STATISTIQUE DESCRIPTIVE

I – Vocabulaire de la statistique descriptive

1) Population

La statistique descriptive est une science qui recueille et analyse des informations sur un ensemble fini, dont le cardinal est souvent très grand.

Définition : L'ensemble étudié s'appelle une population. Les éléments de cet ensemble s'appellent des individus.

La population étant en général très grande, on étudie souvent une partie seulement.

Définition : Un échantillon est une partie de la population. Le cardinal de cette partie s'appelle la taille de l'échantillon.

Dans la suite, Ω désignera la population ou l'échantillon observé. Plus tard (en seconde année), on distinguera les deux car on voudra, à partir d'observations sur l'échantillon, déduire des propriétés de la population entière.

2) Caractère statistique

La question est maintenant : qu'est-ce qu'on étudie sur cette population ?
Exemple : la couleur des yeux, la taille, le poids, le nombre de frères et sœurs, …

Définition : On appelle caractère statistique ou variable statistique toute application X définie sur la population Ω.
Si l'application X est à valeurs dans $\mathbb{R}$, on dira que le caractère est quantitatif. Sinon, on dira que le caractère est qualitatif.

Le premier exemple est qualitatif, alors que les autres sont quantitatifs.

3) Etude du caractère

Les différentes étapes d'une étude statistique sont :

- Recueillir les données.
- Les classer car on les obtient « en vrac ».
- Les représenter graphiquement pour avoir un aspect visuel.
- Analyser ces données, c'est-à-dire les résumer par quelques nombres significatifs.

Pour classer, la première idée est de considérer toutes les valeurs possibles du caractère, donc $X(\Omega)$ et de regrouper tous les éléments ω qui correspondent à la même valeur. Par exemple si l'on observe 10 individus numérotés de 1 à 10 :

i	1	2	3	4	5	6	7	8	9	10
$X(\omega_i)$	2	5	5	8	4	5	4	4	5	8

On renumérote et on va classer sous la forme :

j	1	2	3	4
x_j	2	4	5	8
n_j	1	3	4	2

n_j représente le nombre d'individus dont le caractère prend la valeur x_j. C'est l'effectif de la classe x_j. L'ensemble des couples (x_j, n_j) est une série statistique.

Cependant, dans le cas d'un caractère quantitatif, lorsque les données sont trop nombreuses ou trop proches, on les regroupe en classes qui peuvent être des intervalles de $\mathbb{R}$. On dira que le caractère est quantitatif continu par opposition aux autres qui sont quantitatifs discrets.

II - Variable qualitative

1) Classement des données

Pour une variable qualitative, chaque classe correspond à une valeur du caractère. Le nombre d'individus qui appartiennent à cette classe s'appelle l'effectif de la classe. La somme des effectifs de toutes les classes est l'effectif total de la population.

Exemple : Moyen de transport pour le trajet domicile - travail. Le tableau suivant donne les effectifs de chaque classe. Recopier le tableau et calculer l'effectif total.

Classe	Car - Bus	Auto - Moto	Vélo	A pied	Tram - Métro
Effectif	162	204	18	72	144

L'intérêt d'une étude statistique étant de pouvoir réutiliser les résultats obtenus pour d'autres populations, ce n'est pas l'effectif d'une classe qui importe, mais la proportion d'individus qui appartiennent à cette classe.

Définition : On appelle fréquence de la classe le quotient de l'effectif de la classe par l'effectif total. La somme des fréquences de toutes les classes est égale à 1.

Exemple : la fréquence de la classe « Vélo » est $\frac{18}{600} = 0{,}03$. Il y a 3% des employés qui viennent à vélo. Ajouter au tableau précédent une ligne indiquant les fréquences de chaque classe et vérifier (aux erreurs d'approximation près) que la somme des fréquences vaut 1.

2) Représentations graphiques

La représentation la plus courante est le diagramme circulaire : l'angle du secteur représentant la classe est proportionnel à l'effectif (et donc à la fréquence).
Exemple : l'angle associé à la classe « Vélo » serait de $0{,}03 \times 360° = 10{,}8°$. Faire tout le diagramme circulaire de l'exemple précédent.
Une autre représentation possible est le diagramme en bâtons : la hauteur du bâton représentant la classe est proportionnelle à son effectif.

3) Analyse de la variable statistique

On ne peut définir qu'une seule caractéristique.

Définition : On appelle mode ou classe modale la classe (ou les classes) qui a le plus grand effectif.

Exemple : Déterminer la classe modale de l'exemple précédent.

III - Variable quantitative discrète

1) Classement des données

Pour une variable quantitative discrète, chaque classe correspond aussi à une valeur du caractère, mais qui a une valeur numérique réelle x_i. Le nombre d'individus qui appartiennent à cette classe s'appelle l'effectif n_i de la classe. La somme des effectifs de toutes les classes est l'effectif total de la population : $n = \sum_{i=1}^{p} n_i$ (s'il y a p classes).

La fréquence de la classe est le quotient de son effectif par l'effectif total : $f_i = \frac{n_i}{n}$.

On supposera que les classes sont numérotées par ordre croissant de la valeur du caractère : $x_1 < x_2 < ... < x_p$.

L'effectif n_i est le nombre d'individus ω tels que $X(\omega) = x_i$.

La famille $(x_i, n_i)_{1 \le i \le p}$ est appelée série statistique (discrète).

Exemple : On a relevé les notes obtenues à un devoir. Le tableau suivant donne les effectifs de chaque classe.

Classe x_i	4	5	6	7	8	9	10	11	12	13	14	15
Effectif n_i	2	0	3	4	3	5	7	4	3	2	2	1

Dans cet exemple, il y a 12 classes : $p = 12$. La $5^{\text{ème}}$ modalité (valeur du caractère dans la classe) est $x_5 = 8$ et l'effectif correspondant est $n_5 = 3$: il y a 3 élèves qui ont eu 8 au devoir.
Recopier le tableau, calculer l'effectif total et compléter le tableau en calculant les fréquences.

Définition : On appelle effectif cumulé croissant de la *i*-ème classe : $N_i = \sum_{k=1}^{i} n_k$ et fréquence cumulée croissante : $F_i = \frac{N_i}{n} = \sum_{k=1}^{i} f_k$.

L'effectif cumulé croissant N_i est le nombre d'individus ω tels que $X(\omega) \leq x_i$.
On peut remarquer que $F_p = 1$.

Exemple : $N_5 = 12$, donc il y a 12 élèves qui ont eu une note inférieure ou égale à 8 et $F_5 = 0{,}33$, donc il y a 33% des élèves qui ont eu une note inférieure ou égale à 8.
Compléter le tableau en calculant les effectifs cumulés croissants, ainsi que les fréquences cumulées correspondantes.

2) Représentations graphiques

On se place dans un repère orthogonal et on trace à partir du point de coordonnées $(x_i, 0)$ un segment vertical de hauteur proportionnelle à l'effectif n_i (et donc à la fréquence f_i). On obtient ainsi le diagramme en bâtons des effectifs (et des fréquences). La ligne polygonale qui joint les sommets des bâtons est appelée polygone des effectifs (ou des fréquences).
On définit de même le diagramme en bâtons des effectifs (ou des fréquences) cumulés ainsi que le polygone des effectifs (ou des fréquences) cumulés.
Exemple : Tracer le diagramme en bâtons et le polygone des effectifs, puis sur une autre figure le diagramme en bâtons et le polygone des effectifs cumulés croissants.

3) Analyse de la série statistique

a) Caractéristiques de position

Il s'agit de résumer la série statistique par un nombre qui donne une image de son comportement.
On peut d'abord penser à la valeur prise le plus souvent.

Définition : Le mode est la valeur (ou les valeurs) de la variable pour laquelle l'effectif est maximal. La (ou les) classe modale est la classe correspondante.

Exemple : Calculer le mode de la série précédente.
Le mode donne un renseignement intéressant, mais le simple fait qu'il y en ait plusieurs ne permet pas de l'utiliser valablement.
On peut ensuite penser à la valeur qui partage la population en deux parties égales.

Définition : La médiane est une valeur *m* de la variable telle que le nombre d'individus ω tels que $X(\omega) < m$ soit égal au nombre d'individus ω tels que $X(\omega) > m$.

Détermination pratique : Si l'effectif total de la population est *n*, on classe par ordre croissant les *n* valeurs $X(\omega)$ correspondantes. Si *n* est impair ($n = 2q+1$), la médiane est la valeur de rang $(q+1)$. Si *n* est pair ($n = 2q$), la médiane est la moyenne des valeurs de rang *q* et $(q+1)$.

Exemple : Dans la série précédente, déterminer la parité de *n*, puis la valeur de *q*, puis à l'aide des effectifs cumulés la médiane.
La médiane présente un intérêt certain, mais se prête mal aux calculs théoriques.
C'est finalement la moyenne arithmétique qui est la plus usitée.

Définition : On appelle moyenne de la série statistique $(x_i, n_i)_{1\le i\le p}$ d'effectif total *n* le réel $\bar{x} = \frac{1}{n}\sum_{i=1}^{p} n_i x_i = \sum_{i=1}^{p} f_i x_i$.

C'est la caractéristique la plus représentative. C'est la moyenne arithmétique simple de toutes les valeurs $X(\omega)$ obtenues pour tous les individus ω ou encore la moyenne arithmétique de toutes les valeurs x_i du caractère pondérées par les effectifs ou les fréquences.
Exemple : Calculer la moyenne de la série précédente.

Propriété : Si *a* et *b* sont des réels, $\overline{ax+b} = a\bar{x} + b$.

Démonstration : On pose $Y = aX + b$. Donc pour tout ω, $Y(\omega) = aX(\omega) + b$.
Si $a \neq 0$, pour tout *i*, *Y* prend la valeur $y_i = ax_i + b$ si et seulement si *X* prend la valeur x_i, donc l'effectif de la classe y_i est n_i. Donc :

$$\bar{y} = \frac{1}{n}\sum_{i=1}^{p} n_i y_i = \frac{1}{n}\sum_{i=1}^{p} n_i (ax_i + b) = a\frac{1}{n}\sum_{i=1}^{p} n_i x_i + b\frac{1}{n}\sum_{i=1}^{p} n_i = a\bar{x} + b .$$

Si $a = 0$, *Y* est constante et prend une seule valeur *b*. Donc $\bar{y} = b = a\bar{x} + b$.

b) Caractéristiques de dispersion

Il s'agit de mesurer la répartition de *X* autour de sa moyenne car un seul nombre ne suffit pas à préciser le comportement de la série.
Par exemple, la série étudiée précédemment et les séries suivantes ont même moyenne, mais la répartition des notes est tout à fait différente.

Classe x_i	4	5	6	7	8	9	10	11	12	13	14	15
Effectif n_i	2	3	5	8	1	2	1	0	1	3	4	6

Celle-ci est beaucoup plus dispersée. La suivante est beaucoup plus concentrée.

Classe x_i	8	9	10	11	12
Effectif n_i	9	8	12	6	1

On veut donc mesurer la dispersion de *X*, donc les écarts à la moyenne, c'est-à-dire étudier la variable centrée associée à *X* : $Y = X - \bar{x}$. Il y a diverses manières de mesurer ces écarts. La méthode la plus courante est le calcul de l'écart-type, moyenne quadratique des écarts.

Définition : On appelle variance de la série statistique $(x_i, n_i)_{1\le i\le p}$ d'effectif total *n* le réel $V(X) = \frac{1}{n}\sum_{i=1}^{p} n_i (x_i - \bar{x})^2$ et écart-type le réel $\sigma_x = \sqrt{V(X)}$ (car $V(X) \ge 0$).

Propriétés : 1) $V(X) = \frac{1}{n}\sum_{i=1}^{p} n_i x_i^2 - \bar{x}^2$ 2) $V(aX+b) = a^2 V(X)$ et $\sigma_{ax+b} = |a|\sigma_x$

Démonstration : 1) $V(X) = \frac{1}{n}\sum_{i=1}^{p} n_i (x_i - \bar{x})^2 = \frac{1}{n}\sum_{i=1}^{p} n_i (x_i^2 - 2x_i\bar{x} + \bar{x}^2)$

$V(X) = \frac{1}{n}\sum_{i=1}^{p} n_i x_i^2 - \frac{2}{n}\bar{x}\sum_{i=1}^{p} n_i x_i + \frac{1}{n}\bar{x}^2\sum_{i=1}^{p} n_i = \frac{1}{n}\sum_{i=1}^{p} n_i x_i^2 - 2\bar{x}^2 + \bar{x}^2 = \frac{1}{n}\sum_{i=1}^{p} n_i x_i^2 - \bar{x}^2$.

2) Si $a = 0$, $Y = aX + b$ est constante, égale à b. Il n'y a qu'une classe et $\bar{y} = b$. Donc : $V(Y) = 0$. Donc $V(aX+b) = 0 = a^2 V(X)$.

Si $a \neq 0$, $Y = aX + b$ prend les valeurs $y_i = ax_i + b$ avec l'effectif n_i et $\bar{y} = a\bar{x} + b$.

Donc : $V(Y) = \frac{1}{n}\sum_{i=1}^{p} n_i (y_i - \bar{y})^2 = \frac{1}{n}\sum_{i=1}^{p} n_i [(ax_i + b) - (a\bar{x} + b)]^2$

$V(Y) = \frac{1}{n}\sum_{i=1}^{p} n_i a^2 (x_i - \bar{x})^2 = a^2 \frac{1}{n}\sum_{i=1}^{p} n_i (x_i - \bar{x})^2 = a^2 V(X)$.

Exemple : Calculer les écarts-types des trois séries citées et les comparer.

On démontre que « en général » l'intervalle $[\bar{x} - \sigma_x, \bar{x} + \sigma_x]$ contient environ 68% de la population et que l'intervalle $[\bar{x} - 2\sigma_x, \bar{x} + 2\sigma_x]$ contient environ 95% de la population. On en verra la justification théorique en probabilités plus tard.

IV - Variable quantitative continue

1) Classement des données

On regroupe les valeurs prises par la variable en p intervalles adjacents qui n'ont d'ailleurs pas forcément tous la même largeur : $[a_1, a_2[$, $[a_2, a_3[$, …, $[a_p, a_{p+1}[$ où les a_i sont des réels qui vérifient $a_1 < a_2 < ... < a_p < a_{p+1}$. La i-ème classe $[a_i, a_{i+1}[$ est l'ensemble des individus ω tels que $a_i \leq X(\omega) < a_{i+1}$. Il peut également y avoir des classes « Moins de a » (ensemble des individus ω tels que $X(\omega) < a$) ou « Plus de b » (ensemble des individus ω tels que $X(\omega) \geq b$).

Les définitions des effectifs, de l'effectif total, des fréquences restent les mêmes que pour une série statistique discrète. On notera n_i l'effectif de la i-ème classe, f_i sa fréquence et n l'effectif total.

Par contre, on définit deux types d'effectifs et de fréquences cumulés en vue du calcul de la médiane.

L'effectif cumulé croissant de la i-ème classe $[a_i, a_{i+1}[$ est $N_i = \sum_{k=1}^{i} n_k$ et la fréquence cumulée croissante est : $F_i = \frac{N_i}{n} = \sum_{k=1}^{i} f_k$.

L'effectif cumulé décroissant de la i-ème classe $[a_i, a_{i+1}[$ est $N'_i = \sum_{k=i}^{p} n_k$ et la fréquence cumulée croissante est : $F'_i = \frac{N'_i}{n} = \sum_{k=i}^{p} f_k$.

L'effectif cumulé croissant N_i est le nombre d'individus ω tels que $X(\omega) < a_{i+1}$, tandis que l'effectif cumulé décroissant N'_i est le nombre d'individus ω tels que $X(\omega) \geq a_i$.

On remarque que $N_p = N'_1 = n$ et que pour tout *i*, $N_i + N'_{i+1} = n$.

Et donc : $F_p = F'_1 = n$ et pour tout *i*, $F_i + F'_{i+1} = 1$.

Exemple : Le tableau suivant donne la répartition des âges des 152 ouvriers d'une entreprise.

Classe	Moins de 20	[20,25[	[25,30[	[30,35[	[35,40[	[40,50[	[50,60[	Plus de 60
Effectif	1	7	28	36	45	26	8	1

La troisième classe est [25,30[. L'effectif est $n_3 = 28$: il y a 28 ouvriers qui ont au moins 25 ans et moins de 30 ans. La fréquence est $f_3 = \frac{28}{152} = 0{,}184$: il y a 18,4% des ouvriers dans cette catégorie d'âge. L'effectif cumulé croissant est $N_3 = 36$ et la fréquence cumulée croissante est $F_3 = \frac{36}{152} = 0{,}237$: il y a 23,7% des ouvriers qui ont moins de 30 ans. L'effectif cumulé décroissant est $N'_3 = 144$ et la fréquence cumulée décroissante de la classe est $F'_3 = \frac{144}{152} = 0{,}947$: il y a 94,7% d'ouvriers qui ont au moins 25 ans.

Recopier le tableau précédent et le compléter par des lignes donnant les fréquences, les effectifs et les fréquences cumulées croissants et décroissants.

2) Représentations graphiques

On se place dans un repère orthogonal et on représente chaque classe $[a_i, a_{i+1}[$ par un rectangle dont la base est le segment qui joint les points de coordonnées (a ,0) et (a ,0) et dont l'aire (et non la hauteur) est proportionnelle à l'effectif (et donc aux fréquences). Une telle représentation s'appelle un histogramme.

Remarque : On considère l'aire et non la hauteur pour compenser le fait que les classes n'ont pas toutes la même largeur. Dans l'exemple, les classes [40,50[et [50,60[ont une largeur double des autres classes. Elles seront représentées par des rectangles dont la hauteur sera respectivement 13 et 4. Le plus souvent, une classe de largeur double sera représentée en réalité par deux rectangles accolés de même largeur que les autres classes (par exemple, la classe [40,50[sera représentée par deux rectangles de base 5 et de hauteur 13, comme s'il y avait 13 ouvriers entre 40 et 45 ans et 13 ouvriers entre 45 et 50 ans). On dira que l'on a utilisé des classes unitaires.

Lorsque la classe est « Moins de *a* » ou « Plus de *b* », sa représentation sera faite par un rectangle dont la base aura même largeur que la classe voisine.

Exemple : Dans toute la suite, la classe « Moins de 20 » sera identifiée à une classe de même largeur que [20,25[, c'est-à-dire [15,20[, et donc représentée par un rectangle de base 5 et de hauteur 1, alors que la classe « Plus de 60 » sera identifiée à [60,70[et donc représentée par un rectangle de base 10 et de hauteur 0,5.

Pour construire le polygone des effectifs (ou des fréquences), on considère l'effectif (ou la fréquence) concentré au centre de chaque classe (éventuellement unitaire), c'est à dire en $x_i = \frac{1}{2}(a_i + a_{i+1})$ et on joint les points de coordonnées (x_i, n_i) ou (x_i, f_i).

Exemple : Tracer sur une figure l'histogramme et le polygone des effectifs de la série précédente.

L'effectif cumulé croissant N_i de la classe $[a_i, a_{i+1}[$ représentant le nombre d'individus ω tels que $X(\omega) < a_{i+1}$, on le considère concentré en a_{i+1} et donc le

polygone des effectifs cumulés croissants est obtenu en joignant les points de coordonnées (a_{i+1}, N_i). Même chose pour le polygone des fréquences cumulées croissantes.

L'effectif cumulé décroissant N'_i de la classe $[a_i, a_{i+1}[$ représentant le nombre d'individus ω tels que $X(\omega) \geq a_i$, on le considère concentré en a_i et donc le polygone des effectifs cumulés décroissants est obtenu en joignant les points de coordonnées (a_i, N'_i). Même chose pour le polygone des fréquences cumulées décroissantes.

Exemple : Sur une même figure, tracer les polygones des effectifs cumulés croissants et décroissants de la série précédente.

3) Analyse de la série statistique

a) Caractéristiques de position

On appelle classe modale toute classe correspondant à un effectif maximal et mode le centre de cette classe. Il peut y en avoir plusieurs.

Exemple : Déterminer la classe modale et le mode de la série précédente.

Définition : La médiane est une valeur *m* de la variable telle que le nombre d'individus ω tels que $X(\omega) < m$ soit égal au nombre d'individus ω tels que $X(\omega) > m$.

Cela revient à dire, en supposant une évolution continue des effectifs cumulés, que l'effectif cumulé croissant associé à *m* est égal à l'effectif cumulé décroissant, donc à un effectif $\frac{n}{2}$ puisque la somme des effectifs croissants et décroissants est *n*.

Détermination pratique : On la détermine graphiquement en prenant l'abscisse du point d'intersection des polygones des fréquences cumulées croissantes et décroissantes. Elle se calcule en déterminant d'abord la classe médiane (classe dans laquelle se trouve la médiane), puis en faisant une interpolation linéaire en supposant la répartition uniforme à l'intérieur de cette classe.

Exemple : Dans la série précédente, la classe médiane est [35,40[, puisque c'est dans cette classe que l'effectif cumulé croissant dépasse 76 (moitié de 152). Sur le polygone des effectifs cumulés croissants, on trouve les points $P(35,72)$ et $Q(40,117)$. On cherche sur le segment $[PQ]$ l'abscisse *m* du point *M* d'ordonnée 76. Si l'équation de la droite (PQ) est $y = ax + b$, alors $a = \frac{y_Q - y_P}{x_Q - x_P} = \frac{y_M - y_P}{x_M - x_P}$. Donc :

$\frac{1}{a} = \frac{m-35}{76-72} = \frac{40-35}{117-72}$, donc $m = 35 + (76-72) \times \frac{40-35}{117-72}$, donc $m = 35{,}44$.

Effectuer le même raisonnement sur le polygone des effectifs cumulés décroissants et montrer que l'on trouve la même valeur de *m*.

On peut aussi faire un calcul analogue sur le polygone des fréquences cumulées croissantes (ou décroissantes) pour trouver l'abscisse du point d'ordonnée 0,5 sur le segment correspondant à la classe médiane.

La définition de la moyenne est la même que pour une variable discrète.

Définition : Si l'on suppose l'effectif n_i de la classe $[a_i, a_{i+1}[$ concentré au centre $x_i = \frac{a_i + a_{i+1}}{2}$, la moyenne de la série statistique est : $\bar{x} = \frac{1}{n}\sum_{i=1}^{p} n_i x_i$.

Bien sûr, les classes « moins de » et « plus de » sont toujours supposées identifiées à des classes de la forme $[a_i, a_{i+1}[$.

Exemple : Calculer l'âge moyen des ouvriers de l'entreprise.

b) Caractéristiques de dispersion

Avec les conventions précédentes, les définitions de la variance et de l'écart-type sont les mêmes que pour une variable discrète :

$$V(X) = \frac{1}{n}\sum_{i=1}^{p} n_i (x_i - \bar{x})^2 = \frac{1}{n}\sum_{i=1}^{p} n_i {x_i}^2 - \bar{x}^2 \qquad \sigma_x = \sqrt{V(X)}$$

Exemple : Calculer la variance et l'écart-type de la série précédente.

Pour évaluer la répartition de la série de manière plus fine, on introduit les quartiles (on partage la population en 4 parties de même effectif) et les déciles (on partage la population en 10 parties de même effectif) :

Définition : Soit k un entier égal à 1,2 ou 3. On appelle k-ème quartile de la série statistique la valeur q_k de la variable qui correspond à un effectif cumulé croissant de $k \times \frac{n}{4}$ et à une fréquence cumulée croissante de 0,25k.

Il y a 25% des individus ω de la population tels que $X(\omega) < q_1$, 50% tels que $X(\omega) < q_2$ (donc q_2 est la médiane), 75% tels que $X(\omega) < q_3$.

L'intervalle $[q_1, q_3]$ s'appelle l'intervalle interquartile et représente l'ensemble des valeurs du caractère associées à 50% de la population (en éliminant les individus les moins « significatifs »).

Définition : Soit k un entier compris entre 1 et 9. On appelle k-ème décile de la série statistique la valeur d_k de la variable qui correspond un effectif cumulé croissant de $k \times \frac{n}{10}$ et à une fréquence cumulée croissante de 0,1k.

Il y a $k \times 10\%$ d'individus ω de la population tels que $X(\omega) < d_k$. La médiane est égale au 5-ème décile.

Les quartiles et les déciles se déterminent comme la médiane par interpolation linéaire.