Centre universitaire de Mila

Institut des sciences et technologies

Département de science de la nature et de la vie

Cours de Biostatistiques

2ème année toutes les filières

Enseignant responsable

Semara L.

I. Vocabulaire de biostatistiques

La statistique est le domaine des mathématiques qui ´étudie les outils de recueil, de traitement et d'interprétation des données. Autrement **La statistique** est l'art de collecter, d'analyser et d'interpréter des « données » pour évaluer la « fiabilité » des décisions fondées sur ces « données ».

L'analyse des données est utilisée pour d´écrire les phénomènes études, faire des prévisions et prendre des décisions à leur sujet. En cela, la statistique est un outil essentiel pour la compréhension et la gestion des phénomènes complexes.tel que les phénomènes biologiques.

Biostatistiques : application des statistiques à des problèmes biologiques (essais thérapeutiques, ´épidémiologie, dynamique des populations, analyse du génome, ...)

La statistique descriptive : est l'ensemble des méthodes et techniques permettant de présenter, de d´écrire, de résumer des données nombreuses et variées. Elle utilise pour cela des représentations de données sous forme de graphiques, de tableaux et d'indicateurs numériques (par exemple des moyennes).Elle permet de dégager les caractéristiques essentielles du phénomène ´étudie et de suggérer des hypothèses pour une étude ultérieure plus sophistiquée.

La statistique inférentielle : est l'ensemble des méthodes permettant, à partir d'un échantillon, d'estimer des paramètres d'une population statistique et/ou de tester des hypothèses sur cette population.

Population

En statistique, le terme de population s'applique à tout objet statistique étudié. C'est l'ensemble des « individus » à propos desquels on souhaite pouvoir inférer des décisions. Elle est le plus souvent définie par une propriété portant une ou plusieurs variables.

Exemple : L'ensemble des algériens, L'ensemble des nouveaux nés de mère diabétique, L'ensemble des hommes obèses.

Échantillon

Pour des raisons techniques ou budgétaires, il est impossible de mesurer la variable d'intérêt sur l'ensemble des individus de la population. On effectue alors une observation partielle de cette population à travers un échantillon .Un échantillon est donc, un sous-ensemble de cette population sur lequel on pourra observer la variable d'intérêt, et utiliser ces observations pour inférer des décisions sur un individu quelconque de la population.

L'effectif de l'échantillon ou la taille d'échantillon ou le nombre d'observation est le nombre d'individus constituant l'échantillon (ou inclus dans l'échantillon).

Échantillonnage représentatif

Il existe déférentes procédures pour choisir un ´échantillon. On parle de procédure d'´échantillonnage. Les plus courantes sont l'´échantillonnage aléatoire simple et l'´échantillonnage aléatoire stratifie.

Pour que l'inférence de décisions soit valide, l'échantillon doit être constitué de manière aléatoire et simple, on parle d'un *échantillonnage représentatif*.

Aléatoire : Chaque individu de la population a la même « chance » d'être inclus dans l'échantillon.

Simple : Le fait de retenir un individu dans l'échantillon n'affecte pas la « chance » d'un autre individu d'être également sélectionné.

Idéalement, affecter un numéro à chaque individu, et tirer au sort avec un générateur de nombres aléatoires. En pratique, plus complexe… nous supposerons toujours que les échantillons dont nous parlons ont été « randomisés »

Paramètre

Un paramètre est une grandeur apportant une information résumée sur la variable d'intérêt (exemple: la moyenne). Un paramètre peut être mesuré dans un échantillon et estimé dans la population, à partir des observations de l'échantillon.

Donnée

Une donnée est le résultat de l'observation fait sur d'un individu (unité statistique).

Variable

Chaque individu statistique est donc décrit par un ou plusieurs traits distinctifs ou grandeurs physiques qui le caractérisant. On les appelle variables statistiques.

Nature statistiques des variables

Une variable statistique (ou caractère statistique) est donc ce qui est observé ou mesuré sur un individu statistique. On distingue deux types de variables statistiques :

Variable quantitative

Toute variable qu'un instrument peut mesurer sous forme numérique (chiffre). : Elle concerne une grandeur mesurable. Ses valeurs sont des nombres exprimant une quantité. Les variables quantitatives peuvent -être scindées en deux types.

Variable quantitative continue : elle peut prendre n'importe quelle valeur dans un intervalle donné.

Exemple : Taille, poids, glycémie, hauteur de la plante.

Variable quantitative discrète ou discontinu : elle ne peut prendre qu'un nombre fini ou dénombrable de valeurs (on peut « compter » les valeurs possibles).Une telle variable s'exprime par un nombre entier.

Exemple : Nombre de globules blanc dans un volume de 1 ml, Nombre d'enfants par famille.

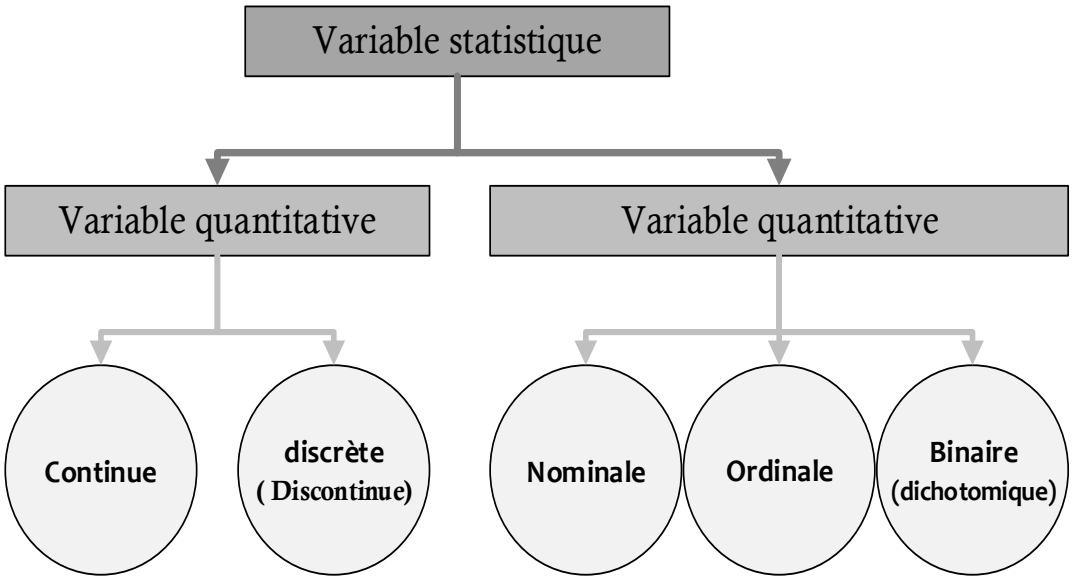
Variable qualitative

Toute variable caractérisée par un attribut qualitatif, et non par une mesure numérique. Ses valeurs sont des '*modalités de réponse*', ou catégories, exprimées sous forme littérale ou par un codage numérique. Les variables quantitatives peuvent -être scindées en trois types.

Variable qualitative nominales : Les modalités de réponse sont des noms qui ne peuvent pas être naturellement ordonnées. Exemple : la couleur des yeux : {noir ; bleu ; vert}, sexe : {homme ; femme}

Variable qualitative ordinales : Les modalités de réponse sont des niveaux qui peuvent être naturellement ordonnées. Exemple : degré de satisfaction des mères accouchant dans une maternité : {très insatisfait ; plutôt insatisfait ; plutôt satisfait ; très satisfait}, stade d'une maladie.

Variable qualitative binaire (dichotomique) : Si la variable elle n'a que deux modalités et de type (présence /absence). Exemple : présence d'un facteur de risque pour une pathologie



Typologie des variables statistiques

II. Statistiques descriptives

2.1. Statistiques descriptives univarie

Distribution d'une variable qualitative

Si la variable est qualitative, on appelle modalités les valeurs possibles de cette variable. L'ensemble des modalités est noté E = {Modalité 1, Modalité 2 . . ., Modalité k}.

La meilleure manière de représenter ces données est d'utiliser l'effectif (fréquences absolues) et les fréquences (fréquences relatives) de chaque modalité de réponse.

Effectif n_i = nombre d'individus pour lesquels la valeur de la variable est V_i, pour chaque valeur de i compris entre 1 et k

Fréquence f_i= On appelle fréquence ou fréquence relative de la modalité le rapport n_i/n (nombre d'individus de la modalité i relativement à l'effectif de l'échantillon n).

Modalité	*Effectif (ni)*	*Fréquence (fi)*
Modalité 1	n_1	f_1
Modalité 2	n_2	f_2
.	.	.
.	.	.
Modalité i	n_i	f_i
Totale	n	1

Si la variable est la couleur des yeux d'un individu, l'ensemble des modalités est E = {bleu, vert, brun, noir}. Si on interroge n = 200 personnes, les données brutes forment un tableau illisible. On peut le résumer en calculant les effectifs de chaque modalité les fréquences de réponse.

Les données brutes incompréhensibles se transforment en un tableau facilement lisible et des représentations graphiques simples.

Tableau de données brutes

Individu	Couleurs des yeux
1	noir
2	vert
3	vert
4	brun
5	bleu
6	vert
..	….
..	….
..	….
199	bleu
200	noir

Tableau statistique de la variable couleur des yeux

Modalité	Effectif (n_i)	Fréquence (f_i)
noir	103	0,515
Vert	25	0,125
brun	58	0,29
bleu	14	0,07
Total	200	1

Représentation graphique d'une variable qualitative

Deux diagrammes permettent de représenter une variable qualitative : le diagramme à secteurs angulaires et le diagramme en barres.

Diagramme en secteurs « camemberts »

Le camembert est un disque partagé en secteurs, chaque secteur représentant une modalité et ayant une surface proportionnelle à la fréquence de cette modalité dans la série statistique.

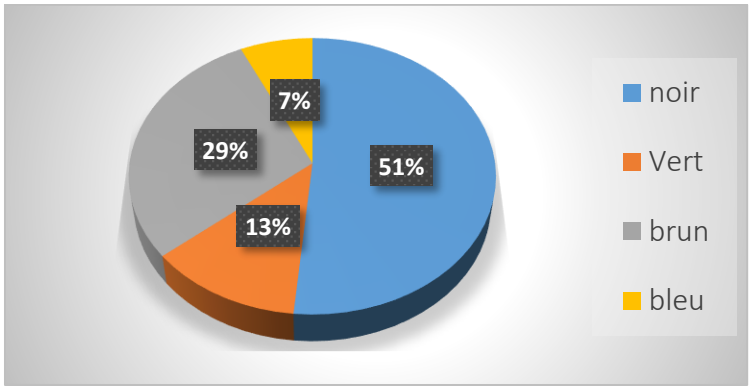


Diagramme en secteur de la variable couleur des yeux

Diagramme en barres

Le diagramme en barre (en bâtons) est un ensemble de rectangles de même largeur séparés par un espace, chaque rectangle représentant une modalité et ayant une hauteur proportionnelle à la fréquence de cette modalité dans la série statistique.

En abscisses, les différentes valeurs possibles. En ordonnées, les effectifs (ou les fréquences). Veiller à choisir l'origine des effectifs à 0 (pour que les surfaces des barres soient proportionnelles à ce qu'on souhaite représenter, ici les effectifs).

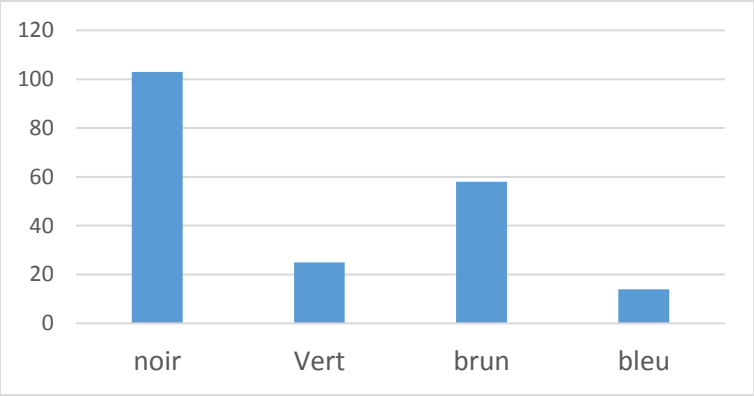


Diagramme en en barres de la variable couleur des yeux

- ✓ Pour les variables qualitatives « non ordinales », préférer la représentation en secteurs.
- ✓ Pour les variables qualitatives « ordinales », les représentations en secteurs ou en barres peuvent se concevoir.
- ✓ L'axe des abscisses se lisant de gauche à droite, il est naturel de placer à gauche la valeur « la plus faible » et à droite à la valeur « la plus grande » (au sens de la relation d'ordre qui caractérise un variable qualitative ordinale).

Distribution d'une variable quantitative

Distribution d'une variable quantitative discrète

À partir de l'observation d'une variable quantitative discrète sur n individus, on peut construire un tableau statistique.

On peut résumer les données brutes des k valeurs distinctes prises par la variable d'intérêt X {x i , i = 1,..,k}

ni est l'effectif associe à la valeur xi c'est-`a-dire le nombre d'individus ayant cette valeur dans l'échantillon ; n est la taille de l´échantillon (nombre total d'individus dans cet échantillon).

fi = ni/n est la fréquence associée `a la valeur xi c'est-`a-dire la proportion d'individus dans l´échantillon ayant cette valeur.

N_i est l'effectif c cumulée en xi c'est-`a-dire le nombre d'individus dans l´échantillon ayant une valeur inférieure ou égale à xi. Le calcul des Ni peut se faire façon récurrente de la manière suivante :

N1 = n1 et Ni = Ni−1 + Fi pour i ∈ {2, · · · i}.

Fi est la fréquence cumulée en xi c'est-`a-dire la proportion d'individus dans l´´échantillon ayant une valeur inférieure ou ´égale `a xi. Le calcul des Fi peut se faire de la même manière que N_i

F1 = f1 et Fi = Fi−1 + fi pour i ∈ {2, · · · i}.

Valeurs	*Effectif (ni)*	*Fréquence (fi)*	*Effectif Cumulées (N_i)*	*Fréquence Cumulées (F_i)*
X_1	n_1	f_1	N_1	F_1
X_2	n_2	f_2	N_2	F_2
.	.	.	.	.
.	.	.	.	.
X_i	n_i	f_i	N_i	F_i
Totale	n	1	-	-

Si la variable d'intérêt est la parité (nombre d'accouchements) pour les mères ayant accouché dans une certaine maternité pendant une année donnée.

Les données brutes sont les n valeurs {x i , i = 1,..,n}

Si n l'effectif de l'échantillon est élevé (par exemple, n = 1000), le tableau correspondant au données est illisible.

Tableau de données brutes

Mère	*Parité*
1	1
2	2
3	4
4	6
5	2
6	5
7	1
..	….
..	….
..	….
..	….
999	2
1000	4

Tableau statistique de la variable parité

Parité	Effectif (n_i)	Fréquence (f_i)	Effectif Cumulées (Ni)	Fréquence Cumulées (Fi)
1	319	0,319	319	0,319
2	234	0,234	553	0,553
3	123	0,123	676	0,676
4	150	0,15	826	0,826
5	98	0,098	924	0,924
6	76	0,076	1000	1
Total	1000	1	-	-

Représentation graphique d'une variable qualitative discrète

Une variable quantitative discrète peut être représentée par le diagramme en bâtons. Le diagramme en bâtons associe `à chaque valeur de la variable un segment vertical de hauteur proportionnelle à la fréquence de cette valeur dans la série statistique.

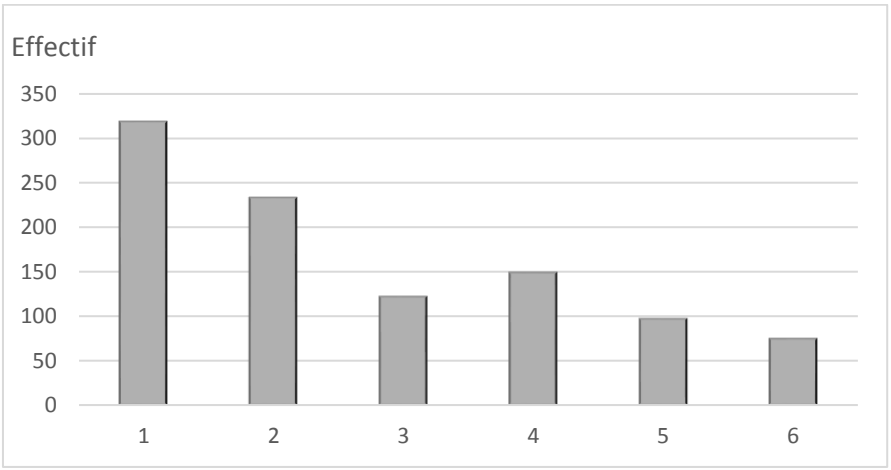


Diagramme en en barres de la variable parité

Distribution d'une variable quantitative continue

À partir de l'observation d'une variable quantitative continue sur n individus (avec n suffisamment grand), on peut d´déterminer k classes statistiques et construire un tableau statistique.

ni est l'effectifs associe `a la classe] ai−1, ai] c'est-`a-dire le nombre d'individus ayant une valeur comprise entre ai−1 (exclus) et ai dans l'´échantillon.

n est la taille de l´échantillon (nombre total d'individus dans cet ´échantillon) ;

fi = ni/n est la fréquence associée `a la classe] ai−1, ai] c'est-`a-dire la proportion d'individus ayant une valeur comprise entre ai−1 (exclus) et ai dans l´échantillon.

N_i est l'effectif c cumulée en ai c'est-à-dire le nombre d'individus dans l´échantillon ayant une valeur inférieure ou égale à a_i. Le calcul des Ni peut se faire façon récurrente de la manière suivante :

N(a1) = N1 et N(ai) = N(ai−1) + Ni pour i ∈ {2, · · · i}.

Fi est la fréquence cumulée en a_i. C'est-`a-dire la proportion d'individus dans l'´échantillon ayant une valeur inférieure ou égale à a_i. Le calcul des Fi peut se faire de la même manière que N_i

F(a1) = f1 et F(ai) = F(ai−1) + fi pour i ∈ {2, · · · i}.

i est le nombre de valeurs distinctes observées dans l´échantillon.

Les bornes de classe vérifient bien évidemment : a0 < a1 < a2 < · · · < ak.

classe	*Effectif (ni)*	*Fréquence (fi)*	*Effectif Cumulées (N_i)*	*Fréquence Cumulées (F_i)*
$]a_0\ a_1]$	n_1	f_1	N_1	F_1
$]a_1\ a_2]$	n_2	f_2	N_2	F_2
.	.	.	.	.
.	.	.	.	.
$]a_{i-1}\ a_i]$	n_i	f_i	N_i	F_i
Totale	n	1	-	-

Il existe quelques formules "toute faites" pour déterminer à l'aveugle le nombre n de classes à partir du nombre N de données :

La règle de STURGE : Nombre de classes = 1+ (3,3 log n, base=10)

La règle de Brooks-Carruthers 5*log (n,base=10)

Amplitude ou l'intervalle de classe : c'est la largeur d'une classe. Pour trouver l'amplitude, on prend la valeur de l'étendue et on divise ce nombre par le nombre de classe

Étendue : C'est la différence entre la plus grande et la plus petite valeur d'une distribution donnée.

Centre de classe

Si la variable d'intérêt est poids du nouveau-né pour les mères ayant accouché dans une certaine maternité pendant une année donnée. Les données brutes sont les n valeurs {x i , i = 1,..,n}

Si n l'effectif de l'échantillon est élevé (par exemple, n = 1000), le tableau correspondant au données est illisible.

Tableau de données brutes

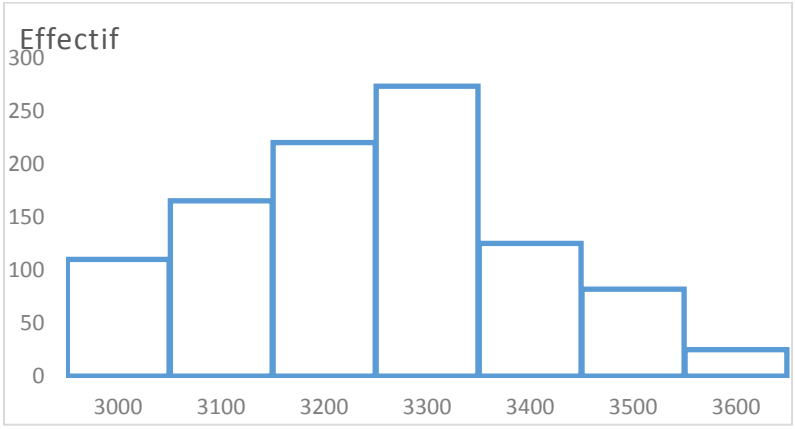
Mère	*poids du nouveau-né*
1	3000
2	3250
3	3320
4	3080
5	3550
6	3700
7	3300
..	....
..	....
..	....
..	....
999	3120
1000	3680

Tableau statistique de la variable poids du nouveau-né

Classe de poids du nouveau-né	*Effectif (n_i)*	*Fréquence (f_i)*	*Effectif Cumulées (Ni)*	*Fréquence Cumulées (Fi)*
] 3000 3100]	110	0,11	110	0,11
] 3100 3200]	165	0,165	275	0,275
] 3200 3300]	220	0,22	495	0,495
] 3300 3400]	273	0,273	768	0,768
] 3400 3500]	125	0,125	893	0,893
] 3500 3600]	82	0,082	975	0,975
] 3600 3700]	25	0,025	1000	1
Total	1000	1	-	-

Représentation graphique d'une variable qualitative continue

Une variable quantitative continue peut être représenté par un l'histogramme. L'histogramme est une juxtaposition de rectangles, chaque rectangle ´étant associe à une classe statistique et ´étant de hauteur est proportionnelle à la fréquence de cette classe.



Histogramme de la variable poids du nouveau-né

Les paramètres des statistiques descriptives

Comment aller encore plus loin dans la « simplification » ou la « réduction » des données d'un échantillon ?

Paramètres de position d'une variable quantitative

Ils visent à résumer la zone des réels où se trouvent les observations faites sur l'échantillon

La moyenne

C'est le centre de gravité de la distribution. Elle s'exprime dans l'unité de la variable et ne se calcule que pour les variables quantitatives.

Soit un échantillon de n valeurs observées x 1 , x 2 , ….,x i ,….,x n d'un caractère quantitatif X,

on définit sa moyenne observée x comme la moyenne arithmétique des n valeurs :

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$$

Si les données observées x i sont regroupées en k classes d'effectif n i (caractère continu regroupé en classe ou caractère discret), il faut les pondérer par les effectifs correspondants:

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{k} n_i x_i \quad \text{avec} \quad n = \sum_{i=1}^{k} n_i$$

La médiane

La médiane, est la valeur du caractère pour laquelle la fréquence cumulée est égale à 0,5 ou 50%. Elle correspond donc au centre de la série statistique classée par ordre croissant, ou à la valeur pour laquelle 50% des valeurs observées sont supérieures et 50% sont inférieures.

Dans le cas où les valeurs prises par le caractère étudié ne sont pas regroupées en classe, est apr »s une ordination des données par ordre croissante ou décroissant

• Si n est impair, la médiane est la valeur $X_{(n+1)/2.}$

• si n est pair, la médiane est une valeur la moyenne des deux valeurs de milieu (la médiane égale à la valeur $(X_{n/2} + X_{(n/2)+1})/2$).

Dans le cas où les valeurs prises par le caractère étudié sont groupées en classe, on cherche

la classe contenant le n^e /2 individu de l'échantillon. En supposant que tous les individus de cette classe sont uniformément répartis à l'intérieur, la position exacte du n^e /2 individu.

$$M_e = x_m + (x_{m+1} - x_m)\left(\frac{\frac{n}{2} - N_i}{n_i}\right)$$

M e= la valeur médiane

x m : limite inférieure de la classe dans laquelle se trouve le n ^{e} /2 individu (classe médiane).

x m+1 : limite supérieure de la classe dans laquelle se trouve le n ^{e} /2 individu (classe médiane).

n i : effectif de la classe médiane

N i. Effectif cumulé inférieur à x m

n : taille de l'échantillon

Le mode

Le Mode d'une série statistique est la valeur du caractère la plus fréquente ou dominante dans l'échantillon. Le mode correspond à la classe de fréquence maximale dans la distribution des fréquences.

Une distribution de fréquences peut présenter un seul mode (distribution unimodale) ou plusieurs modes (distribution bi ou trimodale).

Si la distribution des valeurs est symétrique, la valeur du mode est proche de la valeur de la moyenne arithmétique et de la valeur médiane.

Comparaison des indicateurs de position

	Avantages	Inconvenients
Moyenne arithmétique	- Facile à calculer, - Répond au principe des moindres carrés.	- Fortement influencée par les valeurs extrêmes de la v.a., - Représente mal une population hétérogène (polymodale).
Médiane	- Pas influencée par les valeurs extrêmes de la v.a., - Peu sensible aux variations d'amplitude des classes, - Calculable sur des caractères cycliques (saison, etc.) où la moyenne a peu de signification.	- Se prête mal aux calculs statistiques, - Suppose l'équi-répartition des données - Ne représente que la valeur qui sépare l'échantillon en 2 parties égales.
Mode	- Pas influencée par les valeurs extrêmes de la v.a., - Calculable sur des caractères cycliques (saison, etc.) où la moyenne a peu de signification, - Bon indicateur de population hétérogène.	- Se prête mal aux calculs statistiques, - Très sensible aux variations d'amplitude des classes, - Son calcul ne tient compte que des individus dont les valeurs se rapprochent de la classe modale.

Paramètres de dispersion d'une variable quantitative

Ces paramètres ont pour objectif dans le cas d'un caractère quantitatif de caractériser la variabilité des données dans l'échantillon. Les indicateurs de dispersion fondamentaux sont la variance observée et l'écart-type observé.

La variance observée

Soit un échantillon de n valeurs observées x 1 , x 2 , ….,x i ,….,x n d'un caractère quantitatif X et soit x sa moyenne observée. On définit la variance observée notée s^2 comme le carré des écarts à la moyenne devisé sur le degré de liberté.

$$s^2 = \frac{1}{N-1}\sum_{i=1}^{N}(X_i - \bar{X})^2$$

L'écart-type observé correspond à la racine carrée de la variance observée:

$$s = \sqrt{s^2}$$

La variance est toujours un nombre positif. Sa dimension est le carré de celle de la variable. Il est toutefois difficile d'utiliser la variance comme mesure de dispersion car le recours au carré conduit à un changement d'unités. Elle n'a donc pas de sens biologique direct contrairement à l'écart-type qui s'exprime dans les mêmes unités que la moyenne.

Le coefficient de variation

La variance et l'écart-type observée sont des paramètres de dispersion absolue qui mesurent la variation absolue des données indépendamment de l'ordre de grandeur des données.

Le coefficient de variation noté C.V. est un indice de dispersion relatif prenant en compte ce biais et est égal à :

$$C.V. = \frac{100s}{\bar{x}}$$

Exprimé en pour cent, il est indépendant du choix des unités de mesure permettant la comparaison des distributions de fréquence d'unité différente.

Probabilités et variables aléatoire

Notion de base

Considérons le jeu du lancer d'un dé. Notons Ω l'ensemble de tous les résultats possibles (appelés aussi épreuves ou résultats élémentaires) de cette expérience aléatoire

$$\Omega = \{1, 2, 3, 4, 5, 6\}.$$

On note ω = 3 pour signifier que 3 est le résultat de l'épreuve.

Dans cette expérience aléatoire, on peut s'intéresser à des événements plus complexes qu'un simple résultat élémentaire. On peut, par exemple, considérer l'événement A = "le résultat est un nombre pair" ou l'événement B = "le résultat est un nombre plus grand que 3". On note A l'ensemble de ces événements. Notons que l'on a toujours $A \subset P(\Omega)$, où $P(\Omega)$ est l'ensemble des parties de Ω. Notons que l'inclusion précédente peut être stricte.

On dit que l'événement A s'est réalisé si le résultat de l'expérience ω est tel que $\omega \in A$.

Enfin, on peut donner à chaque événement une pondération ou encore une probabilité. Ainsi, si le dé n'est pas pipé, l'intuition nous dit que la probabilité d'avoir l'événement A ="le résultat est un nombre pair" est 1/2,

$$P(A) = 1/2$$

Tout phénomène aléatoire ou expérience aléatoire fait appel à deux ensembles de type différent :

• Un ensemble Ω, appelé espace fondamental ou univers, qui contient l'ensemble de tous les résultats possibles. Ces derniers sont également appelés épreuves.

•Une famille A de parties (i.e. de sous-ensembles) de Ω. Ces parties sont appelées des événements. On dit que l'événement A s'est réalisé si et seulement si le résultat ω de Ω qui s'est produit appartient à A.

Quelques définitions à connaître

Évènements exclusifs

Des événements sont dits « exclusifs » lorsqu'ils ne peuvent se produire simultanément.

Dans le cas où une expérience se résume à deux événements exclusifs, alors la réalisation de l'un des évé nements implique forcément la non-réalisation de l'autre.

Exemple : fille ou garçon

Évènements non exclusifs

Ils peuvent se produire en même temps.

Exemple : être malade et avoir des douleurs

Évènements indépendants

Deux événements sont dits indépendants lorsque la réalisation de l'un n'influence pas la réalisatio n de l'autre.

Exemple : avoir une bronchite et avoir une voiture.

A bien noter que ces événements peuvent (ou pas) se réaliser en même temps, à ne pas confondre avec l' exclusivité !

Évènements non indépendants :

Deux événements sont dits dépendants lorsque la réalisation de l'un influence la réalisation de l'autre.

Exemple : être allergique et avoir le nez qui coule.

Axiomes élémentaires à connaître également :

Si 2 événements A et B sont exclusifs, alors :

$$P(A\ ou\ B) = P(A + B) = P(A \cup B) = P(A) + P(B)$$

Si ces 2 événements constituent l'ensemble des possibles et qu'ils sont mutuellement exclusifs ;

$$P(A) + P(B) = 1$$

Si on connaît la probabilité de l'un, alors on peut facilement calculer la probabilité de l'autre ;

$$P(A) = 1 - P(B) \text{ ou } P(B) = 1 - P(A).$$

Si 2 événements sont non exclusifs ;

$$P(A \cup B) = P(A) + P(B) - P(A \cap B)$$

Concernant les probabilités conditionnelles, on étudie la probabilité qu'un événement se produise sachant qu'un autre s'est déjà réalisé

$$P_{A/B} = \frac{P(A \cap B)}{P(B)}.$$

Si la probabilité de A sachant que B est réalisé est la même que la probabilité de A sans condition, alors les deux événements sont indépendants

$$P_{A/B} = \frac{P(A \cap B)}{P(B)} = P(A).$$

Dans ce cas

$$P(A \cap B) = P(A) \times P(B).$$

Les lois de probabilité associée à une variable aléatoire

Intuitivement, il s'agit d'un nombre — le concept est bien sûr généralisable — dont la valeur dépend du résultat d'une expérience aléatoire. Souvent, on ignorera l'expérience en question, on ne fera aucune allusion à Ω ou à l'un de ses éléments. La loi d'une variable aléatoire sera alors la probabilité qu'elle prenne une certaine valeur. Cette formalisation permet de déterminer des caractéristiques intéressantes d'une loi.

Lois de probabilité associée à une variable aléatoire discrète

Loi de Bernoulli

On s'intéresse ici à la réalisation ou non d'un ´évènement. Autrement dit, on n'étudie que les expériences aléatoires qui n'ont que deux issues possibles (ex : un patient à l'hôpital survit ou non, un client signe le contrat ou non). Considérons une expérience aléatoire de ce type. On l'appelle une ´épreuve de Bernoulli. Elle se conclut par un succès si l''évènement auquel on s'intéresse est réalisé ou un ´échec sinon. On associe à cette ´épreuve une variable aléatoire X qui prend la valeur. Si l''évènement est réalisé et la valeur 0 sinon. Cette v.a. ne prend donc que deux valeurs (0 et 1).

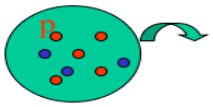
- Loi : $X \sim \mathcal{B}(p) \Leftrightarrow$

$$P(X = x) = p^x q^{1-x}, \quad x \in \{0,1\}$$

- Moments

$$E(X) = p \quad ; \quad V(X) = pq$$

$\mathcal{E}$: Tirage dans une urne de Bernoulli ayant une proportion p de boules rouges. q=1-p



X=nombre de boules rouges

Loi binomiale

Soit X la v.a. qui représente le nombre de succès obtenus lors des n ´épreuves d'un schéma de Bernoulli. Alors on dit que X suit une loi binomiale de paramètres (n, p), notée B (n, p). Cette loi est donnée par

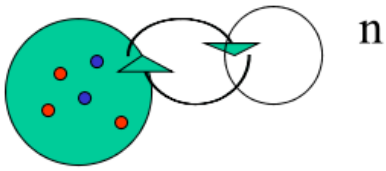
- Loi : $X \sim \mathcal{B}(n,p) \Leftrightarrow$

$$P(X=x) = C_n^x p^x q^{n-x} \ \forall x \in \{0,\ldots,n\}$$

Moments :

$$E(X) = np \quad ; \quad V(X) = npq$$

$\mathcal{E}$: n tirages avec remise dans une urne de Bernoulli ayant une proportion p de boules rouges



X=nombre de boules rouges

Loi de Poisson

On utilise la loi de Poisson pour modéliser par exemple le nombre le nombre de globules rouges dans un ml de sang, le nombre d'accidents du travail dans une entreprise pendant un an...par exemple

- Loi : $X \sim \mathcal{P}(\lambda), \lambda > 0 \Leftrightarrow$

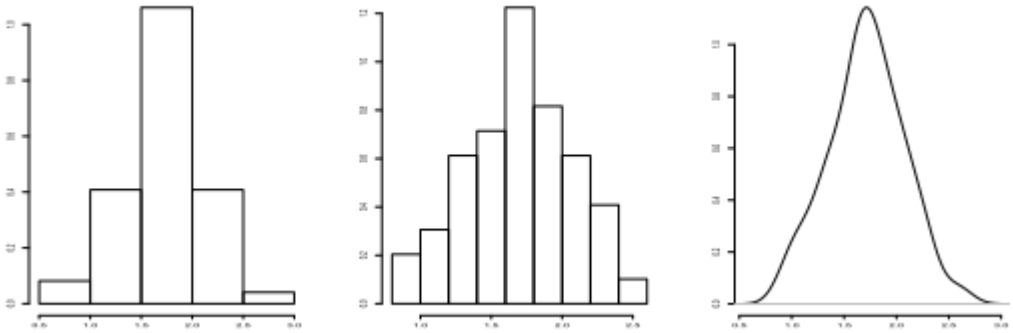
$$P(X=x) = e^{-\lambda}\frac{\lambda^x}{x!}, \quad \forall x \in \mathbb{N}$$

- Moments

$$E(X) = V(X) = \lambda$$

Loi de probabilité associée à une variable aléatoire continue

Un biologie en fait relever régulièrement des paramètre continu (dosage de molécules par exemple ou prise de poids) si on regroupe les valeurs pour tracer un histogramme. Suivant le nombre de classes choisi, on obtient différents tracés. On peut imaginer de raffiner encore en faisant d'autres mesures avec des instruments de mesure plus précis.



On aurait alors une courbe, d'aire 1 comme les histogrammes, qui représente la manière dont sont reparties les valeurs de la v.a. X . Cette courbe est la courbe d'une fonction appelée densité de probabilité Une densité f d´écrit la loi d'une v.a. X en ce sens :

$$\text{pour tous } a, b \in \mathbb{R}, \qquad P[a \leq X \leq b] = \int_a^b f(x)\mathrm{d}x$$

La loi normale et la loi normale centrée réduite

C'est la loi la plus importante. Son rôle est central dans de nombreux modèles probabilistes et dans toute la statistique. Elle possède des propriétés intéressantes qui la rendent agréable à utiliser.

Une v.a. X suit une loi normale (ou loi gaussienne ou loi de Laplace-Gauss) N (0, 1) si sa densité f est donnée par

$$f(x) = \frac{1}{\sqrt{2\pi}} \exp\Big(-\frac{x^2}{2}\Big), \ \textit{pour tout } x \in \mathbb{R}$$

$$\text{si } X \text{ suit une loi normale } \mathcal{N}(0,1), \text{ alors pour tous } a < b,$$

$$P[a \leq X \leq b] = \frac{1}{\sqrt{2\pi}} \int_a^b \exp\Big(-\frac{x^2}{2}\Big)\mathrm{d}x = \varphi(b) - \varphi(a)$$

Pour calculer P[a ≤ X ≤ b] ou P[X ≤ x], on a recours au calcul numérique sur ordinateur ou, plus simplement, à une table (table de la loi normale et la loi normale centrée réduite) qui donne P[X ≤ x] pour tout décimal positif x à deux chiffres après la virgule.

Distributions dérivant du la loi normale

Les distributions que nous allons étudier sont importantes non pas pour représenter des modèles théoriques de séries statistiques comme les précédentes, mais en raison du rôle qu'elles jouent dans les problèmes d'estimation ou de tests que nous verrons par la suite. Pour l'instant leurs définitions peuvent sembler complexes, notamment parce que la notion de « degrés de liberté » n'a pas encore été précisée. Pour le moment, il importe simplement de connaître leur définition et de savoir lire les tables correspondantes.

La distribution du Khi-deux χ2

Cette distribution (qui se prononce khi-deux) est très importante pour tester l'ajustement d'une loi théorique à une distribution expérimentale (test du χ2) et pour déterminer la loi de la variance d'un échantillon.

Si X1,X2,.....,Xn sont n variables aléatoires indépendantes qui suivent toute la loi normale centrée réduite,

alors la quantité $X = X_1^2+X_2^2+.....+X_n^2 = \sum_{i=1}^{n} X_i^2$ est une variable aléatoire distribuée selon la loi de χ^2 à n degrés de liberté

Pour des raisons de commodité, au lieu de donner la table des fonctions de répartition des variables aléatoires χ2 pour les différentes valeurs de n, on donne, en fonction de n (nombre de degrés de liberté) et d'une probabilité α que l'on peut choisir, la valeur χ 2 .

La distribution de Fischer

Cette distribution fut découverte par l'anglais Fisher en 1924 puis tabulée par Snédecor en 1934. Elle interviendra lors des comparaisons des variances de deux échantillons (test d'hypothèse F).

Si χ^2_1 et χ_2^2 sont deux variables aléatoires indépendantes qui suivent toutes les deux une loi de khi-deux de degrés de liberté respectifs n1 et n2. Alors la quantité

$$F = \frac{\chi_1^2 / n_1}{\chi_2^2 / n_2}$$

Est variable aléatoire qui suit la loi de Fischer-Snedecor à n1 et n2 degrés de liberté.

Les valeurs tabulées de la variable F dépendent d'un seuil α que l'on peut choisir et des nombres de degré de liberté n 1 et n 2.

La distribution de Student (pseudonyme de V.S Gosset - 1908)

Soient X et Y deux variables aléatoires indépendantes, la première étant distribuée selon une loi normale centrée réduite N(0,1) et la deuxième selon une loi de khi-deux à n degrés de liberté χ^2_n

La quantité $T = \frac{X\sqrt{n}}{\sqrt{Y}}$ est une variable aléatoire qui suit une loi de Student à n degrés de Liberté

Les valeurs tabulées de la variable T dépendent d'un seuil α que l'on peut choisir et du nombre de degré de liberté n.