REVUE DE STATISTIQUE APPLIQUÉE

J. MOTHES

Principes fondamentaux de la méthode statistique et applications industrielles

Revue de statistique appliquée, tome 2, n° 3 (1954), p. 5-36
<http://www.numdam.org/item?id=RSA_1954__2_3_5_0>

PRINCIPES FONDAMENTAUX DE LA MÉTHODE STATISTIQUE ET APPLICATIONS INDUSTRIELLES

par

J. MOTHES

Ancien Élève de l'École Polytechnique,
Ingénieur au Département Commercial du Gaz de France

L'enquête faite auprès des abonnés de la Revue (cf. Revue de Statistique Appliquée, vol. II, n° 2), a montré que de nombreux lecteurs, encore peu au courant du développement des techniques industrielles de la statistique, souhaitaient trouver dans la Revue quelques articles élémentaires précisant les idées fondamentales en cette matière.

La présente étude a été rédigée en vue de répondre à leurs desiderata. Elle comprend trois parties respectivement consacrées :

— *aux fondements des techniques statistiques ;*
— *au controle de la qualité ;*
— *aux techniques d'analyse des résultats de fabrication.*

Bien entendu, le but de ce travail n'est pas de décrire l'ensemble des techniques statistiques susceptibles d'application dans le domaine industriel mais seulement d'énumérer les principales d'entre elles et d'en définir, de façon élémentaire, les modalités pratiques d'emploi.

Ainsi conçus, les chapitres qui vont suivre — dans ce numéro figurent les trois premiers — visent surtout à convaincre les ingénieurs et cadres de l'efficacité d'une science en plein développement.

INTRODUCTION

En évoquant sans autre explication la puissance de la statistique appliquée aux problèmes industriels, on risque de provoquer quelque surprise chez certains lecteurs. Assez souvent, en effet, le public français ne connaît de la statistique qu'un de ses aspects assez particuliers, à savoir les présentations sous forme de tableaux numériques plus ou moins compliqués et plus ou moins rébarbatifs de résultats de dénombrements opérés sur des ensembles — animés ou inanimés — importants. Pour beaucoup d'industriels, il va de soi que les données concernant l'évolution de la production nationale d'acier, la ventilation — par catégorie de contribuables — des sommes payées au titre de l'impôt sur le revenu, la ventilation — par catégorie de marchandises ou de voyageurs — du trafic des chemins de fer, la décomposition, — par région, par sexe ou par âge, etc... — de la population, sont autant de « statistiques ». Mais il en est peu qui savent que le prélèvement périodique — à la sortie d'une machine — d'un petit nombre de pièces et la mesure des cotes de ces pièces constituent une **opération statistique** primaire, riche de conséquences pour qui veut exploiter, toujours par les **méthodes statistiques**, les résultats des mesures opérées.

On ne saurait donc trop insister ici sur le fait que la statistique, après avoir été longtemps une technique de présentation de résultats enregistrés à l'occasion de dénombrements, est devenue, au cours des cinquante dernières années, tout autre chose. Retenant la définition qu'en a donné le statisticien anglais M. G. KENDALL, on peut dire qu'il s'agit de « la méthode scientifique qui traite des données obtenues en comptant ou mesurant les propriétés des populations de phénomènes naturels ». Et, par le fait même que le statisticien analyse les phénomènes naturels **en vue de prendre des décisions**, le but de la statistique est, en définitive, de permettre à l'homme de se comporter de façon telle que la fréquence des réussites consécutives à ses décisions obéisse à des exigences bien définies et fixées à l'avance.

Cette définition posée, il semble évidemment inutile d'insister sur la diversité du champ d'application de cette branche de la mathématique. Il n'y a, en particulier, rien d'étonnant à la voir s'implanter chaque jour davantage dans le domaine industriel où l'on est amené, en permanence, à procéder à des essais de contrôle, puis à analyser les données ainsi recueillies.

Dans un article très complet (1), M. MORICE, Directeur de la « Revue de Statistique Appliquée », a fait le bilan du développement mondial actuel des applications industrielles de la statistique. Nous souhaitons vivement que nos lecteurs s'y reportent : ils y trouveront, en effet, une documentation fort instructive.

***

L'introduction des techniques statistiques dans l'industrie implique l'adoption, par le praticien, d'un vocabulaire nouveau. Si les éléments de ce vocabulaire — d'ailleurs restreint — doivent être, pour la plupart, définis au cours même de l'exposé, il convient cependant d'attirer dès maintenant l'attention du lecteur sur le fait que de nombreux termes d'usage courant ont pris, en statistique, des significations particulières. En vue d'éviter par la suite des méprises regrettables, il paraît donc utile de préciser dès maintenant le sens attribué par les statisticiens à ces termes d'usage courant.

Matière. — On désigne sous le terme de **matière** tout produit (ou toute substance) défini indépendamment de sa forme, que cette forme lui ait été donnée en raison de ses conditions d'emploi, pour faciliter son transport, ou pour répondre à certains usages commerciaux.

Le mot matière peut indifféremment se rapporter à l'acier constituant un lingot, au ciment emmagasiné dans un wagon, au tissu d'une robe, etc...

(1) « Revue de Statistique Appliquée », volume I, n° 2.

Pièce. — Si l'on parle au contraire d'un produit défini en fonction de sa forme extérieure, on utilise le mot **pièce.**

Un lingot (d'acier), un rail, un sac (de ciment), une ampoule (électrique) sont autant d'exemples de pièces.

Lot. — On désigne sous le terme de **lot** toute quantité bien définie d'une même matière ou tout ensemble, également bien défini, de pièces de même type.

Réceptionnant, par exemple, un certain tonnage de ciment ou un certain nombre de caisses de tubes à pommade (de même modèle), on parle le plus souvent du **lot de ciment** ou du **lot de tubes** donnant lieu à réception.

Prise. — Quant on prélève une petite quantité de matière au sein d'une quantité de matière plus importante, on dit qu'il s'agit d'une **prise.**

Specimen. — Quand, d'autre part, on prélève une pièce dans un ensemble de pièces, on dit qu'il s'agit d'un **specimen.**

Eprouvette. — L'**éprouvette** est une pièce (de type généralement spécifié par des normes techniques) fabriquée à partir d'une prise ou directement prélevée dans un spécimen.

Si, par exemple, à l'aide d'une prise opérée dans un sac de ciment on fabrique, en vue d'en étudier la résistance à la compression, un petit cube, ou si l'on découpe dans un lingot, en vue d'en mesurer la charge de rupture, une petite pièce, on a affaire — dans les deux cas — à des éprouvettes.

Essai. — Toute expérience effectuée sur une prise, un spécimen ou une éprouvette, en vue d'en mesurer, d'en repérer, ou d'en indexer une propriété, est un **essai.**

Caractère. — La propriété (des prises, des specimens ou des éprouvettes) sur laquelle on fixe son attention est généralement désignée sous le terme de **caractère.**

Si, par exemple, en présence de pièces mécaniques d'un certain type, on décide de vérifier une de leurs cotes — disons un diamètre — cette cote, autrement dit le diamètre en question, constitue le caractère examiné.

Le caractère est, selon le cas, **qualitatif** ou **quantitatif.** Revenant aux diamètres ci-dessus envisagés, il est clair, en effet, qu'on peut :

1° les distinguer à l'aide de calibres en « convenables » et « défectueux » (sous-entendu : par rapport aux spécifications du dessin) ;

2° ou bien les mesurer.

S'il est, d'autre part, très facile d'exprimer qu'un caractère quantitatif est variable (il suffit de distinguer les valeurs prises par le caractère en question), il paraît plus délicat d'exprimer la même idée dans le cas d'un caractère qualitatif.

Au cours de cette brochure, on tournera néanmoins la difficulté en faisant intervenir la notion d'**attribut** de tout caractère qualitatif.

Dans le cas, par exemple, de l'examen aux calibres des diamètres ci-dessus envisagés, on liera au caractère « diamètre » les deux attributs « convenable » et « défectueux ».

Individu. — On vient de distinguer les prises, les specimens et les éprouvettes, mais il est souvent commode, en Contrôle des Fabrications, de désigner indistinctement prise, specimen ou éprouvette sous le seul terme d'**individu.** Une telle convention offre, en particulier, l'avantage de simplifier la description des techniques de contrôle.

De par sa définition même, l'individu est ainsi, dans tous les cas, l'unité du produit sur laquelle il convient (en vue de mesurer, de repérer, ou d'indexer le caractère pris en considération) de faire porter l'essai.

Par extension, il est souvent commode d'appliquer le terme d'individu au résultat de l'essai.

Population. — Si, en présence d'un lot, on se propose d'étudier une certaine propriété — autrement dit, un certain caractère — du produit constituant le lot, l'observation de ce caractère implique qu'on fasse porter les essais sur des individus et l'ensemble des individus qu on peut, ou pourrait, tirer de l'ensemble du lot constitue ce qu'on appelle la **population.**

Il paraît important de souligner que la notion de « population » ne s'identifie pas exactement à celle de « lot ». Elle présente, sur cette dernière, l'avantage d'être beaucoup plus précise. Elle n'est, en effet, définie qu'après le choix du caractère sur lequel on entend porter son attention et donc du type des individus à soumettre aux essais. A cet égard, l'exemple suivant souligne la différence existant entre les deux concepts.

Imaginons un lot de lingots. Si l'on s'attache à l'étude du poids des lingots, chaque lingot est un individu et la population s'identifie au lot lui-même. Si, en revanche, on s'attache à l'étude de la résistance à la traction de l'acier constituant les lingots, les individus sont des éprouvettes d'un certain type (à extraire des lingots) et la population est l'ensemble des éprouvettes qu'on peut, ou pourrait, tirer de l'ensemble du lot, ou encore l'ensemble des mesures correspondantes de résistance.

Les distinctions qui précèdent peuvent sembler bien subtiles. Elles sont cependant primordiales. Dans la mesure où elles obligent à réfléchir sur la nature exacte des problèmes de contrôle rencontrés, elles conduisent, en effet, à poser correctement ces problèmes.

Inspection à 100 %. — Quand **tous** les individus composant la population observée sont soumis aux essais, on dit qu'il est procédé à une **inspection à 100 %**.

Très souvent, l'inspection à 100 % ne s'impose nullement. Dans bien des cas, elle n'est d'ailleurs même pas réalisable (essais destructifs).

Inspection sur échantillon. — Quand seule une **partie** des individus constituant la population observée est soumise aux essais, on dit qu'il est procédé à une **inspection sur échantillon.**

On désigne de ce fait, sous le terme d'**échantillon**, l'ensemble des individus effectivement soumis aux essais (et **non** une simple « prise témoin » comme on le fait souvent dans le commerce).

PREMIÈRE PARTIE

FONDEMENTS DES TECHNIQUES STATISTIQUES

CHAPITRE I

PRÉSENTATION ET DESCRIPTION DES FLUCTUATIONS RENCONTRÉES EN PRATIQUE

La première phase de tout contrôle consiste à soumettre aux essais un plus ou moins grand nombre d'individus prélevés dans la population à contrôler, les résultats des essais constituant la documentation à exploiter par la suite.

Réceptionnant, par exemple, un lot de ciment dont on désire contrôler la résistance à l'écrasement, la première phase du contrôle consiste à prélever au sein du lot un certain nombre de prises ; à confectionner — à l'aide des prises opérées — des éprouvettes d'un certain type ; à soumettre ces éprouvettes aux essais requis. L'étude des résultats obtenus doit ensuite permettre de décider si le lot de ciment présente ou non les caractéristiques espérées.

De façon analogue, le contrôle — en cours d'usinage — de pièces métalliques débitées par un tour automatique implique obligatoirement que soit prélevé et examiné périodiquement un certain nombre de pièces, l'étude des résultats enregistrés devant permettre de déceler, sans trop de retard, les déréglages du tour.

Abstraction faite, donc, des difficultés de réalisation pratique des essais, on peut dire que la première phase de tout contrôle pose un problème de présentation rationnelle des résultats d'essais.

I. 1 — DISTRIBUTION DE RÉSULTATS D'ESSAIS.

On résout le plus souvent ce problème en adoptant la présentation qui consiste à grouper ensemble les résultats identiques, autrement dit, qui consiste à faire correspondre à chaque valeur ou attribut du caractère pris en considération, le nombre des individus ayant effectivement présenté cette valeur ou cet attribut.

Le tableau à deux colonnes obtenu en procédant de la sorte définit ce qu'on appelle une **distribution.**

1. — **Cas d'un caractère qualitatif** (Tableau I. 1. 1.) :

Distribution obtenue après inspection aux calibres de 175 pièces métalliques :

Attribut	Nombre d'observations	Fréquence en %
Convenable . . .	142	81,1
Défectueux . . .	33	18,9
Total	175	100.

2. — **Cas d'un caractère quantitatif** (Tableau I. 1. 2.) :

Distribution obtenue après avoir fait brûler, à l'air libre, 100 segments de coton poudre (longs de 1 m. chacun) :

Durée de combustion (secondes)	Nombre d'observations
85	1
86	2
87	6
88	14
89	21
90	23
91	16
92	10
93	5
94	2
Total	100

Remarques :

(1) Quand le nombre total des essais effectués est très réduit, il ne peut évidemment être question de présenter leurs résultats sous forme de distribution, aussi se contente-t-on généralement de les ranger par ordre croissant (ou décroissant).

(2) Au lieu de juxtaposer à chacun des attributs ou à chacune des valeurs du caractère pris en considération le **nombre** d'observations qui lui correspond, il est souvent commode d'exprimer ce nombre en valeur relative. Les rapports ainsi obtenus sont alors désignés sous le terme de **fréquences.**

(3) Lorsque le caractère examiné est quantitatif, la distribution peut être présentée de deux façons différentes. L'appareil de mesure mis en œuvre découpant (obligatoirement) de manière discontinue l'intervalle dans lequel varie l'ensemble des observations effectuées (1), il y a, en effet, lieu de distinguer deux cas selon que le nombre des observations est ou non élevé comparativement au nombre total des subdivisions résultant du découpage en question.

Dans le premier cas, plusieurs résultats correspondant — en général — à chaque subdivision, rien ne s'oppose à ce qu'on utilise la distribution sous la forme enregistrée dès la première mise en ordre. Ayant, par exemple, observé, à la seconde près, les durées de combustion (de 100 segments de coton poudre) reportées au Tableau I. 1. 2, c'est-à-dire dispersées dans un intervalle de 10 secondes, tout traitement supplémentaire des données recueillies s'est révélé inutile.

Il n'en est, en revanche, pas de même dans le second qui conduit à procéder à l'opération dite de **groupage en classes.**

Revenant à l'exemple précédent, imaginons qu'on ait utilisé, pour mesurer les durées de combustion, un chronomètre donnant le millième de seconde. L'intervalle total de variation aurait couvert 10.000 subdivisions et chaque subdivision n'aurait **au plus** compris qu'un très petit nombre d'observations. Telle quelle, la distribution obtenue se serait donc avérée très difficile à manier.

Grouper en classes consiste à diviser l'intervalle total de variation du caractère observé en un nombre raisonnable de classes successives, puis à attribuer à chaque classe le nombre de résultats qui lui correspond.

(1) En dixièmes, centièmes ou millièmes de millimètre par exemple.

Exemple (Tableau I. I. 3.) :

Distribution d'excentricités de pointes de balles de fusil :

Excentricité 1/100 de mm.	Nombre d'observations	Fréquence en %
0 — 10	15	3,3
10 — 20	62	13,8
20 — 30	94	20,9
30 — 40	90	20,0
40 — 50	86	19,2
50 — 60	58	12,9
60 — 70	28	6,2
70 — 80	11	2,4
80 — 90	4	0,9
90 — 100	2	0,4
	450	100

Au demeurant, les deux types de présentation ne diffèrent pas radicalement l'un de l'autre, le second ne consistant, en dernière analyse, qu'à superposer un second groupage à celui qu'opère initialement l'appareil de mesure mis en œuvre. Il convient seulement de noter que tout groupage en classes, s'il est commode, fait cependant perdre un peu d'information. Si l'on dispose, par utilisation d'un chronomètre qui ne donne que la seconde, de 6 évaluations de durées de combustion égales à 87 secondes, il est clair qu'on n'a aucune indication sur la répartition des 6 résultats dans l'intervalle 86,5 — 87,5 secondes, tandis qu'en présence de résultats enregistrés au millième de seconde :

86.574
86.691
86.908
87.057
87.221
87.342

on dispose d'une information dont on accepte **volontairement** de se priver dès qu'on procède au groupage sur 87.

En pratique, il est utile de choisir l'intervalle de classe de façon à ce que l'intervalle total de variation (ou **étendue**) des résultats enregistrés couvre une quinzaine de classes (environ). Il y a intérêt à utiliser des classes d'égale amplitude.

(4) La représentation graphique d'une distribution du type I. I. 2, représentation qui consiste à porter sur l'horizontale les valeurs du caractère et, verticalement les effectifs ou les fréquences correspondants, est désignée sous le nom de diagramme en bâtons.

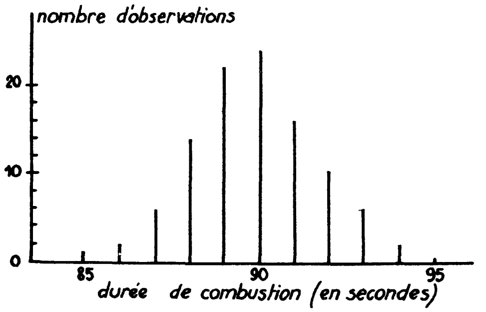

Fig. I. I. — Diagramme en bâtons.

De son côté, la représentation graphique d'une distribution du type I. I. 3, représentation qui consiste à porter sur l'horizontale les intervalles de classe et verticalement, sur ces intervalles de classe, des rectangles de surfaces proportionnelles aux fréquences correspondantes, est désignée sous le nom d'**histogramme.**

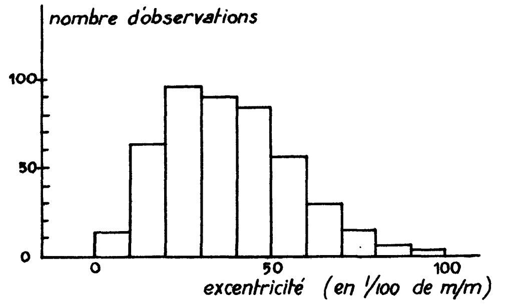

Fig. I. 2. — Histogramme.

On conçoit aisément que si la précision des mesures augmente et si le nombre d'observations est très grand, on peut envisager des classes d'amplitude très petite ; dans ces conditions la ligne en escalier limitant supérieurement l'histogramme tendra vers une courbe continue appelée **courbe des fréquences.**

Il est souvent commode d'envisager une représentation graphique donnant le nombre (ou la fréquence) des observations inférieures (ou supérieures) à une valeur donnée : on obtiendra ainsi le **diagramme cumulatif** tendant lui aussi dans les mêmes conditions vers une courbe continue (courbe de répartition de la distribution).

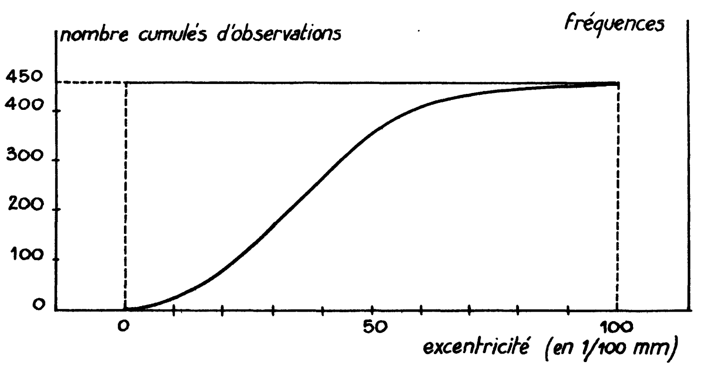

Fig. I. 3. — Diagramme cumulatif.

I. 2. — DESCRIPTION DES DISTRIBUTIONS RELATIVES A DES CARACTÈRES QUANTITATIFS.

Du tableau et du polygone de fréquence I. I. 2 — qui constituent la somme documentaire recueillie à la suite des essais poursuivis sur les segments de coton poudre — il est possible de tirer un aperçu synthétique.

Très intuitivement, on sent qu'on a le droit de « résumer », au moins en première approximation, la distribution observée en disant que les durées de combustion des segments soumis aux essais se sont, dans l'ensemble, groupées aux alentours de 90 secondes. On sent également qu'il convient, si l'on veut préciser davantage ce résumé, de mettre en évidence, d'une part, le fait que les durées de combustion se révèlent en majorité voisines de 90 secondes et, d'autre part, le fait qu'elles se dispersent autour de cette position centrale. Il n'y a donc pas lieu de s'étonner de ce que les statisticiens se soient efforcés de définir des paramètres — on dit également des **caractéristiques** — susceptibles de décrire les distributions sous les deux aspects « **tendance centrale** » et « **dispersion** ». La liste des caractéristiques auxquelles ils sont parvenus étant très longue, nous n'avons d'ailleurs pas la prétention de toutes les passer en revue ici. Il va nous suffire de définir les principales d'entre elles.

1. — Caractéristiques de tendance centrale.

Mode : C'est la valeur du caractère à laquelle correspond le plus grand nombre d'observations.

Médiane : C'est la valeur du caractère autour de laquelle se partagent par moitié les observations.

Moyenne arithmétique : Elle est trop connue pour être définie. De façon générale, si l'on désigne par n_i le nombre d'observations de valeur x_i, la moyenne arithmétique s'écrit :

$$m' = \frac{n_1 x_1 + n_2 x_2 + \dots}{n_1 + n_2 + \dots} = \frac{\Sigma\ n_i\ x_i}{\Sigma\ n_i} .$$

On la désigne aussi, très souvent, en surlignant l'indicatif de la variable à laquelle elle se rapporte. Dans le cas présent, par exemple, cela conduit à remplacer m' par $\overline{x}$.

2. — Caractéristiques de dispersion.

Etendue (ou amplitude ou encore intervalle de variation) : C'est la différence entre la plus forte et la plus faible valeur observée à la suite des essais.

Variance : C'est la somme, divisée par le nombre total d'observations, des carrés des différences à leur moyenne arithmétique des valeurs observées. On la désigne généralement par le symbole σ'^2 (lire : sigma prime deux), ou par le symbole s'^2.

$$\sigma'^2 = \frac{\Sigma\ n_i\ (x_i - m')^2}{\Sigma\ n_i} .$$

Ecart-type : C'est la racine carrée (positive) de la variance évidemment désignée par σ' (lire : sigma prime), ou s' .

Exemple : Les calculs des principales caractéristiques de tendance centrale et de dispersion des distributions de résultats d'essais sont, en général, extrêmement simples, beaucoup plus simples même que ne le laissent supposer les définitions qui précèdent.

Pour les simplifier au maximum, il suffit de tenir compte des remarques suivantes :

1°. — Quant on a affaire à une distribution obtenue après groupage en classes, il convient de remplacer — au cours des calculs — chaque classe par sa valeur centrale ;

2°. — Il y a souvent intérêt à calculer la moyenne, la variance et l'écart-type de la distribution non pas à partir des valeurs x_i directement observées, mais à partir des valeurs $t_i \quad x_i - x_0$ des x_i rapportées à une nouvelle origine x_0.

Dans le cas de la distribution I. I. 2, par exemple, il est commode — plutôt que de mener les calculs à l'aide des valeurs 85, 86, etc... de choisir, au départ, comme nouvelle origine

$$x_0 \quad 90$$

et de travailler sur les valeurs — 5, — 4..., etc...

On repasse ensuite, sans difficulté, de la moyenne t' des t_i à la moyenne réelle m , en utilisant la relation :

$$m' = x_0 + t' .$$

Par ailleurs, la variance et l'écart-type des t_i coïncident avec la variance et l'écart-type des x_i.

3°. — Le calcul de la médiane est immédiat par cumul successif des nombres d'observations correspondant aux différentes valeurs du caractère.

Il suffit, en effet, de retenir comme médiane la valeur du caractère en regard de laquelle le total des observations jusque là cumulées atteint la moitié du total général des observations prises en compte dans la distribution.

Graphiquement, pour un ensemble de n observations, la médiane sera déterminée à partir du diagramme cumulatif comme étant l'abscisse du point dont l'ordonnée est égale à $\frac{n}{2}$ ou à 0,50, selon qu'il s'agit du diagramme des effectifs ou du diagramme des fréquences.

TABLEAU DE CALCUL I. 2. I. — DISTRIBUTION I. I. 2.

Durée de combustion x_i (1)	Nombre d'observations n_i (2)	n_i cumulées (3)	Durée après changement d'origine $t_i = x_i - x_o$ (4)	$n_i\ t_i = n_i\ (x_i - x_o)$ (5)	$n_i\ t_i^2 = n_i\ (x_i - x_o)^2$ (6)=(4)x(5)
85	1	1	— 5	— 5	25
86	2	3	— 4	— 8	32
87	6	9	— 3	— 18	54
88	14	23	— 2	— 28	56
89	21	44	— 1	— 21	21
90	23	67	— 0	0	0
91	16	83	1	16	16
92	10	93	2	20	40
93	5	98	3	15	45
94	2	100	4	8	32
Total	100			— 21	321

Mode : Le mode est égal à 90. C'est, en effet, à cette valeur de x que correspond le nombre d'observations maximum (23).

Médiane : La médiane étant par définition la valeur du caractère autour de laquelle se partagent, par moitié, les observations, correspond ici à la 50-51[e] observation (puisqu'il y en a 100). Or, il ressort de l'examen de la colonne (3) que 23 observations dont les numéros d'ordre, par ordre croissant, vont de 45 à 67, correspondent à $x_i = 90$. C'est donc là la valeur médiane.

Moyenne arithmétique : Par rapport à l'origine $x_o = 90$, la moyenne arithmétique

$$t' = \frac{\Sigma n_i\ (x_i - x_o)}{\Sigma\ n_i}$$

n'est autre que le total de la colonne (5) divisé par le nombre total d'observations

$$t' = \frac{-21}{100} - 0{,}21 ,$$

d'où

$$\bar{x} = m' = 90 - 0{,}21 = 89{,}8$$

Amplitude ou étendue : L'étendue, différence entre la plus forte et la plus faible observation, est évidemment égale à :

$$w' = 94 - 85 = 9$$

Variance : La variance des x_i étant la même que la variance des t_i, on peut écrire

$$\sigma'^2 = \frac{\Sigma\, n_i\, (t_i - t')^2}{\Sigma\, n_i} = \frac{\Sigma n_i\, t_i^{\,2}}{\Sigma\, n_i} - t'^2$$

La somme : $\Sigma\, n_i\, t_i^2$ figurant dans la colonne (6) on a alors :

$$\sigma'^2 = 3{,}21 - (0{,}21)^2 = 3.17.$$

Ecart-type : Dans ces conditions

$$\sigma' = 1.78$$

Remarque : Dans le cas des distributions du type I. I. 3. (valeurs du caractère groupées en classes) les calculs s'effectuent de façon identique en retenant comme valeur x_i de chaque classe, la valeur centrale de cette classe.

Dans ce cas, si on pose :

$$x = x_0 + a\, t,$$

x_0 étant le centre d'une classe quelconque,

a : l'intervalle de classe,

t : une variable auxiliaire définissant le rang (positif ou négatif) d'une classe quelconque par rapport à la classe du centre x_0 prise comme origine,

on aura :

$$\overline{x} = m' = x_0 + a\, t'$$

t' étant la moyenne pondérée des valeurs des t :

$$t' = \frac{\Sigma\, n_i\, t_i}{\Sigma\, n_i}$$

De même

$$\sigma'^2 = \frac{a^2\, \Sigma\, n_i\, (t_i - t')^2}{\Sigma\, n_i} = \frac{a^2}{\Sigma\, n_i} \quad \Sigma\, n_i\, t_i^2 - t'^2$$

TABLEAU DE CALCUL I. 2. 2. — DISTRIBUTION I. I. 3.

Excentricité (1)	Centre de classe x_i (2)	Nombre d'observations n_i (3)	Variable auxiliaire t_i (4)	$n_i\ t_i$ (5)	$n_i\ t_i^2$ (6) = (4) x (5)
0 — 10	5	15	— 4	— 60	240
10 — 20	15	62	— 3	— 186	558
20 — 30	25	94	— 2	— 188	376
30 — 40	35	90	— 1	— 90	90
40 — 50	45	86	0	0	0
50 — 60	55	58	1	58	58
60 — 70	65	28	2	56	112
70 — 80	75	11	3	33	99
80 — 90	85	4	4	16	64
90 — 100	95	2	5	10	50
		450		— 351	1 647

d'où l'on tire $a = 10$ $\qquad x_0 = 45$

$$t' = \frac{351}{450} = -\ 0{,}78 \qquad \overline{x} = 45 - 10 \times 0{,}78 = 37{,}2$$

$$\sigma'^2 = 365{,}8648 \qquad \sigma' = 19{,}13$$

I. 3. — DIVERSITÉ DES CAUSES DE FLUCTUATIONS DES FABRICATIONS.

La mise en ordre — sous forme de distribution — des résultats d'essais ne s'impose évidemment que dans la mesure où les résultats sont entachés de plus ou moins larges fluctuations. Mais c'est là un phénomène très général. En dépit des précautions prises en fabrication, les propriétés des produits usinés ne sont jamais constantes : un exemple (1) que nous emprunterons à M. YRIBARREN illustre d'ailleurs clairement, dans un cas particulier, la diversité des sources possibles de dispersion.

Se référant au cas d'un **tour parallèle classique**, cet auteur les dénombre, en effet, comme suit (2) :

A. — **Provenant de la machine et des montages :**

1. — Reproduction des ondulations des surfaces de glissement ou des chemins de roulement ;
2. — Imperfections des organes cinématiques discontinus ou périodiques, bielles, engrenages ;
3. — Imperfections par défaut d'équilibrage et inerties secondaires des moteurs, pièces ou porte-pièces.

B. — **Provenant du Porte-Pièces :**

a) **Ecarts de manœuvre :**

4. — Déplacements de l'axe de la broche de poupée par rapport à lui-même sous l'influence : de démarrages et freinages, de changements de vitesse de broche, de variations éventuelles dues à la présence de courroies ;
5. — Variations par rapport à l'axe poupée — contrepointe, dues au dégagement et au réengagement entrepointes ;
6. — Jeu de la broche de contrepointe sous l'action du système de blocage de celle-ci.

b) **Ecarts dûs aux contraintes en travail :**

7. — Déplacement de l'axe de la broche de la poupée à l'attaque de l'outil et en passe (influence des jeux de coussinets ou des roulements) ;
8. — Déplacement de l'axe de la broche de contrepointe dans les mêmes conditions.

C. — **Provenant du Porte-outil :**

a) **Ecarts de manœuvre :**

9. — Variations, par rapport à sa position type, de la position réelle de l'outil dans des conditions apparemment identiques de remise à zéro (influence de jeux dans les glissières et les systèmes vis-écrou des chariots transversal et longitudinal) ;
10. — Variations, par rapport à sa position type, de la position réelle de l'outil, lors du déplacement du traînard (influence des jeux et imperfections dans les glissières banc-traînard et contraintes d'entraînement pignon-crémaillère ou vis-mère).

b) **Ecarts dûs aux contraintes en travail :**

11. — Affaissement ou levée des chariots transversal et longitudinal sur leurs glissières ;
12. — Affaissement ou levée du grand chariot longitudinal sur ses glissières ;
13. — Déformations de l'outil en travail ;
14. — Déformations du porte-outil et de la tourelle porte-outil.

M. YRIBARREN dénombre donc, en ce qui concerne la machine-outil la plus courante, quatorze sources de dispersion. Encore ne tient-il compte que des éléments perturbateurs imputables au seul matériel de fabrication. En pratique, les matières premières, les opérateurs, les appareils de mesure mis en œuvre lors des essais, sont autant de causes supplémentaires de fluctuations qu'il est a priori difficile de négliger.

---

(1) Cité par M. METRAL : « La haute précision, impératif catégorique de la machine-outil » (MICROTECNIC, Vol. I, n° 6, 1952).

(2) Dans l'hypothèse, restrictive, de la constance des caractéristiques de l'outil.

CHAPITRE II

QUELQUES SCHÉMAS THÉORIQUES DE FLUCTUATIONS

En présence d'un phénomène naturel, la démarche de l'esprit humain est d'essayer d'en trouver une représentation simplifiée. Les théoriciens se sont donc efforcés d'imaginer des schémas « théoriques » de fluctuations voisins des schémas les plus couramment observés en pratique.

Pour ce faire, ils ont analysé certains mécanismes générateurs de fluctuations et ils ont tout d'abord porté leur attention sur les mécanismes les plus simples : les jeux (de dés, de cartes, etc...). Ensuite, ils ont progressivement envisagé des schémas plus complexes.

Ils sont ainsi parvenus à élaborer une très grande variété de « modèles théoriques » : il nous suffira d'énumérer ici les principaux d'entre eux ; ce sont d'ailleurs les plus utiles en pratique industrielle courante.

II. — I. — LA DISTRIBUTION NORMALE OU DISTRIBUTION DE LAPLACE-GAUSS.

I. — Description.

La distribution normale, également appelée distribution de Laplace-Gauss, a pour représentation graphique la courbe connue sous le nom de « courbe en cloche » dont tout le monde a plus ou moins entendu parler. C'est une distribution « continue » en ce sens que le caractère x auquel elle se rapporte varie de façon continue.

Il lui correspond, comme à toutes les distributions, des caractéristiques de tendance centrale et de dispersion. Les principales d'entre elles sont la **moyenne** et l'**écart-type** : une distribution normale est en effet **entièrement définie** par ces deux paramètres.

Si l'on désigne par m la moyenne et par σ l'écart-type d'une distribution normale et si l'on fait le changement de variable

$$t = \frac{x - m}{\sigma}$$

la distribution se met sous une forme qu'on dit « réduite ».

Dans la table N° 3 (il en existe de plus détaillées) figurent les proportions théoriques d'observations — on dit encore : les fréquences théoriques ou probabilités π (t) — correspondant, pour différentes valeurs de t, à l'intervalle $(-\infty, t)$ (moins l'infini, t). Les fréquences théoriques correspondant à un intervalle quelconque (t_1, t_2) s'en déduisent immédiatement par différence

$$P(t_1, t_2) = \pi(t_2) - \pi(t_1)$$

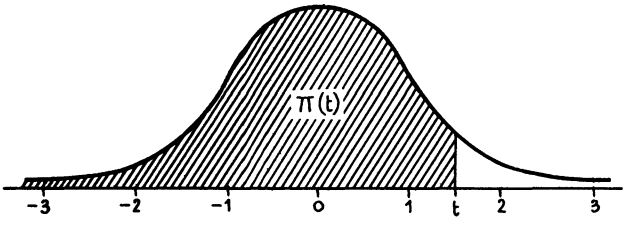

Fig. II. I. — Loi normale (Courbe de distribution.)

La relation existant entre les valeurs de t et les valeurs des fréquences théoriques correspondant à l'intervalle ($-\infty$, t) définit ce qu'on appelle la **fonction de répartition de la distribution normale réduite.**

Remarques importantes.

Si l'on considère une distribution normale de moyenne m et d'écart-type σ , il suffit, pour définir la fréquence théorique correspondant à un intervalle (x_1, x_2), de former :

$$t_1 = \frac{x_1 - m}{\sigma} \qquad\qquad t_2 = \frac{x_2 - m}{\sigma}$$

et de faire la différence entre les valeurs de la fonction de distribution de la loi normale réduite lues en regard de t_1 et t_2, dans la table N° 3.

Exemple :

Etant donné la distribution normale de moyenne m = 19,8 et d'écart-type $\sigma = 2,3$, déterminer la fréquence correspondant à l'intervalle : $x_1 = 21,2$; $x_2 = 22,3$.

On a :

$$t_1 = \frac{x_1 - 19,8}{2,3} = 0,6 \qquad t_2 = \frac{x_2 - 19,8}{2,3} = 1,1$$

Or, dans la table N° 3, on lit, en regard de t_1 et t_2, les fréquences

$$\pi_1 = 0,7257 \qquad\qquad \pi_2 = 0,8643$$

A l'intervalle (x_1, x_2) correspond donc la fréquence

$$\pi_2 - \pi_1 = 0,1386$$

2. — Propriétés :

La distribution normale possède un certain nombre de propriétés qu'il convient de toujours garder présentes à l'esprit, en raison de leur intérêt pratique :

a) — **Forme :**

Elle est de forme symétrique : la fréquence théorique correspondant à tout intervalle (t_1, t_2) est identique à la fréquence correspondant à l'intervalle ($-t_1$ $-t_2$).

b) — **Intervalles privilégiés:**

De l'examen de la table de la distribution normale réduite (cf. table N° 3), il ressort :

— qu'à l'intervalle ($t = -2/3$, $t = 2/3$ correspond la fréquence 50 % environ ;

— qu'à l'intervalle ($t = -1$, $t = +1$) correspond la fréquence 68 % environ ;

— qu'à l'intervalle ($t = -2$, $t = +2$) correspond la fréquence 95 % environ ;

— qu'à l'intervalle ($t = -3$, $t = +3$) correspond la fréquence 997 ‰ environ.

Ces résultats définissent très simplement quatre intervalles particuliers comprenant respectivement la moitié, les deux tiers, les dix-neuf vingtièmes et la quasi totalité d'un ensemble d'observations « théoriques » normalement distribuées. Fournissant une description résumée de la forme de la distribution normale, ils sont très souvent utilisés en pratique lorsqu'il s'agit de comparer sommairement à la normalité la forme d'une distribution effectivement observée.

On notera de plus que la distribution normale est illimitée (t pouvant varier de $-\infty$ à $+\infty$), alors que pratiquement les variables observées varient dans un intervalle fini.

Il n'en résulte généralement aucun inconvénient en raison du fait, signalé ci-dessus, que la quasi totalité des valeurs de la variable réduite t se trouve dans un intervalle limité (la fréquence des observations extérieures à l'intervalle ($t = -4\sigma$, $t = +4\sigma$) étant de l'ordre de un demi cent-millième),

c) — **Droite de Henry :**

La représentation graphique qui consiste à porter en abscisse les valeurs de t et en ordonnée les valeurs de la fonction de répartition π (t) de la loi normale réduite conduit à la courbe en forme de S dessinée Fig. II. 2., mais par une transformation simple, il est possible de passer de cette courbe en S **à une droite.**

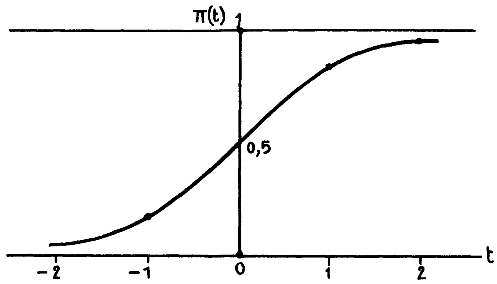

Fig. II. 2. — Loi normale (Courbe de répartition).

Si l'on considère la relation :

$$t = \frac{x - m}{\sigma}$$

relative à une loi normale donnée (m,σ), on voit que les points de coordonnées (x, t) seront alignés.

Si la variable x observée est normalement distribuée, on pourra déduire des observations la fréquence expérimentale π'_i des observations inférieures à une valeur particulière x_i de x.

A cette valeur π'_i, la table N° 3 permet de faire correspondre une valeur t_i de t. Ainsi, par exemple, si 70 % des observations sont inférieures à x_i, on aura, dans l'hypothèse d'une distribution normale,

$$\pi\,(t_i) = \pi'_i = 0{,}70 \longrightarrow t_i = 0{,}53$$

On pourra donc construire, à partir des observations, le lieu des points (x_i, t_i) qui sera une droite si la distribution est normale. En réalité, même si la distribution est rigoureusement normale, les points ne seront qu'approximativement alignés, les observations ne portant, en général, que sur un échantillon limité de cette population.

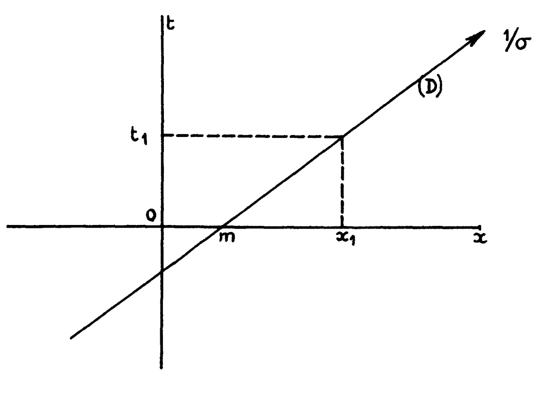

Fig. II. 3. — Droite de Henry.

On remarquera que si l'hypothèse d'une distribution normale est vérifiée, c'est-à-dire si les points (x_i, t_i) sont alignés, la droite D obtenue permet une estimation de m et de σ .

En pratique, il existe dans le commerce des papiers quadrillés dans lesquels l'axe des ordonnées est directement gradué suivant les valeurs de π (t).

L'usage de ces papiers est très commode pour comparer à la normalité la forme d'une distribution effectivement observée.

3. — Importance pratique de la distribution normale.

L'importance pratique de la distribution normale est considérable.

Dans la pratique, en effet, les causes perturbatrices élémentaires à l'origine des fluctuations observées sont, généralement, **nombreuses, d'importance réduite** et telles que leurs effets **s'ajoutent** les uns aux autres or, on démontre que les fluctuations obtenues sous l'effet de causes élémentaires de ce type répondent justement au « modèle » normal.

Une expérience, bien facile à réaliser (1), illustre très clairement le phénomène. Cette expérience consiste, lorsqu'on a à sa disposition une boîte dont le fond est parsemé de clous plantés verticalement, dont un côté est percé de manière à laisser entrer l'extrémité d'un entonnoir et dont le couvercle est fait de matière transparente, à poser la boîte verticalement et, l'entonnoir étant en place, à y faire tomber un assez grand nombre de billes. Les chocs successifs de chacune des billes sur les clous étant autant de causes de fluctuations additives, on observe, en fin d'expérience, une répartition de l'ensemble des billes utilisées assez voisine de la courbe en cloche.

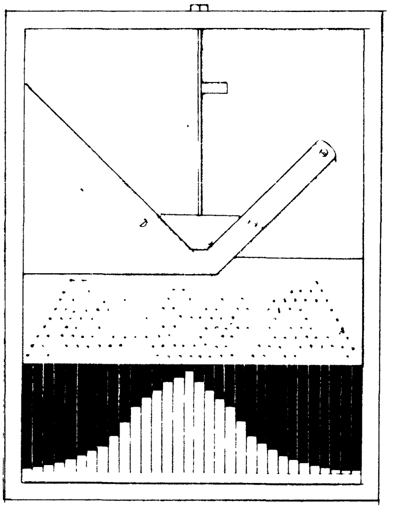

Figure II. — 4

II. 2. — LA DISTRIBUTION BINOMIALE.

I. — Description.

Parmi les mécanismes générateurs de fluctuations, il en est un extrêmement élémentaire : celui des tirages dans une urne contenant des boules de deux catégories.

Si l'on considère, en effet, une urne contenant deux catégories de boules, par exemple des boules noires et des boules blanches — les premières en proportion p, les secondes en proportion I — p — et si l'on procède, dans cette urne, à des tirages successifs de n boules, on constate que la proportion de boules noires observées parmi les n boules prélevées **varie d'un tirage à l'autre**

(1) Un minimum de précautions est cependant nécessaire. L'appareillage ici décrit, pour fonctionner convenablement, doit être construit très soigneusement. Il est recommandé de le doter d'un niveau d'eau de façon à pouvoir assurer sa parfaite horizontalité en cours d'expérience.

autour de p, autrement dit, que le nombre de boules noires prélevées chaque fois varie autour de la valeur np. Ce nombre, dans les conditions de l'expérience envisagée, est donc bien un résultat entaché de fluctuations.

La distribution binomiale qui résulte de l'étude de ce mécanisme définit — pour n et p donnés, c'est-à-dire une fois fixées la composition de l'urne et la taille de l'échantillon — les fréquences théoriques correspondant aux n + 1 valeurs possibles (0, 1, ... n) du nombre k de boules noires susceptibles d'être observé à l'occasion de chaque prélèvement.

Dans la table N° 4 figurent les proportions théoriques (ou probabilités) correspondant, pour différentes valeurs de n et de p, d'une part aux intervalles (o, k) (limites comprises), d'autre part, aux valeurs de k. Dans le premier cas, il s'agit de la **fonction de répartition de la distribution binomiale considérée,** dans le second de la **distribution binomiale** elle-même.

Si l'on désigne par $P_{(o,k)}$ les ordonnées de la fonction de répartition de la distribution et par p_k celles de la distribution elle-même, on a, par définition :

$$p_{k+1} = P_{(o,k+1)} - P_{(o,k)}$$

De même :

$$P(a \leq x \leq b) = P_{(o,b)} - P_{(o,a-1)}$$

On notera qu'il s'agit, cette fois, de la distribution d'une variable discontinue (qui ne peut prendre que les valeurs 0, 1, ... n).

Le graphique de la fonction de répartition est alors un diagramme en escalier.

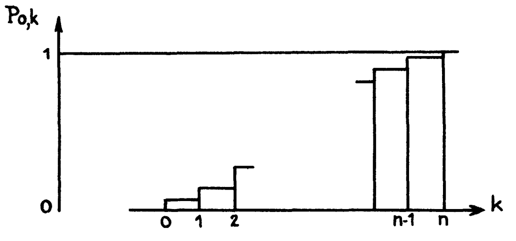

Fig. II. 5. — Distribution binomiale.

REMARQUE IMPORTANTE SUR LES CONDITIONS DE PRÉLÈVEMENT DE L'ÉCHANTILLON

Si l'on considère une urne contenant des boules blanches et des boules noires, les secondes en proportion p, et si l'on procède, dans cette urne, à des tirages de n boules, on vient de dire que le nombre k de boules noires observé dans ces échantillons varie conformément à la loi binomiale correspondant à n et à p.

Cette affirmation n'est cependant exacte qu'à condition d'opérer le prélèvement des échantillons de façon à ce que :

1° — la composition de l'urne reste invariable au cours des tirages ;

2° — tous les éléments de l'urne aient, avant chacun des tirages, autant de chances les uns que les autres d'être prélevés.

La première condition, pour être rigoureusement respectée, implique un « **échantillonnage non exhaustif** », échantillonnage qui consiste à remettre chaque boule tirée dans l'urne avant de prélever la suivante.

La seconde condition définit ce qu'on appelle l'**échantillonnage au hasard.**

Au point de vue pratique, la condition de « non-exhaustivité » n'est pas absolument impérative. **Si la taille n de l'échantillon est faible comparativement au nombre total des boules placées dans l'urne,** on démontre en effet que les résultats continuent à se raccorder correctement à la loi binomiale dans l'hypothèse des tirages effectués sans remises successives des boules extraites dans l'urne.

La condition « d'échantillonnage au hasard » est, au contraire, beaucoup plus impérative. L'anecdote suivante est, à cet égard, très significative.

Lorsque le personnel du Centre de Formation des Ingénieurs et Cadres aux Applications Industrielles de la Statistique définit le programme de son stage d'enseignement élémentaire et décida de présenter sous forme expérimentale à ses élèves les résultats fondamentaux de la statistique, il lui parut suffisant, pour traiter de la distribution binomiale, de prévoir des tirages dans une urne. Il se procura donc un bocal cylindrique, une population de perles blanches et noires et une petite spatule percée de 25 trous d'un diamètre correspondant à celui des perles, puis expérimenta le matériel en répétant un grand nombre de fois l'échantillonnage de 25 perles. Cela fait, il constata que la distribution des nombres k de boules noires successivement observés ne se raccordait absolument pas au résultat théorique attendu.

L'explication de ce phénomène fut assez rapidement trouvée. Le bocal, cylindrique, se prêtait mal à un prélèvement au hasard : pratiquement, la spatule laissait systématiquement de côté les **perles** situées au voisinage de la paroi verticale, c'est-à-dire que l'échantillonnage se faisait dans le tronc de cône C et non dans le cylindre de composition connue (n, p).

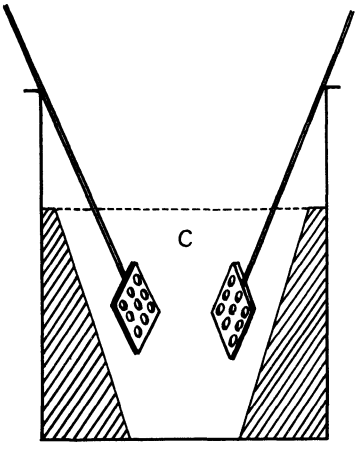

Fig. II. 6. — Urne cylindrique

Pour vaincre cette difficulté, le Centre dut mettre au point un type d'urne particulier assurant, en cours de prélèvement, un brassage automatique de l'ensemble des perles.

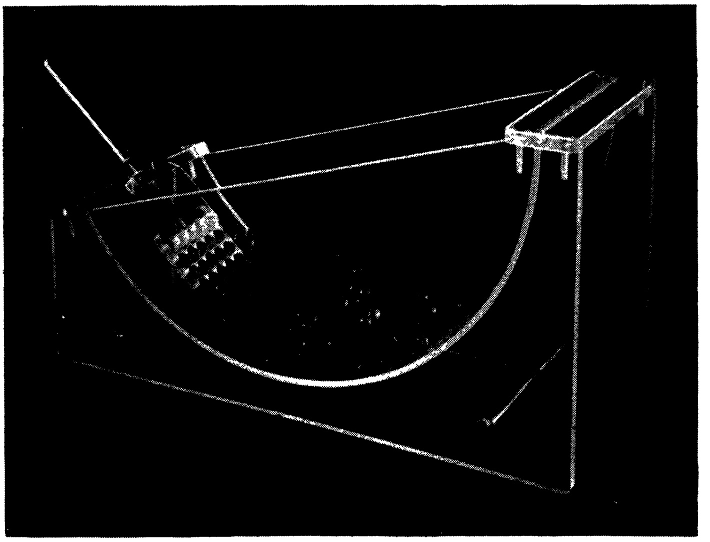

photo Michel CAMBAZARD

2. — Propriétés.

a) **Moyenne et variance.** — La moyenne d'une distribution binomiale — autrement dit : la moyenne du nombre k de boules noires susceptibles d'être observées dans des échantillons de n boules extraits d'une urne contenant des boules noires en proportion p — est égale à :

$$m - np$$

La variance est, d'autre part, égale à :

$$\sigma^2 = np\,(1 - p)$$

Si l'on s'intéresse à la fréquence f k/n du nombre des boules noires tirées dans des échantillons de n boules, on obtiendra ainsi une nouvelle variable aléatoire f dont la distribution binomiale sera caractérisée par une moyenne $\frac{np}{n} = p$ et une variance $\frac{p\,(1 - p)}{n}$

b) **Formes limites.** — Lorsque la taille n de l'échantillon augmente, la proportion p n'étant pas trop faible, la distribution binomiale tend à s'identifier à la distribution normale de moyenne np et de variance $np\ (1 - p)$.

On peut considérer ces conditions comme pratiquement satisfaites lorsque l'on a :

$$np\ (1 - p) > 20$$

Quand la proportion p est au contraire faible, la distribution binomiale a pour « forme limite » la distribution dite de Poisson qui fera l'objet de la section II. 3.

En pratique, p peut être considérée comme « faible » en deçà de 5 %, et n peut être considéré comme « élevé » au delà de 100.

3. — Importance pratique de la distribution binomiale.

L'intérêt pratique de la distribution binomiale est évident. Au lieu de parler d'une urne contenant une certaine proportion p de boules noires, il suffit, en effet, de parler « d'un lot contenant une certaine proportion p de rebuts » pour constater que le schéma théorique décrit permet de définir la fréquence théorique ou probabilité du nombre k de pièces défectueuses susceptibles de figurer dans un échantillon de n pièces extrait du lot en question. On en déduit immédiatement l'importance du rôle joué par la loi binomiale dans un grand nombre de problèmes de jugement sur échantillon.

II. 3. — LA DISTRIBUTION DE POISSON.

1. — Description.

La distribution dite de Poisson peut être considérée comme une forte limite — pour n grand et p petit — de la distribution binomiale.

Elle est entièrement définie par le produit np qui n'est autre que sa moyenne m.

Dans la table N° 5 figurent les proportions théoriques (ou probabilités) p correspondant, pour diverses valeurs de $m = np$ (limites comprises), d'une part aux intervalles (o, k) d'autre part, aux valeurs de k (1).

Bien entendu, les remarques relatives aux conditions de prélèvement de l'échantillon faites au sujet de la distribution binomiale s'appliquent sans changement à la distribution de Poisson.

2. — Importance pratique de la distribution de Poisson.

L'intérêt pratique de la distribution de Poisson est peut-être supérieur à celui de la distribution binomiale. Dans le domaine industriel, les proportions de déchets le plus fréquemment rencontrées étant en général faibles, on est en effet amené à s'y référer dès que les prélèvements effectués sont relativement importants et que l'on s'intéresse au nombre de pièces défectueuses observées dans des échantillons d'effectif constant n.

Rappelons que la proportion p peut être considérée comme « faible » quand elle est inférieure à 5 %, et que la taille n de l'échantillon peut être considérée comme forte dès qu'elle atteint la centaine.

(1) En désignant toujours par k le nombre de boules noires (ou de rebuts) observé dans l'échantillon prélevé, d'effectif constant n.

II. 4. — LA NOTION DE PROBABILITÉ.

En vue de présenter sous une forme aussi concrète que possible les schémas théoriques de fluctuations que constituent les distributions normale, binomiale et de Poisson, nous avons, dans chaque cas, fait correspondre aux différentes valeurs ou aux différents intervalles de valeurs possibles de la variable examinée ce que nous avons appelé des **fréquences théoriques**.

Cette présentation est en tout point identique à la présentation (définie au chapitre I), des distributions observées en pratique, qui consistait à faire correspondre à chaque valeur ou intervalle de valeurs de la variable prise en considération la fréquence des observations égales à cette valeur ou tombant dans cet intervalle. On a toutefois noté, en passant, qu'on désignait généralement les fréquences théoriques sous le terme de **probabilités** ; or, dans la mesure même où le terme de probabilité est devenu d'usage courant, des précisions supplémentaires paraissent s'imposer.

Pour la grande majorité des lecteurs, le mot « probabilité » recouvre une notion intuitive : quand on n'est pas absolument certain qu'un événement doive se produire, mais quand on sent que cet événement a cependant « plus ou moins de chances » de se produire, on dit couramment qu'il est « plus ou moins probable ». On entend fréquemment, dans la vie courante, des phrases telles que :

« Il est probable qu'il pleuvra d'ici ce soir »,

« Il est peu probable que je gagne le gros lot à la Loterie nationale ».

Nous allons voir comment la notion mathématique de probabilité permet de préciser de telles intuitions.

a) Probabilité a priori.

Les premières définitions de la probabilité ont été suscitées par l'étude d'exemples très simples « d'événements incertains ».

C'est ainsi que le jet d'un dé à six faces numérotées de 1 à 6 conduit à un résultat a priori entaché d'incertitude. Le résultat peut, en effet, si le dé et la façon de le jeter ne sont l'objet d'aucun truquage, être indifféremment le 1 ou le 2 ou le 3 ou le 4 ou le 5 ou le 6. Il y a donc a priori 6 éventualités possibles, et une fois que le dé a été jeté, une seule de ces éventualités se trouve être en fait réalisée.

De là à définir la probabilité d'un résultat donné — disons, par exemple, l'arrivée du 5 — comme le rapport de l'éventualité considérée au nombre total des éventualités possibles, il y a un pas qu'il est naturel de franchir. Cela conduit à dire « **qu'il y a une chance sur six** » d'observer le 5 après jet du dé ou, ce qui revient au même « **que la probabilité d'obtenir le 5 est égale à 1/6** ».

Cette façon de raisonner, appliquée au tirage (au hasard et non exhaustif) d'un échantillon de n boules dans une urne contenant des boules noires et blanches, les premières en proportion p, conduit sans grande difficulté à la distribution binomiale.

Pour définir la probabilité des échantillons contenant k boules noires, on calcule le nombre total des séries de n boules qu'il est possible d'extraire de l'urne et, parmi ces séries, le nombre de celles qui présentent k boules noires : le rapport du second de ces nombres au premier fournit la probabilité cherchée.

Bien entendu, les modalités mêmes de ce type de calcul ne permettent de préciser numériquement la probabilité d'un événement que dans des cas très particuliers. Il est, par exemple, impossible de calculer a priori, à sa naissance, la probabilité qu'a un homme de vivre au moins jusqu'à 50 ans. On peut toutefois généraliser la notion de probabilité en adoptant le point de vue qu'on désigne sous le terme de « fréquentiste » et dont la légitimité découle de la très importante « loi des grands nombres ».

b) Loi des grands nombres.

La loi des grands nombres est une loi en vertu de laquelle la « fréquence d'un résultat » devient aussi voisine que l'on veut — et cela avec une probabilité aussi forte qu'on le désire — de la « probabilité a priori de ce résultat » lorsqu'on fait croître indéfiniment le nombre des observations effectuées.

Revenant à l'exemple du dé, la loi des grands nombres signifie qu'en faisant croître indéfiniment le nombre des jets, on est en mesure d'observer — et cela avec une probabilité aussi forte qu'on le souhaite — une fréquence d'arrivée du 5 aussi voisine de 1/6 qu'on le veut.

De façon analogue, elle signifie qu'en répétant un nombre croissant de fois le prélèvement d'un échantillon de n boules dans une urne contenant des boules noires en proportion p , on observera une proportion d'échantillons contenant k boules noires aussi voisine qu'on le voudra de la probabilité a priori correspondante et cela, avec une probabilité aussi forte qu'on le désirera.

De par la loi des grands nombres, on est donc en droit (1), dans tous les cas où l'on ne peut calculer a priori une probabilité, mais où il est possible de faire un nombre croissant d'observations, de définir la probabilité cherchée comme la limite (si elle existe) du système de fréquences observé.

C'est ainsi, par exemple, que l'on pourra parler de la probabilité de décès des individus de 20 ans entre 20 et 21 ans ou de leur probabilité de survie jusqu'à 50 ans.

(1) Sous diverses réserves d'ordre logique qu'il paraît inutile d'aborder dans le cadre de cet exposé.

CHAPITRE III

LE RAISONNEMENT STATISTIQUE

Au cours des sections qui vont suivre seront examinés les principales techniques statistiques couramment appliquables et appliquées dans l'industrie.

Ces techniques mettent en jeu un mode de raisonnement qui est toujours le même et dont on va s'efforcer de préciser les grandes lignes au cours de ce chapitre de transition.

III. I. — NATURE DU RAISONNEMENT STATISTIQUE.

En vue de décomposer en ses principaux éléments ce qu'on peut appeler le processus de raisonnement statistique, il est commode d'en examiner de près des exemples typiques.

Nous emprunterons le premier à la pratique la plus courante, soit le « jugement sur échantillon » sans d'ailleurs préjuger des développements dont cette technique fera, ici même, l'objet.

Le problème est extrêmement simple : en présence d'un lot de marchandises assujetti à diverses spécifications, on se propose de décider, sur échantillon, si le lot répond ou non aux conditions de recette imposées. De façon plus précise, en présence par exemple d'un lot de pièces mécaniques qu'on distingue en « convenables » ou « défectueuses », on se propose de juger sur échantillon si la proportion de déchets présentée par le lot est inférieure ou supérieure à un certain seuil choisi comme limite d'acceptation.

La solution statistique consiste :

1) à prélever un échantillon dans le lot ;

2) à déterminer la proportion de déchets **présentée** par cet échantillon ;

3) à rapprocher le résultat obtenu des résultats auxquels conduit la théorie de l'échantillonnage dans l'hypothèse de prélèvements effectués dans des lots contenant des proportions variables de déchets ;

4) à prendre, compte tenu des indications fournies par ce rapprochement, la décision d'acceptation ou de refus du lot examiné.

Nous emprunterons, d'autre part, le second exemple à la très importante technique de contrôle en cours de fabrication — dite de Carte de Contrôle — qui doit également faire l'objet d'un chapitre ultérieur.

En présence d'un processus quelconque de fabrication, cette technique consiste :

1) à observer pendant un laps de temps donné les résultats de la fabrication ;

2) à mettre en ordre les données ainsi recueillies ;

3) à raccorder la ou les distributions obtenues à une ou plusieurs distributions théoriques bien définies ;

4) à déterminer à partir des propriétés de cette ou de ces distributions théoriques — par la théorie dite de l'échantillonnage — les propriétés des échantillons susceptibles d'être prélevés en cours d'usinage **tant que la fabrication restera stable** puis à observer ensuite si les échantillons prélevés périodiquement possèdent ces propriétés ou dans quelle mesure ils s'en écartent.

La technique envisagée permet, dès qu'un échantillon présente des propriétés inattendues, de suspecter un déréglage.

Dans les deux cas, le mode de raisonnement utilisé, toujours le même, se décompose en quatre phases :

1) Observations ;

2) Mise en ordre des données (ou description) ;

3) Analyse des données ;

4) Prédiction.

C'est là très exactement le type général de décomposition des raisonnements statistiques retenu par l'éminent statisticien britannique R. A. FISHER.

Dans un très remarquable petit ouvrage publié par la Direction des Etudes et Recherches de l'Electricité de France, M. MORLAT adopte ce schéma en lui apportant néanmoins une modification. Il introduit, au stade de l'analyse, les connaissances « a priori » dont on dispose, ce qui le conduit à présenter comme suit la décomposition de R. A. FISHER :

$$\begin{array}{l} \text{Observations} \longrightarrow \text{Description} \searrow \\ \qquad\qquad\qquad\qquad\qquad\qquad\qquad \text{Analyse} \longrightarrow \text{Prédiction.} \\ \text{Connaissances a priori} \nearrow \end{array}$$

Cette modification va beaucoup plus loin qu'on ne l'imagine de prime abord. Dès qu'on fait intervenir, en statistique, des « connaissances a priori », on s'aventure sur un terrain que d'aucuns estiment très dangereux. Les connaissances en question peuvent en effet n'être que des « idées fausses ». Le sujet vaut donc qu'on s'y arrête.

III. 2. — REMARQUES SUR L'INTRODUCTION DANS LES RAISONNEMENTS STATISTIQUES DES « CONNAISSANCES A PRIORI ».

Il existe un petit problème de Calcul des Probabilités dont l'intérêt philosophique est considérable.

On considère deux urnes A et B contenant chacune des perles blanches et noires, la proportion de perles noires étant de 10 % dans l'urne A et de 20 % dans l'urne B. Un opérateur ayant prélevé 10 perles dans l'une des deux urnes — on ignore laquelle — et deux perles noires ayant été observées parmi les 10 perles prélevées, on se demande quelle est la probabilité d'avoir affaire à un échantillon extrait de l'urne A et quelle est la probabilité d'avoir affaire à un échantillon extrait de l'urne B.

Le processus est aisément résoluble **si l'on connaît la règle déterminant le choix, par l'opérateur, de l'urne dans laquelle il opère.**

Si l'on sait, par exemple, que l'opérateur fixe son choix par tirage à pile ou face (pile = Urne A ; face = Urne B ou inversement), on est en mesure, au cours des calculs, d'introduire le fait que la probabilité de tirage dans l'urne A — ou dans l'urne B — est égale à 1/2 et cette « connaissance a priori », ajoutée à la connaissance de la composition des deux urnes, permet de conclure que la probabilité d'avoir affaire à un échantillon extrait de A est de 0,40, la probabilité d'avoir affaire à un échantillon extrait de B étant elle-même de 0,60.

Si, au contraire, on ne sait rien de la règle utilisée par l'opérateur pour déterminer son choix, **on peut être amené** à dire : « Comme dans mon ignorance, rien ne me permet de suspecter une préférence particulière de l'opérateur pour une des deux urnes, je vais attribuer la même probabilité (1/2) au choix de l'une ou de l'autre ».

Dans les deux cas, le calcul est identique, mais il est inutile de réfléchir longuement pour s'apercevoir qu'il existe une différence fondamentale, entre les deux modes de raisonnement.

Dans le premier cas, les données « a priori » introduites au cours des calculs sont **objectives** ; dans le second, elles sont parfaitement **subjectives.** Si l'opérateur, **à notre insu**, a joué aux dés l'urne dans laquelle il a prélevé l'échantillon en convenant de choisir A s'il obtenait les as et de choisir B dans tous les autres cas, les résultats obtenus par introduction de l'hypothèse d'égale probabilité de choix de A et de B n'ont pas de sens.

Cet exemple simple explique la méfiance instinctive des statisticiens pour les « connaissances a priori ». Sans doute reconnaissent-ils **qu'il serait absurde de s'en priver quand elles sont véritablement objectives**, mais ils craignent toujours de voir le praticien utiliser des données seulement prétendues telles sans l'être le moins du monde.

A cet égard, bien des solutions apportées au problème du jugement sur échantillon font intervenir des hypothèses a priori — concernant la distribution possible des proportions de déchets susceptibles d'être présentées par des lots de même nature et de même provenance que le lot soumis à examen — qui ne reposent, en définitive, sur rien d'autre que la seule imagination de leurs auteurs.

III. 3. — PRINCIPE DES TESTS STATISTIQUES.

La démarche statistique met, à tout moment, en jeu des opérations désignées sous le nom de TESTS dont nous allons définir les principes toujours les mêmes, sur des exemples.

Revenons aux différentes phases de la **technique de Carte de Contrôle** sommairement décrites au début de ce chapitre.

La **première opération** à effectuer est, avons-nous dit, d'observer — pendant un laps de temps donné — les résultats de la fabrication du processus considéré. S'il s'agit d'un tour automatique usinant de petits axes métalliques dont le diamètre seul importe, cette opération consiste à prélever un certain nombre de pièces à la sortie de la machine, puis à mesurer leurs cotes.

La **seconde opération** consiste à mettre en ordre, c'est-à-dire à mettre sous forme de distribution statistique, les résultats observés. Elle comprend, bien entendu, le calcul des principales caractéristiques de la distribution. Toutes les indications concernant ces problèmes ont été fournies au chapitre I.

La **troisième opération** — ou analyse des données — est, dans ce cas, une opération de raccordement. Ils 'agit de trouver dans le catalogue des distributions théoriques celle qui semble se raccorder le mieux à la distribution observée.

Arrêtons-nous alors un instant sur cette opération. Même dans l'hypothèse ou nous estimons avoir trouvé une distribution théorique compatible avec la distribution observée (fig. III. I.), il va de soi qu'il existe certains écarts entre les deux distributions. Ces écarts n'ont rien d'étonnant. Nous savons, en effet, (chapitre II) que la distribution théorique correspondant à la distribution observée n'est qu'une distribution limite correspondant à la population infinie de résultats de fabrication que donnerait le processus examiné s'il était capable de fonctionner indéfiniment dans les conditions existant au moment où les observations ont été faites. Pratiquement donc, la distribution observée n'est que le résultat d'un échantillonnage effectué dans une population hypothétique infinie et il est normal d'enregistrer des fluctuations d'échantillonnage.

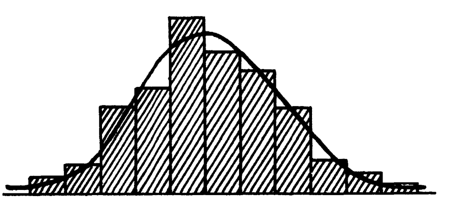

Fig. III. I.

Cela posé, la distribution théorique retenue pour raccordement ne peut être considérée comme correctement choisie que dans la mesure où les fluctuations observées entre elle et la distribution obtenue en pratique **se révèlent compatibles avec les seuls aléas de l'échantillonnage.** Il est donc nécessaire de le vérifier.

Les principales phases de la vérification sont les suivantes :

I) on fait l'hypothèse de la légitimité du raccordement envisagé ;

2) on détermine la « distribution théorique » d'un échantillon de même taille que l'échantillon observé, ce qui conduit à faire correspondre un nombre théorique n_t à tout nombre n d'observations relatif à la valeur (ou à la classe) x de la variable (tableau III. I.) ;

TABLEAU III. 1.

Valeur ou classe de la variable x	Nombre d'observations recueillies n	Nombre théorique d'observations correspondant n_t	$\frac{(n - n_t)^2}{n_t}$
x_1	n_1	n_{t1}	
x_2	n_2	n_{t2}	
x_k	n_k	n_{tk}	
	N	N	$\Sigma \frac{(n - n_t)^2}{n_t}$

3) on choisit une variable z fonction des fluctuations d'échantillonnage (ici : $z = \frac{(n - n_t)^2}{n_t}$) et dans l'hypothèse de la conformité, on détermine la distribution théorique de z (fig. III. 2.) ;

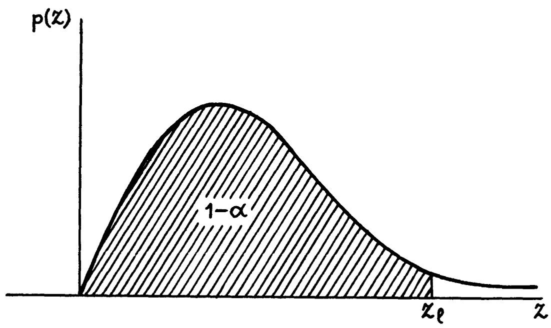

Fig. III. 2.

4) examinant cette distribution et se fixant un seuil de probabilité α, petit, **qu'on décide de négliger**, on détermine la valeur limite z_e, telle que la probabilité, correspondant à l'ensemble des valeurs de z inférieures à z_e, soit 1 — α.

5) cela fait, on raisonne comme suit : « Dans l'hypothèse où le raccordement opéré est justifié et, par conséquent, où les fluctuations mises en évidence entre distribution théorique et distribution observée sont uniquement provoquées par les aléas de l'échantillonnage, on doit — au seuil de probabilité α près — obtenir un z inférieur à z_e. Si donc, dans le cas considéré, le z calculé est supérieur à z_e, il est préférable de conclure que le raccordement en question est mauvais, la probabilité d'une telle valeur de z, dans l'hypothèse de la conformité, étant par trop étroite ».

La **quatrième et dernière phase** de la Carte de Contrôle consiste enfin à déterminer, à partir de la distribution théorique correspondant à la distribution observée — une fois le raccordement effectué — les propriétés des échantillons susceptibles d'être prélevés en cours d'usinage tant que la fabrication restera stable.

Arrêtons-nous également, un instant, sur cette opération. Elle consiste :

1) à faire l'hypothèse de la stabilité future de la fabrication ;

2) à déterminer, dans cette hypothèse, les propriétés des échantillons susceptibles d'être prélevés dans l'avenir : par exemple, la distribution possible de leur valeur moyenne m' ;

3) à définir, à partir de cette distribution (fig. III. 3.), pour un seuil de probabilité α petit et considéré comme négligeable, un intervalle (L_1, L_2) tel que la probabilité d'avoir m' à l'intérieur de (L_1, L_2) soit égale à $1 - \alpha$;

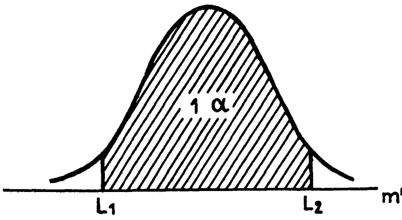

Fig. III. 3.

4) cela fait, à raisonner comme suit : « Dans l'hypothèse de la stabilité de la fabrication, les moyennes des échantillons observés doivent — au seuil de probabilité α près — tomber entre L_1, L_2. Si donc, les moyennes m' des échantillons tombent entre L_1 et L_2, il n'y a aucune raison de suspecter la stabilité de la fabrication. En revanche, si la moyenne d'un échantillon tombe à l'extérieur de (L_1, L_2), il est préférable de conclure au déréglage, la probabilité d'un tel résultat, dans l'hypothèse de la stabilité, étant par trop faible. »

Dans les deux cas, la démarche intellectuelle est toujours la même :

1) on fait une hypothèse ;

2) dans le cadre de cette hypothèse, on définit la distribution théorique d'une caractéristique Z des échantillons susceptibles d'être observés ;

3) on détermine ensuite, par examen de la distribution théorique obtenue, l'intervalle dans lequel doit tomber toute valeur observée de Z si l'on convient de négliger un certain seuil de probabilité α ;

4) ayant alors prélevé un échantillon et calculé la valeur de sa caractéristique Z, de deux choses l'une : ou bien cette valeur tombe dans l'intervalle retenu et dans ce cas on peut dire que le résultat de l'échantillon ne s'oppose pas à l'hypothèse mise en jeu, ou bien cette valeur tombe à l'extérieur de l'intervalle retenu et dans ce cas, il semble préférable de penser que le résultat de l'échantillon infirme l'hypothèse faite.

Ce type de raisonnement, absolument fondamental, définit ce qu'on appelle le TEST STATISTIQUE.

Au cours des chapitres suivants, on en multipliera les exemples.

Observations :

1) On ne saurait trop insister sur le caractère en quelque sorte négatif du test statistique.

Si la valeur de la caractéristique de l'échantillon observé tombe dans l'intervalle voulu, cela ne prouve pas que l'hypothèse faite est juste. Cela prouve seulement que les données recueillies **ne sont pas en contradiction** avec cette hypothèse.

Cette remarque introduit très naturellement, nous le verrons ultérieurement, la notion **d'efficacité des tests statistiques.**

2) Le choix de niveau de probabilité à négliger n'est pas statistique, mais technique. Il est fonction des conditions d'efficacité dans lesquelles on entend se placer. Nous reviendrons également par la suite sur cette question.

Pour le moment, il nous suffit de remarquer que la démarche qui consiste à négliger une petite probabilité en vue de prendre une décision est dans la vie courante le propre de tous les hommes.

Un promeneur décide de sortir sans son parapluie ou son imperméable parce qu'il néglige l'éventualité d'une averse ;

Un chasseur appuie sur la gachette de son fusil parce qu'il estime pouvoir négliger l'éventualité d'un accident par ricochets ou retombée des plombs ;

Un voyageur prend le train sans appréhension parce qu'il néglige la probabilité, faible, d'un déraillement, etc..., etc...