

Chapitre I

Quelques définitions indispensables

1. Population, individu et échantillon

On appelle *population*, toutes les unités observables dans un ensemble défini par une propriété.

C'est, par exemple, la population des feuilles d'un arbre dont la propriété commune est qu'elles appartiennent à cet arbre. Attention, le nombre de feuilles n'est pas une propriété commune : il ne peut donc pas caractériser la population.

On appelle *individu*, une unité quelconque de cette population. Dans l'exemple de arbre, il s'agit d'une feuille. L'ensemble des individus, forme la population.

On appelle *échantillon*, une partie seulement d'une population. Dans l'exemple de l'arbre, ce sont les feuilles d'une branche, celles de la cime ou celles atteintes d'oïdium¹ etc. L'échantillon est constitué de n individus.

Les statisticiens utilisent pour préciser les caractéristiques des populations, des lettres de l'alphabet grec et pour caractériser l'échantillon les lettres de l'alphabet latin.

2. Caractère

On appelle *caractère*, une propriété des individus. Un caractère exprime une information sur une population ou un échantillon. Cette définition du caractère introduit une importante partition dans les études statistiques.

- Les *analyses statistiques à une dimension* quand on travaille sur un caractère. Elles correspondent aux statistiques les plus utilisées.
- Les *analyses statistiques multivariées* quand on travaille simultanément sur plusieurs caractères. On cherche à établir des ressemblances entre les individus au regard des caractères étudiés ou à reconnaître des liaisons entre les caractères.

Les études statistiques sont relativement simples quand on travaille sur un ou deux caractères ; au-delà techniquement et intellectuellement, elles sont beaucoup plus ardues.

- **Caractère quantitatif**

Quand pour connaître un caractère, on effectue des mesures, le caractère est *quantitatif* (ou *numérique*)

¹ L'oïdium du chêne est un champignon qui se développe sur les feuilles. Il leur confère un aspect blanchâtre, cotonneux.

Les résultats sont des nombres entiers ou décimaux, avec lesquels il est possible de faire des opérations. C'est, par exemple, la taille des feuilles ou l'âge des étudiants.

- Lorsque la série de mesures peut prendre toutes les valeurs comprises entre deux bornes, un *caractère quantitatif* est *continu*.

Par exemple, l'âge est exprimé en années, la longueur d'une feuille en millimètres.

- Lorsque les valeurs observées correspondent à un nombre fini de données dans un intervalle fixé, le *caractère quantitatif* est *discret*.

C'est, par exemple, le nombre d'enfants par famille : 0, 1, 2 ... Il est impossible d'observer 1,38 enfants, ou - 2 enfants par famille.

- **Caractère semi-quantitatif**

Un caractère est *semi-quantitatif* quand sa mesure correspond à une échelle arbitraire. Le caractère semi-quantitatif est un cas fréquent en phytopathologie² où les observations expriment une échelle de jugement. Ce sont les résultats obtenus après l'action d'un antifongique³ sur une culture. L'expérimentateur se donne une échelle empirique de 0 à 5. Pour les plantes indemnes la note 0 est attribuée, pour les plantes très touchées, c'est 5 et pour les attaques plus ou moins fortes les chiffres intermédiaires. Cette échelle de notes est aussi très utilisée dans le monde médical pour juger de l'action d'un médicament.

- **Caractère qualitatif ou nominal**

Lorsque les valeurs données au caractère correspondent à un attribut ou à une modalité non mesurable, un caractère est *qualitatif* ou *nominal*.

C'est, par exemple, la couleur des yeux ou des fleurs. Le sexe est, aussi, un caractère qualitatif représenté par une variable à deux classes. Les médecins, ainsi que l'état civil, adoptent le code suivant : 1 correspond au sexe masculin, 2 au sexe féminin.

- **Caractère ordinal**

Lorsque ses valeurs données au caractère expriment un ordre par rapport à une référence, le caractère est *ordinal*.

Le classement alphabétique ou le classement par mérite croissant ou décroissant, la pagination d'un livre sont des exemples de caractères ordinaux.

Dans la pratique, on dispose essentiellement de caractères quantitatifs, le plus souvent discrets ou de caractères qualitatifs.

3. Variable aléatoire

On appelle *variable aléatoire* X ; l'ensemble des valeurs x_1, x_2, x_3, x_n que peut prendre un caractère.

² Phytopathologie : science traitant des maladies des végétaux qu'elles soient fongiques, bactériennes, virales ou animales.

³ Traitement antifongique : traitement contre les champignons généralement pathogènes.

Problèmes d'échantillonnage

En expérimentation-statistique, que ce soit en biologie, en démographie ou dans le milieu industriel, on ne travaille pratiquement jamais sur les populations mais uniquement sur des échantillons. Par exemple, si un responsable de ligne de fabrication de pots de yaourt veut vérifier que la conditionneuse qui remplit 40000 pots chaque heure, fonctionne correctement, il n'étudie pas tous les pots de yaourt mais seulement quelques-uns.

Ceci a des conséquences importantes :

1. L'échantillon doit être représentatif de la population étudiée. Cette notion évidente n'est pas toujours facile à réaliser. L'échantillon est constitué en tirant des éléments dans la population, soit au hasard, soit selon des règles précises et très élaborées d'échantillonnage. Dans le cadre de ce livre, les échantillons sont extraits des populations au hasard, on dit aussi de manière aléatoire. Cette méthode est parfaitement adaptée aux problèmes biologiques et industriels courants. Pour constituer l'échantillon, on extrait de la population, le plus souvent, n individus sans les remettre dans la population après avoir fait la mesure. Ceci évite de reprendre cet individu une deuxième fois, voire une troisième fois. Le tirage est exhaustif. Beaucoup plus rarement, on remet les individus extraits dans la population. Le tirage est alors non exhaustif.

2. On estime le comportement inconnu de la population, au travers de l'échantillon E . Cette estimation peut être une valeur unique, c'est une estimation ponctuelle. Mais le plus souvent, on détermine un intervalle où se situe l'estimation du paramètre inconnu de la population. Ce sera l'intervalle de confiance du paramètre de la population estimé au travers l'échantillon observé.

3. Qui dit estimation sous-entend évidemment erreur possible. Les statisticiens expriment traditionnellement cette erreur dans le résultat présenté par le risque d'erreur. Chaque expérimentateur est libre de choisir le risque d'erreur avec lequel il présente son estimation. S'il choisit de donner son estimation avec un risque d'erreur faible, concrètement il ne prend pas de risque, il joue la sécurité, l'intervalle de confiance du paramètre de la population, estimé au travers l'échantillon observé est grand. Si inversement, il accepte un risque d'erreur élevé car plus audacieux, l'intervalle de confiance du paramètre de la population, estimé au travers l'échantillon observé est plus petit.

4. A partir du moment où une estimation est faite et quel que soit le risque d'erreur choisi, on peut se tromper dans les résultats statistiques élaborés à partir d'un échantillon.

5. La pensée unique en statistiques n'existe pas. En fonction de l'information souhaitée, du nombre d'observations dans le ou les échantillons observés, des lois de distribution des observations et des plans d'expériences adoptés...(tout ceci sera vu au fil des chapitres), les méthodes statistiques sont différentes et il faut savoir choisir le bon test. C'est très certainement la difficulté majeure des statistiques et l'utilisation de l'ordinateur, où il suffit d'appuyer sur une touche pour obtenir un résultat, occulte bien souvent cet indispensable travail préliminaire.

- Une variable aléatoire X est *continue* si l'ensemble des valeurs qu'elle peut prendre est dans un ensemble infini non dénombrable.

Par exemple X peut être l'âge d'une personne prise au hasard, ou son poids, ou sa taille. Si celle-ci est exprimée en centimètres, on peut raisonnablement penser que X est un intervalle compris dans $[0, 250]$.

- Une variable aléatoire X est *discrète* si l'ensemble des valeurs qu'elle peut prendre est un ensemble numérique fini (c'est-à-dire comprenant un nombre fini d'éléments) ou un ensemble infini dénombrable (c'est-à-dire comprenant une infinité de nombres que l'on peut énumérer).

Par exemple, en lançant un dé à six faces, comme on ne peut obtenir qu'une des 6 valeurs $\{1, 2, 3, 4, 5, 6\}$, la variable aléatoire, face de dé, est discrète. Si cette fois, la variable aléatoire est le nombre de lancers nécessaires pour obtenir un 6, $\{1, 2, 3, \dots\}$, puisqu'elle peut comprendre un ensemble infini de nombres, c'est aussi une variable aléatoire discrète.

4. Effectif et classe statistique

On appelle *effectif* d'une valeur du caractère étudié, le nombre de fois qu'apparaît cette valeur.

Dans une série de mesures associée à une variable aléatoire, l'effectif est le nombre d'apparitions de cette modalité dans la série, c'est-à-dire le nombre d'individus de la population pour lesquels le caractère étudié a cette modalité.

Par exemple dans un groupe de travaux pratiques, quatre étudiants ont dix-neuf ans, cinq vingt ans et trois vingt et un ans. Quatre, cinq et trois sont les effectifs d'une valeur de la variable aléatoire.

L'effectif total de la série est le nombre total d'individus de la population étudiée. Dans l'exemple ci-dessus présenté, l'effectif total est douze.

On appelle *classe statistique* chacun des intervalles successifs qui partage l'intervalle total de variation d'une variable statistique quantitative.

Par exemple dans les enquêtes de satisfaction, il est souvent demandé l'âge des participants. Mais l'intervalle total $[0, 120 \text{ ans}]$ étant trop grand, il est souvent scindé en 4 ou 5 nouveaux intervalles qui sont des classes statistiques.

La grandeur de ces intervalles est appelée l'*amplitude*⁴ de la classe statistique et l'effectif total de la série est la somme des effectifs de toutes les modalités ou de toutes les classes. .

⁴ Par tacite habitude, on inclut dans cet intervalle la valeur la plus faible et on exclut la plus haute $[x[$.

Chapitre II

Les données

Quantifier un caractère paraît simple, pourtant ceci exige le plus souvent la mise en œuvre de techniques expérimentales qui, seules, permettent sa mesure qualitative ou quantitative.

Ces mesures peuvent concerner tous les individus indistinctement, c'est le cas le plus fréquent. Parfois, il faut regrouper de manières distinctes tout ou partie des individus d'un ensemble : c'est l'analyse combinatoire.

Les mesures qualitatives sont d'une complexité très variable. Il peut s'agir d'apprécier la couleur d'une fleur, comme en horticulture florale¹ ou d'un sol en pédologie. Il faut, aussi, peut-être, identifier des composés obtenus en chromatographie liquide à haute performance (HPLC) ou avec des spectres moléculaires d'absorption. Ceci nécessite d'acquérir un signal et éventuellement procéder à diverses opérations permettant d'améliorer cette acquisition (transformation, lissage, filtrage...).

Les mesures quantitatives, apparemment plus simples sont souvent complexes et sources d'incertitudes, surtout que rares sont les mesures où n'intervient qu'une lecture de données. En effet, le résultat recherché est une combinaison de mesures qui se cumulent, se soustraient, se divisent ou se multiplient.

Quantitatives ou qualitatives ces mesures sont sources d'erreurs qu'il faut apprécier. Ce chapitre se limite à présenter, d'abord, l'analyse combinatoire, puis le problème de chiffres arrondis et significatifs, la notion de distribution de fréquences, enfin les principaux types d'erreurs faits lors de mesures ainsi que leur propagation dans des calculs.

1. L'analyse combinatoire

L'analyse combinatoire a « *pour objet de dénombrer les manières distinctes ou de grouper tout ou partie des objets d'un ensemble, suivant des lois déterminées* » (Azoulay et Cohen, 1993)².

C'est souvent par l'analyse combinatoire que les étudiants abordent les statistiques... ce qui les en dégoûte à tout jamais! Elle est pourtant indispensable pour traiter les calculs de probabilités équiprobables, comme la probabilité d'avoir toujours six quand on lance un dé plusieurs fois. Mais, objectivement, son utilisation est rarissime en biologie, en industrie ou en médecine. C'est ce manque d'intérêt pratique qui les fait abhorrer. Pourtant pour l'anecdote, elle peut être utile aux étudiants curieux qui s'interrogent sur la probabilité qu'ils ont d'obtenir la moyenne à un QCM ou à ceux intrigués par les jeux de hasard.

Aussi, seuls les grands points fondamentaux, illustrés de quelques exemples, sont abordés dans ce paragraphe.

¹ La couleur des roses obéit à des normes très précises.

² Azoulay et Cohen, Statistique, Edisciences 1993

- **Arrangement**

On appelle *arrangement* A , le nombre de possibilités de ranger k objets choisis parmi n ($k \leq n$).

Le nombre des arrangements a pour formule :

$$A_n^k = n(n-1)(n-2)\dots(n-k+1) \quad (1)$$

Par exemple, avec les trois lettres a, b, c , en les regroupant par deux, on a six solutions : ab, ac, ba, bc, ca et cb .

- **Permutation**

On appelle *permutation* P , l'arrangement ordonné de n objets.

Le nombre des permutations a pour formule :

$$P = n! = 1 \times 2 \times 3 \times \dots \times n \quad (2)$$

Les permutations constituent un cas particulier des arrangements. C'est le cas où $n = p$. Deux permutations ne diffèrent que par l'ordre des objets. Par exemple, avec les trois lettres a, b, c on obtient six solutions : abc, acb, bca, bac, cba et cab .

Il arrive que, parmi les n objets dont on cherche le nombre de permutations, certains d'entre eux, au nombre de r , soient tous semblables. Auquel cas, rien ne distingue les permutations de ces r objets entre eux. Le nombre de permutations est diminué et si r est le nombre des objets communs, la formule devient :

$$P = \frac{n!}{r!} \quad (3)$$

Par exemple, avec les huit lettres du mot « exercice », où la lettre e se trouve trois fois et la lettre c deux fois, le nombre de permutations est égal à : $8! / 3! 2!$ soit 3360.

- **Combinaison**

On appelle *combinaison* C , la manière d'associer k objets parmi n ($k \leq n$), mais sans tenir compte de l'ordre des objets dans le rangement.

Le nombre des combinaisons a pour formule :

$$C_n^k = \frac{A_n^k}{k!} = \frac{n!}{k!(n-k)!} \quad (4)$$

La différence entre arrangements et combinaisons correspond au fait que dans les combinaisons, on ne s'intéresse pas à l'ordre des éléments. Cette fois, par exemple, les trois lettres a, b, c prises deux à deux ne donnent plus que trois combinaisons : ab, bc, ca .

2. Le binôme de Newton et le triangle de Pascal

On appelle *binôme de Newton*, le produit de n facteurs égaux à $(a + b)$ soit $(a + b)^n$.
Le *triangle de Pascal* donne les coefficients du binôme de Newton.

Contrairement aux arrangements, permutations et autres combinaisons, le binôme de Newton³ et son complément le triangle de Pascal⁴, sont fréquemment utilisés en statistiques à l'occasion de la mise en œuvre de la loi binomiale. Les identités remarquables $(a + b)^2 = a^2 + 2ab + b^2$ ou $(a + b)^3 = a^3 + 3a^2b + 3ab^2 + b^3$ peuvent être généralisées grâce à l'analyse combinatoire. C'est le *binôme de Newton*. En effet :

$$(a + b)^n = C_n^0 a^n + C_n^1 a^{n-1} b + C_n^{n-2} a^{n-1} b^2 + C_n^{n-3} a^{n-3} b^3 + \dots + C_n^n b^n$$

Les coefficients correspondant aux combinaisons $C_n^0, C_n^1, C_n^{n-2}, C_n^{n-3} \dots C_n^n$ sont donnés par le *triangle de Pascal* :

Polynôme	Coefficients														
(a+b)							1	↘ ↙					1		
(a+b) ²						1	↘ ↙	2	↘ ↙				1		
(a+b) ³				1	↘ ↙	3	↘ ↙	3	↘ ↙				1		
(a+b) ⁴			1	↘ ↙	4	↘ ↙	6	↘ ↙	4	↘ ↙			1		
(a+b) ⁵		1	↘ ↙	5	↘ ↙	10	↘ ↙	10	↘ ↙	5	↘ ↙		1		
(a+b) ⁶		1	↘ ↙	6	↘ ↙	15	↘ ↙	20	↘ ↙	15	↘ ↙	6	↘ ↙	1	
(a+b) ⁷	1	↘ ↙	7	↘ ↙	21	↘ ↙	35	↘ ↙	35	↘ ↙	21	↘ ↙	7	↘ ↙	1
...			...		...		...		...		...		...		...
(a+b) ⁿ	1	↘ ↙		↘ ↙		↘ ↙		↘ ↙		↘ ↙		↘ ↙		↘ ↙	1

Ce triangle est d'usage facile, puisque pour trouver les valeurs des coefficients il suffit d'additionner les coefficients du rang précédent. Ainsi, si $(a + b)^3 = a^3 + 3a^2b + 3ab^2 + b^3$,
 $(a + b)^4 = a^4 + (3+1)a^3b + (3+3)a^2b^2 + (3+1)ab^3 + b^4 = a^4 + 4a^3b + 6a^2b^2 + 4ab^3 + b^4$.

3. Les nombres arrondis et les chiffres significatifs

En expérimentation-statistique, on travaille sur des nombres. Ils peuvent être divisés en trois catégories :

- Les nombres qui correspondent à une mesure : le poids, la taille, l'âge... Leur précision dépend de l'unité choisie et les valeurs données ne sont pas exactes.

³ Sir Isaac Newton (1642-1727), physicien et mathématicien anglais, auteur des lois de l'attraction universelle.

⁴ Blaise Pascal (1623-1662) mathématicien, physicien et philosophe français, auteur avec Fermat (1601-1665) du calcul des probabilités. Lorsque Pascal construit en 1654 son triangle arithmétique, il est loin d'être le premier à avoir organisé les nombres de cette façon. Le triangle arithmétique est déjà connu en Chine, en Inde ou dans les pays arabes dès le onzième siècle. L'apport de Pascal est d'utiliser pour la première fois le raisonnement par récurrence.

- Les nombres qui codent les catégories d'une variable qualitative : un pour un garçon, deux pour une fille par exemple. Ces nombres ne peuvent être discutés quant à leur signification.
- Les nombres qui résultent d'un dénombrement : ce sont des nombres exacts, au sens que leur valeur n'est pas arrondie. Ils ne créent pas d'incertitude dans les calculs. Il s'agit par exemple des 24 étudiants d'un groupe de travaux dirigés.

• Les arrondis

Intuitivement, on sait bien que tous les chiffres d'une opération ne sont pas utilisés pour présenter un résultat. Soit, par exemple, le calcul : $(6,19 \times 7,87) / 4,31$. La calculatrice donne le résultat avec force décimales : 11,30285383. Pratiquement, tous ces chiffres ne servent pas : le résultat est *arrondi* selon une règle simple. Si quatre décimales sont souhaitées c'est 11,3029 qui est écrit, trois décimales 11,303 et pour deux décimales 11,30. La règle pratique, universellement reconnue, est de tronquer le nombre si le chiffre supprimé est inférieur à 5 et d'ajouter une unité à la dernière décimale conservée si le premier chiffre supprimé est supérieur ou égal à 5.

Attention, c'est le résultat d'une opération qui est arrondi. Mais quand le nombre trouvé doit être repris pour effectuer d'autres opérations, il ne faut pas arrondir les nombres intermédiaires sous peine de voir diverger grandement le résultat final.

• Les chiffres décimaux significatifs

Les *chiffres décimaux significatifs* sont les chiffres inscrits après la virgule et qui donnent une bonne approximation du nombre exact.

Pourquoi choisir deux, trois ou quatre chiffres décimaux pour exprimer le résultat ? Ceci est tout le problème des *chiffres décimaux significatifs*. Avec la multiplication des ordinateurs et des calculatrices, pratiquement tous les calculs sont donnés avec au moins huit chiffres après la virgule, que les étudiants transcrivent, sans état d'âme, dans leurs devoirs. Pourtant, il existe des règles simples et de bon sens pour savoir combien de chiffres décimaux significatifs doit comporter un résultat.

- Dans les calculs où il y a *addition* ou *soustraction* de nombres décimaux, le résultat final ne doit pas avoir plus de décimales que le nombre qui en possède *le moins* :

$$16,19 + 7,87 - 4,3 = 19,8 \text{ et non } 19,76. \text{ Mais } 16,19 + 7,87 - 4,30 = 19,76.$$

- Dans les calculs où il y a *multiplication*, *division*, *puissance*, *racine*..., le résultat ne peut pas comporter plus de décimales que le nombre qui en comporte *le moins* :

$$\sqrt{16,19 \times 7,87} = 11,29 \text{ mais } \sqrt{16,19 \times 7,8} = 11,2$$

- Si une valeur exacte intervient dans une opération (cas des effectifs), elle ne limite pas la précision du résultat. On se réfère aux autres chiffres. Le nombre de décimales du résultat est égal à celui de la valeur ayant le moins de décimales dans l'opération, augmenté du nombre de chiffres de la valeur exacte moins un, soit une décimale quand la valeur exacte est comprise entre 10 et 99, deux décimales entre 100 à 999, trois décimales entre 1000 et 9999 *etc.* Ainsi, dans un calcul de moyenne de l'âge de 11 étudiants, on écrit :

$$\bar{m} = \frac{\sum_{i=1}^n x_i}{n} = \frac{234}{11} = 21,3 \text{ et non } 21 \text{ ou } 21,27.$$