XV – Les statistiques descriptives

1/ Généralités

Les statistiques consistent à extraire de l'information à partir d'un ensemble de données. Contrairement aux probabilités, où l'on modélise une expérience aléatoire afin de déterminer la probabilité de certains événements, on utilise en statistiques les résultats d'une expérience aléatoire pour construire un modèle permettant d'expliquer au mieux ces observations.

1.1) Population statistique

Définition

- *Une **population** (statistique) est un ensemble fini d'éléments soumis à une étude.*
- *Les éléments de la population sont appelés **individus**.*
- *Le nombre total d'individus est appelé **effectif total** ou **taille** de la population.*
- *Une partie de la population est appelée **échantillon** de la population.*

1.2) Caractère et modalité

Définition

*Un **caractère** (ou **variable statistique**) est une donnée que l'on peut observer sur les individus d'une population. Ce caractère est dit **quantitatif** lorsqu'il prend des valeurs numériques mesurables et **qualitatif** sinon.*

Remarque : on ne traite pas de la même manière les caractères quantitatifs et qualitatifs. En particulier, les calculs de moyenne ou de variance n'ont aucun sens pour ces derniers. Dans ce cours, nous nous intéresserons essentiellement aux variables quantitatives.

Définition

*Une **série statistique** est la donnée d'une population et d'un ou plusieurs caractères portant sur cette population. On parle de :*

- ***statistiques univariée** lorsque l'on observe un caractère par individu.*
- ***statistiques bivariées** lorsque l'on mesure deux caractères par individu.*

Remarque : une fois choisi le caractère étudié, les statistiques ignorent les individus au profit des valeurs du caractères. Par exemple, si l'on étudie la taille d'une population, on ne retient pas la taille de telle ou telle personne mais juste le fait qu'un individu a une taille donnée.

Définition

- *Les valeurs (distinctes) prises par un caractère s'appellent les **modalités** de ce caractère. Pour décrire un caractère x, on peut donner la liste de ses modalités et noter $x = (x_1, \dots, x_p)$. On dit alors que x est **discret**.*
- *Lorsque les modalités sont trop nombreuses, on les regroupe en intervalles de valeurs appelés **classes** et dont les longueurs sont appelées **amplitudes**. On dit alors que x est **continu**.*

Remarques :

❶ Les différentes classes doivent former une **partition** de l'ensemble des modalités, ce qui signifie qu'elles sont deux à deux disjointes et que leur réunion vaut cet ensemble. De la sorte, toute modalité appartient à une et une seule classe.

❷ En outre, on suppose souvent que la répartition au sein de chaque classe est **uniforme**, ce qui permet de la remplacer par son centre lors des calculs de moyenne (par exemple).

2/ Statistiques univariées

Dans cette partie, une population d'effectif total N est étudiée à travers un caractère quantitatif discret x dont les modalités $(x_1, x_2, x_3, \cdots, x_p)$ sont rangées dans l'ordre croissant. Les diverses notions exposées s'adaptent facilement au cas d'un caractère continu.

2.1) Effectifs et fréquences

Définition (effectifs)

- *Pour tout $k \in [\![1 ; p]\!]$, on appelle **effectif** de la modalité x_k le nombre n_k d'individus pour lesquels le caractère prend la valeur x_k. Ces données, qui constituent la série statistique de x, sont habituellement rangées dans un tableau à deux lignes modalités / effectifs :*

Modalités	x_1	x_2	x_3	$\dots$	x_p
Effectifs	n_1	n_2	n_3	$\dots$	n_p

- *Pour tout $k \in [\![1 ; p]\!]$, on appelle **effectif cumulé croissant** de la modalité x_k le nombre $n_1 + \cdots + n_k$ d'individus pour lesquels le caractère prend une valeur inférieure ou égale à x_k.*

Remarque : la somme des effectifs vaut l'effectif total, c'est-à-dire $n_1 + n_2 + \cdots + n_p = \mathrm{N}$.

Exemples :

❶ On classe suivant leurs pointures les chaussures vendues dans un magasin en un mois

Pointure	34	35	36	37	38	39	40	41	42	43	44	45	46
Effectif	10	20	30	131	241	122	112	154	290	124	30	20	16
Eff. cumulé	10	30	60	191	432	554	666	870	1160	1284	1314	1334	1350

❷ On classe en fonction de leur nombre d'habitants les villes d'un département donné :

Nombre d'habitants	[0, 100[	[100, 1000[	[1000, 10000[	[10000, 80000]
Effectif	102	143	27	2
Effectif cumulé	102	245	272	274

L'effectif d'une modalité n'est pas une information suffisante en soi car elle ne prend pas en compte l'effectif total. Pour remédier à ce problème, on introduit la notion de fréquence (on conservera les notations de la définition précédente tout au long de la partie 2).

Définition (fréquences)

- *Pour tout $k \in [\![1 ; p]\!]$, on appelle **fréquence** de la modalité x_k le réel $f_k = \dfrac{n_k}{\mathrm{N}}$.*
  C'est la fréquence d'apparition d'individus pour lesquels le caractère prend la valeur x_k.
- *Pour tout $k \in [\![1 ; p]\!]$, on appelle **fréquence cumulée croissante** de la modalité x_k le réel $f_k = \dfrac{n_1 + n_2 + \cdots + n_k}{\mathrm{N}}$. C'est la fréquence d'apparition d'individus pour lesquels le caractère prend une valeur inférieure ou égale à x_k.*

Proposition

La somme des fréquences vaut toujours 1.

2.2) Représentations graphiques

Pour faciliter la visualisation des séries statistiques, les fréquences sont représentées graphiquement sous forme de diagrammes. Voici les plus usuels.

(a) Le diagramme en bâtons

C'est une suite de fins rectangles de hauteurs proportionnelles aux fréquences.

(b) Le diagramme circulaire

C'est un disque découpé en différents secteurs dont les angles au centre sont proportionnels aux fréquences. Les normands parlent aussi de *diagramme camembert*.

(c) L'histogramme

On l'utilise pour les caractères continus. C'est une suite de rectangles dont les bases sont proportionnelles aux amplitudes des différentes classes et dont les **surfaces** (et non les hauteurs) sont proportionnelles à leurs fréquences.

(d) Le polygone des fréquences cumulées croissantes

Il représente la fonction qui à x associe la fréquence d'apparition d'individus pour lesquels le caractère prend une valeur inférieure ou égale à x. C'est une courbe en escalier qui effectue un saut de f_k au niveau de chaque modalité x_k.

Dans le cas d'un caractère continu, on trace une ligne brisée qui monte de f_k le long de la k^{e} classe. On peut déterminer son équation pour calculer précisément les intersections utilisées plus loin.

2.3) Caractéristiques de position

Elles permettent de voir autour de quelles valeurs se concentrent les valeurs prises par le caractère étudié dans la série statistique.

Définition (mode et classe modale)

*On appelle **mode** d'un caractère discret une valeur d'effectif maximal, et **classe modale** d'un caractère continu une classe de rapport effectif/amplitude maximal.*

Définition (moyenne)

*On appelle **moyenne** du caractère x le réel* $\overline{x} = \dfrac{1}{\mathrm{N}}\sum_{k=1}^{p} n_k\, x_k = \sum_{k=1}^{p} f_k\, x_k$.

Dans le cas d'un caractère continu, on prendra les centres des classes à la place des x_k.

Proposition (linéarité de la moyenne)

Soient x et y deux caractères d'une même population tels que $y = ax + b$, *avec a et b deux réels. Alors* $\overline{y} = a\overline{x} + b$.

Remarques :

❶ Si le caractère x vérifie $x_k \in [\,a\,;b\,]$ pour tout k, alors $\overline{x} \in [\,a\,;b\,]$.
De la même manière, si $x_k \leqslant y_k$ pour tout k, alors $\overline{x} \leqslant \overline{y}$.

❷ Si x a une moyenne $\overline{x_1}$ sur une population de taille N_1 et une moyenne $\overline{x_2}$ sur une autre population de taille N_2, alors sa moyenne sur la réunion des deux populations est $\overline{x} = \dfrac{\mathrm{N}_1\overline{x_1} + \mathrm{N}_2\overline{x_2}}{\mathrm{N}_1 + \mathrm{N}_2}$. Ceci permet ainsi d'actualiser une moyenne après l'acquisition de nouvelles données.

Définition (médiane)

- *On appelle **médiane** de la série statistique la valeur* M_e *qui partage la population en deux parties égales : le caractère x prend une valeur strictement inférieure à* M_e *pour la moitié des individus et une valeur supérieure à* M_e *pour l'autre moitié.*
- *Lorsque* $\mathrm{N} = 2n + 1$ *est impair, la médiane est* x_{n+1}.
  Lorsque $\mathrm{N} = 2n$ *est pair, on prend pour médiane la moyenne de* x_n *et* x_{n+1}.

Remarques :

❶ Contrairement à la moyenne, la médiane est insensible aux valeurs extrêmes.

❷ Graphiquement, la médiane est l'abscisse du point d'intersection du polygone des fréquences cumulées croissantes et de la droite d'équation $y = 1/2$ (si cette intersection est un segment horizontal, on prend son milieu).

❸ Dans le cas d'un caractère continu, on appellera de même classe médiane la classe qui partage la population en deux. Elle contient la médiane.

2.4) Caractéristiques de dispersion

On s'intéresse désormais à la dispersion du caractère autour des valeurs de référence.

Définition (étendue)

*On appelle **étendue** du caractère x la différence entre la plus grande valeur et la plus petite valeur de la série statistique. Elle donne une idée de son étalement.*

Définition (variance et écart-type)

- *On appelle **variance** de x le réel positif* $\mathrm{V}_x = \dfrac{1}{\mathrm{N}}\displaystyle\sum_{k=1}^{p} n_k\,(x_k - \overline{x})^2 = \sum_{k=1}^{p} f_k\,(x_k - \overline{x})^2$.
- *On appelle **écart-type** du caractère x le réel positif* $s_x = \sqrt{\mathrm{V}_x}$.

Remarques :

❶ la variance est en fait **la moyenne des écarts quadratiques**. Elle mesure la dispersion de la série statistique autour de sa moyenne et est peu sensible aux rares valeurs extrêmes.

❷ L'écart-type s_x possède la même unité que le caractère x, et il est nul uniquement pour un caractère constant.

Proposition (formule de Kœnig-Huygens)

On a $\mathrm{V}_x = \overline{x^2} - \overline{x}^2$ *: la variance est la moyenne du carré moins le carré de la moyenne.*

Proposition

Soient x et y deux caractères d'une même population tels que $y = ax + b$, *avec* $(a, b) \in \mathbb{R}^2$. *Alors* $\mathrm{V}_y = a^2\mathrm{V}_x$ *et* $s_y = |a|\,s_x$.

On généralise la notion de médiane en découpant la population en 4 puis en 10 parties.

Définition (quartiles)

- *Le **premier quartile** Q_1 est la plus petite valeur de x telle qu'au moins 25 % des valeurs de la série lui soient inférieures ou égales.*
- *Le **deuxième quartile** Q_2 est la médiane.*
- *Le **troisième quartile** Q_3 est la plus petite valeur de x telle qu'au moins 75 % des valeurs de la série lui soient inférieures.*
- *L'**écart interquartile** est la différence $\mathrm{Q}_3 - \mathrm{Q}_1$. C'est la longueur de l'**intervalle interquartile** $[\,\mathrm{Q}_1\,;\mathrm{Q}_3\,]$, qui contient environ la moitié de la population.*

Remarque : graphiquement, le k^{e} quartile Q_k est en fait l'abscisse du point d'intersection du polygone des fréquences cumulées croissantes et de la droite d'équation $y = k/4$.

Définition (déciles)

- *Pour $k \in [\![1 ; 10]\!]$, le* ***k^e décile*** D_k *est la plus petite valeur de x telle qu'au moins $10k$ % des valeurs de la série lui soient inférieures ou égales.*
- *L'****écart interdécile*** *est la différence $D_9 - D_1$. C'est la longueur de l'****intervalle interdécile*** $[D_1 ; D_9]$*, qui contient environ 80 % de la population.*

Remarque :

❶ Graphiquement, le k^e décile D_k est l'abscisse du point d'intersection du polygone des fréquences cumulées et de la droite d'équation $y = k/10$.

❷ L'écart interdécile est un indicateur que l'on préfère à l'étendue car on s'est débarrassés des valeurs extrêmes souvent peu significatives (erreurs de mesure, etc...)

3/ Statistiques bivariées

Dans toute cette partie, une population d'effectif total N est étudiée à travers deux caractères quantitatifs discrets x et y dont les modalités $(x_1, \dots, x_p)$ et $(y_1, \dots, y_q)$ sont rangées dans l'ordre croissant.

3.1) Effectifs et fréquences conjoints

Définition (effectifs conjoints et marginaux)

- *Soient $k \in [\![1 ; p]\!]$ et $\ell \in [\![1 ; q]\!]$. Le couple (x_k, y_ℓ) est une* ***modalité conjointe*** *de (x, y) et son* ***effectif conjoint*** *est le nombre $n_{k,\ell}$ d'individus pour lesquels $x = x_k$ et $y = y_\ell$.*
- *Pour tout $k \in [\![1 ; p]\!]$, on appelle* ***effectif marginal*** *de x_k le nombre $n_{k,\bullet}$ d'individus pour lesquels le caractère x prend la valeur x_k. Il vérifie $n_{k,\bullet} = \sum_{\ell=1}^{q} n_{k,\ell}$.*
- *Pour tout $\ell \in [\![1 ; q]\!]$, on appelle* ***effectif marginal*** *de y_ℓ le nombre $n_{\bullet,\ell}$ d'individus pour lesquels le caractère y prend la valeur y_ℓ. Il vérifie $n_{\bullet,\ell} = \sum_{k=1}^{p} n_{k,\ell}$.*

Remarques :

❶ L'ensemble de ces données constitue la série statistique double de (x, y).

❷ On peut ranger les effectifs conjoints dans un tableau à double entrée, en rajoutant les effectifs marginaux au bout des lignes et des colonnes (ce sont les sommes des effectifs de la ligne ou de la colonne correspondante).

❸ Les effectifs marginaux ont pour somme l'effectif total.

Définition (fréquences conjointes et marginales)

On définit de la même façon la ***fréquence conjointe*** *et la* ***fréquence marginale*** *en divisant les effectifs conjoints et marginaux par l'effectif total.*

3.2) Représentation graphique

Pour visualiser une série statistique double, on utilise un **nuage de points** défini de la manière suivante : chaque modalité (x_k, y_ℓ) est représentée par un disque de centre $C_{k,\ell}(x_k, y_\ell)$ et de surface proportionnelle à la fréquence conjointe $f_{k,\ell}$.

Définition (point moyen)

*Le **point moyen** de la série statistique double de (x, y) est le point de coordonnées $(\overline{x}, \overline{y})$.*

3.3) Covariance et coefficient de corrélation

On cherche maintenant à savoir si les caractères x et y sont liés.

Définition (covariance)

*On appelle **covariance** des caractères x et y le réel*

$$s_{x,y} = \frac{1}{N} \sum_{\substack{1 \leqslant k \leqslant p \\ 1 \leqslant \ell \leqslant q}} n_{k,\ell} (x_k - \overline{x})(y_\ell - \overline{y}) = \sum_{\substack{1 \leqslant k \leqslant p \\ 1 \leqslant \ell \leqslant q}} f_{k,\ell} (x_k - \overline{x})(y_\ell - \overline{y})$$

Remarque : c'est une mesure de l'influence mutuelle de x et y, qui ont tendance à varier dans le même sens lorsqu'elle est positive et dans des sens opposés lorsqu'elle est négative.

Proposition (formule de Kœnig-Huygens)

On a $s_{x,y} = \overline{xy} - \overline{x}\,\overline{y}$: la covariance est la moyenne du produit moins le produit des moyennes.

Proposition

On a $s_{x,x} = V_x$ et $V_{x+y} = V_x + 2s_{x,y} + V_y$.

Définition (coefficient de corrélation)

*On appelle **coefficient de corrélation (affine)** des caractères x et y le réel $r_{x,y} = \dfrac{s_{x,y}}{s_x s_y}$.*

Proposition

- *Le coefficient de corrélation appartient à l'intervalle $[-1\,;1]$.*
- *Il vaut ± 1 si, et seulement si, il existe a et $b \in \mathbb{R}$ tels que $y = ax + b$.*

Remarque : le coefficient de corrélation affine indique si les caractères x et y sont liés.

- S'il est proche de 1 ou -1, ils sont fortement liés ou **corrélés** et le nuage de points se concentre autour d'une droite.
- S'il est proche de 0, ils ne sont pas corrélés et le nuage de points mérite son nom.

Attention cependant, il n'y a pas forcément de lien de causalité entre deux caractères corrélés, qui peuvent tout simplement dépendre d'un tierce caractère.

3.4) Ajustement affine

On suppose dans cette dernière partie que l'on a mesuré les caractères x et y sur les N individus de la population, et que l'on dispose des N couples deux à deux distincts (x_1, y_1), (x_2, y_2), ..., (x_N, y_N).

Les effectifs conjoints $n_{k,l}$ valent dans ce cas 1 si $k = l$ et 0 sinon. De plus, le nuage de points est constitué de points et non plus de disques de tailles variées.

Lorsque le coefficient de corrélation est proche de ± 1, le nuage de points se concentre autour d'une droite ... que l'on va chercher à déterminer.

Définition (droite de régression)

*On appelle **droite de régression linéaire** de y par rapport à x la droite passant par le point moyen $G(\overline{x}, \overline{y})$ et de pente $a = \dfrac{s_{x,y}}{V_x}$.*

Proposition

La droite de régression linéaire a pour équation $Y = aX + b$, avec $a = \dfrac{s_{x,y}}{V_x}$ et $b = \overline{y} - a\overline{x}$.

Théorème

Cette droite est l'unique droite d'équation $Y = aX + b$ qui minimise la somme des carrés des distances verticales entre la droite et les points du nuage, c'est-à-dire la grandeur

$$S = \sum_{k=1}^{N} (y_k - (ax_k + b))^2$$

Remarque : on dit pour cette raison que l'on obtient la droite de régression linéaire à l'aide de la **méthode des moindres carrés**. C'est une bonne approximation du nuage de points lorsque le coefficient de corrélation est grand.

Autres ajustements

On est amenés à croiser d'autres relations entre les caractères x et y, que l'on peut déterminer à l'aide d'un ajustement affine sur d'autres caractères. Par exemple :

❶ Si l'on penche pour une relation de la forme $y = C\exp(ax)$, on étudie alors $(x, \ln y)$.

❷ Si l'on opte pour une relation de la forme $y = Cx^a$, on étudie $(\ln x, \ln y)$.

Tout dépend du modèle qui se cache derrière les caractères.

Exercices

1/ Statistiques univariées

Exercice 1. Déterminer les différents quartiles des séries statistiques ci-dessous :

1) $\{0;2;5;8;9;12;15;18;20\}$
2) $\{2;5;7;10;12;15;17;20;21;22\}$
3) $\{2;5;7;7;7;9;11;12\}$

Exercice 2.

Afin de renouveler le mobilier d'un lycée, le proviseur commande une enquête sur la taille d'un échantillon de 100 élèves. Voici les résultats obtenus en cm :

165	159	158	185	168	170	154	166	164	163
185	169	157	189	164	185	160	163	164	165
158	185	184	177	170	155	190	187	157	173
158	155	178	183	157	179	178	192	150	182
182	159	150	160	178	176	167	164	157	161
170	169	179	171	173	169	187	187	165	154
189	159	156	158	159	159	166	169	187	191
188	168	153	170	155	165	182	156	179	169
177	153	189	188	166	164	171	189	177	153
189	188	166	164	171	189	158	161	176	168

1) Regrouper ces données en classe de 10 cm d'amplitude $[\,150\,;160\,[$, $[\,160\,;170\,[$, etc... et donner le tableau des effectifs.
2) Calculer la moyenne de cette série.
3) Représenter graphiquement la série statistique.

Exercice 3.

Une entreprise fabrique des pièces métalliques dont le diamètre, mesuré en millimètres, est donné dans le tableau suivant :

Diamètre	$[\,4.0\,;4.2\,[$	$[\,4.2\,;4.4\,[$	$[\,4.4\,;4.6\,[$	$[\,4.6\,;4.8\,[$	$[\,4.8\,;5.0\,[$
Effectif	6	24	41	25	4

En supposant que la répartition est uniforme au sein de chaque classe, calculer la moyenne et l'écart-type de cet échantillon.

Exercice 4.
Procéder de même avec l'échantillon suivant :

Longueur	[35 ; 37 [	[37 ; 39 [	[39 ; 41 [	[41 ; 43 [	[43 ; 45 [
Effectif	3	25	50	20	2

2/ Statistiques bivariées

Exercice 5. Pour chacune des séries statistiques doubles ci-dessous :

- tracer le nuage de points ;
- trouver le point moyen du nuage ;
- proposer un ajustement entre les deux variables ;
- établir cet ajustement par le calcul.

1)

X	1	2	3	4	5	6	7	8
Y	7	6	6	4	5	6	2	3

2)

X	1	2	5	7	10	12	15	20
Y	2	3	5	6	7	8	9	10

3)

X	1	2	3	4	5	6	7	8
Y	3	19	69	356	835	832	1455	1738

Exercice 6.

Plus une région est vaste, plus le nombre d'espèce y vivant est grand. Pour modéliser mathématiquement ce phénomène et mesurer la *biodiversité*, on utilise la loi SPAR càd « species-area relationship ». Celle-ci stipule que la surface S de la région étudiée et le nombre N d'espèces présentes dans cette région sont reliées par la formule

$$N = C\,S^z$$

où C et z sont des constantes dépendant de la région. On fait les mesures suivantes :

S	1	2	4	8	16	32	64
N	6	7	8	10	10	13	14

Afin de déterminer C et z, on effectue une régression linéaire sur de ln(N) par rapport à ln(S).

1) Justifier le choix de cette régression linéaire.
2) Tracer le nuage de points correspondant aux variables ln(N) et ln(S).
3) Déterminer la droite de régression linéaire. et la placer sur le dessin.
4) Donner des valeur approchées de C et z.