La Méthode statistique et la Foresterie - III -Distributions statistiques

Léon Schaeffer

III - Distributions statistiques

Variables

Les mesures à effectuer en vue de préciser des grandeurs sont considérées par les statisticiens comme des « *variables* ».

Les variables peuvent être *continues* ou *discontinues*. Les mesures discontinues se rencontrent quand on effectue des dénombrements qui s'expriment par des nombres entiers: telle parcelle renferme tant d'arbres de telle grosseur. Théoriquement, les mesures continues qui se rencontrent chaque fois que l'observation porte sur une longueur, un poids, un temps, devraient être beaucoup plus fréquentes. En fait, nous sommes toujours limités par la précision de nos mesures: aurions-nous le moyen de mesurer le temps au 1/100 de seconde, comme le permettent certains appareils, notre mesure n'en serait pas moins discontinue. D'ailleurs, comme nous sommes amenés, pour faciliter les recherches ultérieures ou simplement pour présenter les résultats, à mettre nos mesures en ordre, à les grouper, un effet de notre classement est de rendre discontinues les variables théoriquement continues. Toutefois, pour les études théoriques, comme on le verra, on est parfois amené à supposer la variable continue.

Le classement s'effectue, suivant les cas, de façons variées. Quand il s'agit de bois abattus, on a souvent recours à deux nombres indiquant la plus grande et la plus petite des mesures appartenant à la classe. C'est ainsi que, sous le régime de la taxation des prix, on avait imaginé les classes suivantes dans les billes à placage d'après leur circonférence sur écorce au milieu, exprimée en centimètres: 160 à 179, 180 à 199, 200 à 239, etc...

Quand il s'agit de la grosseur d'arbres sur pied, on définit habituellement une classe par son point médian, mais on peut aussi préciser quelles sont ses limites réelles ou ses mesures-limites. La classe 25 cm de diamètre a pour limites réelles 22,5 et 27,5 cm. Si on effectue les mesures avec un compas gradué en millimètres et si on réalise ultérieurement leur groupement, la classe 25 cm sera définie par les mesures-limites: 22,5 et 27,4 cm.

Quand on est amené à arrondir un nombre, on n'omet pas d'augmenter d'une unité le dernier chiffre conservé lorsque le premier chiffre supprimé est supérieur ou égal à 5. Par exemple, si on ne doit conserver que la partie entière du nombre 62,5, on adopte 63.

Représentations graphiques

Les observations se prêtent à des représentations graphiques, souvent très parlantes, que ce soit de simples diagrammes de points, des diagrammes en bâtons ou des polygones de fréquence. C'est en fait par des polygones de fréquence qu'on traduit le plus souvent les compositions des peuplements forestiers. La construction en est simple. L'histogramme qui substitue à ce simple tracé une juxtaposition de rectangles accolés de même largeur formant une sorte d'escalier, a l'avantage de faire ressortir la discontinuité des variables, mais il est un peu plus long à dresser si on veut le faire minutieusement (fig. 13, p. 56).

Une simple page de calepin de comptage effectué en forêt constitue un histogramme (voir couverture). On peut à cet égard remarquer que le procédé de pointage le plus usité par les forestiers en France n'est pas nécessairement toujours le plus pratique : il suppose une certaine habileté de la part du pointeur, un bon crayon et un temps pas trop pluvieux. A l'étranger, on a souvent recours à d'autres figures. Les feuilles de pointage utilisées pour le dépouillement des scrutins les soirs de vote en vue d'élections fournissent aussi d'autres modèles. Enfin, au lieu de pointer par 10, on peut aussi le faire par 5, soit qu'on dispose 5 tirets suivant les côtés et l'une des diagonales d'un carré, soit qu'on représente les éléments au moyen de tirets horizontaux, mais en remplaçant un tiret sur 5 par une barre transversale.

Aire limitée. — On peut figurer soit les nombres absolus tels qu'ils résultent directement des dénombrements effectués, par exemple les nombres de tiges existant dans toute la parcelle ou sur un hectare, soit les « fréquences » existant dans chaque classe.

Le terme de « fréquence » est encore employé par quelques auteurs dans le sens d'*effectif absolu,* de *nombre de répétitions,* mais le plus souvent, on entend par là l'*effectif relatif.*

Si, par exemple, la population est de 300 individus, dont 6 se trouvent dans la classe considérée, on dira que la fréquence pour cette classe est de 6/300 = 0,02. La fréquence n'est pas à confondre avec le pourcentage, qui, au cas particulier, serait exprimé par le nombre 2.

Ayant calculé les fréquences relatives et ayant construit l'histogramme correspondant, on obtient des rectangles accolés. La somme des surfaces des rectangles correspond à la totalité des mesures faites ; elle est donc égale à l'unité.

Série statistique

Qu'est-ce qu'une série statistique ? C'est la collection d'un grand nombre de mesures se rapportant à un même fait.

L'expérience montre qu'il existe une valeur moyenne des faits relevés et que par conséquent, les valeurs n'en sont pas absolument quelconques.

Par exemple, il y a une valeur moyenne pour la température annuelle d'un lieu donné, résultant de sa position et affectée par le jeu de tous les phénomènes qui amènent la chaleur, activité du soleil, courants atmosphériques et marins, etc... Les écarts entre les valeurs observées chaque année sont plus ou moins grands, mais non quelconques.

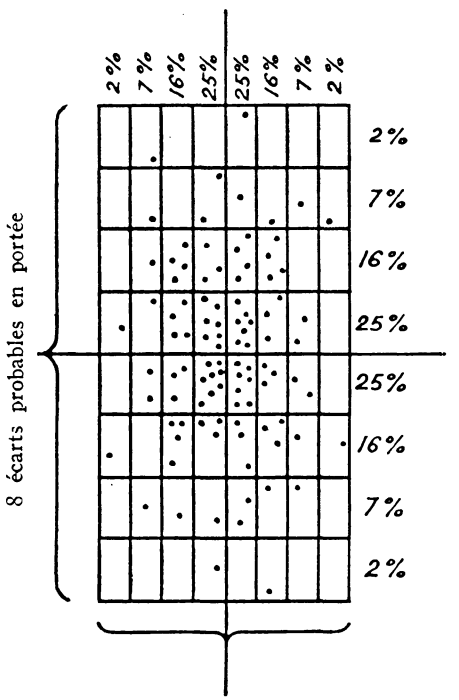

FIG. 5. — Dispersion du tir.

Les petits écarts (ceux qui diffèrent peu de la moyenne) sont les plus fréquents ; si la moyenne annuelle est + 10°, il y aura plus d'années où la température observée sera comprise entre + 9° et + 11° que d'années où cette température observée ira de + 8° à + 9° et de + 11° à + 12° ; et de celles-ci, toujours sur 2° d'écart, plus que d'années où la température observée sera comprise entre + 7° et + 8° et entre + 12° et + 13°.

Si les mesures ont été nombreuses, la fréquence de part et d'au-

tre de la moyenne suit la loi de Gauss, qu'on peut énoncer en la simplifiant comme font les artilleurs (fig. 5).

Sur 100 écarts de part et d'autre de la moyenne, il y en a 50 positifs et 50 négatifs : la moyenne est au milieu de l'échelle des écarts avec le maximum de fréquence. En négligeant les écarts exceptionnellement grands, qui sont très rares (environ 8 pour 1 000), si l'on divise l'intervalle total des écarts observés en 8 parties égales, 4 positives, 4 négatives, les fréquences dans chacune de ces divisions seront systématiquement, à peu près :

2 7 16 25 25 16 7 2 Total : 100

Plus les mesures sont nombreuses, plus la répartition des écarts se rapproche de cette loi, si la série est *homogène.*

Une *série homogène* est caractérisée par la répartition de ses écarts autour d'un maximum de fréquence, parce que cette répartition même témoigne de l'existence d'une cause principale stable. Par exemple, comme ci-dessus, la température moyenne témoigne d'une position particulière du point à la surface de la terre. Les points d'arrivée des balles sur une cible témoignent de ce qu'on a visé le but.

Il y a encore des *séries non homogènes,* où la répartition des fréquences d'écarts ne montre pas un maximum distinct ou en fait deviner plusieurs. Par exemple, la statistique de la taille des conscrits de toute la France est moins nette que celle d'un département seul, parce qu'il s'y fait un mélange de plusieurs races ; et l'on sait bien que les Basques n'ont pas la même taille moyenne que les Flamands.

Loi binomiale et loi normale

Toutes les distributions rencontrées dans la pratique ne suivent pas nécessairement la loi de Gauss. Il ne manque pas de « distributions non-gaussiennes ». Quelques mots sur la « distribution binomiale » ne seront pas déplacés ici : la façon la plus commode d'aborder le sujet est sans doute de le faire en jouant à pile ou face.

Quand nous utilisons une seule pièce de monnaie, nous avons évidemment autant de chances de la voir tomber côté face que côté pile et l'éventualité face F a la même probabilité 1/2 que l'éventualité côté pile P.

Mais jetons en l'air à la fois 2 pièces semblables. Une fois retombées, elles pourront présenter soit 0, soit 1, soit 2 faces. Mais la probabilité de ces 3 éventualités n'est pas la même, car s'il n'y a qu'une façon d'avoir 0 ou 2 faces, il y en a 2 d'avoir 1 face : avec une pièce ou avec l'autre. En regard des éventualités : 0, 1 et 2 F, nous pouvons donc écrire que les probabilités seront entre elles comme les nombres : 1, 2 et 1.

Au lieu de 2 pièces, jetons-en 3 à la fois. Le nombre des éventualités augmente : nous pouvoir avoir 0, 1, 2 ou 3 faces. Il n'y a qu'une façon d'obtenir 0 ou 3 faces, mais une seule face peut être obtenue avec l'une ou l'autre des 3 pièces.

Il en serait de même de l'éventualité inverse : une « pile » contre deux « face ».

Le schéma suivant résume les différentes combinaisons possibles :

	0 F	1 F	2 F	3 F
1re pièce	P	F P P	P F F	F
2e pièce	P	P F P	F P F	F
3e pièce	P	P P F	F F P	F

Au total, au regard des éventualités	0	1	2	et	3 F
nous inscrivons les probabilités (sur 8)	1	3	3		1

Augmentons le nombre n de pièces. Avec 4 pièces, nous obtenons les éventualités :

0 1 2 3 4 F

et les probabilités :

1 4 6 4 1

(sur 16).

En continuant, on constate que les probabilités s'obtiennent très facilement en construisant le triangle suivant, dit « triangle de Pascal » :

Nombre de pièces										Total
1	1	1								2
2	1	2	1							4
3	1	3	3	1						8
4	1	4	6	4	1					16
5	1	5	10	10	5	1				32
6	1	6	15	20	15	6	1			64
7	1	7	21	35	35	21	7	1		128
8	1	8	28	56	70	56	28	8	1	256
9	1	9	36	84	126	126	84	36	9	1	512

Un nombre d'une ligne quelconque se déduit de l'addition de deux nombres situés dans la ligne précédente : celui qui le surmonte exactement et celui qui est à gauche.

Tous ces nombres se trouvent dans le développement du binôme élevé à la n^{me} puissance et les distributions qui se font suivant la même règle sont désignées sous le nom de « *distributions binomiales* ».

On les rencontre en biologie chaque fois que, comme au jeu de pile ou face, on se trouve en présence d'une alternative, que ce soit une question d'action d'un traitement ou d'hérédité : l'intervention a réussi ou échoué, tel phénomène s'est produit ou non, un individu est mort ou a survécu, un petit pois est lisse ou ridé, un animal est mâle ou femelle, etc...

Les distributions binominales sont donc extrêmement importantes à étudier, mais elles donnent lieu à des calculs laborieux, dès que l'exposant n est un peu grand. Heureusement, en même temps que n grandit, le graphique représentatif se régularise jusqu'à tendre vers une courbe continue aux propriétés bien établies : la courbe en cloche qui est l'expression graphique de la loi de Gauss. Une distribution conforme à la loi de Gauss est dite « *normale* ». A la limite, la loi binomiale s'identifie avec la loi normale.

A titre d'exemple, considérons le cas où $n = 9$ (dernière ligne du tableau précédent) et construisons le diagramme en bâtons, puis réunissons les sommets des bâtons par une courbe. Elle présente avec la courbe normale de grandes analogies.

Tout d'abord, elle est symétrique. La moyenne est évidemment à la jonction des classes 4 et 5.

Eliminons ensuite les deux coups extrêmes. Nous obtenons, en arrondissant à l'unité la plus voisine, pour les 510 écarts restant le classement en 8 échelons, tel que nous l'avons exposé plus haut.

N° des échelons	1	2	3	4	5	6	7	8
Fréquence %	2	7	16	25	25	16	7	2

Les pourcentages des artilleurs qui sont déduits de la loi normale se retrouvent très sensiblement dans le cas de la loi binomiale ici analysé. Cette analogie entre la distribution binomiale et la distribution normale justifie l'abandon qu'on fait souvent de la première au profit de la seconde chaque fois qu'il peut résulter de cette substitution une simplification des calculs.

Les paramètres de la courbe normale. — La distribution de Gauss se prête à des manipulations algébriques relativement simples : un de ses principaux avantages est qu'elle peut se caractériser complètement par deux paramètres d'un emploi commode : la « moyenne » et l' « écart-type ».

La moyenne est tout simplement la moyenne arithmétique des mesures. Elle joue un grand rôle dans tous les calculs des statisticiens. Comme, en matière de calcul de probabilité, le principe est qu'une mesure en vaut une autre, c'est la moyenne qui réunit le maximum de probabilité et finalement c'est elle qu'on adopte. Elle est souvent désignée par le signe $\bar{x}$.

En deçà et au delà de la moyenne, les mesures font des écarts et il est intéressant de caractériser de façon simple leur répartition. On y parvient au moyen de l' « écart-type ».

L'écart le plus caractéristique ne peut évidemment se calculer en faisant la moyenne des écarts, puisque, si ces derniers sont pris avec leur signe, du fait de leur répartition symétrique de part et d'autre de la moyenne, la moyenne des écarts est nulle.

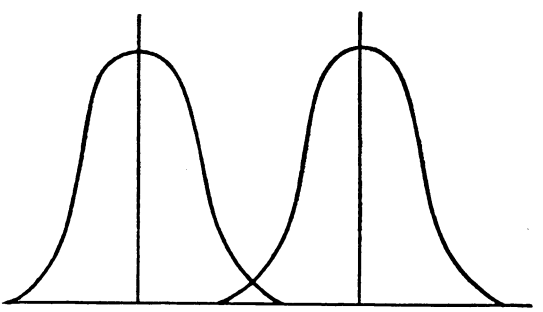

FIG. 6. — Même dispersion, moyennes différentes.

On fait quelquefois la moyenne des écarts pris en valeur absolue et on peut ainsi caractériser un « écart moyen ». Mais en fait l'écart moyen est difficile à manier dans les calculs, à cause du signe « valeur absolue » qu'il contient.

On facilite les calculs en élevant les écarts au carré et du même coup on fait disparaître la différence de leur sens qui produisait la compensation. La moyenne des carrés des écarts constitue la *variance* et la racine carrée de la variance est l'*écart-type* désigné habituellement par la lettre σ.

Les deux courbes de la figure 6 diffèrent seulement par la valeur de la moyenne, mais la dispersion des observations de part et d'autre de la moyenne est la même dans les deux cas.

Au contraire, les deux courbes de la figure 7 représentent deux distributions, ayant même moyenne, mais inégalement aplaties. Plus l'écart-type est considérable, plus l'aplatissement est marqué. L'écart type a en outre une signification géométrique bien nette. En effet, la courbe de Gauss présente des points d'inflexion situés de part et d'autre de la moyenne à une distance mesurée par l'écart-type.

Ecart-type réduit. — Toutes les courbes normales peuvent être considérées comme dérivant d'une courbe normale unique.

Pour concevoir cette courbe, il nous faut partir d'une courbe normale quelconque et faire subir aux axes de coordonnées des déplacements et des changements d'échelles.

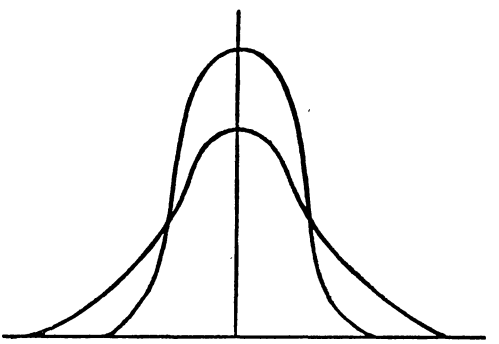

FIG. 7. — Même moyenne, dispersions différentes.

Tout d'abord, nous faisons subir à l'axe oy une translation de façon à ce qu'il devienne un axe de symétrie. Ensuite, nous adoptons comme unité de longueur la valeur de l'écart-type. On a donc une courbe telle que la moyenne est 0 et la variance $\sigma^2 = 1$.

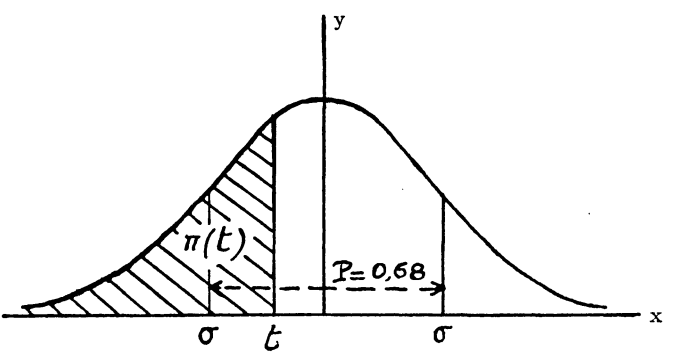

FIG. 8. — Courbe normale réduite.

Enfin, comme il s'agit de courbes de fréquence, il faut que l'aire de la courbe reste toujours égale à l'unité. On y parviendra en amplifiant l'échelle des ordonnées précisément dans le rapport où on aura réduit celle des abscisses.

La courbe normale ainsi obtenue porte le nom de courbe normale réduite. L'écart à la moyenne, chiffré en nombre d'écarts-types, devient l'*écart réduit*.

La figure 8 nous représente cette courbe dont tous les éléments : ordonnée, surface, ont été de longue date calculés par les statisticiens.

Il est entre autres souvent utile de savoir quelle est la surface comprise entre la gauche et une ordonnée donnée. Le tableau indique ainsi la valeur de la surface hachurée de la figure.

t	π (t)	t	π (t)	t	π (t)	t	π (t)
— 3,00	0,00135	— 1,40	0,0808	0,00	0,5000	1,60	0,9452
— 2,90	0,0019	— 1,30	0,0968	0,10	0,5398	1,70	0,9554
— 2,80	0,0026	— 1,20	0,1151	0,20	0,5793	1,80	0,9641
— 2,70	0,0035	— 1,10	0,1357	0,30	0,6179	1,90	0,9713
— 2,60	0,0047	— 1,00	0,1587	0,40	0,6554	2,00	0,9772
— 2,50	0,0062	— 0,90	0,1841	0,50	0,6915	2,10	0,9821
— 2,40	0,0082	— 0,80	0,2119	0,60	0,7257	2,20	0,9861
— 2,30	0,0107	— 0,70	0,2420	0,70	0,7580	2,30	0,9893
— 2,20	0,0139	— 0,60	0,2743	0,80	0,7881	2,40	0,9918
— 2,10	0,0179	— 0,50	0,3085	0,90	0,8159	2,50	0,9938
— 2,00	0,0228	— 0,40	0,3446	1,00	0,8413	2,60	0,9953
— 1,90	0,0287	— 0,30	0,3821	1,10	0,8643	2,70	0,9965
— 1,80	0,0359	— 0,20	0,4207	1,20	0,8849	2,80	0,9974
— 1,70	0,0446	— 0,10	0,4602	1,30	0,9032	2,90	0,9981
— 1,60	0,0548	— 0,00	0,5000	1,40	0,9192	3,00	0,99865
— 1,50	0,0668			1,50	0,9332		

Les indications de ce tableau permettent de tirer des conclusions d'une grande valeur pratique.

Nous remarquons tout d'abord que la surface située à gauche de l'origine est de 0,50, ce qui est bien évident puisque la courbe est symétrique et que la surface totale est de 1.

Considérons maintenant les abscisses + 1 et — 1. Nous lisons respectivement en face d'elles les surfaces 0,8413 et 0,1587.

Entre l'ordonnée — 1 et l'ordonnée + 1, on a donc une surface égale à la différence de ces deux nombres, soit 0,6826.

Nous en conclurons que pour une courbe normale quelconque, nous aurons entre $-\sigma$ et $+\sigma$ une surface de 68 % de l'aire totale, ce qui revient à dire que normalement 68 % des mesures présentent par rapport à la moyenne un écart en plus ou en moins égal à l'écart-type.

Le même calcul nous montre de — 2,0 à + 2,0 une différence de

surface égale à 0,9772 — 0,0228 = 0,9544. Enfin, de — 3,0 à + 3,0, on a : 0,9987 — 0,0014 = 0,9973.

En résumé, si on tire au hasard une mesure, il y a probabilité de 68 % pour que son écart à la moyenne soit égal ou inférieur à une fois l'écart-type, de 95 % pour que cet écart ne dépasse pas 2 fois l'écart-type, enfin de 99,7 % pour que cet écart ne dépasse pas 3 fois l'écart-type.

EXEMPLE. — Supposons que nous avons effectué toute une série de sondages à la tarière dans des arbres sur pied en vue de déterminer leur « temps de passage ». Les mesures ainsi effectuées sont réparties autour d'une moyenne arithmétique M, qui se trouve être de 16 ans. Le calcul des carrés des écarts permet d'autre part de calculer l'écart-type de cette distribution qui se trouve être de 6 ans (1).

Nous passons en coupe et, à cette occasion, nous éliminons les arbres à croissance ralentie. Quel sera le temps de passage des arbres restants ?

1re application. — Admettons que l'enlèvement supprime systématiquement les arbres qui ont les temps de passage les plus longs, et porte sur 25 % de l'effectif. Il nous faut savoir quel est le temps de passage des 75 % restants. Autrement dit, où se situe sur 100 arbres classés par ordre de temps de passage croissants celui qui occupe la place 75/2 = 37,5. Le tableau de la page 40 nous indique que ce sera entre 0.3 σ et 0,4 σ très près de 0,33 σ ou 1/3 σ, Puisque σ = 6 ans, nous concluons à un abaissement de 2 ans du temps de passage. Il faut retrancher 2 ans de M pour obtenir la nouvelle moyenne, qui est 16 — 2 = 14 ans.

2e application. — Si l'enlèvement avait porté sur 20 % seulement, il faudrait chercher quel est le temps de passage des 80 % restants, en d'autres termes, où se trouve l'arbre de rang 0,40. On trouve 0,25 σ, on en conclut que la nouvelle moyenne est améliorée de 6/4 = 1,5 an et devient donc 14,5 ans.

3e application. — Demandons-nous enfin quel est le temps de passage du meilleur tiers. Ceci revient à rechercher l'abscisse correspondant à la surface 1/6, soit 0,16. Nous lisons dans le tableau que si π (t) = 0,16, t est très voisin de — 1. Le meilleur tiers a un temps de passage inférieur de σ = 6 ans au temps de passage moyen qui est de 16 ans. Leur temps de passage moyen est de 16 — 6 = 10 ans.

L. SCHAEFFER.

(1) Il s'agit ici simplement d'une application numérique. Les courbes de temps de passage sont généralement symétriques, mais il n'est pas sûr qu'elles soient gaussiennes.