Avant-propos.

L'importance des statistiques n'est plus à démontrer dans les domaines les plus divers. Dans les sciences dites dures, les outils statistiques sont largement répandus dans les applications techniques mais sont plus diversement présents dans l'enseignement classique. Depuis de nombreuses années, la référence à ces outils est présente dans les programmes, les articles dans les revues liées à l'enseignement se multiplient. Outre les difficultés liées à la maîtrise par les acteurs (les statistiques sont peu enseignées dans les cursus classiques), la pratique avec les étudiants exige souvent du temps, à la fois pour une collecte appréciable de données et une discussion complète des hypothèses et conclusions. Cet ouvrage expose les outils statistiques mais aussi les concepts et méthodes à la base des traitements les plus répandus. Dans certains chapitres exposant les techniques élémentaires, nous avons essayé de suivre une progression pour que ceux qui découvrent en partie ce domaine puissent le faire à partir de considérations simples. Mais ce choix est aussi d'ordre didactique. Pensant en particulier aux enseignants, nous avons voulu proposer des raisonnements progressifs utilisables dans un enseignement de ces outils. N'ayant pas voulu restreindre cet ouvrage à un niveau donné, il en résulte une certaine hétérogénéité dans la difficulté des thèmes abordés. Certains passages seront certainement perçus comme trop élémentaires, d'autres au contraire ne seront utilisables que dans un enseignement plus spécialisé.

L'exposé des techniques a été complété par des exemples d'application. Il existe de nombreux logiciels spécialisés pour ce type de traitement mais, outre l'aspect financier, ils peuvent dérouter par la multitude de tests présentés. C'est pourquoi nous avons associé à l'ouvrage des classeurs Excel® illustrant les outils décrits dans le texte. Ces classeurs sont de deux types. Certains sont des outils de démonstration illustrant certains aspects conceptuels importants. L'utilisateur a peu d'initiatives, ce sont les résultats qui sont intéressants à commenter et à relier aux concepts concernés. Certains classeurs proposent en particulier des simulations permettant de retrouver les valeurs données dans les tables statistiques classiques mais aussi d'en déduire des résultats plus détaillés. Les autres sont des outils qui permettent de traiter des cas concrets présentés sur le plan théorique. Ils peuvent en particulier être utilisés par des étudiants pour avoir une réponse quantitative rapide et disposer de plus de temps pour répondre aux questions qu'ils peuvent se poser sur les résultats expérimentaux qu'ils ont collectés. Pour plus d'efficacité, tous les calculs sont gérés par des macros VBA. Dans tous les cas, ces classeurs peuvent évidemment être améliorés et adaptés pour en faire des outils de travail dédiés à des utilisations personnalisées. Ecrits avec Excel® 97, ils sont utilisables avec les versions ultérieures.

Pour des raisons de sécurité, ces classeurs ne sont pas joints à l'ouvrage mais disponibles sur Internet dans le site :

http : //www.cetice.u-psud.fr / complements/stats/r-journeaux.html

L'ouvrage comporte 12 chapitres.

Le chapitre I donne quelques rappels de probabilités et présente les lois statistiques qui seront utilisées plus loin.

Le chapitre II recense les principaux concepts et problème posés par la mesure d'une grandeur, en particulier en Sciences dures. Le chapitre III présente le traitement d'une série de mesure du point de vue des incertitudes des types A et B. Le chapitre IV traite le problème de la propagation des incertitudes.

Les chapitres V et VI présentent les techniques de régression. La régression linéaire est présentée avec les diverses variantes liées à l'importance des incertitudes. Les autres régressions sont ensuite exposées, avec en particulier le cas très général de la régression à solutions non linéaires qui élargit le domaine d'utilisation de cette technique. La technique classique de la variance est présentée et confrontée à une méthode plus récente et reconnue "exacte". Cette dernière technique, encore peu répandue, a fait l'objet de nombreuses publications dans les revues anglo-saxonnes et mérite d'être plus connue.

Le chapitre VII met en place les concepts et méthodes de conduite des tests statistiques qui seront détaillés ensuite.

Les chapitres VIII à X recensent les principaux tests en fonction des questions posées. Le chapitre VIII traite de la corrélation. Le chapitre IX présente les tests de comparaison appliqués aux variances et moyennes dans plusieurs cas de figure. Les techniques d'analyse de la variance sont présentées dans les cas simple et double. Enfin le chapitre X donne quelques exemples de tests d'ajustement à une loi de probabilité ou à un modèle mathématique, ce qui permet de compléter les résultats des régressions. La technique du chi2 impliquée dans ces tests est enfin présentée dans le cas plus général du tableau de contingence. Cet élargissement à des domaines moins spécifiquement " sciences dures " est également présent dans les chapitres VIII et IX avec des tests portant sur les rangs utilisables en particulier en sciences humaines, donc en didactique.

Les deux derniers chapitres reviennent aux statistiques descriptives mais avec une utilisation plus orientée vers les sciences humaines. Ils sont dédiés à tous les enseignants qui sont amenés à analyser des résultats de tests, à comparer des groupes d'étudiants, à évaluer des évolutions pour un groupe etc… Ils sont en particulier utilisables par les chercheurs en didactique.

Ce travail est le résultat de nombreuses années de discussion et d'essais avec des étudiants à plusieurs niveaux, avec la collaboration de nombreux collègues de lycée ou d'université. Qu'ils soient ici tous remerciés pour la richesse et la variété de leurs apports tant sur le plan scientifique qu'humain.

Chapitre I

Probabilités et Statistiques

Les statistiques sont constituées par l'ensemble des moyens permettant d'analyser et de donner du sens à un ensemble de données chiffrées. La conclusion de ces études va souvent être constituée par un nombre restreint de résultats quantitatifs et/ou qualitatifs. Ceci vient du fait que les données à analyser ont des caractéristiques particulières :

- elles sont très nombreuses ; leur lecture ne permet donc pas d'en dégager rapidement des conclusions, des tendances.
- chacune d'entre elles est le fruit d'une variabilité qui interdit de prévoir exactement le résultat trouvé. Cette propriété est liée à la notion de hasard comme il sera précisé plus loin.
- chacune d'entre elles n'a de sens que par rapport à l'ensemble.

Par exemple, si on détermine la taille des garçons de *12* ans à l'échelle d'un pays, l'ensemble des résultats sera avantageusement remplacé par quelques nombres permettant de qualifier cette population : taille moyenne, tailles extrêmes, forme de la répartition. La taille d'un individu n'est donc pas une grandeur pertinente pour résumer les caractères principaux de la population, elle ne prend son sens que par sa participation à l'ensemble.
Les traitements statistiques vont porter sur les valeurs prises par des variables attribuées à un caractère. Dans l'exemple précédent, le caractère taille est caractérisé par la variable longueur mesurée par un appareil adapté au problème posé. La variable est continue si toutes les valeurs sont possibles, au moins dans un certain domaine. Elle est discontinue si seules certaines valeurs sont possibles. Il est important de noter que le caractère continu d'une variable n'a de sens qu'au plan théorique. Au plan pratique, le résultat de la mesure est toujours un résultat plus ou moins tronqué. Par exemple, au niveau macroscopique, l'intensité d'un courant est considérée comme une variable continue (cela ne le serait plus au plan microscopique : on voit que le niveau d'observation est lui aussi important). Si on mesure cette intensité avec un appareil numérique, la nature même de l'instrument va donner à la grandeur un caractère discontinu évident. Si on effectue la mesure avec un appareil à aiguille, la continuité du résultat est illusoire : les limites des capacités d'interpolation sont telles que les résultats possibles de la mesure sont sensiblement les mêmes qu'avec un appareil numérique de performance moyenne. C'est la possibilité, au moins théorique, d'augmenter la résolution des ampèremètres qui confère à la variable intensité un caractère continu au niveau macroscopique.

Un autre caractère important des variables traitées par les statistiques est leur lien avec la notion de hasard. La taille d'un individu est le résultat d'un nombre important de facteurs. Pour un observateur qui est incapable d'inventorier ces causes, la grandeur taille est donc liée en partie au hasard. On dira également que la taille d'un individu possède un caractère aléatoire, c'est à dire que sa valeur est tributaire d'un grand nombre de facteurs ayant agis en quantité et en qualité d'une façon qui n'est pas prévisible pour l'observateur. Ceci n'est pas contradictoire avec le déterminisme qui régit les phénomènes physiques. Le résultat d'une mesure en physique possède une part d'aléatoire parce qu'il est impossible de contrôler tous les facteurs qui conduisent à ce résultat, en particulier les facteurs humains si l'expérimentateur joue un rôle dans l'opération. En sciences humaines, les mesures sont aussi le résultat d'un grand nombre de facteurs agissant de façon inconnue comme si le hasard était en jeu. Cette caractéristique explique pourquoi la théorie des probabilités va être à la base des traitements statistiques.

1. Population et échantillon

Les valeurs prises par une variable peuvent être nombreuses, ce nombre peut même devenir infini. L'ensemble de toutes ces valeurs possibles constitue une population. Pour des raisons de temps, d'économie, il est souvent impossible d'atteindre la population entière. Dans ces conditions, on s'intéressera à un nombre limité d'observations : on effectuera ce qu'on appelle un *échantillonnage*. L'étude de l'échantillon ainsi constitué permettra alors de faire des hypothèses sur la population dont il est issu. Mais ces conclusions devront être regardées avec une certaine prudence : les calculs statistiques ne donnent pas des résultats certains mais seulement probables ; ils ne donnent pas des valeurs exactes mais des plages ou des intervalles de probabilité importante ; ils ne fournissent pas des conclusions sûres mais seulement plausibles. Toute décision résultant d'une étude statistique est donc susceptible d'être mauvaise, c'est à dire que l'expérimentateur prend un risque souvent chiffrable de se tromper mais est prêt à le prendre à cause d'un enjeu qui le justifie. Par exemple, l'efficacité d'un médicament n'est jamais parfaitement établie. Malgré le risque de l'utiliser alors qu'il est peut être sans influence, il sera pourtant retenu s'il est le seul susceptible de faire progresser la lutte contre une maladie.

Ces techniques peuvent être utilisées pour le traitement des résultats de mesure en Physique et dans toute science expérimentale. En effet, le résultat d'une mesure peut être considéré comme l'altération de la valeur exacte (on reviendra plus loin sur cette notion délicate) par tout un faisceau d'erreurs diverses faisant intervenir l'appareil de mesure, l'observateur, la méthode, le phénomène lui-même etc... Le caractère aléatoire est parfois évident. Par exemple, la recherche de la position d'une image en optique nécessite une appréciation de sa netteté qui varie pour un même observateur. Si on effectue 10 pointés, on obtient des résultats non identiques qui peuvent être considérés comme un échantillon extrait d'une population beaucoup plus grande, voire même infinie, qu'on peut approcher en répétant l'expérience un grand nombre de fois. En revanche, si on mesure l'intensité dans un circuit, le caractère aléatoire du résultat est moins évident. Il apparaît si on remarque que ce résultat a été obtenu avec cet appareil parmi tous les ampèremètres possibles. Si on fait la mesure avec dix appareils différents, on obtient en général des résultats différents qui sont un échantillon de ce qu'on aurait pu obtenir avec tous les

appareils disponibles. Le résultat unique est donc bien le résultat du hasard lié au caractère arbitraire de l'appareil choisi pour faire la mesure.

2. Rappels de probabilité

2.1. Généralités.

Un événement est l'occurrence d'un fait résultant de la mise en place d'une épreuve par l'expérimentateur. Au jeu de dé, l'événement "le six est sorti" est un constat qui résulte de l'épreuve lancer du dé. Le résultat de l'événement peut souvent se traduire par la valeur particulière que prend une variable pour chaque réalisation de l'événement. Par exemple, la mesure d'une longueur conduit à un nombre x_i qui est la valeur prise par la variable "longueur de l'objet" pour cette opération. On dit encore que x_i est une réalisation particulière de l'opération de mesurage pour l'épreuve en question.
Le résultat de l'épreuve peut être non prévisible par une loi déterministe, soit que le phénomène soit effectivement non déterministe (émission radioactive), soit que la complexité des phénomènes et la méconnaissance de tous les paramètres et conditions rendent l'étude irréaliste. Tout se passe comme si le hasard régissait le résultat et on dit alors que la variable atteinte est aléatoire. La variable aléatoire *X* conduit à des résultats x_i pour chaque réalisation de l'épreuve.

La probabilité est, au sens moderne du terme, une théorie mathématique basée sur des axiomes décrivant des concepts fondamentaux. Ces concepts et axiomes ne sont pas évidemment arbitraires mais sont choisis de façon qu'ils conduisent à un outil utilisable pour répondre à des problèmes pratiques. Nous renvoyons à des ouvrages spécialisés pour cette approche.
Il est souvent plus facile, même si cette approche pose des problèmes, d'aborder les probabilités par une voie plus concrète et accessible avec un niveau mathématique restreint. C'est ainsi que la probabilité d'un événement peut se définir par le nombre de fois où cet événement s'est produit rapporté au nombre total de cas également possibles (on notera le caractère ambigu de cette définition car il s'agit de savoir comment on définit deux cas également possibles!). Par exemple, avec un dé non pipé (il n'y a pas de raisons qu'une face soit privilégiée), obtenir le *1* (*x=1*) est une possibilité parmi *6*, donc la probabilité de l'obtenir est *1/6*. Le nombre aléatoire *x* est donc le résultat de la mesure et possède une probabilité d'occurrence fonction de *x* qui peut être connue expérimentalement ou de façon théorique selon les cas. On note ce résultat sous la forme détaillée $P(X=x)=p$: la probabilité que *X* prenne la valeur *x* est *p*.
Si la variable est continue, il faut revoir cette notion de probabilité. Pour cela, on raisonne sur un intervalle compris entre *x* et $x+dx$ et on appelle $dP(x)=P(x<X<x+dx)$ la probabilité que le résultat de la mesure soit situé dans cet intervalle. Si la quantité $\frac{dP(x)}{dx}$ tend vers une valeur limite quand $dx \to 0$, cette limite $p(x)$ est appelée densité de probabilité.
Soit la variable aléatoire *X* et p_k la probabilité de trouver la valeur x_k. Si on fait la somme des termes p_k sur tous les événements possibles, on trouve la probabilité de

trouver un événement parmi tous, et cette probabilité vaut *1* puisqu'on est sûr de trouver quelque chose. D'où $\sum_{k=1}^{k=n} p_k = 1$, *n* étant le nombre de valeurs possibles de la variable *X*.

Pour une variable continue, le résultat est transposable et on a : $\int p(x)dx = 1$ (I. 1) **,** l'intégrale étant étendue à tout le domaine où *p(x)* n'est pas nulle.

On introduit également la fonction de distribution $F(x)$ définie par la relation :

$$F(x) = \int_{-\infty}^{x} p(x)dx \quad (I.\ 2)$$

C'est donc la probabilité de trouver la valeur de la variable inférieure à *x*. Si $x \to \infty$, l'intégrale vaut 1, donc la fonction de distribution tend asymptotiquement vers 1.

2.2. Le cas de couples de variables.

Soient deux variables aléatoires *X* et *Y* correspondant aux deux événements A et B et conduisant aux réalisations x_i *et* y_i.

On peut considérer que le couple (*XY*) traduit le fait suivant : on a trouvé *x* ***et*** *y* , ce qui correspond à l'intersection des deux événements A et B. On peut définir sur ce nouvel ensemble une probabilité $p_{X,Y} = p(X = x \text{ et } Y = y) = p(XY)$. Si les deux variables sont indépendantes, cela signifie que la probabilité de trouver une valeur x_i n'est pas influencée par la valeur trouvée y_i. On définit cette indépendance par la relation :

$$p_{X,Y} = p(X = x)p(Y = y) = p_X\, p_Y \quad (I.\ 3)$$

Chacune des lois de probabilité agit donc de façon indépendante.

On peut également chercher la probabilité de trouver soit x *soit* y *donc* p(X=x ou Y=y), *soit encore* p(X+Y) *traduisant la réunion de deux événements. Si les événements sont exclusifs ou incompatibles (obtenir "1"* et *"2" à un jet de dé), alors :*

$$p(X + Y) = p_X + p_Y \quad (I.\ 4).$$

Si les deux événements sont compatibles, alors : $p(X + Y) = p_X + p_Y - p_{X.Y}$ *(I. 5)*

2.3. Grandeurs associées aux lois de probabilité.

Soient les valeurs discrètes $x_1, x_2 \ldots\ldots x_n$ prises par la variable *X* avec les probabilités $p_1, p_2 \ldots\ldots p_n$. L'espérance mathématique est définie par : $E(X) = \sum_{i=1}^{i=n} p_i x_i$ *(I. 6).*

Pour une distribution continue caractérisée par une loi de probabilité *p(x),* l'espérance est donnée de façon analogue par : $E(X) = \int xp(x)dx$ *(I. 7),*

la somme étant étendue au domaine où *p(x)* est non nulle.

Dans le cas discret, si la probabilités de chaque événement est la même pour tous ces événements, la probabilités p_i est indépendante de *i* et vaut $p_i = \frac{1}{n}$. L'espérance mathématique vaut alors :

$$E(X) = \sum_{i=1}^{i=n} \frac{1}{n} x_i = \frac{1}{n} \sum_{i=1}^{i=n} x_i \quad (I.\ 8).$$

On retrouve l'expression connue de la moyenne arithmétique dont l'espérance mathématique est une généralisation. Par la suite, on parlera souvent de moyenne au lieu d'espérance.
Si chaque événement est équiprobable et si l'événement *i* se produit n_i fois, on peut écrire :

$E(X) = \frac{1}{n} \sum_{1}^{n} n_i x_i = \sum_{1}^{n} f_i x_i$ avec $f_i = \frac{n_i}{n}$ fréquence d'occurrence de l'événement *i*.

L'espérance mathématique possède des propriétés intéressantes pour les traitements statistiques.
- l'espérance mathématique d'une constante est égale à cette constante : *E(a)=a*.
- si *X* est une variable aléatoire et *a* une constante, on peut écrire :

$$E(aX) = \sum p_i (ax_i) = a \sum p_i x_i = aE(X) \quad (I.\ 9)$$

La démonstration dans le cas continu se fait de la même manière.

On montre de même que :

$$E(X + a) = \sum p_i (a + x_i) = \sum a p_i + \sum p_i x_i = a \sum p_i + E(X)$$

$$E(X + a) = a + E(X) \quad (I.\ 10) \qquad car \ \sum p_i = 1.$$

-si *X* et *Y* sont deux variables aléatoires, la théorie des probabilités donne les résultats suivants :

- *E(X+Y)=E(X)+E(Y)* dans le cas général..

- *E(XY)= E(X)E(Y)* *(I. 11)* si les variables sont indépendantes.

Soit la variable aléatoire *X-E(X)* et cherchons son espérance mathématique.
E[X-E(X)]=E(X)-E[E(X)]=E(X)-E(X)=0 (car *E(X)* est une constante).
cette nouvelle variable aléatoire d'espérance nulle est dite variable centrée.

Les moments d'ordre supérieur.

On peut définir un moment d'ordre *k* par la relation :

$$E(X^k) = \int p(x) x^k dx \quad (I.\ 12) \qquad ou \qquad \sum p_i x_i^k \quad (I.\ 13)$$

pour des variables discrètes. On peut remarquer que l'espérance mathématique est le moment d'ordre *1*, et que le moment d'ordre *k* peut être considéré comme l'espérance mathématique de la $k^{ème}$ puissance de la variable.

Il est souvent intéressant de travailler sur la variable centrée *X-E(X)*. Le moment d'ordre *1* est alors toujours nul et le moment d'ordre *2*, très utilisé, est appelé la variance :

$$Var(X) = E[X - E(X)]^2 \qquad (I.\ 14).$$

L'écart-type $\sigma = \sqrt{Var(X)}$ est également très utilisé en statistiques.

Quelques propriétés de la variance.

Soit *a* une constante.

$$Var(X + a) = E[X + a - E(X + a)]^2 \text{ avec } E(X + a) = a + E(X)$$

donc : $Var(X + a) = E[X - E(X)]^2 = Var(X)$ *(I. 15)*

Une translation par une constante ne modifie donc pas la variance.

$$Var(aX) = E[aX - E(aX)]^2 = E[aX - aE(X)]^2 = a^2 E[X - E(X)]^2$$

$$Var(aX) = a^2 Var(X) \quad (I.\ 16)$$

Autre expression de la variance.

$$Var(X) = E[X - E(X)]^2 = E[X^2 + (E(X))^2 - 2XE(X)] = E[(X)^2] - [E(X)]^2$$

qu'on traduit souvent par l'expression : la variance est égale à la moyenne des carrés diminuée du carré de la moyenne.

Soient deux variables aléatoires *X* et *Y*. On sait que $E(X + Y) = E(X) + E(Y)$.

Calculons la variance de *X+Y*.

$$Var(X + Y) = E[X + Y - E(X + Y)]^2 = E[X - E(X) + (Y - E(Y)]^2$$

$$= E[X - E(X)]^2 + E[Y - E(Y)]^2 + 2E[[X - E(X)][Y - E(Y)]]$$

On reconnaît dans les deux premiers termes les variances de *X* et *Y*. Le dernier terme est le double de l'espérance mathématique du produit de deux variables aléatoires centrées. On l'appelle la covariance des deux variables $Cov(X,Y)$. Si les variables *X* et *Y* sont indépendantes, les variables centrées le sont aussi et on sait alors que l'espérance du produit est le produit des espérances, donc :

$$E[[X - E(X)][Y - E(Y)]] = E[X - E(X)]E[Y - E(Y)] = 0$$

car chaque variable est centrée, donc d'espérance nulle.

On aboutit donc au résultat :

$Var(X + Y) = Var(X) + Var(Y) + 2Cov(X,Y)$ *(I. 17)* dans le cas général.

$Var(X + Y) = Var(X) + Var(Y)$ *(I. 18)* quand les variables sont indépendantes.

La covariance se présente donc comme le degré de liaison entre les deux variables. On retrouvera plus loin l'importance de ce terme.

On peut écrire cette covariance de la façon suivante :

$$Cov(X,Y) = E[[X - E(X)][Y - E(Y)]] = E[XY + E(X)E(Y) - XE(Y) - YE(X)]$$

$$= E(XY) + E(X)E(Y) - E(X)E(Y) - E(Y)E(X)$$

$$Cov(X,Y) = E(XY) - E(X)E(Y) \quad (I.\ 19)$$

qu'on traduit souvent par l'expression : la covariance est égale à la moyenne des produits diminuée du produit des moyennes.

3. Théorèmes importants

3.1. L'inégalité de BIENAYME-TCHEBYCHEV.

Si une variable aléatoire X admet une variance $Var(X) = \sigma^2$, alors on peut démontrer que :

$$P(|X - E(X)| > a) < \frac{Var(X)}{a^2} \quad (I.\ 20)$$

On peut aussi l'exprimer en fonction de l'écart-type σ, en posant $a = b\sigma$:

$$P(|X - E(X)| > b\sigma) < \frac{1}{b^2} \quad (I.\ 21)$$

Ceci traduit le fait que la probabilité de trouver x à une certaine distance du centre (donné par l'espérance mathématique), l'unité étant σ, est d'autant plus faible que l'on s'éloigne de ce centre (b grand). Mais cela signifie aussi qu'à b donné (probabilité donnée), on se retrouve d'autant plus près du centre que σ est faible. La variance ou l'écart-type sont donc des *indices de dispersion* puisqu'ils donnent une idée de la région où se concentrent les valeurs. Par exemple, si $b=2$, on a $\frac{1}{b^2} = 0{,}25$, ce qui signifie que moins de *25%* des valeurs sont à l'extérieur de l'intervalle $E(X) \pm 2\sigma$, soit plus de *75%* des valeurs sont dans ce même intervalle. Cette inégalité, qui est valable quelle que soit la loi de probabilité pour X, est évidemment pessimiste pour certaines lois comme on le verra pour la répartition gaussienne (environ *95%* des valeurs dans le même intervalle $E(X) \pm 2\sigma$).

3.2. La loi faible des grands nombres.

Elle concerne la somme S_n de n valeurs prises par n variables aléatoire X_i indépendantes ayant même distribution caractérisée par une espérance μ et une variance σ^2. On a alors :

$$\lim_{n\to\infty} P\left(\left|\frac{S_n}{n} - \mu\right| < \varepsilon\right) = 1 \quad (I.\ 22) \quad \text{pour tout } \varepsilon \text{ positif.}$$

Ce résultat est très important. Il signifie que si on extrait au hasard n valeurs d'une population caractérisée par μ et σ^2, la moyenne des n valeurs x_i indépendantes, qui vaut $\frac{S_n}{n}$, est d'autant plus proche de l'espérance μ que le nombre d'épreuve est grand. Il justifie le fait que la moyenne pour des variables indépendantes issues de la même population peut être considérée comme un estimateur de μ, et ceci d'autant mieux que n est grand (voir le chapitre III pour l'utilisation dans le traitement des mesures).

3.3. Le théorème limite central.

C'est un théorème essentiel dont les conséquences pratiques sont très utiles dans le traitement statistique. Il concerne ici encore la somme S_n des *n* réalisations de variables aléatoires ayant même loi de probabilité, avec une espérance μ et une variance σ^2.

Si on introduit la quantité : $Z_n = \frac{S_n - n\mu}{\sigma\sqrt{n}}$ ***(I. 23),*** on obtient une nouvelle variable aléatoire dont la loi de probabilité tend vers une limite quand $n \to \infty$. Cette loi de probabilité limite est la loi normale réduite.

Si on raisonne sur la variable $\frac{S_n}{n}$ qui est la moyenne *m* des *n* valeurs, le théorème précédent devient :

la variable $\frac{m-\mu}{\frac{\sigma}{\sqrt{n}}}$ suit une loi de Gauss réduite quand $n \to \infty$.

Cela signifie que la variable *m* suit une loi normale de même moyenne que la population originelle, et dont la variance est *n* fois plus faible que la variance de la population initiale.

Ce théorème reste vrai quand les variables n'ont pas la même distribution et qu'elles ne sont pas indépendantes pourvu que certaines conditions soient remplies, en particulier que les variances soient définies. Cela signifie en pratique que si un phénomène est la résultante d'un grand nombre de facteurs aléatoires, indépendants et d'amplitudes voisines, alors une grandeur attachée à ce phénomène a de bonnes chances d'avoir une loi de probabilité voisine d'une loi normale. C'est d'ailleurs une des raisons pour laquelle on lui a donné historiquement ce nom. Un problème de vocabulaire se pose pour ce terme qui vient de l'anglais "central limit theorem". On trouve souvent une traduction mot à mot " théorème central-limite ". Si on prend le vocabulaire des probabilités, on rencontre de nombreux théorèmes valables seulement à la limite, en particulier quand une quantité tend vers l'infini, d'où la notion de théorème limite, le mot central renvoyant au fait que la loi suivie est centrée. Mais on trouve aussi le terme de théorème central de la limite, ce qui semble privilégier le fait que ce théorème a une place centrale par ses conséquences. On trouve enfin le terme de théorème de tendance normale, ce qui est sûrement plus explicite mais moins utilisé. Nous utiliserons le terme de théorème limite central (ou de la limite centrale), le plus utilisé et souvent abrégé par TLC. Nous verrons toutes les conséquences qui pourront en être tirées quand on appliquera les méthodes statistiques aux résultats de mesure (voir en particulier le chapitre II pour la théorie des erreurs de Laplace).

Un cas particulier est important : celui où la variable de base suit elle-même une loi normale. Le théorème limite central est vrai *même pour n fini*. Cela signifie que la moyenne de *n* variables issues de cette population normale (μ,σ) est elle-même normale de variance $\frac{\sigma^2}{n}$ quelle que soit la valeur de *n*.

Ce théorème peut s'illustrer par simulation, en particulier par Excel®. Nous renvoyons au paragraphe 6 pour la présentation de ce programme.

4. Lois de probabilité importantes

4.1. La loi de Laplace-Gauss ou normale.

C'est la loi qui sera le plus souvent évoquée dans cet ouvrage. On a déjà pressenti son importance avec le théorème limite central. On verra qu'elle conditionnera de nombreux tests statistiques dans les derniers chapitres.
Elle donne la loi de probabilité pour une variable continue par la densité :

$$p(x) = \frac{1}{\sigma\sqrt{2\pi}} exp(-\frac{(x-\mu)^2}{2\sigma^2}) \quad (I.\ 24)$$

Le terme $\sigma\sqrt{2\pi}$ est une constante de normalisation pour que la somme des probabilités soit égale à *1*.
Les calculs donnent deux résultats importants :
- l'espérance mathématique de la variable aléatoire suivant cette loi est μ.
- la variance de la variable est σ^2.

Si on fait le changement de variable $u = \frac{x-\mu}{\sigma}$, la loi devient :

$$p(u) = \frac{1}{\sqrt{2\pi}} exp(-\frac{u^2}{2}) \quad (I.\ 25)$$

On a alors la loi normale réduite dont la tabulation permet, par le changement inverse, de trouver les caractéristiques de la loi générale.

La loi réduite a des propriétés intéressantes qui sont résumées ici et sur la figure I.1 :
- elle possède un point d'inflexion pour *u=1* et *–1*.
- *68%, 95% et 99,7%* des valeurs possibles sont respectivement contenues dans les intervalles *(-1,1), (-2, 2) et (-3,3)*.

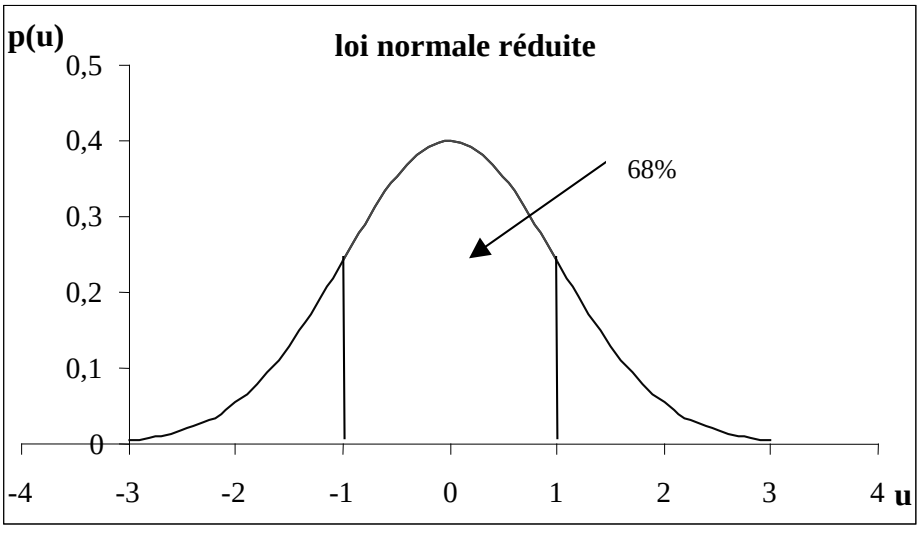

Fig.I. 1

4.2. Les lois binomiale et de Poisson.

Nous renvoyons à des ouvrages spécialisés pour plus de détails sur ces deux lois classiques[1]. Utiles dans certains domaines (par exemple en Physique nucléaire), elles seront peu évoquées ici. Seule la loi binomiale sera rencontrée dans un test.
La loi binomiale intervient quand un événement ne peut donner que deux résultats, par exemple *X* avec la probabilité *p* et *Y* avec la probabilité *1-p*. Si on réalise l'événement *n* fois et qu'on trouve *x* fois le résultat *X*, on a alors : $P(x=k)=p^k(1-p)^{n-k}C_n^k$ *(I. 26)*
On dit que *x* suit une loi binomiale. C'est une loi discontinue qui tend vers une loi normale quand *n* tend vers l'infini. Sa moyenne est *np* et sa variance *np(1-p)*. Un cas particulier, qu'on utilisera plus loin, est le cas *p=0,5*.

La loi de Poisson intervient quand une variable aléatoire peut prendre toutes les valeurs entières k *avec la probabilité :* $P(x=k)=\frac{e^{-\lambda}\lambda^k}{k!}$ *(I. 27),*

λ est le paramètre caractéristique de la loi. La loi de Poisson ainsi définie possède une moyenne et une variance λ. C'est une loi fondamentale dans les phénomènes radioactifs et les théories des files d'attentes. C'est ce qui se passe quand un événement possède une probabilité faible mais constante de se produire pendant un intervalle de temps donné. Le nombre *k* d'événements pendant un intervalle de temps plus grand suit alors une loi de Poisson.

4.3. La loi du CHI2 (χ^2).

Soient *n* valeurs x_i aléatoires indépendantes issues d'une loi normale réduite. La somme $\sum x_i^2$ suit une loi dite du CHI2 de Pearson. Elle fait également intervenir la quantité *n* appelé nombre de degrés de liberté qui prendra son vrai sens plus loin. On peut dire pour l'instant que ce nombre est égal au nombre de valeurs mises en jeu *diminué* des contraintes sur ces grandeurs.
Si la variable est extraite d'une loi normale (μ,σ), le changement de variable ci-dessus $u=\frac{x-\mu}{\sigma}$ permet de dire que la somme : $\chi^2=\sum u_i^2=\frac{\sum(x_i-\mu)^2}{\sigma^2}$ *(I. 28)* suit une loi du CHI2.
Cette loi n'est pas symétrique et elle dépend de *n*. Quand le nombre *n* augmente, la loi du CHI2 ressemble de plus en plus à la loi normale sans la rejoindre (χ^2 est un nombre positif alors que la distribution normale va de $-\infty$ à $+\infty$). C'est pour cela que, pour *n>30*, les tables de χ^2 ne sont plus chiffrées. Il faut alors faire la transformation :

$z=\sqrt{2\chi^2}-\sqrt{2n-1}$ *(I. 29)*

et appliquer le fait que *z* suit une loi normale réduite.

La figure I.2 illustre ces évolutions (*ddl=n*).

On peut aussi visualiser la courbe donnant la probabilité pour qu'une grandeur *x* expérimentale dépasse la valeur du CHI2 pour le nombre de degrés de liberté donné

[1]Voir par exemple SPIEGEL M.R. Théorie et applications de la statistique McGraw Hill 1984

(figure I.3). C'est l'aire sous la courbe ci-dessus (figure I.2) comprise entre CHI2 et l'infini. Par exemple, si on trouve *x=10* et avec *5* degrés de liberté, la probabilité trouvée est légèrement inférieure à *8%* mais n'est que de *1%* pour *2* degrés de liberté. On verra les applications de ces résultats dans les tests, mais on peut tout de suite faire remarquer que plus cette probabilité est faible, plus la valeur de *x* trouvée est anormalement grande et conduira à rejeter une hypothèse.

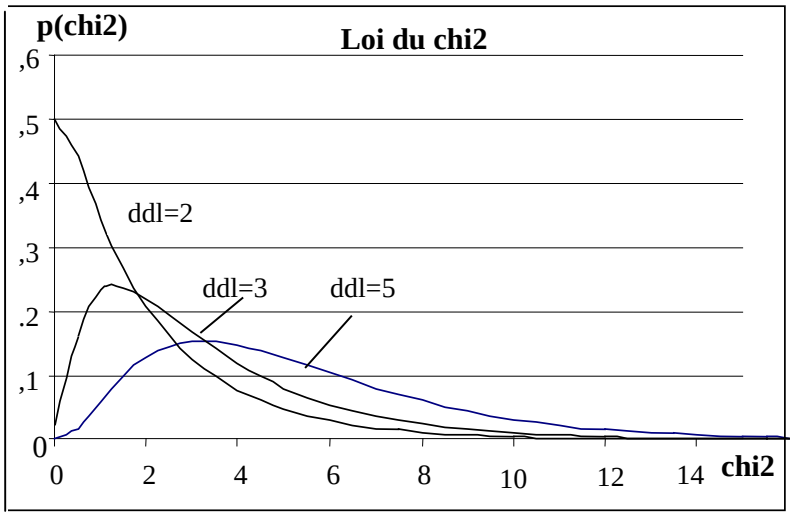

Fig.I. 2

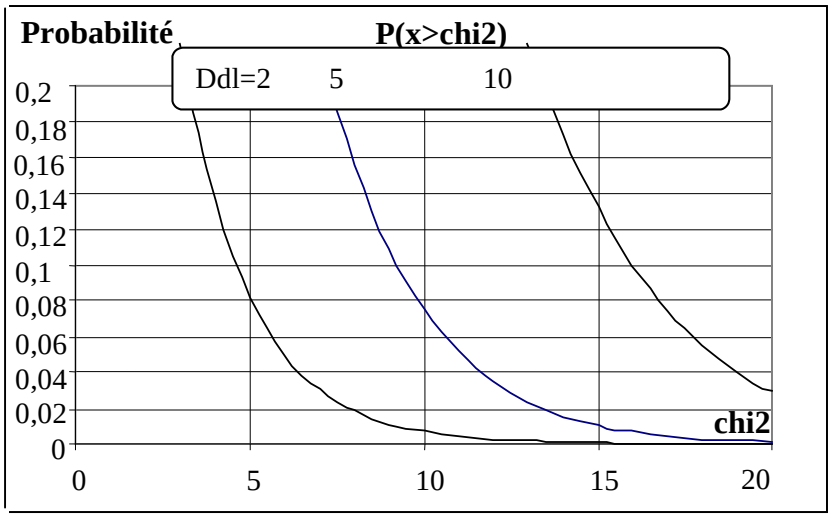

Fig.I. 3

4.4. Loi de Student.

La définition la plus générale introduit deux variables *X* (suit une loi normale centrée réduite) et *Y* (suit une loi de CHI2 de *n* degrés de liberté). La quantité :

$\frac{X}{\sqrt{\frac{Y}{n}}}$ *(I. 30)* suit une loi de Student à *n* degrés de liberté. Cette loi peut être appliquée avec *Y* loi de CHI2 associée à *X* . Soient x_0 une réalisation de *X* et x_i les *n* réalisations de *X*. La valeur de *t* est alors $t = \frac{x_0}{\sqrt{\frac{\sum x_i^2}{n}}}$ *(I. 31)*

Le dénominateur est une variance (moyenne nulle) estimée à partir des *n* valeurs de *x*.
Une forme plus utilisée dans la pratique consiste à considérer la moyenne *m* et la variance estimée s^2 d'un échantillon de taille *n* extrait d'une population normale (μ,σ). Ces *n* valeurs fournissent les estimations *m* de la moyenne et $s^2 = \frac{1}{n-1}\sum(x_i - m)^2$ de la variance (on verra au chapitre III la notion d'estimateur et la signification du terme *n-1* au dénominateur).

La quantité : $t = \frac{m-\mu}{\frac{s}{\sqrt{n}}}$ *(I. 32)* suit une loi de probabilité dite de Student. Ici, le nombre de degrés de liberté devient égal à *n-1* parce qu'on a remplacé l'écart-type inconnu par son estimation. On rencontre pour la première fois cette notion de contrainte : il a fallu estimer un écart-type inconnu.

Cette distribution est symétrique et a la forme d'une courbe en cloche comme la distribution de Gauss, mais d'autant plus aplatie que *n* est faible. Par ailleurs, cette loi tend vers la loi de Gauss quand *n* tend vers l'infini comme le suggère le théorème limite central. Il y a donc autant de distributions que de valeurs de *n*. Le graphe suivant (figure I.4) donne ces lois de densité pour Gauss (trait continu épais) et Student (*t* est remplacé par *u* pour Student) pour les degrés de liberté (ddl) *3* (trait continu mince) et 15 (trait discontinu). Les courbes étant symétriques, on n'en a représenté que la moitié.

4.5. La loi F de Snedecor.

Dans le cas le plus général, la loi *F* est le rapport de deux variable CHI2 associées à deux échantillons n_1 et n_2. Comme pour la loi de Student, on peut en donner une définition plus simple.
On suppose que l'on extrait deux échantillons (tailles n_1 et n_2) d'une distribution normale de moyenne μ et de variance σ^2.

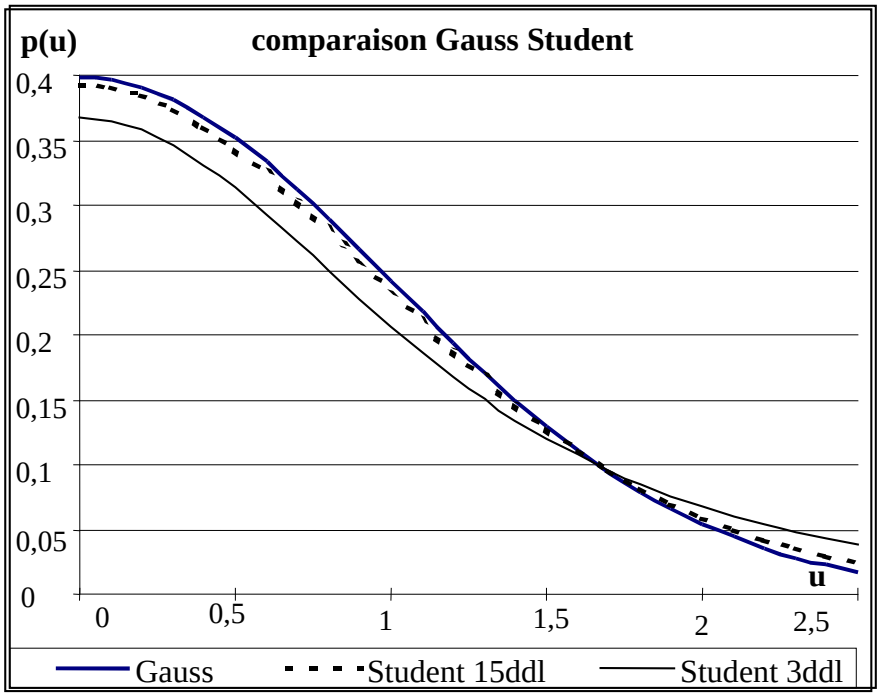

Fig.I. 4

Chaque échantillon permet d'obtenir une variance estimée s_1^2 et s_2^2. La quantité :

$$F = \frac{s_1^2}{s_2^2} \quad \textit{(I. 33)}$$

suit une loi de probabilité connue (dite de Fischer –Snedecor) qui dépend de deux paramètres n_1 et n_2. Il y a ici deux degrés de libertés $n_1 - 1$ et $n_2 - 1$, ce qui conduira à des tables plus complexes comme on le verra plus loin dans l'analyse de la variance. Comme dans le cas de la loi de Student, on retranche *1* pour obtenir le nombre de degrés de liberté puisqu'on a estimé les deux variances.

4.6. Deux distributions intéressantes.

En dehors des lois de probabilité classiques, on peut imaginer d'autres lois utiles pour certaines applications.

La loi rectangulaire.

Cette loi correspond à une probabilité constante dans un intervalle donné. On peut prendre le cas particulier d'une variable centrée, la translation permet de passer à une loi quelconque puisque seule la moyenne subit cette translation.
Soit une fonction constante dans l'intervalle (*-a*,*a*) de valeur *k*. Pour que cette fonction soit une loi de probabilité, il faut que

$$\int_{-a}^{a} f(x)dx = 1 = \int_{-a}^{a} k dx = 2ak \text{ soit } k = \frac{1}{2a}.$$

La loi de probabilité est donc de la forme : $p(x) = \frac{1}{2a}$ *(I. 34)* dans l'intervalle donné.

La moyenne est donnée par :

$$\int_{-a}^{a} xp(x)dx = \frac{1}{2a}\left[\frac{x^2}{2}\right]_{-a}^{a} = 0$$ ce qui était évident par symétrie.

La variance est donnée par :

$$\int_{-a}^{a} x^2 p(x)dx = \frac{1}{2a}\left[\frac{x^3}{3}\right]_{-a}^{a} = \frac{a^2}{3}$$ soit un écart-type : $\sigma = \frac{a}{\sqrt{3}}$ *(I. 35)*

Cette répartition peut facilement être simulée par ordinateur. Excel® possède par exemple une fonction mathématique *alea()* qui génère un nombre aléatoire entre *0* et *1*. Si on veut générer un nombre appartenant à une répartition rectangulaire centrée sur x_0 et de largeur *2a*, il faut faire un changement de variable. On obtient d'abord un nombre entre *–1* et *+1* par le changement *2*(alea()-0,5)*, puis un nombre entre *–a* et *+a* par la transformation :
2a(alea()-0,5)*, enfin le nombre cherché par la translation x_0 :

$x = x_0 + 2a * (alea() - 0{,}5)$

Dans les macros écrites en Visual Basic, l'ordre *alea()* correspondant est *Rnd*.

Loi triangulaire.

On prend encore la loi symétrique par rapport à l'origine et comprise entre *-a* et *a*. Dans l'intervalle (*-a,0*), la relation est de la forme $p(x)=\alpha x+\beta$, avec *p(-a)=0* et *p(0)=k*.
On en déduit facilement $\beta=k$ et $\alpha = \frac{k}{a}$.

Pour l'intervalle (*0,a*), un calcul analogue conduit à $\beta=k$ et $\alpha = -\frac{k}{a}$.

Pour que cette fonction représente une loi de probabilité, il faut en plus que l'intégrale sur l'intervalle soit égale à *1*, soit *0,5* sur le demi-intervalle par raison de symétrie. On calcule par exemple :

$$\int_0^a p(x)dx = 0{,}5 = \int_0^a (-\frac{k}{a}x + k)dx = k\left[-\frac{x^2}{2a} + x\right]_0^a = 0{,}5ka$$ soit $k = \frac{1}{a}$.

Entre *0* et *a*, la loi de probabilité est donc : $p_1(x) = -\frac{1}{a^2}x + \frac{1}{a}$ *(I. 36)*

et $p_2(x) = \frac{1}{a^2}x + \frac{1}{a}$ *(I. 37)* dans l'autre intervalle.

Le calcul de la moyenne ne pose pas de difficultés et donne évidemment *0*.

Le calcul de la variance est donné par :

$$\int_0^a x^2 p_1(x)dx + \int_{-a}^{0} x^2 p_2(x)dx = 2\int_0^a x^2 p_1(x)dx$$ à cause de la symétrie.

On a donc : $\int_0^a x^2 p_1(x)dx = \int_0^a x^2(-\frac{1}{a^2}x + \frac{1}{a})dx = \frac{1}{a}\left[-\frac{x^4}{4a} + \frac{x^3}{3}\right]_0^a = \frac{1}{12}a^2$