v_k
Fréquences	f_1	f_2	f_3	$\dots$	f_k

Où, dans le cas continu :

Valeurs	$[e_1; e_2[$	$[e_2; e_3[$	$[e_3; e_4[$	$[\dots[$	$[e_k; e_{k+1}[$
Fréquences	f_1	f_2	f_3	$\dots$	f_k

Les fréquences sont toujours comprises entre 0 et 1 :

$$0 \leq f_i \leq 1$$

Table de fréquences

La somme des fréquences est toujours égale à 1, c'est-à-dire à 100% si on exprime les valeurs en pourcentage. En effet :

$$\sum_{i=1}^{k} f_i = f_1 + f_2 + f_3 + ... + f_k \quad (1)$$

$$= \frac{n_1}{N} + \frac{n_2}{N} + \frac{n_3}{N} + \ldots + \frac{n_k}{N} \quad (2)$$

$$= \frac{1}{N} \mp (n_1 + n_2 + n_3 + \ldots + n_k) \quad (3)$$

$$= \frac{N}{N} \quad (4)$$

$$= 1 \quad (5)$$

Table de fréquences

La somme des fréquences est toujours égale à 1, c'est-à-dire à 100% si on exprime les valeurs en pourcentage. En effet :

Table de fréquences

La somme des fréquences est toujours égale à 1, c'est-à-dire à 100% si on exprime les valeurs en pourcentage. En effet :
La signification de la fréquence est la proportion, par rapport au nombre total des observations, des individus pour lesquels la variable statistique prend la valeur v_i ou appartient à la classe c_i.

Table de fréquences

La somme des fréquences est toujours égale à 1, c'est-à-dire à 100% si on exprime les valeurs en pourcentage. En effet :
La signification de la fréquence est la proportion, par rapport au nombre total des observations, des individus pour lesquels la variable statistique prend la valeur v_i ou appartient à la classe c_i.
L'ensemble des couples (v_i, f_i) constitue une distribution en fréquences par opposition à la distribution en effectifs).

Table de fréquences

La somme des fréquences est toujours égale à 1, c'est-à-dire à 100% si on exprime les valeurs en pourcentage. En effet :
La signification de la fréquence est la proportion, par rapport au nombre total des observations, des individus pour lesquels la variable statistique prend la valeur v_i ou appartient à la classe c_i.
L'ensemble des couples (v_i, f_i) constitue une distribution en fréquences par opposition à la distribution en effectifs).

Table de fréquences

La somme des fréquences est toujours égale à 1, c'est-à-dire à 100% si on exprime les valeurs en pourcentage. En effet :
La signification de la fréquence est la proportion, par rapport au nombre total des observations, des individus pour lesquels la variable statistique prend la valeur v_i ou appartient à la classe c_i.
L'ensemble des couples (v_i, f_i) constitue une distribution en fréquences par opposition à la distribution en effectifs).
A noter que les effectifs et les fréquences sont proportionnels : on passe de l'un à l'autre en multipliant ou en divisant par le même nombre N.

Table de fréquences

On reprend les données concernant le nombre de pièces des résidences principales. Le nombre total d'observations est N = 27 347 626. On obtient donc les proportions en divisant par N :

Table de fréquences

On reprend les données concernant le nombre de pièces des résidences principales. Le nombre total d'observations est N = 27 347 626. On obtient donc les proportions en divisant par N :

Table de fréquences

On reprend les données concernant le nombre de pièces des résidences principales. Le nombre total d'observations est N = 27 347 626. On obtient donc les proportions en divisant par N :

Nombre de pièces	Effectifs	Fréquences
1	1571 903	5.75%
2	3417233	12.50%
3	5723944	20.93%
4	6914989	25.29%
5	5315838	19.44%
6	4403719	16.10%